Use of Whole Genome Sequencing in Understanding Transmission Dynamics of Tuberculosis

Jones, Rhys C.

Use of Whole Genome Sequencing in Understanding Transmission Dynamics of Tuberculosis

Rhys Compton Jones

Submitted to Swansea University in fulfilment of the requirements for the Degree of

Doctor of Philosophy in Medical and Healthcare Sciences

Swansea University

2018

Abstract

Mycobacterium tuberculosis is the leading cause of death from an infectious disease worldwide. An understanding of tuberculosis transmission dynamics in outbreak settings is vital for its control. The advent of affordable whole genome sequencing (WGS) has provided scope for superior resolution of tuberculosis outbreaks, compared to previous methods. However, the challenge lies in standardising the vast quantities of resulting data in a structured manner which lends itself to easy comparison of isolates. Gene-by-gene Multi-Locus Sequence Typing (MLST) methods of analysing WGS data, as opposed to Single Nucleotide Polymorphism (SNP) mapping, have shown promise in providing a uniform platform for outbreak resolution.

WGS was performed on clinical isolates from three *M. tuberculosis* outbreaks in South West Wales. Molecular typing by MIRU-VNTR and epidemiological investigation had resulted in conflicting conclusions. Outbreak analysis and phylogenetic typing of all isolates was carried out using the WGS gene-by-gene MLST analysis method of core genome MLST (cgMLST) and traditional WGS SNP mapping. Where DNA quality was unsatisfactory, an ancient DNA library preparation was used successfully. Provean and BEAST software analysis provided physiological information and ancestral dating respectively on outbreak isolates. WGS successfully resolved all three outbreaks, with cgMLST providing clear conclusions across each outbreak. Traditional SNP mapping provided greater resolution than cgMLST in one outbreak. Ancestral dating also proved useful in understanding the outbreaks.

Phylogenetically, the dataset was dominated by Euro-American lineage strains, providing the first snapshot of tuberculosis diversity within Wales. Provean analysis identified physiological features in isolates worthy of future research.

In summary, WGS was successfully used to resolve three *M. tuberculosis* outbreaks across South West Wales and correlated better with the epidemiological data than molecular typing by MIRU-VNTR had done. The study highlighted the dominance of the Euro-American lineage within the outbreaks and included the first use of ancient DNA library preparation in a clinical outbreak.

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed……………………………………… Date……………

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed……………………………………… Date……………

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed……………………………………… Date……………

Ariannwyd Rhys Jones gan Gynllun Ysgoloriaethau Ymchwil y Coleg Cymraeg Cenedlaethol.

Rhys Jones was funded by the Coleg Cymraeg Cenedlaethol Research Scholarship Scheme.

Table of Contents

Abstract....2

DECLARATION 3

Acknowledgments....11

List of Figures 13

List of Tables 15

List of Abbreviations....18

Publications....20

Manuscripts Submitted: 20

Manuscripts Pending:....20

Chapter 1....21

Introduction 21

1.1: Mycobacteria 22

1.2.1: Mycobacterium tuberculosis....24

1.2.2: *M. tuberculosis* lineages and strains....25

1.3: Tuberculosis in history....27

1.4: Current status of tuberculosis disease worldwide....28

1.5: Tuberculosis in Wales....30

1.5.1: Tuberculosis in Wales – Overview of demography.....31

1.6: Pathogenesis of tuberculosis....34

1.7: Treatment of tuberculosis....34

1.7.1: Tuberculosis Vaccination....35

1.8: Control measures and strain typing 36

1.8.1: Tuberculosis control....36

1.8.2: Strain typing 37

1.8.3: Spoligotyping....38

1.8.4: IS6110 RFLP 39

1.8.5: VNTR and MIRU-VNTR....40

1.8.6: MIRU-VNTR 41

1.8.7: Limitations of MIRU-VNTR typing....42

1.9: Whole Genome Sequencing 43

1.9.1: Whole Genome Sequencing of *M. tuberculosis*....44

1.9.2: Genome plasticity....44

1.9.3: WGS for *M. tuberculosis* outbreak resolution 45

1.9.4: Limitations of traditional WGS methods 46
1.9.5: Application of gene-by-gene WGS multi-locus sequence types (MLST) 46
1.10: Aims of the study 48
Chapter 2..49
Materials and Methods 49
2.1: Sample collection and epidemiological information 50
2.2: Sequencing and assembly 50
2.3: Core genome MLST analysis 51
2.3.1: Analysis using the established Ridom SeqSphere cgMLST scheme...........................51
2.3.2: Analysis using an in-house cgMLST scheme 51
2.3.3 Extended 3646 gene-by-gene MLST 51
2.4: Ancestral dating 52
2.5: Whole Genome Sequence Single Nucleotide Polymorphism Mapping 52
2.5.1: SNP bar-coding 53
2.5.2: Extraction of individual SNPs 55
2.5.3: Functional analysis of non-synonymous SNPs 55
2.6: *In silico* spoligotyping 55
2.7: Principal Genetic Grouping (PGG) 56
2.8: SNP Cluster Grouping (SCG) 57
2.9: Construction of cgMLST phylogenies for phylogenetic assignment 57
2.10: Sub-lineage genotyping 58
2.10.1: Beijing genotyping...58
2.10.2: Latin American Mediterranean (LAM) genotyping 60
2.10.3: Haarlem genotyping 60
2.10.4: X family genotyping...60
2.10.5: Lineage 1 genotyping of BK22 61
2.11: Statistical analysis for lineage association 61
2.12: Gel electrophoresis 62
2.13 Ancient DNA extraction and library preparation protocol 63
2.13.1: Ancient DNA extraction protocol 63
2.13.2: Ancient DNA library preparation and sequencing of DNA 64
2.14: Production of Geographical map..64
2.15: Statistical analysis of in-house and established cgMLST variation...............................64
Chapter 3..65
Phylogenetic analyses of *M. tuberculosis* isolates from across South West Wales. 65

3.1: Introduction 66

3.1.1: M. tuberculosis strains and lineages 66

3.1.2: *M. tuberculosis* major lineages 67

3.1.3: *M. tuberculosis* sub-lineages 69

3.1.4: Robust genetic markers for phylogenetic classification *M. tuberculosis* 71

3.1.4.1: Principal Genetic Grouping 72

3.1.4.2: SNP cluster grouping 72

3.1.4.3: Robust SNP barcode method 74

3.1.5: Novel use of a core genome MLST method for phylogenetic purposes 75

3.1.6: Aims and objectives 75

3.2: Methods 76

3.2.1: Principal genetic grouping (PGG) 76

3.2.2: SNP cluster grouping (SCG) 76

3.2.3: Major lineage and sub-lineage assignment 76

3.2.4: Construction of cgMLST phylogenies 76

3.2.5: Statistical analysis for lineage association 77

3.2.6: Sub-lineage Genotyping 77

3.2.6.1: Beijing genotyping 77

3.2.6.1.1: Advanced Beijing Genotyping 77

3.2.6.2: Latin American Mediterranean (LAM) genotyping 77

3.2.6.3: Haarlem genotyping 78

3.2.6.4: X family Genotyping 78

3.2.6.5: Lineage 1 genotyping for BK22. 78

3.3: Results 79

3.3.1: PGG analysis 79

3.3.2: SNP cluster group (SCG) analysis 81

3.3.3: Robust SNP bar-coding 84

3.3.4: cgMLST association in the Welsh isolates 89

3.3.5: Statistical analysis of lineage association 92

3.3.6: Supporting genotype information 93

3.3.6.1: Beijing Genotyping 93

3.3.6.2: Latin American Mediterranean (LAM) genotyping 96

3.3.6.3: Haarlem genotyping 96

3.3.6.4: X family genotyping 98

3.3.6.5: Lineage 1 Genotyping 100

3.3.7: Phylogenetic composition of outbreak isolates..102

3.4: Discussion..103

3.5: Conclusions and future perspectives...109

Chapter 4..111

Analysis of an outbreak of *M. tuberculosis* in the Llwynhendy area of South West Wales using Whole Genome Sequencing. 111

4.1: Introduction 112

4.1.1: Epidemiology of Cases..112

4.1.2: Chapter aims 118

4.2: Methods...119

4.2.1: Sample collection and gathering of epidemiological information 119

4.2.2: DNA sequencing 119

4.2.3: Core genome MLST 119

4.2.4: Whole Genome Sequence SNP mapping...120

4.2.5: Functional analysis 120

4.3 Results...121

4.3.1 Initial Sequencing results...121

4.3.1.1: cgMLST 122

4.3.1.2: Neighbour joining analysis of the outbreak complex 124

4.3.2 Analysis of individual allelic differences 126

4.3.3: Traditional SNP mapping..127

4.3.4: Functional analysis of the allelic differences within the Llwynhendy outbreak isolates. 129

4.4: Discussion..132

4.5 Conclusion...136

Chapter 5..137

Resolution of an outbreak of *Mycobacterium tuberculosis* in the Neath Port Talbot area of South Wales using Whole Genome Sequencing. 137

5.1 Introduction 138

5.1.1 Epidemiology and Typing of Neath Port Talbot Outbreak Cases 138

5.1.2 Proposed transmission chain...140

5.1.3: Aims...143

5.2 Methods...144

5.2.1: Sample collection and epidemiological information 144

5.2.2: DNA Sequencing...144

5.2.3: Phylogenetic filtering of original dataset...144

5.2.4: Core genome MLST (cgMLST) analysis 144

5.2.4.1: Established cgMLST scheme 144

A cgMLST scheme 144

5.2.4.2: In-house core genome MLST (cgMLST) scheme 145

5.2.5: SNP analysis 145

5.2.5.1: Extraction of individual mutations 145

5.2.5.2: Prediction Functional Genomic 145

5.2.6: WGS SNP mapping 145

5.2.7: Ancestral dating 146

5.2.8: Statistical calculation of average variation between cgMLST schemes 146

5.3 Results 147

5.3.1 Outbreak resolution using the cgMLST scheme 147

5.3.2 Mutation breakdown 149

5.3.3 Functional Analysis 150

5.3.4 Outbreak resolution using an in-house core genome MLST scheme 152

5.3.5 Comparison between the in-house and established cgMLST schemes 154

5.3.6 Whole Genome Sequence SNP mapping 155

5.3.7 Ancestral dating 159

5.4: Discussion 161

5.5 Conclusion 167

Chapter 6 168

The study of a *M. tuberculosis* outbreak in Gorseinon, South Wales, using a novel ancient DNA library build and sequencing protocol. 168

6.1: Introduction 169

6.1.2: Background 171

6.1.2.1: Gorseinon public-house associated outbreak (2003-2004 and 2007) 171

6.1.2.2: Townhill outbreak (2006-2008) 174

6.1.3: Aims 177

6.1.4: DNA sequencing problems 178

6.2: Methods 180

6.2.1: Standard Sequencing and assembly 180

6.2.2: Ancient DNA sequencing protocol 180

6.2.3 Core Genome MLST 180

6.2.4: Extended gene–by-gene MLST. 180

6.2.5: WGS SNP mapping 181

6.2.6 In silico spoligotype 181

6.2.7: Ancestral dating 181

6.3: Results 182

6.3.1: Standard sequencing 182

6.3.2: Ancient DNA sequencing 183

6.3.3: *In silico* spoligotyping 184

6.3.4: cgMLST analysis 185

6.3.5: Extended gene-by-gene analysis 187

6.3.5.2: Whole Genome SNP mapping 188

6.3.6: Ancestral dating 191

6.4: Discussion 192

6.5: Conclusion 197

Chapter 7 199

General Discussion 199

Bibliography 210

Appendix 234

Acknowledgments

I would like to say to Dr Angharad Davies, my 1st supervisor, that I greatly appreciate the support you have given me and the effort you expended during the research and the highs and lows of writing this Thesis. Angharad, had you not been as flexible and free thinking as you have been I am not sure I ever would have won a grant to work in Copenhagen using ancient DNA methods, which provided us with our unique results. I hope that our work together has been as much of a pleasure to you as it has been to me. I would like to thank Dr Llinos Harris for being a cornerstone throughout my whole PhD experience. Your honesty, work ethic and quality advice gave me confidence through the whole process. I would also like to thank Dr Thomas Wilkinson for his academic guidance and positive nature throughout my time at Swansea University. To Professor Tom Humphrey, your presence itself had a surreal reassurance and I don't think I can thank you enough for the advice and comfort you provided me during my PhD.

I must also acknowledge the Coleg Cymraeg Cenedlaethol for funding this studentship and in addition for providing me with support throughout the whole time period. Thanks are also due to St David's Medical Foundation which provided a seed-corn grant to the project. I must also thank The European Microbiology Organisation, Professor Tom Gilbert and Dr Marcella Sandoval Velasco and Copenhagen University for putting faith in my hypothesis that ancient DNA sequencing methods could have clinical utility.

I would also like to pay a special thanks to Dr Lisa Williams, although originally not part of my supervisory group, Lisa you have had an immense impact on my PhD from scientific advice to being my emotion councillor. You're a great scientist but an even better person and I can't thank you enough, I'm sure we will remain friends long after this thesis. I must also thank Dr Matt Hitchings for being my guidance through the process of sequencing and scripting, your help has been priceless.

Mr Dan John, I don’t think I’ll ever have a work colleague that I enjoy working with more, thanks a lot for your support and for keeping me sane during the whole course.

I would like to offer my special thanks to all the members of the Microbiology and Infectious Disease Group in Swansea University for their friendship and scientific encouragement. To Heather Chick, Dr Leonardos Mageiros, Dr Susan Murray and Dr Gethin Thomas thank you very much for everything.

To my family, I would like to thank you for supporting me through the entire six and a half year journey it’s taken to make me a Doctor, I am sorry if I have driven you mad at times.

Finally, I would like to dedicate this Thesis to my parents, Sian and Neil Jones. Without your support I never would have made it this far, I love you very much.

List of Figures

Figure 1.1: A Geographical map extracted from the Public Health Wales Annual tuberculosis report……..33

Figure 3.1: Phylogeny highlighting the PGG assignment of 57 Welsh isolates........................80

Figure 3.2: The number of isolates in each SCG..81

Figure 3.3: A neighbour joining phylogeny showing the SCG profile for 57 *M. tuberculosis* isolates (including the H37Rv reference genome)………………………………………..………………83

Figure 3.4: A phylogenetic tree of the 59 Welsh *M. tuberculosis* isolates based on SNPs at 60 loci adapted to assign isolates to lineages and sub lineages………………………………….............85

Figure 3.5: Number of isolates representing each sub-lineage present within this collection of 59 Welsh *M. tuberculosis* isolates...86

Figure 3.6: A subset of the phylogeny based on the cgMLST association of 57 isolates between 66 Welsh isolates and 179 isolates that were assigned a lineage in a previous study.. 91

Figure 3.7: Neighbour joining phylogeny tree based on 48 SNPs containing Welsh isolates LL9, BK25 and BK21 and 26 well characterised Beijing isolates from the Mestre *et al*. (2011) study.. 95

Figure 4.1: A map showing the locations of where potential cases associated with the Llwynhendy outbreak resided across South West Wales..114

Figure 4.2: Diagram showing the epidemiological relationship between each isolate with a confirmed or potential epidemiological link to the outbreaks..116

Figure 4.3: A minimum spanning tree showing the number of allelic differences across the cgMLST scheme...123

Figure 4.4: Neighbour joining tree constructed using the core genome MLST scheme. The figure contains only the outbreak complex subset defined in Figure 4.3........................... 125

Figure 4.5: A neighbour joining tree based on the WGS data obtained from the conserved signature indels (CSI) phylogeny analysis of the outbreak complex subset 127

Figure 5.1: Diagram showing the epidemiological relationship between each isolate with a confirmed or indirect epidemiological link to the outbreak... 142

Figure 5.2: A minimum spanning tree of 17 cases constructed using Ridom SeqSphere software and based on a 2891 cgMLST scheme...148

Figure 5.3: A minimum spanning tree generated by in-house cgMLST and based on 2562 genes present with high quality scores across each isolate within the dataset..153

Figure 5.4: A graph showing the average number of allelic differences between each isolate in outbreak 1 and 2 defined by the published and in-house cgMLST schemes....................154

Figure 5.5: (a) Neighbour- joining phylogeny produced by traditional WGS SNP mapping described in section 5.2.7 (b)Corresponding Neighbour-joining phylogeny produced by cgMLST correlating to the original minimum spanning tree result in figure 5.2..................156

Figure 5.6: A phylogenetic tree based on the ancestral reconstruction and dating of the most recent common ancestor (TMRCA) of both NPTA and NPTB outbreaks.............................159

Figure 6.1: A diagrammatic view of the epidemiological connections between all isolates associated with the public house outbreak in Gorseinon..173

Figure 6.2: A map showing the distances between the 9 Gorseinon and 3 Townhill outbreak cases...175

Figure 6.3: A diagrammatic view of the epidemiological connections between the Townhill isolates...177

Figure 6.4: cgMLST-based phylogenies of the 11 successfully sequenced isolates..............186

Figure 6.5: A minimum spanning tree of the 11 isolates, based on the cgMLST scheme and 755 accessory genes...188

Figure 6.6: (a) Minimum spanning tree results from WGS SNP mapping of the Gorseinon/Townhill Outbreak...189

Figure 6.6: (b) Neighbour Joining analysis results from WGS SNP mapping of the Gorseinon/Townhill Outbreak...189

Figure 6.7: An ancestral dating phylogeny based on the consensus sequence of outbreak 1 and the three outbreak 2 isolates..191

Figure PH1: The figure above shows an un-rooted phylogeny of all isolates originally included in the cgMLST analysis and is the origin of the subset phylogeny presented in figure 3.6..234

Figure LL1: A minimum spanning tree based on 1466 core genes successfully aligned to the cgMLST scheme...246

List of Tables

Table 1.1: The primary host of currently defined MTBC members and three further ill-defined ones of the complex which appear at the bottom of the table. 23

Table 1.2: List of the established *M. tuberculosis* typing methods, with corresponding comments relating to the features, advantages and disadvantages of each method............38

Table 2.1: A table extracted from the Coll *et al* (2014) publication which highlights the correlating phylogenetic lineage and sub lineage to specific SNP at designated loci.........53-54

Table 2.2: The combinations of amino acids seen at the loci gyrA and katG and the possible PGGs 56

Table 2.3: List of the nine loci derived from Alland *et al*. (2007) used for assigning isolates into one of six SNP cluster groups...57

Table 2.4: A table showing the 48 loci across 22 genes analysed for Beijing genotyping........59

Table 3.1: A matrix showing the correlation between the SCG and spoligotype families..73

Table 3.2: Percentage presence of the 60 loci analysed in the 59 isolates 88

Table 3.3: SNP analysis of the 3 well-characterised Beijing genotype markers in the 3 Welsh isolates assigned by SNP bar-coding as Beijing isolates...94

Table 3.4: Table showing the presence or absence of a SNP at position 309 of gene Rv0129, for the isolates designated as LAM sub-lineage isolates..96

Table 3.5: Table showing the Haarlem polymorphisms within genes mtgC, ogt and ung in 14 Welsh isolates..97

Table 3.6: Presence of X family specific SNPs at loci Rv3221c and Rv2330 in the 16 Welsh isolates assigned as X family genotypes by SNP bar-coding 99

Table 3.7: Analysis of 17 SNPs specific to lineage 1 genotypes...101

Table 3.8: Composition of the outbreak-related isolates with regards to their PGG, SCG and sub-lineage assignments 102

Table 4.1. Overview of the basic epidemiology and MIRU-VNTR results on the Llwynhendy tuberculosis outbreak cases 117

Table 4.2: The percentage of good targets present in each isolate following sequencing...121

Table 4.3: Distance matrix showing the number of allelic differences between each of the isolates represented in the outbreak complex defined in Figure 4.3 124

Table 4.4: A matrix highlighting each allelic difference between each isolate and the source isolate (LL1) 126

Table 4.5: The number of SNPs present between each isolate defined as being part of the outbreak complex in Figure 4.3..128

Table 4.6a: A detailed matrix based on the allelic differences between LL1, LL2, LL3, LL5 and NPTB6 130-131

Table 4.6b: Mutations observed within the Rv3594 gene of LL3 131

Table 5.1: A descriptive summary of each outbreak case and their MIRU-VNTR patterns, generated by the PHW molecular unit...139

Table 5.2: Synonymous and non-synonymous mutations, by gene, between the consensus sequences of outbreaks 1 and 2...150

Table 5.3: Functional analysis on the non-synonymous mutation genes between outbreak 1 and outbreak 2...151

Table 5.4: A correlating distance matrix to figure 5.4, produced by traditional WGS SNP mapping described in section 5.2.7...157

Table 6.1: Epidemiological data for each isolate of the 13 isolates included within this dataset..170

Table 6.2: A table showing the MIRU-VNTR profiles across 15 loci seen for each of the 13 isolates within this dataset 172

Table 6.3: Basic sequencing results for the 13 Gorseinon and Townhill outbreak isolates..183

Table 6.4: Sequencing results from 9 Gorseinon and Townhill outbreak isolates that failed to sequence using the standard sequencing protocol, obtained using the ancient DNA protocol..184

Table 6.5: In silico spoligotyping results for each of the Gorseinon and Townhill-associated outbreak isolates 185

Table 6.6: A table showing the number of SNPs between each isolate found in the WGS SNP mapping phylogeny shown in Figure 6.4a 190

Table PH1: Below is a conclusion table showing the phylogenetic classification`s across all methods described in chapter 3 for each isolate...236-239

Table PH2a and PH2b: Tables showing the full SNP pattern of 48 loci identified in the Mestre *et al* (2011) study and used to construct figure 3.7...240-243

Table PH3a: A distance matrix based on the number of polymorphisms across the 48 defined in the Mestre *et al*, 2011 study..244

Table PH3b: A subset of the distance matrix seen above, highlighting the number of SNP`s, across the 48 previously defined loci, between only those isolates most closely related to the Welsh isolates LL9, BK21 and BK25...245

Table LL1: Distance matrix showing the number of allelic differences between all isolates defined in Figure 4.3..247

Table PH4: A table showing the percentage probability that each isolate is associated with a given lineage...248-249

List of Abbreviations

µg: micro gram

µl: micro litre

aDNA: Ancient DNA.

BEAST: Bayesian Evolutionary Analysis Sampling Trees.

BEST: Blunt End Single Tube

BIGSdb: Bacterial Isolate Genome Sequence database

BK: Background isolate

Bp: Base pair

BWA: Burrows Wheel Aligner

CBN: Conformal Bayesian Network

cgMLST: Core Genome MLST.

CSF: Cerebral Spinal Fluid

CSI: Conserved Signature Indels.

DNA: Deoxyribonucleic acid

EDTA: Ethylene diamine tetra acetic acid

ETR: Exact Tandem Repeats

GO: Gorseinon

HCl: Hydrogen Chloride

IS6110 RFLP: Insertion Sequence 6110 Restriction Fragment Length Polymorphism.

iTol: International Tree of Life

Kb: Kilo base

LL: Llwynhendy outbreak

MDR: Multi Drug resistance

XDR: Extensive Drug resistance

MIRU-VNTR: Mycobacterial Interspersed Repeat Unit Variable Number Tandem Repeats.

Ml: millilitre

MLST: Multi Locus Sequence Typing

mM: milliMolar

MPTR: Major Polymorphic Tandem Repeat

MTBC: Mycobacterium Tuberculosis Complex

NaCl: Sodium Chloride

NPT: Neath Port Talbot

NPTA: Neath Port Talbot outbreak A

NPTB: Neath Port Talbot outbreak B

PCR: Polymerase Chain Reaction

PGG: Principal Genetic Group

PH: Phylogenetics

PHE: Public Health England

PHW: Public Health Wales

Provean: Protein Variation Effect Analyzer

SCG: SNP Cluster Group

SDS: Sodium dodecyl sulphate

SNP: Single Nucleotide Polymorphism

SNV: Single Nucleotide Variant

SNV: Single Nucleotide Variant

SRA: Short Read Archive

TAE: Tris Acetate EDTA

TH: Townhill

UPGMA: Unweighted Pair Group Method with Arithmetic Mean

UV: Ultra Violet

WCM: Wales Centre for Mycobacteriology

WGS: Whole Genome Sequencing.

WHO: World Health Organization

Publications

Manuscripts Submitted:

“Analysis of a clinical outbreak of tuberculosis using palaeogenomic methodologies for successful genome reconstruction from Mycobacterium tuberculosis boilates”. Submitted to Thorax May 2018.

Manuscripts Pending:

“A snapshot of the phylogenetic diversity *of M. tuberculosis* in Wales using novel and traditional WGS bioinformatic analysis”.

“The resolution of three clinical Tuberculosis outbreaks, using Whole Genome sequencing and a standardized bioinformatic platform, whereby MIRU-VNTR and traditional contact tracing contrasted”.

Chapter 1

Introduction

1.1: Mycobacteria

Mycobacteria are a genus within the family Mycobacteriaceae and the order Actinomycetales (King *et al.*, 2017, Ryan and Ray., 2004). They are aerobic, non-motile, non-spore forming and are rod shaped. Their size varies from 0.1 to 0.6 µm in width and 1.0 to 10 µm in length (Ryan and Ray., 2004). They grow at temperatures that range between 30 and 45°C, and the speed at which they grow also varies. Some take over 7 days to form visible colonies (slow growers), while others do so more rapidly (rapid growers) when sub-cultured on Lowenstein-Jensen media (Ryan and Ray., 2004).

Mycobacteria can be further divided into one of three major groups: *Mycobacterium tuberculosis* complex species (MTBC), non-tuberculosis mycobacterial (NTM) species and *Mycobacterium leprae* (Health, 2010, Ryan and Ray., 2004, Parish and Brown, 2009), the causative agent of leprosy (Young *et al*., 1985, Ryan and Ray, 2004). Human tuberculosis is a predominantly (though not exclusively) pulmonary disease caused by members of the MTBC (Ryan and Ray., 2004, Frothingham and Meeker-O'Connell., 1998, Van Soolingen *et al*., 1997), which contains eight established members capable of causing tuberculosis in humans or animals. The complex contains three human-specific members, *M. tuberculosis, M. canetti* and *M. africanum* as shown in Table 1.1 (Van Soolingen *et al*., 1997, Sharma *et al*., 2016, Niemann *et al*., 2004, Ryan and Ray., 2004). The rest infect animal hosts primarily but share the potential to infect humans (Ryan and Ray., 2004). The most prominent of these is *M. bovis,* which is the causative agent of bovine tuberculosis, a disease which is currently responsible for a large economic burden within the UK livestock industry, amounting £500 million in the last decade with the most recent report showing the annual cost rising to £34.1 million in 2013, as opposed to £28.6 million in 2008, paid in compensation across England alone (Grange., 2001, DEFRA., 2013). *M. bo*vis can infect humans, causing zoonotic tuberculosis (De la Rua-Domenech R., 2006). Prior to the advent of milk pasteurisation and strict cattle culling programmes between 1935 and 1950, *M. bovis* was endemic across the UK with reports showing around 2,500 individuals were dying from the pathogen annually (De la Rua-Domenech R., 2006).

Table 1.1: The primary host of currently defined MTBC members and three further ill-defined ones of the complex which appear at the bottom of the table. The latter three have limited evidence of zoonotic transfer to humans (Van Ingen *et al*., 2012, World Organisation of Animal Health, 2016)

Member	**Primary Host**
Mycobacterium tuberculosis	Human
Mycobacterium africanum	Human
Mycobacterium bovis	Cattle
Mycobacterium canetti	Human
Mycobacterium caprae	Deer
Mycobacterium mungi	Mongoose
Mycobacterium pinnipedi	Pinnipeds (seals and sea lions)
Mycobacterium microti	Voles
Mycobacterium suricattae	Meerkats
Dassie bacillus	Rock hyrax
Oryx bacillus	Bovidae including Oryx, Gazelles and Deer

Non-Tuberculosis Mycobacterium (NTMs) account for over 150 mycobacterial species and are ubiquitous in the natural environment, being present in soil, ground water, drinking water and even pasteurised milk and cheese, although the latter are rare events (Greenwood., 2012, Spahr and Schafroth., 2001). They are generally less pathogenic in humans. However, certain species including *M. abcessus and M. kansassi* can be pathogenic (Nguyen.,1997, Greenwood., 2012), especially in individuals who are immuno-compromised or those with concurrent pulmonary disease such as cystic fibrosis (Jeong *et al*., 2004, Nguyen., 1997, Greenwood., 2012). Clinical manifestations of disease caused by an NTM may be pulmonary but species do also cause skin, lymphatic, soft tissue and disseminated infections (Uslan *et al*., 2006, Greenwood., 2012). Although causing less of a disease burden than the MTBC species, numbers of NTM infections in immuno-compromised individuals, such as those suffering from AIDS and cystic fibrosis, is substantial (Greenwood., 2012).

Mycobacteria are characterised by having a cell wall that is thick, hydrophobic, waxy and contains long chain fatty acids, called mycolic acids (Barry *et al*., 1998). Each mycolic acid molecule contains between 60-90 carbon atoms but the exact number varies between mycobacterial species (Barry *et al*., 1998). The presence of mycolic acids contributes to the robustness of the mycobacteria, conferring increased resistance to chemical damage, dehydration, hydrophilic antibiotics and various biocides (Greenwood, 2012, Ryan and Ray, 2004a).

1.2.1: Mycobacterium tuberculosis

Despite all members of MTBC having the ability to cause disease, *M. tuberculosis* is the one that represents the primary causative agent of tuberculosis in humans, especially in more developed countries such as the UK whereby *M. tu*berculosis accounts for over 98% of all tuberculosis cases (Brosch *et al*., 2002, Huard, R. C *et al*., 2003, Public Health England., 2017).

M. bovis and M. africannum represent the two other MTBC agents that commonly cause tuberculosis, although their impact is largely restricted to areas of Africa and in reality are a significantly lesser burden in relation to the global epidemic of tuberculosis (WHO., 2016, WHO., 2017, Public Health England., 2014, Public Health England., 2017). The number of human cases caused by *M. bovis* varies by geographic region, with 10,000 of the 13,400 worldwide deaths from *M. bovis* in 2015 occurring on the African continent (World Organisation of Animal Health., 2016). In more developed countries, *M. bovis* is not a prominent human pathogen, causing only 0.9% of human tuberculosis cases in the UK in 2016 (Public Health England., 2016, World Organisation of Animal Health., 2016, WHO., 2016). The majority of *M. africanum* cases are restricted to West African individuals and are thus not as prominent globally, causing around 1.6% of all tuberculosis cases in 2015 (WHO., 2016) and only 1.4% of culture-confirmed ones in the UK as of 2014 (Public Health England., 2014).

1.2.2: *M. tuberculosis* lineages and strains

M. tuberculosis strains can be categorised into 1 of 7 phylogenetic major lineages, based originally on large sequence polymorphism classification (Gagneux *et al*., 2006, Gagneux and Small., 2007) and subsequent WGS single nucleotide polymorphism (SNP) mapping (Comas *et al*., 2013, Reiling *et al*., 2013, Firdessa *et al*., 2013, Yimer *et al*., 2015).

The 7 phylogenetic major lineages are split into phylogeographically related groups: Lineage 1 - Indo-Oceanic, Lineage 2 -East-Asian, Lineage 3 -East African-Indian, Lineage 4 -Euro-American, Lineage 5-West Africa 1, Lineage 6-West Africa 2 and Lineage 7 -Horn of Africa (Gagneux and Small., 2007). Previous studies have shown that the different lineages (and sub-lineages within them) of *M. tuberculosis* have different clinical phenotypes with regards to transmissibility, drug resistance, host interaction, latency and vaccine efficacy (Reed *et al.*, 2009, Reiling *et al*., 2013, Firdessa *et al*., 2013). Identifying which lineages are present within an area is important with regards to understanding the epidemiological background of tuberculosis in that area, especially with global human migration blurring the phylogeographic restrictions of the lineages. Lineages can be grouped into those that are evolutionarily modern and those that are ancient (Gagneux and Small., 2007). Lineages 1, 5 and 6 are thought to represent more ancient lineages that coexisted with humans prior to Neolithic expansion of anatomically modern *Homo sapiens* (Comas *et al*., 2013). Strains from these ancient lineages have been shown to initiate a greater immune response and are more likely to become latent than their modern relatives (Portevin *et al*., 2011). In contrast, the modern strains of Lineages 2, 3 and 4 appear to possess more virulent phenotypes with higher transmissibility, more readily causing active disease and propensity for drug resistance, and are more geographically dispersed (Sarkar *et al*., 2012, Gagneux., 2012). In particular, Lineage 2 (East Asian) and Lineage 4 (Euro-American) contain strains, such as the Beijing and Haarlem genotypes respectively, which are notorious for their association with tuberculosis outbreaks and are over-represented amongst drug resistant cases (Bifani *et al*., 1996, Mardassi *et al*., 2005, Marais *et al*., 2006,).

The W/Beijing genotype is a strain found within Lineage 2, which has been regularly associated with large *M. tuberculosis* outbreaks (Thwaites *et al*., 2008, Bifani *et al*., 1996). In particular, the W/Beijing strains are known to have a higher propensity for acquisition of drug resistance genes, causing the development of multi-drug and extensively-drug resistant strains (Bifani *et al*., 1996, Thwaites *et al*., 2008). They have been shown to disseminate rapidly and cause more severe disease phenotypes. This was particularly evident in a large outbreak caused by the strain in New York in the 1990s (Bifani *et al*., 1996, Thwaites *et al*., 2008). The W/Beijing genotype has the ability to produce a unique phenolic glycolipid that attenuates the host's immune response (Nicol *et al*., 2005). Meningeal tuberculosis caused by W/Beijing strains also has a more rapid progression, with shorter symptom duration when compared with meningeal infection caused by other lineage strains (Nicol *et al*., 2005). W/Beijing strain infections are associated with low levels of cerebro spinal fluid (CSF) leukocytes, which is an independent risk factor for death or severe disability from meningeal tuberculosis (Thwaites *et al*., 2008).

The Haarlem genotype is found within Lineage 4 and has also been associated with large outbreaks of *M. tuberculosis*. This genotype accounts for 25% of tuberculosis cases in Europe, as well as central America and the Caribbean, where research suggests its presence may be correlated with post-Columbus European colonisation (Cubillos-Ruiz *et al*., 2010). Its significance is based on its global dissemination and its predominance in multi-drug resistant outbreaks (Mardassi *et al*., 2005, Marais *et al*., 2006, Khanipour *et al*., 2016). A study in Tunisia showed Haarlem strains cause 22-fold more multi-drug resistant cases than non-Haarlem Euro-American strains (Mardassi *et al*., 2005). Haarlem genotype strains have been found to have an increased clonal expansion rate compared to non-Haarlem strains which, it has been postulated, allows them to spread quickly whilst also being multi-drug resistant (Mardassi *et al*., 2005, Tessema *et al*., 2013).

1.3: Tuberculosis in history

Tuberculosis represents a disease that has plagued human kind throughout its recorded history for millennia. The earliest records of a tuberculosis-like disease were written in 2000-3000 BC in the Chinese Huang Ti Nei-Ching and the legal texts of Babylonian King Hammurabi (Herzog., 1998, Bynum., 2012). A consumptive disease, characteristic of tuberculosis has also been described in other famous historical documents and figures such as the Ebers Paparus of ancient Egypt (circa 1500 BC) (Cave and Demonstrator., 1939), Homer's Iliad (circa 800 BC), the documents of Hippocrates during Greek antiquity (410-400 BC) and reports by Aristotle (384-322 BC) and Galen (174 AD) (Herzog., 1998, Frith., 2014). Anatomical evidence for tuberculosis cooperative history with hominids dates back as far as *Homo erectus* in the middle Pleistocene (circa 490,000 years BC) (Hershkovitz *et al*., 2008), with molecular diagnosis of tuberculosis being reported in skeletons of an early eastern Mediterranean settlement (9000 years BC) (Hershkovitz *et al*., 2008). However, it was not until 1679 that we see the appearance of the term 'tubercles in phthisis' when Sylvius de la Boë described phthisis of the lung as tubercular granulosa disease (Frith., 2014). The term consumption then became synonymous with the disease, being used as a lay term for phthisis between the 17th and 19th centuries, until tuberculosis was used by Johann Lukas Schönlein in 1839 (Frith., 2014). Tuberculosis became a major pandemic across Europe during the 18th and 19th centuries (Dormandy., 1999), killing a quarter of the total adult population, having a mortality rate of 1000 per 100,000 per year during those centuries (Daniel., 2006). The disease severely affected the European continent, causing 1 in 6 deaths in France by 1918, and killing over 4 million people in England and Wales from 1851-1910 (Segen., 1992 Daniel., 2006). During this era, the disease was commonly known as "the robber of youth" due to one third of those between 15-34 years old being killed by tuberculosis (Bynum., 2012). In 1882, Robert Koch provided the first major breakthrough in tuberculosis research. His work identified the tubercle bacilli as responsible for the disease (Frith., 2014). The confirmation of the causative agent then led to further key developments in the diagnosis and treatment of tuberculosis, such as the Pirquet and Mantoux tuberculin skin tests in 1907; the *Bacillus* Calmette-Guérin vaccine in 1921; and eventually to Selman Waksman's Streptomycin discovery in 1943, which represented the first successful antibiotic used to treat Tuberculosis (Frith., 2014).

1.4: Current status of tuberculosis disease worldwide

In 2016, tuberculosis killed more people globally than any other infectious agent and is currently the ninth leading cause of death worldwide (WHO., 2017). In 2016, tuberculosis disease caused the deaths of 1.3 million people worldwide and contributed to a further 400,000 deaths in co-infected HIV patients (WHO., 2017). Around 10.4 million new cases were reported in 2016, with 5.9 million men suffering from tuberculosis, followed by 3.5 million women and 1.4 million children (WHO., 2017). The current tuberculosis burden is larger than earlier thought, with around 4.9 million previously 'missing' cases being documented in 2015, the majority in the South East Asia region (World Health Organisation: Jim Yong Kim., 2017). Poor surveillance is to blame for these 'missing cases', with India reporting a 34% increase in disease notification after the private health sector was included in the national tuberculosis programme there in 2015 (World Health Organisation: Jim Yong Kim., 2017). Countries such as Bangladesh, Myanmar and Indonesia also heavily under-report cases (World Health Organisation: Jim Yong Kim., 2017). Six countries account for around 60% of all new cases, namely India, China, Indonesia, Nigeria, South Africa and Pakistan (WHO., 2016). Tuberculosis is still largely a disease of poverty, with Africa harbouring the most severe burden relative to its population, with 281 cases per 100,000 people - double the global average of 133 (WHO., 2015). Global progress is largely dependent on the prevention of tuberculosis within these countries (WHO., 2016).

Europe represents the region with the lowest incidence of tuberculosis disease and much of this burden is borne by eastern European countries (European Centre for Disease Prevention and Control/WHO Regional Office for Europe., 2017). However, it is calculated that tuberculosis kills around 7 people every hour across the European Union with around

32,000 deaths and 323,000 cases per year (European Centre for Disease Prevention and Control/WHO Regional Office for Europe., 2017), equating to a rate of between 35.5 per 100,000 individuals (European Centre for Disease Prevention and Control/WHO Regional Office for Europe., 2017). The UK reported an incidence of 9.4 cases per 100 000 population in its most recent national report (Public Health England., 2017). In addition, the rate of multi-drug resistance(MDR) is higher in European cases than elsewhere in the world (European Centre for Disease Prevention and Control/WHO Regional Office for Europe., 2017). The most recent reports have suggested that there is also an increase in the number of extensively drug resistant (XDR) cases across eastern European countries, potentially making up 1 in every 4 MDR cases there (WHO., 2017¸ European Centre for Disease Prevention and Control/WHO Regional Office for Europe., 2017). The financial burden tuberculosis imposes across the European Union now stands at an estimated €5.9 billion per annum (European Centre for Disease Prevention and Control/WHO Regional Office for Europe., 2017) and treatment of multi-drug resistant strains contributes heavily to this.

1.5: Tuberculosis in Wales

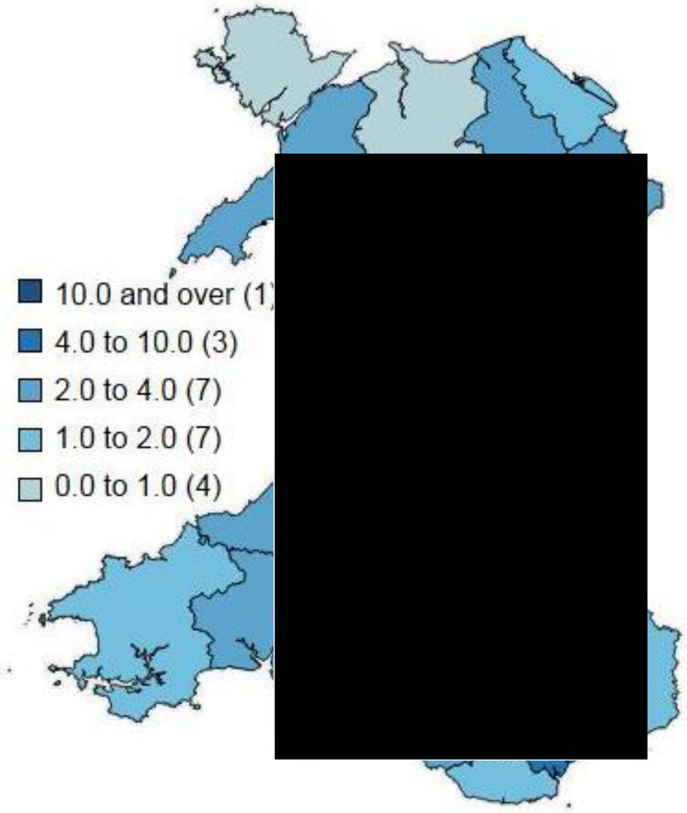
During the industrial revolution, the onset of the major coal mining and steel industries, and into the early 20th century, tuberculosis was a major health issue in Wales. Statistics from the start of the 20th century show England and Wales to have an incidence of over 1000 in every 100,000 individuals (Frith., 2014, Jackson *et al*., 2016), far greater than the rate in countries considered high-incidence today, such as South Africa (Frith, 2014, Jackson *et al*., 2016). However, *M. tuberculosis* infection was exacerbated by the living conditions during the industrialisation of the UK in the late 19th and early 20th centuries, which was a major contribution to the tuberculosis disease burden. *M. bovis* infections were also a major contributor to the tuberculosis burden in that time, with unpasteurised cow's milk being a major source of the disease, causing 65,000 deaths between 1912 and 1937 (Wilson., 1943). In 1913, of all the counties in the UK, the five with the highest incidence of tuberculosis were in Wales (Michael., 2008). Tuberculosis was so common in Wales that it was considered part of someone's genetic profile and thought to be an inherited disorder (Michael., 2008). Today tuberculosis incidence in Wales is low, with around 116 to 200 cases per year over the past decade (Public Health England., 2014, Public Health Wales., 2014, Public Health Wales., 2016). This correlates to an incidence of between 3.7-4.5 cases per 100,000 population (Public Health Wales., 2014, Public Health Wales., 2016), compared to 9.4 per 100 000 for the UK as a whole in the most recent UK-wide report (Public Health England., 2017). Although the overall incidence of tuberculosis in Wales is lower than that seen in England and the UK as a whole, the overall UK rate masks major differences between regions as London has about 40% of all UK cases (Public Health England., 2016). In addition, in terms of deaths, the 2014 UK-wide report noted that Wales was the region with the highest proportion of deaths in drug-sensitive cases, at 11.7% (16 out of 137 cases in 2012) (Public Health England., 2016). The incidence of tuberculosis in an area is often heavily affected by immigration from high TB-burden countries (Public Health England., 2016). In areas of high immigrant populations, the level of incidence of tuberculosis is higher than in areas with lower immigrant populations (Public Health England., 2016). The Public Health England 2016 tuberculosis report stated, that in England the rate of tuberculosis was 15 times higher in non-UK born individuals, and in 2015 made up 73% of all cases in England (Public Health England., 2016).

In Wales, the structure of the NHS Trust responsible for health protection, Public Health Wales, links microbiology, public health and epidemiology into one organizational team, with an active tuberculosis programme group which meets regularly. Wales is uniquely placed to provide a model for other such programmes, with an ideal setting for studying the disease. The relatively stable population enables detailed analysis of links between cases, which is not possible with more transient populations and fractured services. Thus, it is possible to obtain very detailed epidemiological information about each case of tuberculosis. Despite the stable population and low incidence of tuberculosis, outbreaks still occur in Wales and since 2005; the number has risen in the South West region of Wales.

1.5.1: Tuberculosis in Wales – Overview of demography.

The incidence of tuberculosis is heavily affected by international immigration from countries with a high tuberculosis-burden (Public Health England., 2016). Wales has a relatively stable population, with international migration fluctuating between small net influxes and effluxes since the early 1990`s (Statistics Bulletin., 2013). In 2016, 50% of all tuberculosis cases in Wales occurred in people born outside the United Kingdom (UK), with UK born individuals accounting for 44% of reported cases and 6% of cases being of unknown birth place (Public Health Wales., 2017). Thus, a substantially lower proportion of cases are born outside the UK than in England where the figure is 74% (Tuberculosis in England.,2017, Public Health England report., 2018). Most cases known to be born outside the UK originated from South Asia and Sub-Saharan Africa (Public Health Wales., 2017). A total of 80 cases were culture-confirmed in 2016, and of these 78 were due to infection by *M.tuberculosis,* one was caused by *M. bovis* and one by *M. africanum.* Two cases (2.5%) of MDR-TB were reported among culture-confirmed cases in 2016, higher than the English proportion of 1.5% phenotypically tested and confirmed MDR-TB in the same year, although the Welsh numbers are small, meaning proportions should be interpreted with caution (Public Health Wales., 2017). The rate of any resistance to one or more first line drug was 5% - again absolute numbers are small. No extensively drug resistant cases were reported in Wales in 2016 (Public Health Wales., 2017).

Within Wales, the highest rates of tuberculosis are found in the South-east, in Cardiff and Newport, which have incidence rates of tuberculosis of 15.4 per 100,000 population and 6.6 per 100,000 population respectively (Public Health Wales., 2017), see Figure 1.1. Tuberculosis incidence within Wales also correlates with poverty, as would be expected. The rate of tuberculosis in those living in the most deprived fifth Wales was 7.0 per 100,000 population, whilst in the wealthiest fifth of the country there was a rate of only 1.1 per 100,000 population (Public Health Wales., 2017). Like the rest of the United Kingdom, tuberculosis in Wales is associated with social risk factors such as drug abuse, alcohol abuse, homelessness or imprisonment with these risk factors having more of an impact on the tuberculosis demographic in Wales than is seen in England. Around 14% of all cases in 2016 reported at least one risk factor, higher than the figure for England (11.1% of cases) (Public Health Wales., 2017, Public Health England., 2017). Seven percent of all cases reported a history of or current drug abuse, 4% reported a history of or current abuse of alcohol, 7% were reported in the homeless population and 9% had a history of or were currently in prison (Public Health Wales., 2017). The Welsh cohort had increased proportion of cases with the social risk factors of drug abuse, homelessness and imprisonment in comparison with England, where these social risk factors accounted for 4.2%, 4.5% and 4% respectively.



[redacted]

Figure 1.1: A Geographical map extracted from the Public Health Wales Annual tuberculosis report for 2016 (Public Health Wales., 2017). The map is split up into the different counties of Wales and the key depicts the colours relevant to the rate of tuberculosis per 100,000 individuals.

1.6: Pathogenesis of tuberculosis

M. tuberculosis is transmitted to individuals through inhalation of droplet nuclei. Following inhalation these lodge within the terminal air spaces of the lung (Glickman and Jacobs., 2001), allowing them to be ingested by alveolar macrophages (Glickman and Jacobs., 2001). Following ingestion into macrophages, the tubercle bacilli may either be destroyed by the host's immune system or survive and replicate within the alveolar macrophages. This will lead to an active infection or a latent state as a consequence of the activity of the host's immune system (Glickman and Jacobs. 2001). Only around 5-10% of individuals progress to active disease state following initial infection. The remaining 90-95% harbour the infection in a latent state (Asante-Poku *et al*., 2015). Only around 10% of individuals carrying the latent form of tuberculosis infection will go on to develop active disease (Flynn and Chan., 2001). However, in immuno-compromised individuals, such as those co-infected with HIV, the likelihood of progression to active disease is greater (Flynn and Chan, 2001). It is thought that a third of the world's population carry the latent form of the disease, which acts as a major reservoir supporting the disease`s continuous presence, making eradication difficult (Flynn and Chan, 2001).

1.7: Treatment of tuberculosis

The first treatments for tuberculosis were based on surgical methods such as artificial pneumothorax and plombage, both of which involved the physical collapse of the infected area of the lung (Kendall., 1915). The discovery of streptomycin in 1943 caused surgical procedures to become less used and was the dawn of the anti-tuberculosis drug era (Thomas E Herchline., 2017). Currently, the empiric treatment for tuberculosis recommended by the World Health Organisation involves the prescription of 4 first-line drugs: isoniazid, rifampicin, pyrazinamide and either streptomycin or ethambutol (Thomas E Herchline, 2017., WHO 2010., WHO 2017., Milburn *et al*., 2010). The recommended treatment period for drug susceptible tuberculosis is 6 months. For the last four months, rifampicin and isoniazid can be used alone if the strain is completely susceptible (Thomas E Herchline., 2017., WHO 2017). Treatment of drug-resistant strains is much more complex, toxic and expensive and is outside the scope of this thesis.

1.7.1: Tuberculosis Vaccination

A vaccine for *M. tuberculosis*, named the Bacillus Calmette–Guérin (BCG) vaccine, has existed in medical practice since 1921, has been rolled out as part of the WHO's Expanded Programme on Immunization (EPI) since 1974 and is the sole vaccine available to prevent tuberculosis. The vaccine is a live-attenuated *M. bovis* strain which has lost its ability to confer disease in humans but still confers immunity to *M. tuberculosis* albeit with varying efficacy across different populations. The BCG vaccine has proven successful in preventing TB meningitis in children, having an efficacy of 70-80% of preventing the disease (Colditz *et al*., 1994, Rodrigues *et al*., 20 11). However, the efficacy of the vaccine for preventing adult pulmonary tuberculosis ranges between 0-80% (Colditz *et al*., 1994, Rodrigues *et al*., 20 11). Within the UK its efficacy is estimated to be between 60-80% in protecting against pulmonary tuberculosis disease (Colditz *et al*., 1994). In countries closer to the equator the efficacy of the BCG vaccine has been reported as 0% (Rodrigues *et al*., 2011). Due to its variable effect in preventing adult pulmonary disease, efforts have been made to develop novel vaccines. However, as of 2018 no such breakthroughs have been made. The need to develop new tuberculosis vaccines has been reported by the WHO as an essential part of the end TB strategy (WHO, 2017).

1.8: Control measures and strain typing

The spread of multi-drug resistant and extensively drug resistant strains, along with the lack of novel vaccines and anti-tuberculosis drugs means prevention through the application of successful control programmes is critical to public health globally. Early diagnosis and epidemiological investigation through systematic screening of contact and high-risk groups has become one of the key pillars of the end-tuberculosis strategy developed by the World Health Organisation, which aims to end the global tuberculosis epidemic by 2050 (WHO., 2015).

1.8.1: Tuberculosis control

The strategy for tuberculosis control in the community is based primarily on passive and active case-finding, contact tracing and molecular strain typing techniques. The contribution of both traditional contact tracing and molecular strain typing allows epidemiological investigations to identify the most appropriate control programmes. Passive case-finding relies on infected individuals noticing symptoms of disease and report themselves to local practitioners, meaning they will not be documented until they are already symptomatically ill (Zachariah, R *et al* ., 2003, Den Boon, S *et al* ., 2008). In contrast, active case-finding involves healthcare workers actively searching for cases to get patients treated as quickly as possible. However, active case-finding is more expensive and requires skilled healthcare workers and infrastructure, features that are barriers in developing countries (Zachariah, R *et al* ., 2003, Den Boon, S *et al* ., 2008). In the UK, both passive and active case-finding are practiced. The use of mobile radiography units successfully led to a decrease in the incidence of tuberculosis in the UK between 1940-1960 (Levitt., 2003, Watson *et al*., 2007) and the units are still useful to target hard-to-reach populations such as the homeless (Story *et al*., 2012, Watson *et al*., 2007). Pre-entry screening of migrants from countries with a high incidence of tuberculosis (> 40 cases per 100,000 population) was credited for a decrease in active pulmonary tuberculosis among non-UK born individuals (Public Health England, 2014), and is a recently used form of active case-finding.

1.8.2: Strain typing

For effective control measures, not only within an outbreak, accurate strain typing is vital (Small *et al*., 1993), and is important both in mycobacterial research and clinical practice (van Embden *et al*., 1993). In research, strain typing provides information which elucidates the evolution, diversity and phylogenetic features of different *M. tuberculosis* strains (van Embden *et al*., 1993). In clinical practice it is important in epidemiological investigations as it provides accurate information on the extent and nature of the outbreak, and for applying the most suitable control measures (Crampin *et al*., 2010, Witney *et al*., 2016). Clinically it is also important for detecting laboratory cross-contamination events, which occur occasionally in laboratories handling isolates of *M. tuberculosis,* and for distinguishing re-infection from a relapse of a previous TB infection (Crampin *et al*., 2010, Small *et al*., 1993).

Strain typing can be achieved through several different techniques. Early approaches used phenotypic typing such as phage typing and drug susceptibility profiling (Schürch and van Soolingen, 2012, Schuitemaker, 1968). These methods were low in discriminatory power although useful in the days before molecular methods were available (Schuitemaker, 1968). Phage typing is still reported to be useful in detecting laboratory cross-contamination (Schürch and van Soolingen, 2012). The advent of molecular strain typing techniques such as spacer oligonucleotide typing (spoligotyping), IS6110 Restriction fragment length polymorphism typing (IS6110 RFLP) and Mycobacterial Interspersed Repetitive Unit Variable Number Tandem Repeat typing (MIRU-VNTR) greatly increased the discrimination of strain typing and substantially increased the success of control programmes due to increased accuracy of outbreak investigations (Table 1.2); (Schürch and van Soolingen., 2012, Small *et al*., 1993).

Table 1.2: List of the established *M. tuberculosis* typing methods, with corresponding comments relating to the features, advantages and disadvantages of each method.

Typing method	Comments
Spoligotyping	Older method which has been modernised. Still useful as a secondary discrimination method to be used with other methods such as MIRU-VNTR for strain confirmation. Reproducible, rapid and allows significant phylogenetic and evolutionary data to be gleaned digitally which MIRU-VNTR and RFLP do not.
IS6110 RFLP	A Very high discrimination level, has long been described as the gold standard for tuberculosis strain typing. Slow, laborious, difficult to modernise and share results between independent labs accurately. Could be becoming redundant.
MIRU-VNTR	Far more rapid than IS6110 as it is PCR-based and can be done directly on the sample. More effective in rapidly acting on outbreaks. Using 16-24 loci MIRU-VNTR discrimination equals RFLP. Better discrimination than RFLP in strains with low IS6110 copy numbers.
WGS	Provides the highest possible level of resolution. Provides evolutionary information and data which cannot be achieved by other methods. More phylogenetic power and more discriminatory than all other methods. Currently difficult to standardize.

1.8.3: Spoligotyping

Established in the 1990s, spoligotyping is a PCR-based reverse-hybridisation blotting method, for genotyping strains of *M. tuberculosis* (Kamerbeek *et al*., 1997, March *et al*., 1996). It allows rapid amplification and analysis of strain-dependant polymorphisms found in the spacer regions of a direct repeat locus (DR-locus) contained in the MTBC genome. Spoligotyping has been used to assay the genetic diversity of this locus for use in the clinical laboratory, evolutionary and population genetics (Brudey *et al*., 2006). Because it is PCR-based and requires few copy numbers, spoligotyping allows for simultaneous detection and strain typing of *M. tuberculosis* from a clinical sample containing the bacterium within one day (Augustynowicz-Kopec *et al*., 2007). The main advantages over the other methods listed in Table 1.2, are its speed and its easily presented results in a digital format that can then be used to create a database (Brudey *et al*., 2006), thus making communication between independent laboratories practical. In addition, spoligotyping is well-established in assigning individual *M. tuberculosis* strains to phylogeographic clades that are submitted to globally accessible databases such as SpolDB4 (Brudey *et al*., 2006). The DR locus is phylogenetically informative and spoligotyping can distinguish between lineages and sub-lineages within the *M. tuberculosis* genome (Abadia *et al*., 2011, Filliol *et al*., 2006). However, in comparison to other molecular typing procedures, spoligotyping has a lower resolution and underestimates the clonal diversity of *M. tuberculosis* when used alone (Augustynowicz-Kopec *et al*., 2007).

Despite being deemed unreliable when used as a single molecular typing tool, studies have highlighted that traditional spoligotyping can be used to support the primary typing methods such as IS6110 RFLP and MIRU-VNTR. Most recently, spoligotyping has been incorporated into the use of data gleaned from WGS of *M. tuberculosis* strains. Spoligotype patterns can now be identified *in silico*, meaning spoligotype data can be obtained from sequence data following WGS, thus done at no extra cost allowing them to be grouped rapidly into phylogenetically informative spoligotype clades without the need to carry out a separate procedure (Xia *et al*., 2016).

1.8.4: IS6110 RFLP

IS6110 restriction fragment length polymorphism (RFLP) is a genotyping technique, which has long been regarded as the gold standard for *M. tuberculosis* complex genotyping, because it has been the method with the highest discriminatory power (Christianson *et al*., 2010). IS6110 is an insertion sequence in the *M. tuberculosis* chromosome belonging to the IS3 insertion sequence family of *Enterobacter* and is present in all MTBC species but not in the more distantly related mycobacteria (Thierry, D *et al*, 1990). This insertion element can range in copy number between strains from zero up to 25 copies. They vary in their sites of insertion, and this feature is essential in the ability of the method to discriminate between strains of

M. tuberculosis. RFLP involves digesting these insertion elements using restriction enzymes and visualising the resulting fragments by probe hybridisation. The pattern of fragments defines the different strains (Park *et al*., 2000). IS6110 RFLP is highly discriminatory and, as stated, is the gold standard in differentiating strains, such as in outbreak investigations. However, the use of IS6110 in phylogenetic and evolutionary studies is somewhat limited, because the insertion elements are restricted to two thirds of the genome around the DR locus. Also, within the genome there are IS6110 hotspots in the DR and the *ipl* loci. These factors have led some authors to question the method`s utility as a reliable marker for population genetics and phylogenetics studies of *M. tuberculosis* (Fang *et al*., 1998).

IS6110 RFLP has other important limitations. It is labour intensive, time consuming and more expensive than a PCR based method such as spoligotyping (Christianson *et al*., 2010). It is also a gel-based method and, as such, is susceptible to interpretive errors (Christianson *et al*., 2010). There is no PCR amplification step and thus it requires prior culture to provide sufficient biomass to get enough DNA for testing (Christianson *et al*., 2010). This makes it a slow process and a more rapid PCR-based method has often been used initially, using IS6110 RFLP as a confirmatory test (Ruddy *et al*., 2004). There is also the problem of comparing IS6110 patterns between independent labs or even ones that are separated in time due to differing experimental conditions, operator interpretation of the result, and also changes in the software used to analyse the patterns (Das *et al*., 2005). Therefore, despite its discriminatory ability, the results are not as easy to use, as the digitalised ones obtained from spoligotyping and MIRU-VNTR, which allow for continuously updated data to be distributed as standards. Thus as discriminatory as it is, this method is not the answer to rapid response to control outbreaks; rather its power is in providing confirmation of strain types (Das *et al*., 2005).

1.8.5: VNTR and MIRU-VNTR

Variable Number of Tandem Repeat (VNTR) typing is a PCR-based technique (Supply *et al*., 2006, Supply *et al*., 1997) which was originally devised as an epidemiological typing method by Frothingham and Meeker-O`Connell in 1998, and involved exact tandem repeat sequences

(ETRs) (Supply *et al*., 2006) located at five different chromosomal loci: ETR-A, ETR-B, ETR-C, ETR-D, ETR-E and ETR-F (Frothingham and Meeker-O'Connell, 1998, Supply *et al*., 2006). These loci show substantial length polymorphism ranging from 53-79 base pairs and differ in the regions of DNA they reside in (Supply *et al*., 2006). The genotype of an isolate is represented by a five-digit allele profile specifying the number of repeat elements at each of these loci. The allele profile is determined by the size of the amplified products. The variations within these sequences can affect expression of certain genes within strains providing further features unique to that strain (Frothingham and Meeker-O'Connell, 1998b). Loci used in VNTR strain typing may also be Major Polymorphic Tandem Repeat (MPTR) ones, which are different from ETRs as they are found in multiple genomic clusters and are imperfectly repeated units. ETRs are found at a unique locus and have a unique spacer sequence (Hermans *et al*., 1992), whilst the MPTR loci consist of 15 base-pair repeats with a single consensus sequence. However, there is significant sequence variability between the adjacent repeats making them useful as markers to determine both species and strains within the *M. tuberculosis* complex.

1.8.6: MIRU-VNTR

Since the release of the whole genome sequence of the H37Rv strain *of M. tuberculosis,* 41 MIRU-VNTR loci have been identified, consisting of both ETRs and MPTRs. These together are referred to as Mycobacterial Interspersed Repetitive Units (MIRUs), now known as MIRU-VNTR (Supply *et al*., 2006). The MIRU sequences contain between 40 - 100 base-pairs and are scattered randomly throughout the chromosome of the *M. tuberculosis* genome (Supply *et al*., 2006). MIRU-VNTR is a PCR based method which amplifies the repeated sequences of the MIRU-VNTR defined loci. The size of the PCR product correlates with the number of repeats present, and the pattern of these results allows a numerical genotype to be assigned for each isolate (Van Soolingen *et al*., 2001). An original set of 12 MIRU-VNTR loci, representing 12 mini-satellite-like regions of the *M. tuberculosis* chromosome, was described in a 2001 study by Mazars *et al* and provided a level of resolution similar to that of the established IS6110 RFLP typing (Mazars *et al*., 2001). The original 12 loci proposed were then expanded to 15 and have now been further supplemented to a 24-locus test, with a discriminatory power comparable to that of IS6110 (Supply *et al*., 2006, Christianson *et al*., 2010). Due to it being PCR-based, the MIRU-VNTR method is rapid and although not quite as discriminatory, the application of 15 and 24 loci MIRU-VNTR typing in conjunction with supporting spoligotyping data has shown discrimination close to that of IS6110 RFLP (Christianson *et al*., 2010, Jonsson *et al*., 2014). In fact, MIRU-VNTR may have higher discrimination when the IS6110 copies within a given isolate are below 6 (Jonsson *et al*., 2014). The MIRU-VNTR method allows easy construction of a digital global database allowing for efficient epidemiological studies of tuberculosis worldwide, with the MIRU-VNTR plus online database being a widely-used platform (Weniger *et al*., 2010). MIRU-VNTR digitalised platforms have been adopted in combination with digitalised spoligotyping data, providing an international basis for large-scale genotyping at a high-throughput and globally accessible level (Weniger *et al*., 2010). For these reasons MIRU-VNTR replaced IS6110 in the first decade of this century as the international standard for *M. tuberculosis* molecular typing, with IS6110 reserved as a confirmatory procedure. Thus, MIRU-VNTR allows large scale global surveillance and large-population based analysis due to its speed, simplicity and ease of comparison between laboratories worldwide.

1.8.7: Limitations of MIRU-VNTR typing

Various studies have highlighted the shortcomings of MIRU-VNTR typing in resolving outbreaks and sometimes contradicted contact tracing data (Gardy *et al*., 2011, Walker *et al*., 2013a, Walker *et al*., 2013b, Kohl *et al*., 2014, Walker *et al.*, 2015, Takiff and Feo., 2015, Witney *et al*., 2016). MIRU-VNTR typing does not account for the heterogeneity of isolates included within the same outbreak that is caused by microevolution (Walker *et al*., 2013). A previous study highlighted that *M. tuberculosis* can harbour a level of within-host heterogeneity that rivals the variation seen between patients within a given outbreak (Pérez-Lago *et al*., 2013). This heterogeneity is caused by the microevolution of the mycobacteria within the host over time (Pérez-Lago *et al*., 2011, Pérez-Lago *et al*., 2013). MIRU-VNTR, along with IS6110 and spoligotyping, is unable to account for this microevolution and thus provides inaccuracies when used to trace the route of transmission within an outbreak, or determining whether isolates are part of an outbreak, when they may harbour minor differences (Takiff and Feo, 2015). In addition, MIRU-VNTR does not have the ability to identify super-spreaders within an outbreak. Super-spreaders are cases within an outbreak that can disproportionally infect secondary cases. The presence of super-spreaders presents public health challenges and may dramatically affect control programmes related to the outbreak (Stein, 2011). The identification of super-spreaders in other diseases has shown to be beneficial in reducing the number of future cases through halting transmission from the super-spreader (Stein, 2011). The use of MIRU-VNTR typing also provides no information for the identification of the source case of outbreaks. Key information such as potential super-spreaders and source cases are dependent on contact tracing investigation with no molecular typing support.

1.9: Whole Genome Sequencing

The advent of high-throughput and affordable WGS, whereby the complete DNA of an organism is sequenced and analysed, presents an opportunity to provide the ultimate level of molecular resolution to infectious disease outbreaks (Le and Diep, 2013). Through WGS it is possible to identify genetic markers, referred to as single nucleotide polymorphisms (SNPs), that can be used to distinguish two separate isolates from each other, with the number of polymorphisms indicating how distantly related to each other given isolates are. Certain SNPs are as associated with, drug resistance, increased transmissibility and immune evasion, thus have the potential to be used as markers to aid treatment optimisation and the understanding of an outbreak’s physiology (Takiff and Feo., 2015, Walker *et al* 2015). Through analysis of drug resistance-associated SNPs within certain pathogens it has been possible to develop assays to aid treatment programmes. An example is the GeneXpert© diagnostic assay for identifying rifampicin resistance in *M. tuberculosis* (Rie *et al*., 2010). The test allows identification of *M. tuberculosis* and the presence of rifampicin resistance within 90 minutes (Rie *et al*., 2010).

Through sequencing the genome of organisms within an outbreak, it is possible to understand the relationship between the isolates using these SNPs as genetic markers (Takiff and Feo, 2015).

Previous studies have highlighted that, when used in conjunction with traditional epidemiological investigation, WGS allows elucidation of transmission dynamics, identification of super-spreaders and source cases in outbreaks with various pathogens. Examples include outbreaks of *Escherichia coli* and *Vibrio cholerae* (Rohde *et al*., 2011, Chin *et al*., 2011). WGS also provided the first conclusive evidence for person-to-person transmission of *M. abscessus,* as the study showed that two separate outbreaks were caused by the same resistant clone transmitted between cystic fibrosis sufferers (Le and Diep, 2013).

1.9.1: Whole Genome Sequencing of *M. tuberculosis*

The commonly used laboratory strain H37Rv was the first MTBC genome to be sequenced completely and published in 1998 (Cole *et al*., 1998). This was a milestone in tuberculosis research that provided unprecedented information on the biology, metabolism and evolution of the pathogen. The sequencing of this first strain paved the way for the extensive search for SNPs within the genomes of all MTBC strains.

1.9.2: Genome plasticity

Sequencing revealed that the *M. tuberculosis* genome was 4.6 million base pairs in size, has around 4000 genes and has a high G+C content which is consistent across the genome (Cole *et al*., 1998). The genome encompasses a circular chromosome, whereby >90% of the genomes coding capacity is used (Cole *et al*., 1998). *M. tuberculosis* is rich in repetitive DNA especially insertion sequences, duplicated housekeeping genes and polymorphic multi-gene families (Cole *et al*., 1998). Initial comparative sequencing studies demonstrated that the sequence diversity of *M. tuberculosis* is much lower than that seen for many other bacteria such as the *Streptococcus pneumoniae* PMEN1 or *E. coli* O157:H7 strains (Musser *et al.*, 2000, Mellman, *et al*., 2011, Koser *et al.*, 2011). Population genetic studies have shown that *M. tuberculosis* has a highly clonal population structure (Hirsh *et al*., 2004, Supply *et al*., 2003, Feil and Spratt, 2001). WGS of other MTBC members revealed that MTBC is highly conserved with only 2,437 SNPs separating *M. tuberculosis* strain H37Rv from the *M. bovis* strain AF2122/97. Horizontal gene transfer has not been documented in *M. tuberculosis*, which again contributes to its low genomic diversity as a species (Gagneux and Small, 2007). Genome plasticity is also relevant to acquisition of drug resistance within *M. tuberculosis* strains as unlike some other pathogenic bacteria; *M. tuberculosis* does not contain plasmids, which often confer resistance in Gram negative organisms (Trauner *et al*., 2014). Rather, the evolution of resistant strains is driven by sequential acquisition and accumulation of random point mutations on the chromosome directly (Trauner *et al*., 2014) which are subject to selection pressure in the presence of anti-tuberculous chemotherapy.

The monomorphic nature of *M. tuberculosis* populations means standard multi-locus sequencing of housekeeping genes has no value in providing useful discrimination of strains within an outbreak (Jolley and Maiden, 2010). Thus, IS6110 RFLP, spoligotyping and MIRU-VNTR which exploit polymorphic regions within the genome, have been preferred to assess strain diversity. However, further analysis has shown that sequence diversity does exist within *M. tuberculosis* and has highlighted the utility of large sequence polymorphisms (LSP) as well as SNPs for stable genetic markers in strain typing (Gagneux and Small., 2007).

1.9.3: WGS for *M. tuberculosis* outbreak resolution

Multiple studies have shown that WGS provides a level of outbreak resolution above and beyond that provided by other molecular typing methods, such as MIRU-VNTR (Comas *et al*., 2009, Gardy *et al*., 2011, Kohl *et al*., 2014, Walker *et al*., 2013a, Walker *et al*., 2013b, Witney *et al*., 2016). Studies have confirmed, through novel WGS (Walker *et al*., 2013b, Pérez-Lago *et al*., 2013), that intra-patient micro-evolution can affect the validity of both IS6110 and MIRU-VNTR (Al-Hajoj *et al*., 2010), and impact on the detection of transmission events within an outbreak (Pérez-Lago *et al*., 2011, Walker *et al*., 2013b). The use of WGS for outbreak resolution accounts for micro-evolution and allows the tracing of SNPs within an outbreak (Witney *et al*., 2016). When used with epidemiological data, WGS allows for the identification and confirmation of previously suspected super-spreaders. As *M. tuberculosis* reproduces by binary fission, if a super-spreader is present then WGS data produces a star-like phylogenetic topology (Walker *et al*., 2013b).

1.9.4: Limitations of traditional WGS methods

Parallel WGS of multiple bacterial genomes now provides the ability to produce epidemiologically useful data at a relatively low cost (Jolley and Maiden, 2010). Many previous studies are based on WGS SNP mapping of multiple isolates, using often in-house un-standardised pipelines (Walker *et al*., 2013b, Gardy *et al*., 2011). The portability and standardisation of WGS SNP mapping analysis is one of the main barriers to its widespread use. However, within tuberculosis research, the results of these SNP mapping based studies have allowed for the estimation of a stable mutation rate for *M. tuberculosis*, and led to the development of a 12 SNP threshold for isolate separation (Walker *et al*., 2013b). In other words, it is considered that differences exceeding 12 SNPs between isolates, indicates that they are not directly related. SNP mapping parameters have led to the development of online databases, such as the Centre for Genomic Epidemiology (Kaas *et al*., 2014) which provides an openly accessible and standardised pipeline for SNP mapping of selected isolates (Kaas *et al*., 2014). However, SNP mapping is still hampered by the need for a high-quality reference sequence and complete sequence data. It cannot be used with partial sequence data, which is often outputted using current sequencing platforms (Sheppard *et al*., 2012, Maiden *et al*., 2013).

1.9.5: Application of gene-by-gene WGS multi-locus sequence types (MLST)

WGS provides vast quantities of data which are difficult to standardise (Maiden *et al*., 2013, Sheppard *et al*., 2012, Jolley and Maiden, 2010). Gene-by-gene WGS MLST methods have attempted recently to provide a standardised, globally relevant and accessible databases for the application of WGS data to outbreak resolution and tuberculosis phylogeny (Jolley and Maiden, 2010, Junemann *et al*., 2013). The use of a gene-by-gene MLST method allows the comparison of assembled genomes from multiple isolates without the need for a high-quality reference genome. Gene-by-gene MLST uses alleles at selected genes as the unit of comparison rather than the whole genome nucleotide sequence used by traditional SNP mapping analysis (Maiden *et al*., 2013).

Gene-by-gene MLST methods also allow for the development of pre-defined MLST schemes. These allow selective comparison across certain sets of genes, providing information of interest such as MLST specific to core genomes, accessory genomes or ribosomes (Jünemann *et al*., 2013, Jolley *et al*., 2012). This has been shown to be useful in the resolution of a tuberculosis outbreak, where only the core genes were selected for analysis (Kohl *et al*., 2014). This allowed the exclusion of *ppe/pe* genes known to be highly polymorphic and have the ability to skew the true picture of isolate relationships within an outbreak (Jünemann *et al*., 2013, Kohl *et al*., 2014). Comparison of coding regions also allows for the exclusion of large indels which occur frequently in non-coding regions (Coll *et al*., 2014). The occurrence of small insertions and large deletions have been reported as being 5 and 17 times higher respectively in non-coding regions than in coding regions (Coll *et al*., 2014). Therefore, the inclusion of only coding regions provides resolution that is based on more stable genomic regions in comparison to WGS SNP mapping which is based on WGS analysis. WGS gene-by-gene MLST typing has been used successfully in previous investigations into *Campylobacter*, *Staphylococci*, and *Salmonella* outbreaks (Maiden *et al*., 2013, Sheppard *et al*., 2012). A previous study highlighted how the WGS gene-by-gene based method successfully resolved an outbreak of meningococcal disease at a UK university (Jolley *et al*., 2012). The approach produced three different MLST schemes (conventional, extended and ribosomal MLST). The outbreak was resolved completely. The investigation demonstrated that multiple closely-related but distinct strains were present in asymptomatic and disease states, with two strains causing disease and one responsible for the outbreak itself (Jolley *et al*., 2012). Both WGS gene-by-gene MLST and WGS SNP analysis gave comparable results in terms of phylogenetic information.

With the introduction of WGS gene-by-gene MLST-based software such as the Bacterial Isolate Genome Sequence Database (BigsDB) (Jolley and Maiden., 2010) and Ridom SeqSphere (Jünemann *et al*., 2013), the resolution of an outbreak is possible in a standardised manner, without the need for a high quality reference genome and the results can be rapidly submitted to a global database. The Ridome SeqSphere software has developed and published a core genome MLST, based on 2891 core genes found across the MTBC genome (Jünemann *et al*., 2013, Kohl *et al.*, 2014). Application of this cgMLST scheme allows for the multiple alignments, across the 2891 core genes, of *M. tuberculosis* isolates of interest and outputs the number of allelic differences between the given isolates. Such data can then be used to clarify genomic relationships between isolates, such as whether two isolates are directly related or not. This cgMLST scheme has resolved an outbreak of *M. tuberculosis* to an adequate level for epidemiological conclusion although it provided less resolution than a WGS SNP mapping method used in the same study (Kohl *et al.*, 2014).

1.10: Aims of the study

This thesis describes the analysis of WGS data from a collection of clinical isolates of *M. tuberculosis* from South West Wales, collected between 2004 and 2011. The collection includes isolates from three separate *M. tuberculosis* outbreaks, which occurred in the region during this time-period. Of interest is the fact that in some cases epidemiological conclusions were at odds with the molecular MIRU-VNTR results. Thus, the aim of this thesis was to use WGS-based methods on the isolates to aid outbreak resolution and provide a level of this that would either support or refute the MIRU-VNTR typing conclusions. Finally, the study also aimed to provide further information on the presence of potential super-spreaders and outbreak source cases which could not be achieved through use of MIRU-VNTR. The study also aims to reveal how further applications, based on WGS data, could glean additional phylogenetic and physiological information on outbreak isolates.

Chapter 2

Materials and Methods

2.1: Sample collection and epidemiological information

DNA from 80 *Mycobacterium tuberculosis* isolates collected between 2004 and 2011 was obtained from the Wales Centre for Mycobacteriology (WCM), Public Health Wales, Llandough Hospital, Cardiff. Forty of the isolates were directly associated with outbreaks from the Gorseinon, Townhill, Neath Port Talbot and Llwynhendy regions of South West Wales, according to both MIRU-VNTR typing and epidemiological investigations. The MIRU-VNTR typing had been carried out by the Public Health Wales Molecular Unit, Cardiff and based on 15 loci: ETRA, ETRB, ETRC, ETRD, ETRE, MIRU2, MIRU10, MIRU16, MIRU20, MIRU23, MIRU24, MIRU26, MIRU27, MIRU39 and MIRU40. A further 40 isolates from the same areas of South West Wales, apparently not belonging to any outbreak, were selected at random from the WCM's collection. Epidemiological information for each isolate had been obtained through face-to-face interviews with a senior public health nurse from Public Health Wales, as part of the routine management of these cases.

2.2: Sequencing and assembly

The 80 isolates were cultured using a BACTEC™ MGIT™ 960 System (Becton Dickinson Diagnostic Systems, Sparks, MD, USA), and the resulting bacterial suspension boiled at 110°C for 35 minutes and centrifuged. The boilates then underwent genomic DNA sequencing using a MiSeq benchtop sequencer (Illumina, San Diego, CA, USA). Sequencing libraries were prepared using Nextera XT library preparation kits (Version 3, Illumina) and paired-end 500 bp reads generated with the MiSeq run kit (Version 3, Illumina). The resulting paired-end reads were quality filtered with the Trimmomatic tool software (version 0.32; Anthony *et al* 2014) using a sliding window approach of 5 bases and a quality score of Q20. The contigs/genomes were assembled using SPAdes genome assembler (version 3.9.0; Bankevich *et al*., 2012). The SRA sequences for the 179 lineage-defined isolates were also assembled using the SPAdes genomic assembler.

2.3: Core genome MLST analysis

To investigate the genetic relationship between the sequenced isolates, all assembled genomes were uploaded onto the Ridom SeqSphere software version 4.1.9 (Ridom; Münster, Germany) and core genome MLST (cgMLST) analysis carried out for each outbreak.

2.3.1: Analysis using the established Ridom SeqSphere cgMLST scheme

Each isolate sequence was aligned to the Ridom SeqSphere *M. tuberculosis* cgMLST scheme of 2891 core genes, which is based on the H37Rv reference genome (GenBank accession number NC_000962.3; Cole *et al*., 1998), and is previously defined for alignment and subsequent genomic analysis (Kohl *et al*., 2014). Successful alignments to the cgMLST were defined as "good targets" by the Ridom SeqSphere software, and full cgMLST analysis was carried out on isolate sequences that conferred >90% "good targets". Inclusion as outbreak associated isolates was based on a threshold of 12 allelic differences defined by Kohl *et al*. (2014).

2.3.2: Analysis using an in-house cgMLST scheme

An in-house cgMLST scheme was developed using the "cgMLST target definer" application within Ridom SeqSphere (Ridom; Münster, Germany). Firstly, the genome sequences of the relevant isolates were aligned to the H37Rv reference genome and only genes present across all isolates were included in the in-house cgMLST analysis.

2.3.3 Extended 3646 gene-by-gene MLST

Ridom SeqSphere software also has an accessory genome MLST scheme consisting of 755 *M. tuberculosis* genes of an accessory nature (Kohl *et al*., 2014, Jünemann *et al*., 2013). For the extended gene-by-gene MLST analysis, a 3646 gene scheme was developed, that consisted of the 2891 core genes (section 2.4.1) and an additional 755 accessory genes.

2.4: Ancestral dating

Single nucleotide polymorphisms across selected isolates were extracted using the "extract SNVs (Single nucleotide variants) from target groups" application within the Ridom SeqSphere software. Single nucleotide variant (SNV) is a synonym for Single nucleotide polymorphism (SNP). The default setting for SNP extraction was applied with indels and SNPs within 10 bps of each other being removed from the analysis. Extracted SNPs were then concatenated using ClustalX (version 2; Larkin *et al*, 2007). To obtain ancestral dating results, concatenated sequences were then submitted to the BEAST ancestral dating software (version 1.8.2) under default parameters (Drummond *et al*., 2012, Ansari and Didelot, 2016), and a strict molecular clock of 0.5 mutations per genome per year was used as this is the mutation rate previously described for *M. tuberculosis* (Walker *et al* , 2013).

2.5: Whole Genome Sequence Single Nucleotide Polymorphism Mapping

For single nucleotide polymorphism (SNP) mapping, the whole genome sequences (WGS) of isolates relevant to each analysis were uploaded to the conserved signature indels (CSI) phylogeny web-server (version 1.3; https://cge.cbs.dtu.dk/services/CSIPhylogeny/). The SNP mapping parameters were set to the default setting of: minimum depth at SNP positions: 10x, minimum relative depth at SNP positions: 10%, minimum distance between SNPs: 10 bp, minimum SNP quality: 30, minimum read: 25, mapping quality minimum Z-score: 1.96 (Kaas *et al*., 2014). The threshold for direct transmission in WGS SNP mapping was defined as 12 SNPs (Walker *et al*, 2013).

2.5.1: SNP bar-coding

Firstly, isolates were aligned to the H37Rv reference genome using Burrows-Wheeler Alignment (BWA, version 0.7.17;Li and Durbin, 2009). SAMtools version 1.3.1 (Li *et al*., 2009) was then used to call SNPs from each of the 60 designated loci previously described by Coll *et al* (2014), with the omission of the two *M. bovis* loci at H37Rv reference genome positions 1882180 and 2831482. Isolates were then classified based on the pattern of SNPs (SNP barcode) at the designated loci using the previously defined phylogenetic classifications listed in table 2.1 below (Coll *et al*., 2014). SNP bar-coding classification provided major lineage, sub-lineage and spoligotyping family phylogenetic assignments.

Table 2.1: A table extracted from the Coll *et al* (2014) publication which highlights the correlating phylogenetic lineage and sub lineage to specific SNP`s at designated loci (Coll *et al*., 2014).

Reference allele/Polymorphism	Genome position*	Phylogenetic assignment
G/A	615938	Indo-Oceanic
G/A	4404247	Indo-Oceanic
G/A	3021283	Indo-Oceanic
G/A	3216553	Indo-Oceanic
G/A	2622402	Indo-Oceanic
G/A	1491275	Indo-Oceanic
C/A	3479545	Indo-Oceanic
C/T	3470377	Indo-Oceanic
G/A	497491	East-Asian
C/T	1881090	East-Asian (non-Beijing)
G/A	2505085	East-Asian (Beijing)
C/T	797736	East-Asian
C/T	4248115	East-Asian
G/A	3836274	East-Asian
G/T	346693	East-Asian
C/A	3273107	East-African-Indian
G/A	1084911	East-African-Indian
G/C	3722702	East-African-Indian
C/G	1237818	East-African-Indian
G/A	2874344	East-African-Indian
T/C**	931123	Euro-American
G/A	62657	Euro-American (X-type/Haarlem)
C/T	514245	Euro-American (X-type)
C/T	1850119	Euro-American (X-type)
T/G	541048	Euro-American (X-type)

C/T	4229087	Euro-American (X-type)
A/G	891756	Euro-American (Haarlem)
C/T	107794	Euro-American (Haarlem)
G/C	2411730	Euro-American
A/C	783601	Euro-American (Ural)
C/A	1487796	Euro-American
T/C	1455780	Euro-American (TUR)
C/G	764995	Euro-American (LAM)
C/A	615614	Euro-American (LAM)
G/A	4316114	Euro-American (LAM)
C/G	3388166	Euro-American (LAM)
G/A	403364	Euro-American (LAM)
G/A	3977226	Euro-American (LAM)
G/A	4398141	Euro-American (LAM)
C/T	1132368	Euro-American (LAM)
C/A	1502120	Euro-American (LAM)
G/A	4307886	Euro-American
G/A	4151558	Euro-American
G/A	355181	Euro-American (S-type)
G/C	2694560	Euro-American
G/A	4246508	Euro-American
G/T	1719757	Euro-American
G/A	3466426	Euro-American
G/C	4260268	Euro-American (Uganda)
G/A	874787	Euro-American
G/C	1501468	Euro-American
G/C	4125058	Euro-American
C/G	3570528	Euro-American
C/T	2875883	Euro-American (Cameroon)
C/G	4249732	Euro-American (mainly T)
G/A	3836739	Euro-American (mainly T)
G/T**	1759252	Euro-American (H37Rv-like)
C/A	1799921	West-Africa 1
C/G	1816587	West-Africa 2
G/A	1137518	Lineage 7
** In these two cases the reference allele is the one specific to the lineage		
* based on reference NC_000962.3		

2.5.2: Extraction of individual SNPs

Individual SNPs within the 2891 core genes from the isolate cases in question were extracted using the “extract single nucleotide variants (SNVs) from target groups” application in Ridom SeqSphere (Junemann *et al.*, 2013). The coding regions corresponding to the location of each SNP were extracted using the Ridom SeqSphere software, and BioEdit version 7.0.5 (Hall, 1999) was used to manually identify whether the identified polymorphisms harboured an amino acid change, thus being non-synonymous, or not (synonymous).

2.5.3: Functional analysis of non-synonymous SNPs

Non-synonymous SNPs were extracted using the method described in section 2.5.2, with indels included in the extraction and synonymous SNPs ignored. The amino acid sequences for each gene with details on each polymorphism (codon position and amino acid alteration), were submitted to the Provean protein functional prediction online tool (version 1.1, http://provean.jcvi.org/seq_submit.php), with deleterious predictions defined as values of -2.5 or less (Choi and Chan, 2015), thus identifying if any changes in functionality had occurred due to the non-synonymous SNP.

2.6: *In silico* spoligotyping

Each isolate sequence was submitted to the Python based SpolTyping (version 2.0) *in silico* software for prediction of spoligotype pattern (Xia *et al.*, 2016). Resulting octal and binary patterns were then submitted to the SitVit database (http://www.pasteur-guadeloupe.fr:8081/SITVITDemo/trouverSouchesParSpoligo.jsp) for determination of international typing assignment and assignment to globally recognised spoligtoype clades (Demay, C *et al.*, 2012, Yeboah-manu *et al.*, 2011, Yeboah-manu *et al.*, 2016). This allows for correlation of in-house spoligotyping results with strains of *M. tuberculosis* strains worldwide in a standardized manner (Demay, C *et al*, 2012).

Additionally, spoligotype patterns were submitted to the TB-lineage online tool online (http://tbinsight.cs.rpi.edu/run_tb_lineage.html) which assigns isolates to recognized international phylogenetic lineages based on their spoligotype octal or binary sequences obtained from the SpolTyping software mentioned above (Shabbeer *et al.*, 2012). The TB-lineage online tool employed the Conformal Bayesian Network (CBN) parameters which employ a hierarchical Bayesian network based on PCR-based spoligotype biomarkers, to assign isolates into phylogenetic lineages (Aminian, M *et al*., 2014).

2.7: Principal Genetic Grouping (PGG)

Gene sequences for *gyr*A and *kat*G were extracted from the WGS of each isolate using Ridom SeqSphere and analysed manually using BioEdit V7.0.5 (Hall, 1999) to identify the presence of PGG-defining amino acids at codons 95 and 463, PGG informative sites within genes *gyr*A and *kat*G as previously documented by Streevatsan *et al* (1997). Based on the composition of amino acids at these loci, each isolate was then assigned a PGG as defined in Table 2.2 (Grimes *et al.*, 2009).

Table 2.2: The combinations of amino acids seen at the loci *gyr*A and *kat*G and the possible PGGs (Sreevatsan *et al.*, 1997).

Loci		Principal Genetic Group
*gyr*A95	*kat*G463	
Threonine	Leucine	1
Threonine	Arginine	2
Serine	Arginine	3

2.8: SNP Cluster Grouping (SCG)

Each isolate was assigned a SNP cluster group based on the nucleotides present at nine specific loci shown in Table 2.3 (Alland *et al*., 2007). Firstly the isolates were aligned to the H37Rv reference genome using Burrows-Wheeler Alignment (BWA, version 0.7.17;Li and Durbin, 2009). SAMtools version 1.3.1 (Li *et al*., 2009) was then used to call SNPs from the 9 specific loci defined in table 2.3. Phylogenetic analysis was carried out on only those isolates with each of the nine loci present.

Table 2.3: List of the nine loci derived from Alland *et al.* (2007) used for assigning isolates into one of six SNP cluster groups.

	Base at SNP position in H37Rv:								
SCG	**1977**	**54394**	**74092**	**105139**	**144390**	**232574**	**311613**	**913274**	**2154724**
1	G	G	C	C	G	G	T	G	A
2	G	G	C	A	G	G	T	C	A
3a	G	G	C	C	G	G	T	C	A
3b	G	G	C	C	G	G	T	C	C
3c	G	G	C	C	G	T	T	C	C
4	G	G	C	C	A	T	T	C	C
5	G	A	C	C	G	G	T	C	C
6a	A	A	C	C	G	G	T	C	C
6b	A	A	C	C	G	G	G	C	C

2.9: Construction of cgMLST phylogenies for phylogenetic assignment

A set of 179 isolates assigned a lineage in the Comas *et al*. (2014) study were downloaded from NCBI (Coordinators, N.R., 2016), and assembled using SPAdes (section 2.2). The resulting 179 genomes were then analysed alongside the sequenced Welsh isolates, making a total of 236 isolates. All the isolates were then submitted to the cgMLST scheme described in section 2.3.1. The resulting phylogeny comparison was determined using Unweighted Pair Group Method with Arithmetic Mean (UPGMA) trees produced by the Ridom seqSphere software (Ridom; Münster, Germany), and further annotated and modified using iTol version 4 (Letunic and Bork, 2006).

2.10: Sub-lineage genotyping

Sub-lineage genotyping was carried out by analysis of SNPs that are specific to different phylogenetic groups, which are described in more detail below (Rad *et al.*, 2003, Mestre *et al*., 2011, Comas *et al*., 2009, Cubillos-Ruiz *et al*., 2010, Alix *et al*., 2006). SNPs were initially identified through extraction of gene sequences relevant to certain genotypes (defined in sections 2.10.1, 2.10.2, 2.10.3, 2.10.4 and 2.10.5 below) from each isolate using the sequence extraction application within Ridom SeqSphere and detected manually using BioEdit Version 7.0.5.

2.10.1: Beijing genotyping

Isolates LL9, BK21 and BK25 were assigned to the Beijing sub lineage by the SNP barcode method (described in section 2.5.2 and shown in figure 3.4). Each isolate assigned to the Beijing sub-lineage, in chapter 3 figure 3.4, was extracted from the dataset and analysed for SNPs associated with the Beijing sub-lineage which are described in a previous study (Rad *et al*., 2003) Genes *mut*T2, *mut*T4 and *ogt* were analysed for polymorphisms as follows: the *mut*T2 mutation is due to an amino acid change at codon position 58, whilst the *mut*T4 and *ogt* mutations are due to nucleotide polymorphisms at positions 142 and 36, respectively (Rad *et al*., 2003). Further Beijing genotyping was carried out through analysis of a selected barcode of 48 SNPs across 22 genes listed in table 2.4 below and defined in a previous study (Mestre *et al*., 2011). The 48 SNPs were extracted as described in section 2.10 and the resulting SNPs were concatenated using ClustalX version 2 (Larkin *et al*., 2007), and then used to produce a phylogenetic tree UPGMA using the iTol software (Letunic and Bork, 2006). A distance matrix based on the 48 SNPs was constructed using MEGA7 (version 7.0) (Kumar *et al* 2016).

Table 2.4: A table showing the 48 loci across 22 genes analysed for Beijing genotyping. The nucleotide positions defined below are relevant to the within gene position.

Gene	Nucleotide position
ligD	485
	1030
	1038
	1738
radA	456
	557
	827
recF	734
	807
recX	23
	175
	458
dnaQ	227
	263
	483
	631
recR	130
	267
recG	853
uvrC	496
	865
	1164
	1301
ruvB	843
ligB	230
	271
recD	360
	416
	831
tagA	537
	385
uvrD1	1384
dnaZX	274
	291
nei	229
nth	5
	101
	365
alkA	31
	34
ligC	630
	938
mutT2	172
ogt	36
	110
mutT4	142
	297
rv2979	41

2.10.2: Latin American Mediterranean (LAM) genotyping

The Rv0129c gene sequence was extracted for each LAM sub-lineage isolate (Comas *et al*., 2009), and the presence of a polymorphism at nucleotide position 309 detected manually as described in section 2.10. The nucleotide at position 309 can be one of either a G or an A with the latter defining isolates as LAM isolates.

2.10.3: Haarlem genotyping

The *mtg*C, *ogt* and *ung* genes were extracted for each Haarlem sub-lineage assigned isolate (Alix *et al*., 2006, Cubillos-Ruiz *et al*., 2010, Dou *et al*., 2008) and the relevant polymorphisms described below were detected manually as described in section 2.10. The *mtgC* marker is based on an amino acid change at codon position 182, whilst the *ogt* and *ung* mutations are at nucleotide positions 44 and 501, respectively.

2.10.4: X family genotyping

The Rv3221c and Rv2330 gene sequences were extracted for each X family sub-lineage assigned isolate (Comas *et al*., 2009), and the relevant polymorphisms described below were manually detected as described in section 2.10. X family related polymorphism are found at position 30 within gene Rv3221c, and at nucleotide position 426 in gene Rv2330.

2.10.5: Lineage 1 genotyping of BK22

Lineage 1 genotyping of BK22 was based on polymorphisms at the well-characterised loci Rv0005, Rv0006, Rv0410c, Rv0934, Rv1996, Rv2462c, Rv3132c, Rv3221c, Rv0006, Rv0164, Rv0288, Rv0410c, Rv1009, Rv1996, Rv2030c, Rv2031c and Rv3261, and detected manually as described in section 2.10. Lineage 1 isolates can be sub-divided into either Manila or non-Manila strains. The Manila strains are identified due to the presence of polymorphisms at loci Rv0006, Rv0164, Rv0288, Rv0410c, Rv1009, Rv1996, Rv2030c, Rv2031c and Rv3261 (Comas *et al*., 2009).

2.11: Statistical analysis for lineage association

The probability of correct lineage assignment was determined for each isolate included in the cgMLST phylogeny analysis using Microsoft Excel (MS Excel, 2013). Prior to inclusion of Welsh isolates, the within lineage genomic distances of the lineage defined isolates were analysed. The mean and standard deviation of the internal genomic distances between isolates of each lineage were calculated through analysing the number of allelic differences between each isolate within each lineage. The distance distribution for each isolate (within a given lineage) from the lineage mean was then calculated through analysis of the normal distribution of allelic differences seen across all isolates within each lineage. By using the distributions calculated from each lineage we can calculate the probability that each given Welsh isolate falls into each of the lineages included in this study. By comparing this value, for each Welsh isolate across each lineage, we can measure which lineage is more credible as the group the Welsh isolate belongs to. Confident assignments were designated as those whereby the value for probability was 100%. Microsoft Excel (MS Excel, 2013) was used to carry out these calculations and to produce the corresponding graphs.

2.12: Gel electrophoresis

Gel electrophoresis was used to determine if DNA was present in the boilate samples. 1 % (w/v) agarose gels in TAE (40mM Tris acetate, 1mM EDTA) were made with SYBR™ Safe DNA Gel Stain (Invitrogen, Paisley, UK) and submerged in 1 x TAE electrophoresis buffer. DNA samples were mixed with one fifth their volume of 6 × DNA loading buffer (Promega, Southampton, UK) and loaded into slots in the gel formed by insertion of the appropriate comb prior to gel solidification. The gel was electrophoresed using a Bio-Rad horizontal gel electrophoresis tank at 75 volts for 40 minutes and visualised using a UV transilluminator. The size and concentration of DNA fragments were estimated by co-electrophoresing 10 µl GeneRuler^{TM} 1kb DNA ladder as size standards (Promega, Southampton, UK).

2.13 Ancient DNA extraction and library preparation protocol

2.13.1: Ancient DNA extraction protocol

DNA was extracted using an adapted Tris-phenol chloroform extraction protocol (Sambrook and Russell, 2006). The *M. tuberculosis* boilates were transferred into Fastprep lysing matrix E tubes (MP-Biomedicals, Santa Ana, California, United States). 500µl of breaking buffer (2% Triton-X100, 1% SDS, 100mM NaCl, 10mM Tris-HCl pH8, 100mM EDTA pH8) and 500 µl Phenol: Chloroform: Isoamyl alcohol (Sigma-Aldrich, St. Louis, MO, United States) were added to each tube. The tubes were then homogenized in a Thermo Savant FastPrep 120 Cell Disrupter Fastprep machine (GMI, Ramsey, MN, USA) at speed setting 6, for 5 x 20 second bursts with cooling on ice between cycles. Each tube was then centrifuged for 10 minutes at 16,000 X g in an Eppendorf 5415R centrifuge (Eppendorf Ltd, Stevenage, UK). The aqueous upper phase of the suspension was then transferred into a clean microcentrifuge tube. Two volumes of 700 µl ice cold absolute ethanol (Thermo Scientific, Waltham, MA, USA) was added to each tube and mixed, then centrifuged for 10 min at 16,000 X g. The supernatant was removed and 700 µl of 70% ethanol added to the tubes, avoiding disruption of the pellet. Ethanol was then decanted and the remaining pellet was then dried through incubation at 70 ^{O}C for 10 minutes on a Techne® digital Dri-Block® heater (Sigma-Aldrich, St. Louis, MO, United States). The pellet was then re-suspended in 50ul de-ionised water containing 10ug/ml RNase A (Sigma-Aldrich, St. Louis, MO, United States). The amount of DNA present in the samples was then quantified using an Agilent Bioanalyser 2100 instrument (Thermo Scientific, Waltham, MA, USA). Following extraction, the samples were fragmented for 20 cycles of 30 seconds within a Diagenode bioruptor 300.

2.13.2: Ancient DNA library preparation and sequencing of DNA

The DNA extracted in Section 2.13.1 was then subjected to the novel BEST aDNA library building protocol (Carøe C *et al*, 2017). The samples were pooled in equimolar concentrations and then sequenced on an Illumina HiSeq platform (Illumina, San Diego, CA, USA) in 80 bp single-end read mode at The Danish National High-Throughput DNA Sequencing Centre, University of Copenhagen, Denmark. The resulting single-end reads were quality filtered with the Trimmomatic tool and contigs assembled as described in section 2.2.

2.14: Production of Geographical map

The geographical map(s) were constructed using the online National Geographic MapMaker programme (https://mapmaker.nationalgeographic.org/#/).

2.15: Statistical analysis of in-house and established cgMLST variation.

The number of allelic differences between each isolate and the remaining isolates within the dataset was taken from both the established and in house core genome MLST results. The average number of differences each isolate had across the whole dataset (variation) was calculated using MS Excel (2013). These values were then used to determine the average amount of variations seen across the dataset for both cgMLST schemes. An unpaired T-test was performed using GraphPad Prism (version 7.00), to calculate the P Value for whether the amount of variation seen in the results of the in house cgMLST scheme and Established cgMLST scheme is different.

Chapter 3

Phylogenetic analyses of *M. tuberculosis* isolates from across South West Wales.

3.1: Introduction

3.1.1: M. tuberculosis strains and lineages

Evidence for different *M. tuberculosis* strains conferring different disease phenotypes and levels of virulence has been documented since the early part of the twentieth century (Gagneux and Small, 2007). A study in 1948, using a guinea pig model, showed that a *M. tuberculosis* strain from India was less virulent than one from a British case. This result has since been supported by various other studies (Mitchison *et al*., 1960, Mitchison *et al*., 1961 Singh, 1964, Collins and Smith, 1969). Further research also showed, based on aerosol infection of guinea pigs, that the less virulent Indian strains were also less infectious than the British ones (Balasubramanian *et al*., 1992, Williams *et al*., 2005). Evidence for strain variation effects on outbreak phenotypes was put forward in the 1990s when an outbreak in Kentucky, USA was caused by the CDC1551 strain (North *et al*., 1999, Williams *et al*., 2005,). Initially suspected of increased virulence (North *et al*., 1999, Williams *et al*., 2005), analysis showed that while it was more transmissible than the reference strain, its pathogenicity was not increased(Bishai *et al*., 1999, Manca *et al*., 1999). More recently, the HN878 strain, found in outbreaks in Los Angeles, USA has shown a consistently hyper-virulent phenotype across various experimental models (Gagneux and Small, 2007). Despite the well-recognised differences in disease characteristics between strains, the monomorphic genomic nature of *M. tuberculosis* is recognised in comparison with other pathogenic bacteria. However, large sequence polymorphisms (LSP) and single nucleotide polymorphisms (SNP) for the analysis of sequencing data, has revealed how different MTBC members, different *M. tuberculosis* lineages and sub-lineages can have significantly different pathobiology, virulence, functional genomic variation and physiological features (Gagneux and Small., 2007, Homolka *et al*., 2012) Thus, the genomic diversity of tuberculosis has been historically underestimated prior to the use of LSP and SNP based genomic analysis (Gagneux and Small., 2007).

3.1.2: *M. tuberculosis* major lineages

Within *M. tuberculosis* seven major lineages have been recognised globally (Gagneux *et al*., 2006, Hershberg *et al*., 2008, Homolka *et al*., 2012, Comas *et al*., 2013, Yimer *et al*., 2015). Each lineage has a different characteristic with regards to its evolutionary status, transmissibility, drug resistance, host interaction, latency and vaccine efficacy (Thwaites *et al*., 2008). Lineages can be grouped into those that are evolutionarily modern and ones that are ancient (Gagneux and Small., 2007). *M. tuberculosis* lineages 2, 3 and 4 are considered modern and 1, 5 and 6 are considered ancient, with lineage 7 thought to be intermediate (Comas *et al*., 2013, Firdessa *et al*., 2013). The more ancient lineages 1, 5 and 6 coexisted with humans prior to Neolithic expansion of anatomically modern *Homo sapiens* populations (Comas *et al*., 2013). Ancient lineages have been shown to exist in pre-Neolithic human populations which rarely exceeded 20 individuals (Blaser and Kirschner., 2007, Comas *et al*., 2013). It is thought that ancient strains were adapted to this low population density demographic, to sustain a continuous infectious reservoir within smaller populations, a feature consistently seen across bacterial species that infect populations of low density (Blaser and Kirschner, 2007). During the post-Neolithic era, with its higher-density human populations, the more modern lineages (2, 3 and 4) have dominated, evolving to become more representative of a crowd disease than the more ancestral and readily latent ancient lineage*s* 1, 5 and 6 (Comas *et al*., 2013).

The ancient lineages cause a more marked immune response (Reiling *et al*., 2013) and are also more likely to become latent than their modern relatives, progressing more slowly to active disease (Portevin *et al.,* 2011, Comas *et al*., 2013). This observation is supported by the fact that infection rates with ancient strains are in decline, in contrast to the increased dissemination of the modern lineage*s s*uch as lineage 2 Beijing strains and lineage 4 Euro-American strains (Comas *et al*., 2013). Despite the ancient lineage 1 being found in predominantly South East Asia, the current hypotheses is that the lineage arose in Africa around 68,000 years ago and followed human migration out of the continent (Wirth *et al*., 2008, Reed *et al*., 2009, Comas *et al.*, 2013). Modern lineage*s* appear more virulent and readily able to cause active disease, elicit a less marked immune response, have higher rates of transmission and increased replication rates and are more geographically dispersed (Gagneux *et al*., 2006, Hershberg *et al*., 2008, Homolka *et al*., 2012, Comas *et al*., 2013, Yimer *et al*., 2015). Lineage 4 isolates have been shown to be less capable of extra-pulmonary TB disease, including meningeal disease, than other lineage*s* (Portevin *et al*., 2011) and are less likely to be associated with HIV co-infection (Caws *et al.*, 2006, Middelkoop *et al*., 2009). Lineage 3 has been shown to be phylogenetically close to lineage 2 (Portevin *et al*., 2011, Comas *et al*., 2013), and is predominantly restricted to India. However, its presence is significant amongst the Indian population within the UK and has caused outbreaks (Sarkar *et al*., 2012). In comparison with the other modern lineage*s*, lineage 3 has shown significant association with extra-pulmonary tuberculosis (Rindi *et al*., 2012). However, it also appears that it may be less virulent than the other modern lineage*s* in terms of rate of replication, having the lowest rate of intracellular growth within human macrophages of the modern lineages (Sarkar *et al*., 2012).

3.1.3: *M. tuberculosis* sub-lineages

Each *M. tuberculosis* lineage can be divided into sub-lineages, which can then be further divided into specific strains and genotypes. Sub-lineage*s* also have phenotypic variations, including differences in their virulence and ability to cause disease (Anderson *et al*., 2013). In particular, lineages 2 (East Asian) and 4 (Euro-American) contain strains, such as Beijing (also known as W/Beijing) and Haarlem genotypes respectively, which are notorious for their association with tuberculosis outbreaks and over-representation amongst drug resistant cases (Bifani *et al*., 1996, , Mardassi *et al*., 2005, Marais *et al*., 2006, De Jong *et al*., 2008). Lineage 2 can be divided into non-Beijing strains and Beijing strains, with Beijing strains also being split into modern and ancient sub-lineages (Thwaites *et al*., 2008). The modern Beijing sub-lineage*s* appear more virulent (Mestre *et al*., 2011). The lineage 2 Beijing strains have a higher propensity to acquire drug resistance than lineage 4 Euro-American strains (Thwaites *et al*., 2008, Ford *et al*., 2013). The Lineage 2 Beijing strains are considered the most successful ones over recent years, as they have disseminated globally (Wirth *et al*., 2008) and have been associated regularly with large *M. tuberculosis* outbreaks (Bifani *et al*., 1996, Thwaites *et al*., 2008). The Beijing strains are known to have a higher propensity for acquisition of drug resistance, causing the development of multi-drug and extensively-drug resistant strains within the lineage (Bifani *et al*., 1996, Thwaites *et al*., 2008). They have been shown to disseminate rapidly and cause more severe disease. This was particularly evident in a large outbreak of the strain in New York in the 1990s (Bifani *et al*., 1996, Thwaites *et al*., 2008). The Beijing genotype has the ability to produce a unique phenolic glycolipid that attenuates the host’s immune response (Nicol *et al*., 2005). Meningeal tuberculosis caused by Beijing strains has a more rapid progression, with shorter symptom duration when compared with meningeal infection with other lineage strains (Nicol *et al*., 2005). Beijing strain infection is associated with low levels of CSF leukocytes, which is an independent risk factor for death or severe disability from meningeal tuberculosis (Thwaites *et al*., 2008).

Lineage 4 represents the most diverse and geographically widespread lineage of *M. tuberculosis*, with each sub-lineage having different phenotypic characteristics (Rasigade *et al*., 2017, Comas *et al*., 2014, Stucki *et al*., 2016). Major sub-lineage*s* within lineage 4 include X, T, Latin American Mediterranean (LAM) and Haarlem (Fitzgibbon *et al*., 2013). Of the lineage 4 sub-lineage*s*, the presence of Haarlem strains is of particular interest as they are thought have a higher propensity to cause outbreaks of increased severity and drug resistance in comparison with other lineage 4 sub-lineage*s* (Mardassi *et al*., 2005). LAM sub-lineage strains have also been associated with widespread epidemics and drug resistance (David *et al*., 2012). The Haarlem genotype accounts for 25% of tuberculosis cases in Europe, as well as central America and the Caribbean, where research suggests its presence may be correlated with post-Columbus European colonisation (Cubillos-Ruiz *et al*., 2010). Its significance is based on its global dissemination and its predominance in multi-drug resistant outbreaks (Marais *et al*., 2006, Khanipour *et al*., 2016, Mardassi *et al*., 2005). A study in Tunisia showed Haarlem strains caused 22-fold more multi-drug resistant cases than non-Haarlem Euro-American strains (Mardassi *et al*., 2005). Haarlem genotype strains have been found to have an increased clonal expansion rate compared to other ones that has been postulated to allow them to spread quickly whilst also being multi-drug resistant (Tessema *et al*., 2013, Mardassi *et al*., 2005). Identifying the presence of such sub-lineage*s* within a Welsh dataset would be of interest to public health and outbreak control. Identification of strains known to harbour drug resistant qualities is also in line with the recent public health England movement to genomics for tuberculosis control (Votintseva *et al*., 2017).

3.1.4: Robust genetic markers for phylogenetic classification *M. tuberculosis*

Previous typing methods, based on MIRU-VNTR profiling and spoligotyping, have allowed the classification of isolates into phylogeographically related clades and families and have led to the development of readily available databases such as SpolDB4 (Brudey *et al*., 2006, Gagneux and Small., 2007) and MIRU-VNTR plus (Weniger *et al*., 2010). These databases provide a large amount of information on strains from across the globe; however, the methods are based on sequence information extracted from polymorphic repetitive regions. Due to the nature of these regions, convergent evolution occurs whereby strains show the same profile but are actually unrelated (Gagneux and Small., 2007, Müller *et al*., 2014). The advent of WGS comparative analysis has showed that on the whole genome level, *M. tuberculosis* displays a substantial genetic divergence (Gagneux and Small., 2007). This has been exploited for the identification of robust genetic markers for phylogenetic strain assignment. Such markers are found as either Large Sequence Polymorphisms (LSPs), used for the original lineage classification of isolates by Gagneux in 2006, or SNPs (Gagneux *et al*., 2006, Gagneux and Small., 2007). SNPs are reliable, phylogenetically informative markers as the low sequence variation and lack of horizontal gene transfer makes independent recurrent mutations unlikely and one does not encounter the convergent evolution problem of the previously mentioned methods (Warren *et al*., 2002, Gutacker *et al*., 2002, Gutacker *et al*., 2006, Gagneux and Small., 2007). Lack of horizontal gene transfer also prevents the re-acquisition of large sequence regions within strains that have previously lost them, thus increasing the robustness of LSP (Gagneux and Small., 2007). LSP analysis requires the use of micro-arrays while SNP markers can be found and analysed *in silico* from WGS (Gutacker *et al*., 2006). The use of both methods has shown more potential in providing genetic markers for production of deep phylogenies in comparison with spoligotyping and MIRU-VNTR profiling (Warren *et al*., 2002, Gutacker *et al*., 2006, Gagneux and Small., 2007). The *in silico* application of WGS data means multiple sequences from various strains can be analysed and compared within the same study (Gagneux and Small., 2007). It is worth noting that despite the benefits of using robust genetic markers such as SNPs and LSP, the SpolDB4 database has proven useful, showing clear correlations between spoligotype-defined clades within lineage*s* and sub-lineage*s* defined by LSP and SNP genetic groupings (Baker *et al*., 2004, Brudey *et al*., 2006).

3.1.4.1: Principal Genetic Grouping

Principal genetic groups (PGGs) classify isolates into one of the three PGGs based on non-synonymous variants at the *katG* and *gyr*A genes at codons 463 and 95 respectively as summarised in Table 2.1, section 2.9 (Sreevatsan *et al*., 1997b). The current consensus is that PGG1 consists of isolates that are ancestral to all the others represented in PGG2 and PGG3 (Sreevatsan *et al*., 1997), a notion supported by spoligotype analysis, which further indicates that PGG3 isolates are descended from isolates of PGG2 (Soini *et al*., 2000). Studies have shown that PGG2 and PGG3 isolates comprise modern lineage*s* but more specifically modern Euro-American lineage*s* (Grimes *et al*., 2009). PGGs correlate with patterns of drug resistance, and PGG2 and PGG3 contain more susceptible strains compared to PGG1 (Arjomandzadegan *et al*., 2012).

3.1.4.2: SNP cluster grouping

SNP cluster grouping (SCG) is a method used to classify isolates into phylogenetically relevant groups (Filliol *et al*., 2006, Alland *et al*., 2007). The original method was based on a 212 SNP analysis, resulting in the division of *M. tuberculosis* isolates into six phylogenetically distinct groups: SCG-1-6 with a seventh group containing all *M. bovis* isolates (Filliol *et al.*, 2006). Alland *et al.* (2007) adapted this method to comprise 9 loci from *M. tuberculosis,* (*M. bovis* was excluded, Table 2.2) allowing further discrimination of SCGs 3 and 6 into five subgroups: SCG-3a, 3b, 3c, 6a and 6b. Each of the SCGs are geographically distinct and have correlating spoligotype families as shown in Table 3.1.

Table 3.1: A matrix showing the correlation between the SCG and spoligotype families, adapted from Filliol *et al.* (2006) and Alland *et al*. (2007).

SCG Group	SCG-1	SCG-2	SCG-3a	SCG-3b	SCG-3c	SCG-4	SCG-5	SCG-6a	SCG-6b
Spoligotype	EAI	Beijing	CAS	H, T, LAM	X	X	H	T, S, LAM, H, X	T, S, LAM, X, H

EAI=East African Indian, CAS=Central Asian, LAM=Latin American Mediterranean, H=Haarlem, T= T family, S=S family, X= X family.

Each SCG has been also been shown to correlate with a specific PGG. Therefore by using both methods discrepancies in the PGG or SCG assignments can be identified and isolates can be ascribed evolutionary timescales (Filliol *et al*., 2006). Filliol *et al* (2006) elucidated that SCG1, SCG-3a and SCG-2 all fall into PGG1, each representing isolates which were the earliest to diverge from the most recent common ancestor with *M. bovis;* PGG2 isolates include SCG-3b, SCG-3c, SCG-4 and SCG-5; and isolates from PGG3 comprise SCG-6a and SCG-6b, the most recently diverged groups. Over time SCGs and PGGs have also been correlated with the LSP and SNP analysis defined lineage*s* and sub-lineage*s* already described (Gagneux and Small., 2007).

3.1.4.3: Robust SNP barcode method

Although SNPs represent phylogenetically informative markers for the assignment of isolates into phylogenetic lineages and sub-lineages, extraction of SNPs from the whole genomes of multiple isolates produces a large amount of data lacking in the level of standardisation offered by methods such as MIRU-VNTR and spoligotyping. Both these methods have well-established, standardised and accessible online databases (Weniger *et al*., 2010, Brudey *et al*., 2006b) allowing for the phylogenetic analysis of *M. tuberculosis* and comparison with strains from across the globe. Coll *et al.* (2014) developed a robust SNP barcode method analysing 60 loci, capable of assigning *M. tuberculosis* isolates into major lineages and sub-lineages. The method increases the level of resolution in comparison to the more primitive PGG and SCG SNP sets. In addition, this robust SNP barcode provides correlation with well-known spoligotype families and results can be compared with a globally established database. A key feature of the SNP barcode method is that it provides results based on a standardised set of genetic markers that are often difficult to attain from standard WGS analysis due to the comparatively low sequence variability in *M. tuberculosis* isolates and impracticality of comparing large amounts of WGS data in a reproducible manner (Hamolka *et al*., 2012).

3.1.5: Novel use of a core genome MLST method for phylogenetic purposes

Currently, the lack of WGS data standardisation is one of the barriers to its widespread usage (Kohl *et al.,* 2014, Walker *et al*., 2013b). In addition, the output of partial sequence data from current sequencing platforms and the need for robust high-quality reference sequences has often meant that SNP mapping analysis, such as that done in previous studies (Walker *et al*., 2013a, Walker *et al*., 2013b), is less ableto analyze sequence data that is incomplete, which is often outputted by sequencing platforms (Maiden *et al*., 2013, Sheppard *et al.,* 2012). However, recently the development of WGS gene-by-gene MLST methods and software such as the Bacterial Isolate Genome Sequence database (BIGSdb) (Jolley and Maiden, 2010) and Ridom SeqSphere (Junemann *et al*., 2013) have resulted in a more standardised and automated approach than traditional WGS SNP mapping (Maiden *et al.*, 2013, Sheppard *et al*., 2012). For example, Ridom SeqSphere allows isolate sequences to be aligned and compared in a standardised manner using a globally defined core genome MLST (cgMLST) scheme (Junemann *et al*., 2013, Kohl *et al*., 2014). To date, this method has only been used for providing clinical resolution of tuberculosis outbreaks (Kohl *et al*., 2014), and has not been used to analyse the phylogenetic composition of a *M. tuberculosis* isolate dataset.

3.1.6: Aims and objectives

The objective of the work in this chapter was to analyse phylogenetically 80 Welsh *M. tuberculosis* isolates by PGG, SCG, SNP *bar-coding* and gene-by-gene based cgMLST. Each method will allow the isolates to be assigned to relevant phylogenetic groups, lineages and sub-lineages in comparison to global *M. tuberculosis* isolates. The first aim was to provide a snapshot of their diversity in the context of global *M. tuberculosis* phylogenetics, something which has not been done previously. Secondly, an aim was to provide the first assessment of the use of cgMLST gene-by-gene analysis for the phylogenetic classification of *M. tuberculosis* isolates and to whether the phylogenetic structures match that seen by more conventional but less standardized SNP mapping methods. Lastly, the aim was to assess any correlation between phylogenetic classifications, according to each method and outbreak status.

3.2: Methods

3.2.1: Principal genetic grouping (PGG)

Isolates were assigned into one of three principal genetic groups (PGGs) based on the amino acid combination present at loci *gyrB* and *katG* codons, codons 95 and 493 respectively. The Principal Genetic Group phylogeny was produced using Ridom SeqSphere software, and for further details on this method see section 2.7 and Table 2.2.

3.2.2: SNP cluster grouping (SCG)

Each Welsh isolate was assigned a SNP Cluster Group (SCG) based on the composition of nucleotides present at 9 specific loci defined previously (section 2.8, Table 2.3). Isolates could be assigned to one of 6 major SCGs (SCG1-6). Isolates assigned to SCG3 or SCG could also be further resolved into sub-groups (SCG3a, 3, 3b, 3c, 6a and 6b) as described in Table 2.3. The SNP cluster group phylogeny was produced using Ridom SeqSphere software.

3.2.3: Major lineage and sub-lineage assignment

Using the pipeline described in section 2.5.1, the 60 SNP loci were extracted from the genomes of the Welsh isolates, whereby sequence data was of sufficient quality to analyze. The isolates were then assigned into major lineages, sub-lineages and correlating spoligotype families based on the 60 SNP barcode as described in section 2.5.1 and table 2.1.

3.2.4: Construction of cgMLST phylogenies

A cgMLST phylogeny was produced using the Ridom SeqSphere software, based on a total of 236 isolates as described in section 2.9. Of the 236 isolates, the lineage of 179 has been defined previously (Comas *et al.*, 2013), with the remainder of the isolates being Welsh.

3.2.5: Statistical analysis for lineage association

For all isolates successfully included in the cgMLST phylogeny analysis, the statistical probability that the lineage associations made were reliable was calculated (section 2.11). Confident associations were deemed as those with 100% probability that the association between each Welsh isolate and the phylogenetic lineage was true.

3.2.6: Sub-lineage Genotyping

Sub lineage genotyping was carried out by analysis of SNPs at positions that are specific to different phylogenetic groups, as described in section 2.10. To support lineage and sub lineage assignments made by both SNP-bar-coding and cgMLST, nucleotide data were extracted on genetic markers that represent different sub-lineage groups. Clade-specific SNPs relevant to the lineage 1, Beijing, Haarlem, LAM and X clades specifically were extracted from the relevant gene sequence and examined manually as described in section 2.10. As the T family sub-lineage is a large family generically that is less defined, no polymorphisms were used to define these isolates.

3.2.6.1: Beijing genotyping

Each isolate assigned as a Beijing sub-lineage by SNP bar-coding was analysed for Beijing sub-lineage relevant SNPs. As described in section 2.10.1, genes *mutT2*, *mutT4* and *ogt* were analysed for polymorphisms at given codon and nucleotide positions.

3.2.6.1.1: Advanced Beijing Genotyping

For advanced Beijing genotyping, phylogenetic analysis of each isolate was subjected to analysis of 48 SNPs across 22 genes as outlined in section 2.10.1. The 48 SNPs were extracted from the three Beijing-related isolates identified.

3.2.6.2: Latin American Mediterranean (LAM) genotyping

Gene locus Rv0129c was extracted and the nucleotide at position 309 examined for the presence of an A instead of G at this position (section 2.10.2). Isolates with this mutation were defined as harbouring the LAM genotype.

3.2.6.3: Haarlem genotyping

Isolates deemed as having the Haarlem sub lineage according to a robust SNP barcode were examined for specific polymorphisms at loci *mtgC*, *ogt* and *ung* (section 2.10.3).

3.2.6.4: X family Genotyping

X family related polymorphisms were examined at nucleotide positions 30 within gene Rv3221c, and 426 in gene Rv2330, as described in section 2.10.4.

3.2.6.5: Lineage 1 genotyping for BK22.

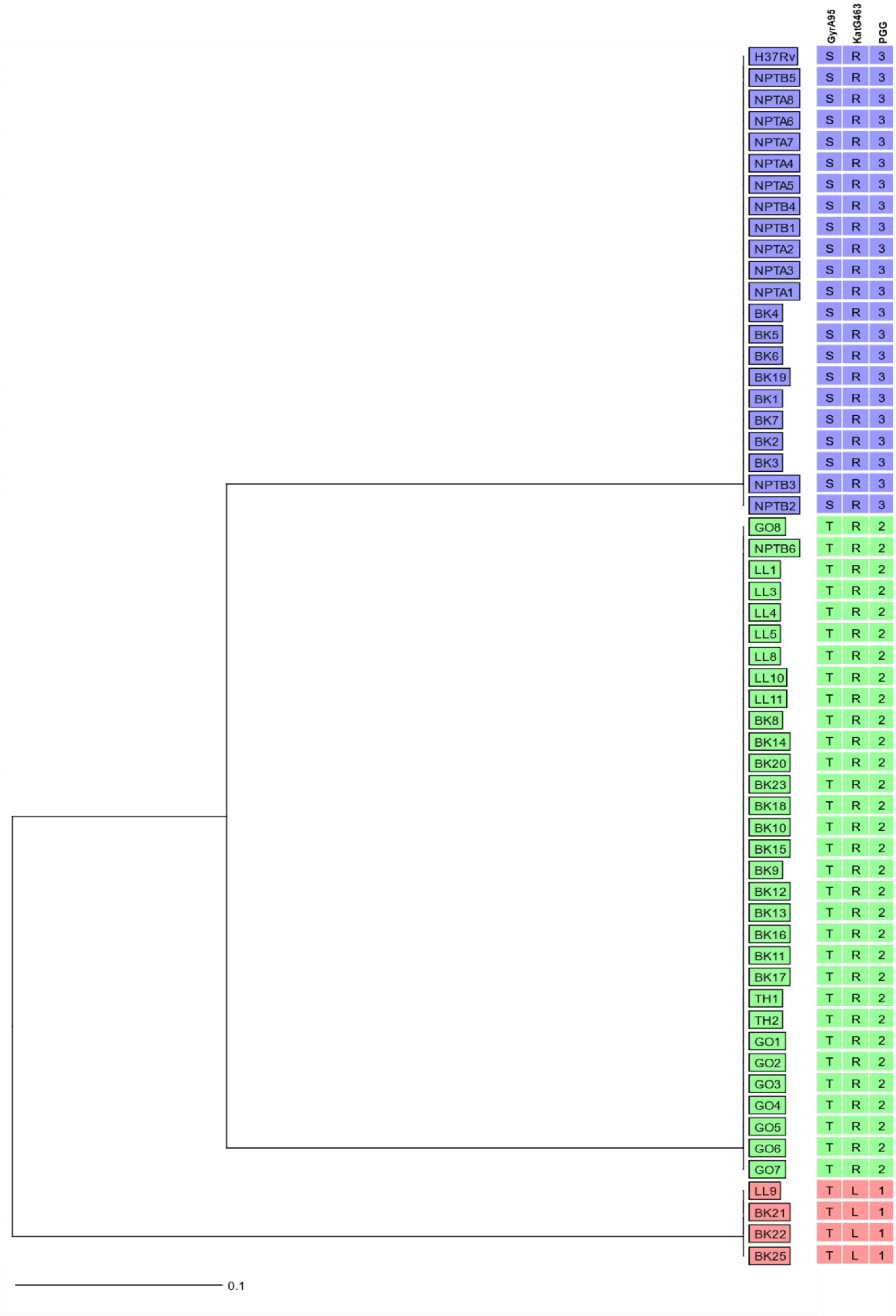
Seventeen well characterised loci, defined in section 2.10.5, were examined for lineage 1 specific polymorphisms, and a subset of these was examined for polymorphisms associated with the Manila genotype found within lineage 1 isolates. The 17 gene sequences, in which the polymorphisms lie, were extracted from BK22 and the 17 relevant polymorphisms were detected manually as described in section 2.10.

3.3: Results

Of the 80 Welsh isolates analysed (section 2.2), only 66 were successfully sequenced. Therefore the DNA sequence data from these isolates were analysed, as well as the reference strain H37Rv sequence.

3.3.1: PGG analysis

Fifty-seven Welsh isolates could be assigned to a PGG based on the sequence data (see section 3.2.1). The analysis identified 31 isolates as PGG2, and 23 as PGG3, including the H37Rv genome (Figure 3.1). Only four isolates clustered within PGG1. Isolates BK24, BK26, GO9, BK27, LL2, LL7, NPTB7, BK28, and BK29 could not be assigned to a PGG due to lack of sequence data at codons 95 and 493 in genes *gyrB* and *katG*.



	GyrA95	KatG463	PGG
H37Rv	S	R	3
NPTB5	S	R	3
NPTA8	S	R	3
NPTA6	S	R	3
NPTA7	S	R	3
NPTA4	S	R	3
NPTA5	S	R	3
NPTB4	S	R	3
NPTB1	S	R	3
NPTA2	S	R	3
NPTA3	S	R	3
NPTA1	S	R	3
BK4	S	R	3
BK5	S	R	3
BK6	S	R	3
BK19	S	R	3
BK1	S	R	3
BK7	S	R	3
BK2	S	R	3
BK3	S	R	3
NPTB3	S	R	3
NPTB2	S	R	3
GO8	T	R	2
NPTB6	T	R	2
LL1	T	R	2
LL3	T	R	2
LL4	T	R	2
LL5	T	R	2
LL8	T	R	2
LL10	T	R	2
LL11	T	R	2
BK8	T	R	2
BK14	T	R	2
BK20	T	R	2
BK23	T	R	2
BK18	T	R	2
BK10	T	R	2
BK15	T	R	2
BK9	T	R	2
BK12	T	R	2
BK13	T	R	2
BK16	T	R	2
BK11	T	R	2
BK17	T	R	2
TH1	T	R	2
TH2	T	R	2
GO1	T	R	2
GO2	T	R	2
GO3	T	R	2
GO4	T	R	2
GO5	T	R	2
GO6	T	R	2
GO7	T	R	2
LL9	T	L	1
BK21	T	L	1
BK22	T	L	1
BK25	T	L	1

Figure 3.1: Phylogeny highlighting the PGG assignment of 57 Welsh isolates and H37Rv. Red: PGG1, green: PGG2, blue: PGG3. Letters refer to the amino acids present at each locus: T = Threonine, R = Arginine, L = Leucine, S = Serine. The scale bar highlights the genetic divergence relevant to branch length measured in units of amino acid differences per site across the *gyrA* and *katG* loci.

3.3.2: SNP cluster group (SCG) analysis

Division of isolates into SCGs based on SNPs at 9 loci allowed further resolution of the information obtained from the PGG clustering analysis (see section 3.2.2). Only isolates with sequence data present for the whole set of 9 loci were analysed. The two predominant SCGs were 6a and 3b with 16 and 15 isolates clustering to these sub-groups respectively (Figure 3.2). Sub-groups SCG-3c and SCG-6b contained clusters of 7 and 4 isolates respectively, whilst 8 isolates clustered as SCG4. SCG-5, SCG-2 and SCG-1 contained 3, 2 and 1 isolates respectively. Sub-group SCG-3a was the only SCG not found in the dataset. Isolates BK29, BK17, NPTB7, LL10, BK28, BK21, GO9, BK19 and BK26 did not yield sequence data for all nine loci and were not analysed but further details on them can be found in the Appendix table PH1.

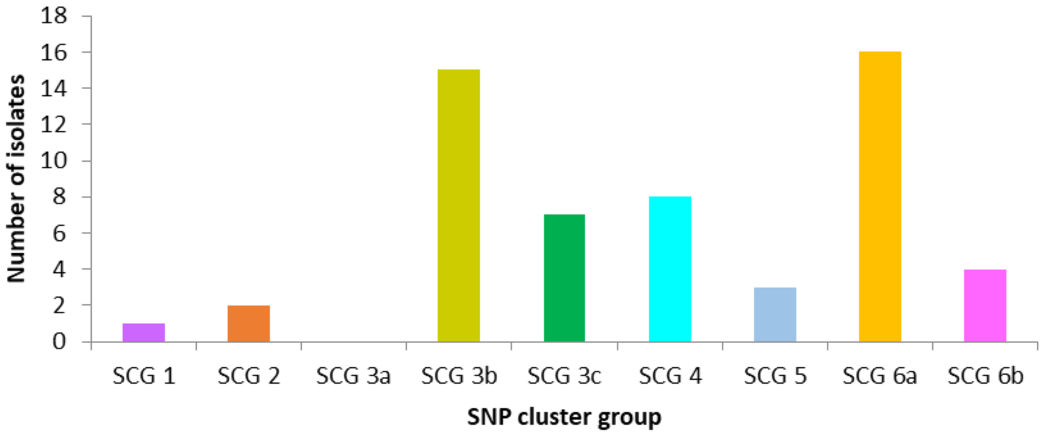
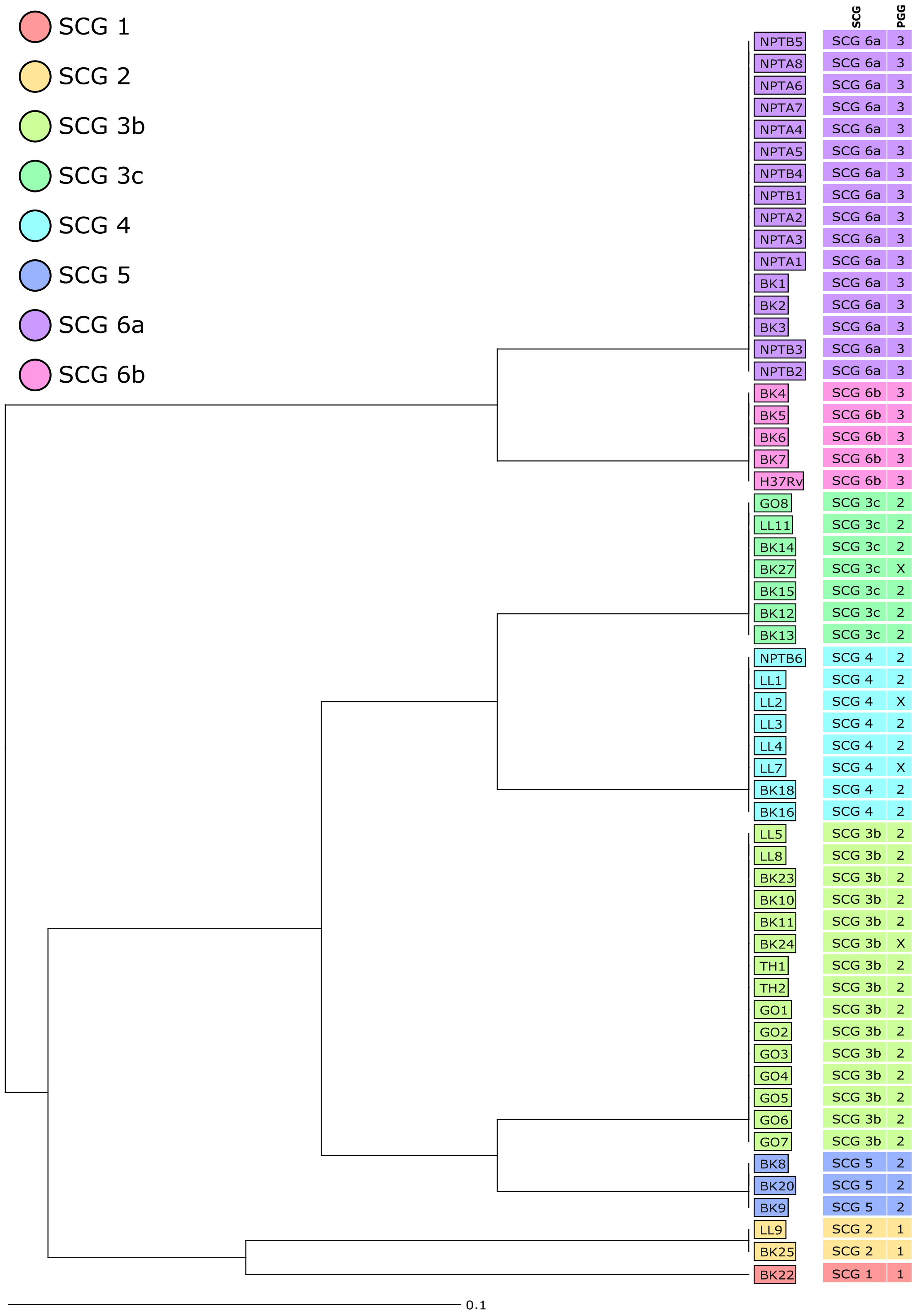


Figure 3.2: The number of isolates in each SCG, for the 57 isolates where nucleotide information at the 9 defined loci was obtained.

More diversity was seen in the clustering of isolates by SCG than with PGG. The SCG phylogeny split into two clear clades; clade 1 contained SCG-6a and SCG-6b and clade 2 contained SCG-1, SCG-2, SCG-3b, SCG-3c, SCG-4 and SCG-5 (Figure 3.3). Clade 2 was also more diverse than clade 1 as SCG-1 and SCG-2 branched off before further divergence was seen with SCG-3b, SCG-3c, SCG-4 and SCG-5. When PGG results were compared with SCG results, it was found that clade 1 contained all the PGG3 isolates and Clade 2 all PGG1 and PGG2 isolates (Figure 3.3). The phylogeny also revealed that the PGG2 isolates divided into four different SCG groups (SCG-3b, SCG-3c, SCG-4 and SCG-5), highlighting that there is further phylogenetic diversity within this dataset than seen with PGG alone (Figure 3.1). The results also indicated that within clade 2, SCG-3c and SCG-4 share a closer relationship with each other than they do with isolates of SCG-3b and SCG-5, and vice versa.

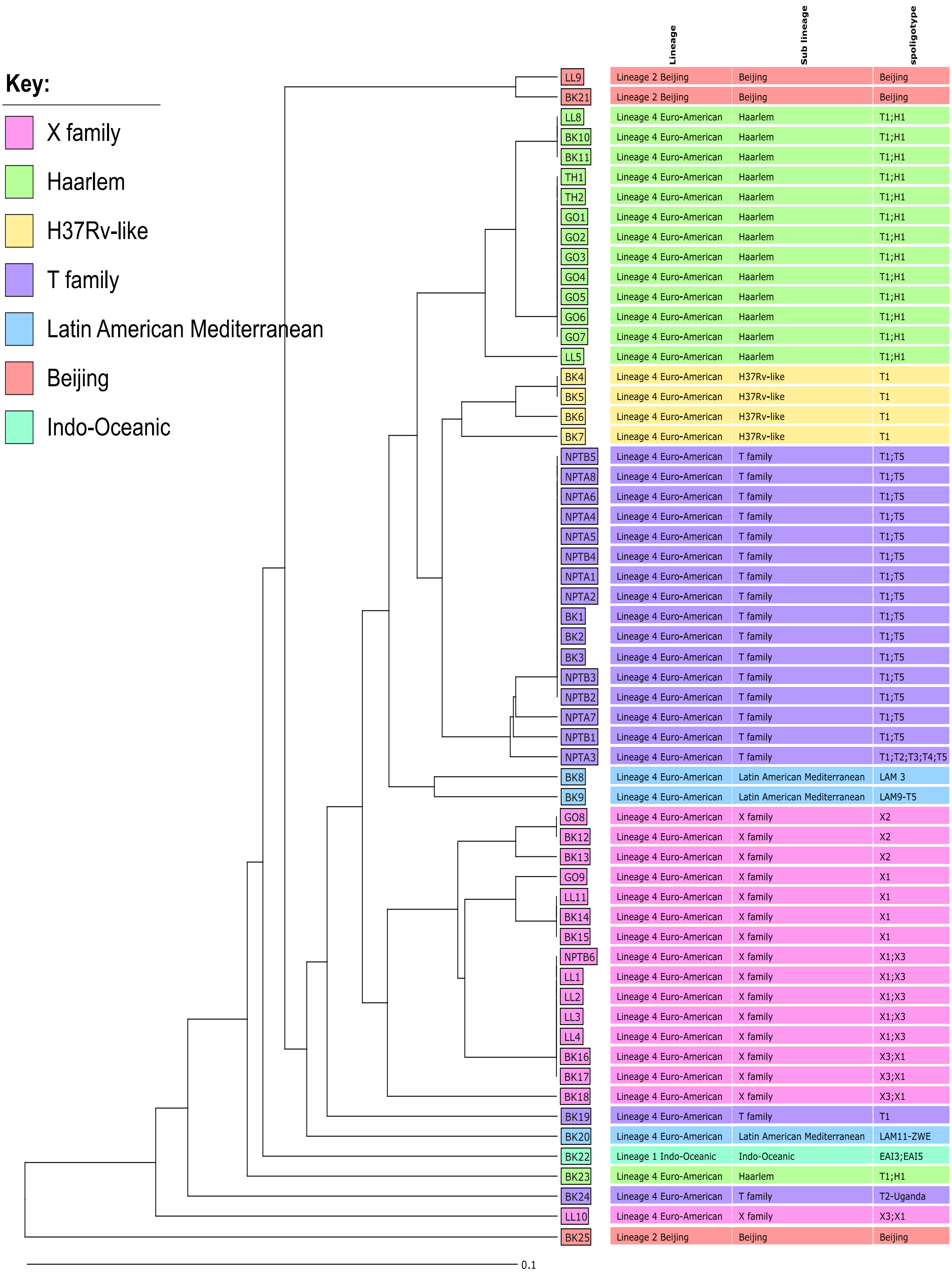


	SCG	PGG
NPTB5	SCG 6a	3
NPTA8	SCG 6a	3
NPTA6	SCG 6a	3
NPTA7	SCG 6a	3
NPTA4	SCG 6a	3
NPTA5	SCG 6a	3
NPTB4	SCG 6a	3
NPTB1	SCG 6a	3
NPTA2	SCG 6a	3
NPTA3	SCG 6a	3
NPTA1	SCG 6a	3
BK1	SCG 6a	3
BK2	SCG 6a	3
BK3	SCG 6a	3
NPTB3	SCG 6a	3
NPTB2	SCG 6a	3
BK4	SCG 6b	3
BK5	SCG 6b	3
BK6	SCG 6b	3
BK7	SCG 6b	3
H37Rv	SCG 6b	3
GO8	SCG 3c	2
LL11	SCG 3c	2
BK14	SCG 3c	2
BK27	SCG 3c	X
BK15	SCG 3c	2
BK12	SCG 3c	2
BK13	SCG 3c	2
NPTB6	SCG 4	2
LL1	SCG 4	2
LL2	SCG 4	X
LL3	SCG 4	2
LL4	SCG 4	2
LL7	SCG 4	X
BK18	SCG 4	2
BK16	SCG 4	2
LL5	SCG 3b	2
LL8	SCG 3b	2
BK23	SCG 3b	2
BK10	SCG 3b	2
BK11	SCG 3b	2
BK24	SCG 3b	X
TH1	SCG 3b	2
TH2	SCG 3b	2
GO1	SCG 3b	2
GO2	SCG 3b	2
GO3	SCG 3b	2
GO4	SCG 3b	2
GO5	SCG 3b	2
GO6	SCG 3b	2
GO7	SCG 3b	2
BK8	SCG 5	2
BK20	SCG 5	2
BK9	SCG 5	2
LL9	SCG 2	1
BK25	SCG 2	1
BK22	SCG 1	1

Figure 3.3: A neighbour joining phylogeny showing the SCG profile for 57 *M. tuberculosis* isolates (including the H37Rv reference genome). The phylogeny harbours two clades, labelled Clade 1 and Clade 2. The PGG assigned to each isolate is shown in the furthermost right column. X in the PGG column denotes isolates that could not be assigned a PGG group. The scale bar highlights the genetic divergence relevant to branch length measured in units of nucleotide differences per site across 60 loci defined in section 3.2.2.

3.3.3: Robust SNP bar-coding

SNP bar coding was carried out on 59 *M. tuberculosis* isolates that had over 90% sequence data needed for the 60 loci SNP barcode analysis (see section 3.2.2). The results in Figure 3.4 show the lineage, sub-lineage and correlating spoligotype family data for the 59 analysed isolates. In terms of major lineages, lineage 4 (Euro-American) dominated the dataset, accounting for 55 of the 59 isolates. The remaining 4 isolates were of lineage 2 (n = 3) and lineage 1 (n = 1).

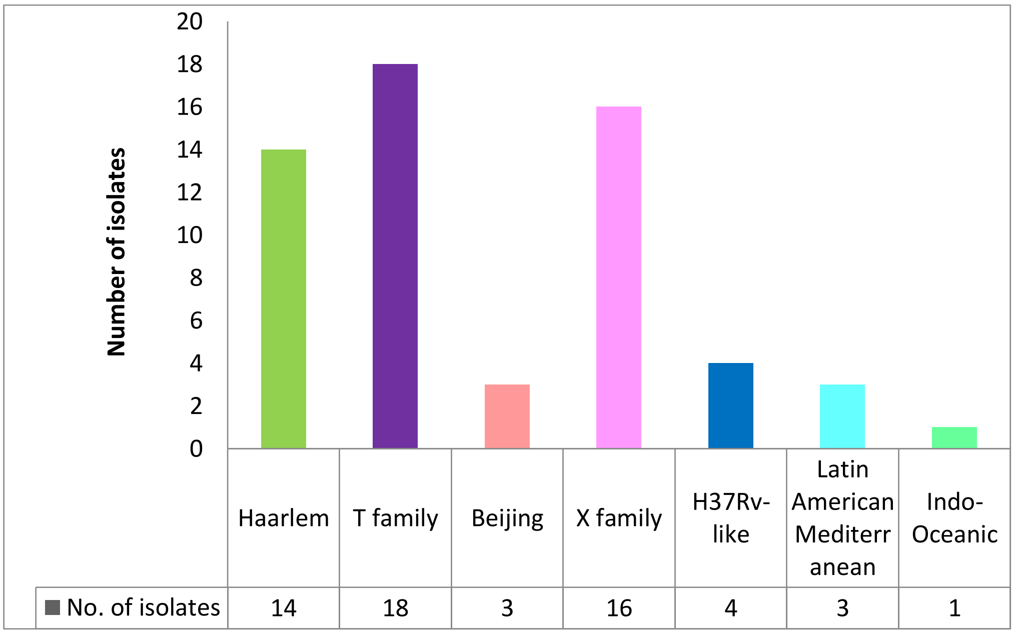


	Lineage	Sub lineage	spoligotype
LL9	Lineage 2 Beijing	Beijing	Beijing
BK21	Lineage 2 Beijing	Beijing	Beijing
LL8	Lineage 4 Euro-American	Haarlem	T1;H1
BK10	Lineage 4 Euro-American	Haarlem	T1;H1
BK11	Lineage 4 Euro-American	Haarlem	T1;H1
TH1	Lineage 4 Euro-American	Haarlem	T1;H1
TH2	Lineage 4 Euro-American	Haarlem	T1;H1
GO1	Lineage 4 Euro-American	Haarlem	T1;H1
GO2	Lineage 4 Euro-American	Haarlem	T1;H1
GO3	Lineage 4 Euro-American	Haarlem	T1;H1
GO4	Lineage 4 Euro-American	Haarlem	T1;H1
GO5	Lineage 4 Euro-American	Haarlem	T1;H1
GO6	Lineage 4 Euro-American	Haarlem	T1;H1
GO7	Lineage 4 Euro-American	Haarlem	T1;H1
LL5	Lineage 4 Euro-American	Haarlem	T1;H1
BK4	Lineage 4 Euro-American	H37Rv-like	T1
BK5	Lineage 4 Euro-American	H37Rv-like	T1
BK6	Lineage 4 Euro-American	H37Rv-like	T1
BK7	Lineage 4 Euro-American	H37Rv-like	T1
NPTB5	Lineage 4 Euro-American	T family	T1;T5
NPTA8	Lineage 4 Euro-American	T family	T1;T5
NPTA6	Lineage 4 Euro-American	T family	T1;T5
NPTA4	Lineage 4 Euro-American	T family	T1;T5
NPTA5	Lineage 4 Euro-American	T family	T1;T5
NPTB4	Lineage 4 Euro-American	T family	T1;T5
NPTA1	Lineage 4 Euro-American	T family	T1;T5
NPTA2	Lineage 4 Euro-American	T family	T1;T5
BK1	Lineage 4 Euro-American	T family	T1;T5
BK2	Lineage 4 Euro-American	T family	T1;T5
BK3	Lineage 4 Euro-American	T family	T1;T5
NPTB3	Lineage 4 Euro-American	T family	T1;T5
NPTB2	Lineage 4 Euro-American	T family	T1;T5
NPTA7	Lineage 4 Euro-American	T family	T1;T5
NPTB1	Lineage 4 Euro-American	T family	T1;T5
NPTA3	Lineage 4 Euro-American	T family	T1;T2;T3;T4;T5
BK8	Lineage 4 Euro-American	Latin American Mediterranean	LAM 3
BK9	Lineage 4 Euro-American	Latin American Mediterranean	LAM9-T5
GO8	Lineage 4 Euro-American	X family	X2
BK12	Lineage 4 Euro-American	X family	X2
BK13	Lineage 4 Euro-American	X family	X2
GO9	Lineage 4 Euro-American	X family	X1
LL11	Lineage 4 Euro-American	X family	X1
BK14	Lineage 4 Euro-American	X family	X1
BK15	Lineage 4 Euro-American	X family	X1
NPTB6	Lineage 4 Euro-American	X family	X1;X3
LL1	Lineage 4 Euro-American	X family	X1;X3
LL2	Lineage 4 Euro-American	X family	X1;X3
LL3	Lineage 4 Euro-American	X family	X1;X3
LL4	Lineage 4 Euro-American	X family	X1;X3
BK16	Lineage 4 Euro-American	X family	X3;X1
BK17	Lineage 4 Euro-American	X family	X3;X1
BK18	Lineage 4 Euro-American	X family	X3;X1
BK19	Lineage 4 Euro-American	T family	T1
BK20	Lineage 4 Euro-American	Latin American Mediterranean	LAM11-ZWE
BK22	Lineage 1 Indo-Oceanic	Indo-Oceanic	EAI3;EAI5
BK23	Lineage 4 Euro-American	Haarlem	T1;H1
BK24	Lineage 4 Euro-American	T family	T2-Uganda
LL10	Lineage 4 Euro-American	X family	X3;X1
BK25	Lineage 2 Beijing	Beijing	Beijing

Figure 3.4: A phylogenetic tree of the 59 Welsh *M. tuberculosis* isolates based on SNPs at 60 loci adapted to assign isolates to lineages and sub lineages. Isolates are coloured according to their sub lineage assignment, see key in figure 3.4 above for details. The scale bar indicates the genetic divergence relevant to branch length measured in units of nucleotide differences per site across 60 loci. The corresponding spoligotype families are also depicted in the further most right column.

The major lineage*s* could then be divided into 7 sub-lineage*s* and the proportion of each is presented in Figure 3.5. These included the Lineage 4 Euro American sub lineages; Haarlem, T family, X family, H37Rv-like and Latin American Mediterranean (LAM) sub lineages, in addition to the Beijing and Indo-oceanic sub lineages of lineages 2 and 1 respectively.

Of these, the T, X family Euro-American and Haarlem sub-lineage*s* dominated the dataset with 18 (30%), 16 (27%) and 14 (24%) isolates in each respectively.



	Haarlem	T family	Beijing	X family	H37Rv-like	Latin American Mediterranean	Indo-Oceanic
■ No. of isolates	14	18	3	16	4	3	1

Figure 3.5: Number of isolates representing each sub-lineage present within this collection of 59 Welsh *M. tuberculosis* isolates.

Of the T family isolates, 13 of the 18 showed a completely clonal pattern across the 60 SNPs (NPTA1, NPTA2, NPTA4, NPTA5, NPTA6, NPTA8, NPTB2, NPTB3, NPTB4, NPTB5, BK1, BK2 and BK3), correlating to the T1 and T5 spoligotypes. The remaining five (BK24, BK19, NPTA3, NPTB1 and NPTA7) all contained T family-specific polymorphisms but did not cluster clonally as the other did. Of these five non-clustering isolates, only NPTA3 had sequence data for the complete set of 60 loci analysed (Table 3.2). Compared to the clonal isolates, NPTA3 differed at only one locus, specifically genomic position 3836739, and is the reason it could not be resolved into a specific T family spoligotype. Isolates NPTA7, BK24, BK19 and NPTB1 did not have sequence data for the complete set of 60 loci (Table 3.2). It is possible that the lack of sequence data at certain loci is responsible for the divergence seen for these isolates from the clonally-related T family isolates within the phylogeny (Figure 3.4).

Table 3.2: Percentage presence of the 60 loci analysed in the 59 isolates. Isolates with less than 90% of sequence data correlating to the 60 SNP barcode were omitted from this analysis.

Isolate	Loci present (%)	Sub lineage	Isolate	Loci present (%)	Sub lineage
BK1	100	T family	GO6	100	Haarlem
BK2	100	T family	GO7	100	Haarlem
BK3	100	T family	GO8	100	X family
BK4	100	H37Rv-like	GO9	98	X family
BK5	100	H37Rv-like	LL1	100	X family
BK6	98	H37Rv-like	LL2	100	X family
BK7	97	H37Rv-like	LL3	100	X family
BK8	98	LAM	LL4	100	X family
BK9	100	LAM	LL5	100	Haarlem
BK10	100	Haarlem	LL8	100	Haarlem
BK11	100	Haarlem	LL9	100	Beijing
BK12	100	X family	LL10	90	X family
BK13	98	X family	LL11	100	X family
BK14	100	X family	NPTA1	100	T family
BK15	100	X family	NPTA2	100	T family
BK16	100	X family	NPTA3	100	T family
BK17	100	X family	NPTA4	100	T family
BK18	97	X family	NPTA5	100	T family
BK19	97	T family	NPTA6	100	T family
BK20	100	LAM	NPTA7	98	T family
BK21	98	Beijing	NPTA8	100	T family
BK22	98	Indo-Oceanic	NPTB1	98	T family
BK23	97	Haarlem	NPTB2	100	T family
BK24	93	T family	NPTB3	100	T family
BK25	90	Beijing	NPTB4	100	T family
GO1	100	Haarlem	NPTB5	100	T family
GO2	100	Haarlem	NPTB6	100	X family
GO3	100	Haarlem	TH1	100	Haarlem
GO4	100	Haarlem	TH2	100	Haarlem
GO5	100	Haarlem			

Twelve of the 16 X family isolates had 100% sequence data and could be split into three clonally related clusters correlating to spoligotypes: X1 (LL11, BK14, BK15), X2 (GO8, BK12) and X3 (NPTB6, LL1, LL2, LL3, LL4, BK16 and BK17). These clusters differ at locus positions 1850119, 541048 and 4229087 respectively. Despite lacking a full complement of sequence data for the 60 loci, the other four X family isolates, GO9, BK13, BK18 and LL10 all contained polymorphisms for the X family. GO9 and BK13 were assigned to the spoligotypes X1 and X2 groups respectively, whilst BK18 and LL10 diverged from the rest of the X1/X3 isolates as seen in Figure 3.4. Fourteen (24%) isolates clustered as Euro-American Haarlem sub-lineage. Each Haarlem-assigned isolate could be split into those harbouring the polymorphism at loci 62657 and 891756 (LL5), those harbouring polymorphisms at genome positions 891756 and 107794 (TH1, TH2, GO1, GO2, GO3, GO4, GO5, GO6 and GO7) and those harbouring polymorphisms at all three loci: 62657, 891756 and 107794 (LL8, BK10, BK11 and BK23). A further four isolates (BK4, BK5, BK6 and BK7) had a SNP barcode identical to that of H37Rv and were thus classed as H37Rv-like strains, correlating to spoligotype T1. Three of the remaining isolates (LL9, BK21 and BK25) were classified into lineage 2 Beijing isolates and the final isolate (BK22) to lineage 1 Indo-Oceanic.

3.3.4: cgMLST association in the Welsh isolates

Fifty seven of the 66 Welsh isolates originally sequenced had the necessary sequence quality required for cgMLST analysis by the Ridom SeqSphere software and were incorporated into a phylogeny that included 179 isolates previously characterised into lineages by Comas *et al* (2013), see section 3.2.4. Figure 3.6 highlights the association of 57 Welsh isolates with previously lineage-defined ones from the Comas *et al* (2013) study. The figure shows only a subset of the isolates included in the original analysis, with the original cgMLST including all 236 isolates being too large to present, with clear labelling, within this study. The original cgMLST, see Appendix figure PH1, showed Welsh isolates only clustered with isolates of lineages 1, 2 and 4 and thus only those isolates were included in the subset phylogeny shown in figure 3.6.

Lineage 4 associations dominate the dataset, with a total of 53 Welsh isolates clustering with the previously defined lineage 4 isolates. All but one outbreak-associated isolate (with prefixes LL, NPTA, NPTB or GO) clustered with the lineage 4 isolates with only LL9 showing a separate association with lineage 2 isolates. The datas*et al*so included a minority representing lineages 2 (3 isolates) and 1 (1 isolate). The results showed a direct correlation to the SNP bar-coding results, which also highlighted a dominance of lineage 4 strains with a minority of lineages 2 and 1 within the dataset (Figure 3.4). The 3 isolates (LL9, BK21, and BK25) that clustered with lineage 2 ones also typed as lineage 2 by SNP bar-coding (Figure 3.4); the same was true for the lineage 1 isolate. To conclude, cgMLST analysis showed an association of 57 Welsh *M. tuberculosis* isolates with ones that had been previously lineage defined in the previous publication by Gagneux *et al* (2013).

The full phylogeny, including lineages 3, 5, 6 and 7, was also constructed and is shown in Appendix Figure PH1. The phylogeny results show that lineage 5 and 6 isolates were the most ancestral, clustering most closely with the *M. canetti* isolates used to root the phylogeny. Following lineages 5 and 6, lineage 1 isolates harboured the most ancestral position within the phylogeny, branching off before the rest of the *M. tuberculosis* lineages (lineages 2, 3, 4 and 7). Lineage 7 harboured an intermediate position between lineage 1 and lineages 2, 3 and 4 in the phylogeny. Lineage 4 branches off from lineage 2 and 3, suggesting that it was ancestral to lineages 2 and 3. Lineage 2 and 3 then go on to branch off from each other.

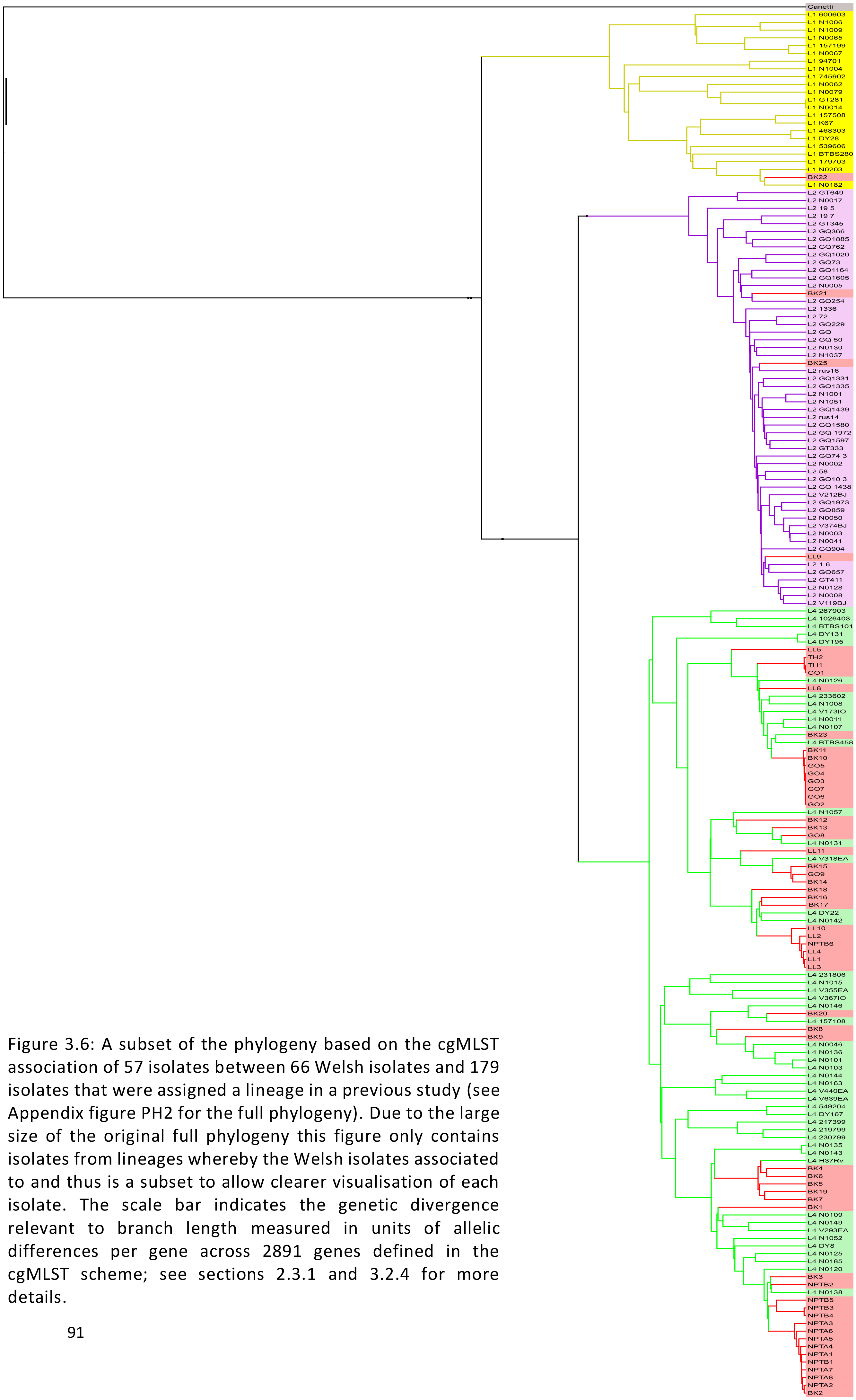


Figure 3.6: A subset of the phylogeny based on the cgMLST association of 57 isolates between 66 Welsh isolates and 179 isolates that were assigned a lineage in a previous study (see Appendix figure PH2 for the full phylogeny). Due to the large size of the original full phylogeny this figure only contains isolates from lineages whereby the Welsh isolates associated to and thus is a subset to allow clearer visualisation of each isolate. The scale bar indicates the genetic divergence relevant to branch length measured in units of allelic differences per gene across 2891 genes defined in the cgMLST scheme; see sections 2.3.1 and 3.2.4 for more details.

3.3.5: Statistical analysis of lineage association

For each of the 57 Welsh isolates, the probability that a given one was related to one of the four major tuberculosis lineage*s* (1, 2, 3 and 4) was calculated (see section 2.11 and 3.2.5 for details). Statistically there was 100% confidence in the lineage assignment of 50 isolates by cgMLST association with 179 lineage-defined isolates. Seven isolates could not be assigned a lineage with 100% confidence (see section 2.11 and 3.2.5 for details). These included: isolates BK25, BK21, BK19, BK6, NPTA6, GO9 and LL10. Further details on the confidence statistics for each isolate can be found in appendix table PH4.

To conclude, cgMLST analysis ascribed 50 out of 57 Welsh isolates to certain lineages with 100% confidence. The cgMLST analysis was not able to ascribe isolates BK25, BK21, BK19, BK6, NPTA6, GO9 and LL10 to a given lineage with 100% confidence.

3.3.6: Supporting genotype information

3.3.6.1: Beijing Genotyping

According to SNP bar-coding, three Beijing lineage isolates were found within the collection, LL9, BK21 and BK25 (Figure 3.4). PGG, cgMLST association and SNP bar-coding identified isolate LL9 as PGG1 (Figure 3.1), lineage 2 (Figure 3.6), and Beijing sub-lineage (Figure 3.4). PGG and SNP bar-coding assigned isolates BK21 and BK25 as PGG1 and Beijing sub-lineage. However, despite clustering amongst lineage 2 isolates in the original phylogeny by cgMLST (Figure 3.6), neither could be assigned with statistical confidence to lineage 2 (Appendix table PH4). LL9 and BK25 were also classified as SCG2 isolates (Figure 3.3). Due to the lack of a full SCG SNP profile, BK21 either belongs to SCG2 or SCG3a, see Appendix table PH1 for more details. To further elucidate their lineage 2 and Beijing sub lineage status, the Beijing genotype genetic markers, *mut*2, *mut*4 and *ogt* genes, were investigated (see sections 3.2.6.1). Both LL9 and BK25 harboured all the relevant SNPs associated with the Beijing genotype (Table 3.3). BK21 did not harbour the *mut*T2 arginine (R) at codon position 58, or the A nucleotide at position 36 within the *ogt* gene, both of which are associated with Beijing isolates, but did harbour the *mut*T4 gene polymorphism for Beijing genotype assignment (Table 3.3). Analysis of established Beijing genetic markers in Table 3.3 supports the original assumption by SNP bar-coding (Figure 3.4) that isolates BK21 and BK25 are Beijing genotype isolates and supports their cgMLST association with lineage 2 as postulated by their clustering positions in figure 3.6.

Table 3.3: A SNP analysis of the 3 well-characterised Beijing genotype markers in the 3 Welsh isolates assigned by SNP bar-coding as Beijing isolates.

Beijing genotype Single nucleotide polymorphisms				
Isolates	***mut*T2**	***mut*T4**	***ogt***	**SCG**
LL9	R	G	A	SCG2
BK21	G	G	G	SCG 2/3a
BK25	R	G	A	SCG2
Polymorphism: Ref>Beijing	G > R at codon position 58	C > G at nucleotide position 142	G > A at nucleotide position 36	

To further investigate the lineage 2 Beijing status of isolates LL9, BK21 and BK25, further assessment, using 48 well characterised Beijing strain-specific SNPs, was carried out (section 3.2.6.1.1). All 3 isolates contained numerous lineage 2 Beijing specific polymorphisms, which are shown fully in Appendix table PH2. The analysis provided further evidence confirming isolates LL9, BK21 and BK25 as being lineage 2 Beijing ones. In addition, the analysis across 48 loci provided further resolution into which Beijing strains showed the closest genomic relationship with isolates LL9, BK21 and BK25. Within the phylogeny analysis, both LL9 and BK25 showed a clonal relationship across the 48 SNPs, branching off together and sharing a recent branching event with BmyC20 and BmyC21 Beijing strains (Figure 3.7). Further analysis also showed the Beijing strain BmyC10 was the closest relative across the 48 defined SNPs to LL9 and BK25 (only 2 SNP differences in both cases; Appendix table PH3a and PH3b). However, BK21 differed in its phylogenetic position from LL9 and BK25. BK21 was descended from a more ancestral branching event within the phylogeny, sharing a recent branching point and clustering with BmyC7 and Bmyc25 (Figure 3.7). Further analysis identified BmyC25 as BK21s closest relative amongst the known Beijing strains analysed (sharing only 2 SNP differences, see Appendix Table PH3a and PH3b).

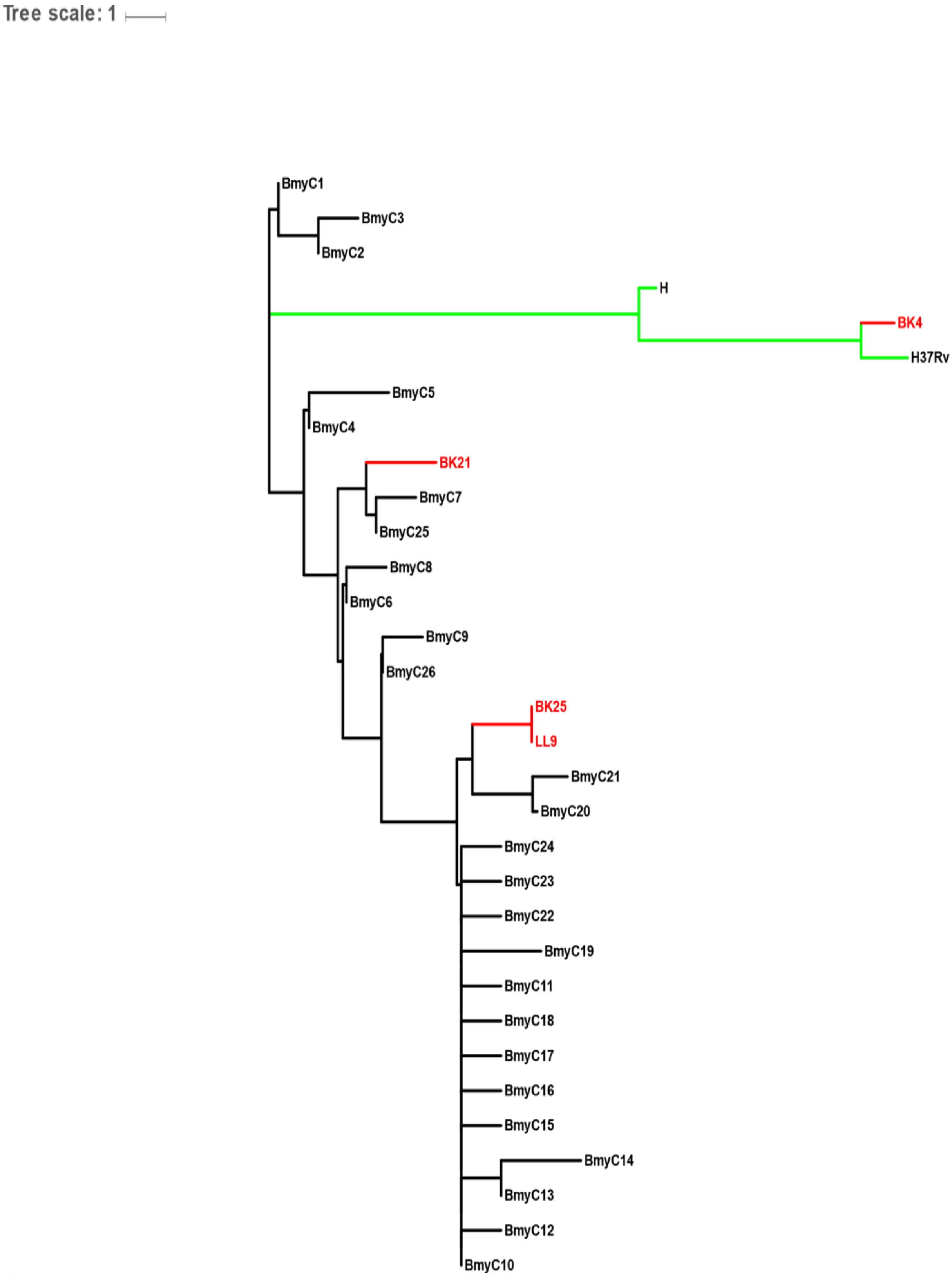


Figure 3.7: Neighbour joining phylogeny tree based on 48 SNPs containing Welsh isolates LL9, BK25 and BK21 and 26 well characterised Beijing isolates from the Mestre *et al*. (2011) study. The scale bar represents the genetic distance between isolates which is measured in units of nucleotide differences across 48 SNP positions. Three isolates known to not be Beijing strains were included as controls, a known Haarlem strain (H), a non-Beijing Welsh isolate (BK4) and H37Rv, the reference strain.

3.3.6.2: Latin American Mediterranean (LAM) genotyping

Three isolates within the dataset, BK8, BK20 and BK9, were assigned LAM sub-lineage status by SNP bar-coding (Figure 3.4). These harboured the LAM-specific polymorphism (G>A) at nucleotide position 309 within Rv0129c (Table 3.4) and had been assigned previously to the SCG5 cluster group (Figure 3.3). See section 3.2.6.2 for more details.

Table 3.4: Table showing the presence or absence of a SNP at position 309 of gene Rv0129, for the isolates designated as a LAM sub-lineage isolates. The correlating SCG is shown in the right column.

Supporting Genotype information for LAM isolates			
Isolate	Rv0129c: Nucleotide at position 309 (G/A)	LAM isolate	SCG
BK8	A	Yes	5
BK20	A	Yes	5
BK9	A	Yes	5

3.3.6.3: Haarlem genotyping

Fourteen isolates were assigned Haarlem sub-lineage status by SNP bar-coding (Figure 3.4). The Haarlem polymorphisms within genes *mtg*C, *ogt* and *ung* were examined across the 14 isolates, see section 3.2.6.3. All but one harboured each of the three Haarlem-specific polymorphisms (Table 3.5). LL5 showed no Haarlem-specific polymorphism and thus its assignment as a Haarlem strain was not supported by genotyping.

Table 3.5: Table showing the Haarlem polymorphisms within genes *mtg*C, *ogt* and *ung* in 14 Welsh isolates.

Supporting Genotype information for Haarlem family isolates				
Isolate	***mtgC***	***ogt***	***ung***	**SCG**
LL5	Arg	C	G	SCG 3b
LL8	His	G	A	SCG 3b
BK23	His	G	A	SCG 3b
BK10	His	G	A	SCG 3b
BK11	His	G	A	SCG 3b
TH1	His	G	A	SCG 3b
TH2	His	G	A	SCG 3b
GO1	His	G	A	SCG 3b
GO2	His	G	A	SCG 3b
GO3	His	G	A	SCG 3b
GO4	His	G	A	SCG 3b
GO5	His	G	A	SCG 3b
GO6	His	G	A	SCG 3b
GO7	His	G	A	SCG 3b
Reference > Haarlem mutation	Arg (CGC) > His (CAC) at codon position 182	C > G at nucleotide position 44	G > A at nucleotide position 501	

3.3.6.4: X family genotyping

Sixteen X family sub-lineage isolates were identified within the Welsh isolates by SNP bar-coding. SNP cluster grouping was not uniform across the X family-assigned isolates, with six being of the SCG-3c group and 10 being of the SCG-4. The results in Table 3.6 show SCG-3c and SCG-4 isolates do not differ in terms of their X family polymorphisms despite being from different SNP cluster groups. See section 3.2.6.4 for more details.

Table 3.6: Presence of X family specific SNPs at loci Rv3221c and Rv2330 in the 16 Welsh isolates assigned as X family genotypes by SNP bar-coding.

X clade supporting information			
Isolate	**Rv3221c (n = 30)**	**Rv2330 (n = 426)**	**SCG**
GO8	A	T	SCG 3c
GO9	A	T	SCG 3c
NPTB6	A	T	SCG 4
LL1	A	T	SCG 4
LL2	A	T	SCG 4
LL3	A	T	SCG 4
LL4	A	T	SCG 4
BK14	A	T	SCG 3c
BK18	A	T	SCG 4
BK15	A	T	SCG 3c
BK12	A	T	SCG 3c
BK13	A	T	SCG 3c
BK16	A	T	SCG 4
BK17	A	T	SCG 4
LL10	A	T	SCG 4
LL11	A	T	SCG 4
X family mutation (Reference> X family)	G > A	C > T	

3.3.6.5: Lineage 1 Genotyping

Only one isolate, BK22, was assigned to lineage 1. Further genotyping confirmed BK22 as a lineage 1 non-Manila strain (Table 3.7), as it harboured six out of eight lineages 1 non-Manila related polymorphisms and no lineage 1 Manila-related polymorphisms. See section 3.2.6.5 for more details.

Table 3.7: Analysis of 17 SNPs specific to lineage 1 genotypes. The analysis was split into two sections, one showing the polymorphisms relevant to lineage 1 non-Manila strains and second showing polymorphisms correlating to lineage 1 Manila strains. Polymorphisms are highlighted in red.

Single nucleotide polymorphisms for lineage **1 genotyping**				
Non-Manila	**Gene**	**Nucleotide position**	**Mutation**	**BK22**
	Rv0005	**990**	**G > C**	**G**
	Rv0006	**1151**	**C > T**	**T**
	Rv0410c	**1842**	**G > A**	**A**
	Rv0934	**1022**	**C > T**	**T**
	Rv1996	**52**	**C > T**	**T**
	Rv2462c	**1086**	**T > C**	**C**
	Rv3132c	**1680**	**G > C**	**C**
	Rv3221c	**85**	**G > A**	**G**
Manila	**Gene**	**Nucleotide position**	**Mutation**	**BK22**
	Rv0006	**1959**	**G > C**	**G**
	Rv0164	**415**	**C > A**	**C**
	Rv0288	**28**	**G > A**	**G**
	Rv0410c	**2117**	**T > C**	**T**
	Rv1009	**724**	**G > A**	**G**
	Rv1996	**157**	**G > C**	**G**
	Rv2030c	**1137**	**G > A**	**G**
	Rv2031c	**426**	**C > T**	**C**
	asRv3261	**15**	**T > C**	**T**

3.3.7: Phylogenetic composition of outbreak isolates

The phylogenetic composition of outbreak isolates was analysed. PGG`s 2 and 3 accounted for 58 and 42% of the outbreak isolates respectively, whilst 3 SCG`s were present amongst the outbreak related isolates (Table 3.8). SCG-6 and SCG-3 dominated the dataset with 43 and 40% of the outbreak isolates respectively, with SCG-4 making up 17% of the dataset. The T family sub-lineage dominated the outbreak isolates with 39%, followed by the Haarlem sub-lineage which accounted for 33% and the X family which accounted for 27% of the isolates.

Table 3.8: Composition of the outbreak-related isolates with regards to their PGG, SCG and sub-lineage assignments.

Composition of outbreak isolates (%)								
PGG			SCG			Sub-Lineage		
PGG1	PGG2	PGG3	SCG6	SCG3	SCG4	Haarlem	T family	X family
0	18(58)	13(42)	15(43)	14(40)	6(17)	11(33)	12(39)	9(27)

3.4: Discussion

This chapter has provided the first insight into the phylogenetic diversity within a collection of *M. tuberculosis* isolates from South West Wales. Major lineage assignment was carried out successfully on 59 of the 66 isolates by the use of an established and robust SNP barcode method (Coll *et al*., 2014). The Welsh dataset was dominated by the Euro-American lineage (lineage 4), which accounted for 55 of the 59 isolates, and included a variety of different Euro-American sub-lineages, including the Haarlem genotype. PGG, SCG and strain-specific genotyping supported the lineage and sub-lineage assignments within this study, providing the first correlations of these methods with the established SNP barcode method. This study found that a gene-by-gene cgMLST method has the potential to provide reliable phylogenetic assignment of isolates and could provide an alternative method to traditional SNP mapping. However, the cgMLST results need to be supported by others to obtain reliable phylogenetic assignments. The predominance of the Euro-American lineage observed was not unexpected as this lineage is known to predominate across Europe (Gagneux *et al*., 2006, Ojo *et al*., 2010 Fenner *et al*., 2012, Fitzgibbon *et al*., 2013) and the proportion of Euro-American lineage strains here is similar to Public Health England data for TB cases in indigenous people (those born in each respective country) across the whole of the UK and Ireland (Ojo *et al*., 2010, , Fenner *et al*., 2012, Fitzgibbon *et al*., 2013, Public Health England., 2014). This study also identified 2% of the isolates as lineage 1 and 5% as lineage 2, again correlating with data for the indigenous population of the UK (Public Health England., 2014) and Ireland (Fitzgibbon *et al*., 2013, Ojo *et al*., 2010). A low prevalence of lineage 2 Beijing strains has also been found in Swiss and Swedish studies (Ojo *et al*., 2010, Ghebremichael *et al*., 2010, Fenner *et al*., 2012, Fitzgibbon *et al*., 2013).

Based on a previously published method (Coll *et al*.,2014), the SNP bar-coding results provided major and sub-lineage assignments and showed that the sub-lineage composition of the Welsh dataset was diverse, which correlated well with other studies (Ojo *et al*., 2010, Fenner *et al*., 2012, Fitzgibbon *et al*., 2013). T, X and Haarlem sub-lineage*s* accounted for 30%, 27% and 24% of the isolates respectively, whilst the Latin American Mediterranean, Beijing, H37Rv-like and Indo-Oceanic strains accounted for 5, 5, 7 and 2% respectively. A similar South West Ireland study also highlighted a high proportion of X family strains (26%) and low levels of LAM (4.1%) and Beijing (5.9%) strains (Ojo *et al*., 2010). However, the Welsh data described in this chapter showed a much higher proportion of the T and Haarlem sub-lineage*s* than reported in the South West Ireland study (Ojo *et al*, 2010). Also, in contrast to these results, another more recent Irish study found that LAM isolates predominated, including amongst outbreak ones (Fitzgibbon *et al.*, 2013). LAM isolates are most commonly found in the populations of Brazil, Africa and certain regions of the Mediterranean (David *et al.*, 2012, Luiz Claudio Oliveira Lazzarini., 2012). Thus, it would be interesting to further investigate the presence of the LAM genotype in Wales and to assess whether the genotype is related to a certain target group, which could aid future control programmes.

The outbreak cases within this dataset were dominated by T, Haarlem and X Sub-lineage*s*.

Previous studies have highlighted that the Euro-American lineage correlates with PGG2 and PGG3 (Gagneux *et al.*, 2007, Reed *et al.*, 2009, Zeinab *et al.*, 2009, Rindi *et al.*, 2014), and lineage*s* 1, 2 and 3 have been shown to be associated with PGG1 (Gagneux *et al.*, 2007, Reed., 2009, Ramazanzadeh *et al.*, 2009, Rindi *et al.*, 2014). The results here are consistent with these previous studies, as each Euro-American lineage isolate was categorised as either PGG2 or PGG3, with Lineage 1 and 2 isolates classified as PGG1. The T Family and H37Rv-like isolates correlated only with PGG3, whilst the Haarlem, X family and LAM sub-lineage isolates grouped with PGG2. Analysis of the PGG data provided the first insight into the evolutionary makeup of the dataset. The evolution from PGG1 to PGG3 is sequential (Gutierrez *et al*., 2006), meaning PGG1 isolates represent ancient isolates, PGG2 isolates an intermediate group and PGG3 the most modern group (Sreevatsan *et al*., 1997., Millán-Lou *et al*., 2013). In this study, isolates of the Haarlem, X and LAM sub-lineages were assigned to PGG2. The ill-defined T family (PGG3) represents the more modern strains, at least within the dataset used in this study, thus providing data that are an extension to results presented in previous studies (Sreevatsan *et al*., 1997, Millán-Lou *et al*., 2013) as they did not describe an association between certain sub-lineages and PGGs. Isolates of PGG3 within this study clustered clonally and made up 42% of the outbreak isolates. Thus, these observations agree with the Millán-Lou *et al*. (2013) study, but contrast with the earlier work of Sreevatsan *et al* (1997) that PGG3 isolates are capable of being associated with clustered cases of *M. tuberculosis* and do not represent purely sporadic ones. SCG and sub-lineage-specific genotyping were also applied to this study. SCGs have been shown previously to correlate with spoligotype families, which in turn correlated with the sub-lineage assigned by the robust SNP barcode used here (Filliol *et al.*, 2006, Gagneux *et al.*, 2007). The results revealed a predominance of SCG-3 and SCG-6 isolates, with SCG-3b and SCG-6a being the most common. Unlike for PGG, the SCG analysis highlighted a large degree of divergence within the Euro-American lineage, consistent with the diversity seen in the SNP barcode result. The SCG analysis highlighted a predominance of SCG-3b (15 isolates) and SCG-6a (16 isolates) in particular. Interestingly, all SCG-3b isolates were identified as Haarlem lineage strains by SNP bar-coding, thus agreeing with previous work that showed that SCG3b links with the Haarlem spoligotype (Filliol *et al.*, 2006). In addition, Haarlem-specific genotyping confirmed that at least 14 isolates within this study can be confidently assigned to the Haarlem sub-lineage of *M. tuberculosis*. All LAM assigned isolates grouped as SCG-5, which correlated well with previous evidence (Filliol *et al.*, 2006). The X and T family dominated the dataset, with all X family sub-lineage isolates being assigned to SCG-3c or SCG-4, as previously described (Filliol *et al.*, 2006, Gagneux and Small., 2007). X family-specific genotyping supported the findings of the SNP bar-coding results and found a specific association between particular X family spoligotypes, SCG-3c and SCG-4, that has not been shown previously. Essentially the study discovered that X family isolates harbouring the X1 or X2 spoligotype SNP barcode, exclusively correlated with the SCG-3c group, whilst, X family isolates harbouring the X1 or X3 (X1; X3) spoligotype were exclusively SCG-4 isolates (See Appendix Table PH1).

Four isolates were identified as not Euro-American lineage ones. Three represented lineages 2, SCG-2, and one represented lineage 1, SCG-1. Both SCG-1 and SCG-2 have been described previously as correlating with PGG1 (Filliol *et al*., 2006). Unfortunately, the SCG of one isolate, BK21, could not be fully determined due to the lack of sequence data; it could either be a SCG-2 or SCG-3a strain. However, BK21 was assigned to PGG1 and further analysis found it had a SNP barcode pattern associated with the lineage 2 Beijing sub-lineage and cgMLST analysis clustered it with Lineage 2 and it also contained Beijing-specific polymorphisms thus making it likely to have been part of SNP cluster group 2. Hence, these isolates demonstrated how the use of multiple methods can aid phylogenetic classification of isolates.

In the context of the outbreaks included in this study, identifying the PGG and SCG groups of outbreak isolates and the spoligotype families and major lineages they associate to provides the first insight into the characteristics of the strains that compose the outbreak isolates. The dominance of the PPG2 and PGG3 groups provided the first indication that the outbreaks within this study were composed of strains of a more modern nature and are associated to the Euro-american major lineage. In addition, the SCG analysis highlighted that the isolates of these outbreaks are associated to particular spoligotype clades. The SCG results supported the Euro-american association first indicated by the PGG results and provided further information, showing that outbreak isolates correlated to the spoligotype families T, X and Haarlem. The SCG analysis showed that the outbreak isolates within this database included those associated to the Haarlem genotype which is known to include virulent strains (Mardassi *et al*., 2005, Marais *et al*., 2006, Khanipour *et al*., 2016). These robust classical SNP set analyses provided the first pieces of evidence that the outbreaks within this study were composed of physiologically modern Euro American lineage tuberculosis strains and indicated that a portion of outbreak isolates are associated to a genotype of known virulence (Haarlem genotype).

As stated previously, gene-by-gene MLST methods have proved useful in clinical outbreak resolution and epidemiological investigations of human pathogens such as MRSA and *Campylobacter*, as well as *M. tuberculosis* itself (Sheppard *et al.*, 2012, Maiden *et al.*, 2013). Specifically, the Ridom SeqSphere gene-by-gene cgMLST scheme has been used previously to look at tuberculosis outbreaks (Kohl *et al.*, 2014), and heralds the advent of a portable, standardised database platform for the use of WGS data in tuberculosis research. However, the method has not been used previously for classification of *M. tuberculosis* into well-defined phylogenetic lineages. The cgMLST analysis provided confident assignments for 50 out of 57 Welsh *M. tuberculosis* isolates according to statistical analysis, and statistical analysis found that 48 of them had 100% probability of being Euro-American lineage 4 isolates, correlating well with the robust SNP barcode results. The association of isolates BK25 and BK21 to lineage 2 was not statistically supported, despite their clustering positions in the cgMLST phylogeny indicating an association with lineage 2 (figure 3.6). However, the PGG, SCG and SNP barcode results all supported the assignment of BK21 and BK25 to lineage 2 and further genotyping confirmed the presence of lineage 2 Beijing specific SNP`s within them. Therefore, the evidence suggests that currently, further results are needed to confirm the associations proposed by the cgMLST analysis and further work is needed to develop the cgMLST method and the way it is interpreted if it is to be used in a widespread manner for phylogenetic assignments as a standalone method. This study also presents the first phylogenetic tree containing all seven major *M. tuberculosis* lineages constructed based on a gene-by-gene method (Appendix figure PH1), as opposed to SNP mapping-based methods (Homolka *et al*., 2012, Gagneux *et al*., 2012, Comas *et al*., 2013). The structure of the resulting phylogeny broadly matched that seen for the same isolates (Comas *et al.,* 2013), as in that study, lineages 1, 5 and 6 harboured the most ancestral positions in the phylogeny. Furthermore, cgMLST showed that lineage 7 isolates harboured an intermediate position between the “ancient lineages” (1, 5, and 6), and the modern lineages (2, 3 and 4), supporting its intermediate evolutionary position described previously (Comas *et al*., 2013, Yimer *et al*., 2015). Consistent with Comas *et al* (2013), lineage*s* 2 and 3 isolates shared a closer relationship with each other than with lineage 4 isolates. Therefore, despite using a separate set of genomic data, the evolutionary positions of each lineage according to cgMLST was consistent with other studies of a similar nature which used in-house SNP mapping pipelines for the construction of their phylogenies (Wirth *et al*., 2008b, Gagneux *et al.*, 2012, Firdessa *et al*., 2013, Stucki and Gagneux., 2013, Comas *et al*., 2013).

To confirm the lineage and sub-lineage results, specific genotyping markers for the Beijing, Haarlem, X family, LAM and lineage 1 sub-lineages were also investigated and supported the SNP bar-coding, PGG and SCG results. Specifically, strain-specific genotyping supported all LAM, Haarlem and X family assignments made by both SNP bar-coding. The results also showed that isolates LL9, BK21 and BK25 harboured Beijing strain-specific polymorphisms, supporting their assignment as lineage 2 isolates by SNP bar-coding.

Additionally, strain-specific genotyping showed isolates LL9 and BK21 to be closely related to the Bmyc10 Beijing family of strains, which are the predominant group of globally disseminated Beijing strains (Mestre *et al.*, 2011). BK25 was identified as more closely related to the ancient Beijing genotype Bmyc25, which again is disseminated globally and renowned as the causative agent of virulent outbreaks of *M. tuberculosis* (Caminero *et al.*, 2001). Thus, two separate Beijing strains are present within this dataset, with BK25 potentially being of high virulence and worth further investigation.

3.5: Conclusions and future perspectives

This chapter has described the use of classical PGG, SCG, robust SNP bar-coding, novel cgMLST analysis and strain-specific genotyping to analyse the phylogenetic make-up of *M. tuberculosis isolates* from across South West Wales. The isolate collection was found to be dominated by Euro-American lineage isolates, with lineage*s* 1 and 2 also being present, but in low numbers. The isolates consisted of a diverse collection of Euro-American sub-lineages, which were not clearly dominated by a single group. T family, X family and the Haarlem family made up a large proportion of the dataset, with the Haarlem isolates being particularly prevalent within the outbreak-assigned cases. The dominance of the Euro-American lineage, as established by the SNP barcode, was supported by all analyses, including the novel cgMLST method. Although successfully providing assignment, with 100% confidence, to 50 isolates within this study, the cgMLST could not provide confident assignment to 7 isolates which were assigned to lineages and sub-lineages by SNP car-coding.

This study provided only a snapshot of the *M. tuberculosis* diversity seen within Wales. Extending the number of samples would provide further data on the phylogenetics of *M. tuberculosis* across Wales in future. In addition, *M. tuberculosis* samples should be taken from across the country. The discovery of numerous Haarlem sub-lineage strains, and some Beijing strains also, was an interesting finding and further investigation into the epidemiology of such isolates in Wales would be worthwhile.

In conclusion, this study was the first to use a gene-by-gene method for phylogenetic analysis of *M. tuberculosis* isolates, clearly demonstrating the potential of cgMLST to provide a more convenient alternative to SNP mapping methods for phylogenetic analysis and global tuberculosis surveillance, especially as WGS of *M. tuberculosis* becomes more widespread in the future. Further work into perfecting the method for phylogenetic purposes is required to reduce the anomalies found in this study. In summary, the use of multiple WGS-based methods has provided an accurate and reliable snapshot of the phylogenetic diversity seen within South West Wales and demonstrated a novel use for cgMLST *M. tuberculosis* typing.

Chapter 4

Analysis of an outbreak of *M. tuberculosis* in the Llwynhendy area of South West Wales using Whole Genome Sequencing.

4.1: Introduction

Between 2009 and 2011, 11 cases of *M. tuberculosis* infection were reported in the Llwynhendy region of South West Wales. Each isolate was genotyped by MIRU-VNTR at the Public Health Wales Molecular Unit, Cardiff, using standard methods (section 2.1). Of the 11 isolates, nine had an identical MIRU-VNTR pattern and were deemed to be part of an outbreak related to a local public house. The remaining two isolates were originally excluded from the outbreak investigation as they differed in MIRU-VNTR type from the outbreak strains, although one did have strong epidemiological links to the outbreak, with the other being linked based only on geography and chronology.

4.1.1: Epidemiology of Cases

On 11th August 2010 the landlord of a Llwynhendy public house was diagnosed with pulmonary tuberculosis (LL1). The landlord had had symptoms of active disease since January 2010 but did not seek medical attention until later that year. He was part of a pool team that travelled to various public houses within the nearby Llanelli and Gorseinon areas, which were also included in the epidemiological investigation. Screening of contacts during August 2010 showed that the landlord's wife (LL2) had contracted clinical tuberculosis, with a positive culture obtained in October 2010. She was a teaching assistant at a local primary school and, in addition, worked as a part-time chef at the public house. Public health authorities continued to investigate recent cases that might relate to the public house ones. Further investigations and contact tracing identified a neighbour (LL4) of the public house with tuberculosis and the mother-in-law of the landlord. The mother-in-law regularly socialised and holidayed with the landlord and his wife but did not frequent the public house and no further epidemiological links were found with anyone else within the local area.

Retrospective analysis identified another potential case connected with the public house outbreak. LL5 was an elderly lady diagnosed with tuberculosis in March 2010, with onset of symptoms dating back to November 2009, who regularly visited the public house, but did not live in its immediate area (LL5 is labelled green in figure 4.1). In addition, she also visited public houses in the Gorseinon region just east of Llwynhendy, where another outbreak of tuberculosis was being investigated (Chapter 6). A further two cases (LL6, LL7) were reported in late 2010. Both had visited the Llwynhendy public house around the time of the initial outbreak. Epidemiological investigations also identified another case (LL9) diagnosed in May 2010 who resided close to the public house. This individual was a recent immigrant who had only arrived in the UK in 2009. Another individual (LL10) from the Burry Port area, approximately 6 miles west of Llwynhendy (Figure 4.1), was diagnosed with tuberculosis in February 2011. In this case, the individual died from tuberculosis the same month, thus the epidemiological investigation was limited, but it was able to establish that the individual did commute to public houses within the Llanelli and Llwynhendy area. A final pub-related case (LL11) was identified in September 2011. Epidemiological investigation showed that LL11 lived near the public house during the outbreak, prior to moving to a different part of the region. However, the individual still visited the public house once a week.

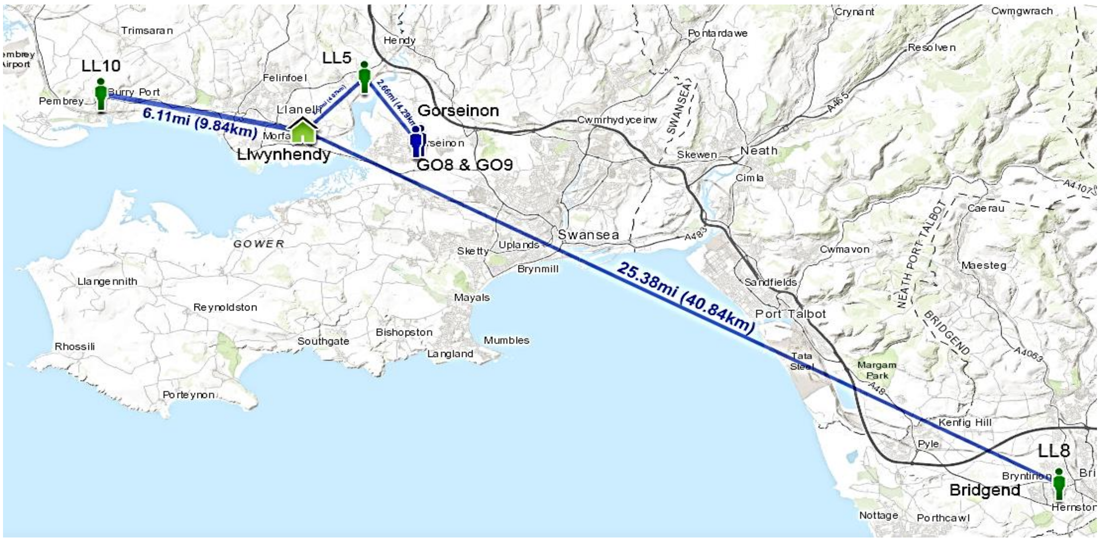


Figure 4.1: A map showing the locations of where potential cases associated with the Llwynhendy outbreak resided across South West Wales. Cases LL1, LL2, LL3, LL4, LL7 and LL11 all resided next to the public house are represented by a house symbol; and LL5, LL8 and LL9 are shown as a green person. The 2 cases shown in blue were associated with the nearby Gorseinon outbreak. mi=miles, km=kilometres. Figure constructed using National Geographic MapMaker (https://mapmaker.nationalgeographic.org/#/).

All the isolates described above were typed using MIRU-VNTR and an identical pattern was obtained for LL1, LL2, LL3, LL4, LL5, LL6, LL7 and LL11 (Figure 4.2). In addition, the MIRU-VNTR typing matched an isolate from an individual (LL8), diagnosed in November 2010 that lived in Bridgend, 40 miles east of Llwynhendy (Figure 4.1). Despite the identical MIRU-VNTR pattern, epidemiological investigations could find no link between LL8 and the Llwynhendy outbreak. The recent immigrant LL9's isolate also did not match the outbreak strains despite the close geographic association, so the epidemiological team concluded that this was an unrelated case brought into the area from the patient's country of origin. LL10, the Burry Port case, also did not match the main outbreak according to MIRU-VNTR. Two Gorseinon outbreak cases (GO8 and GO9, discussed further in Chapter 6), were thought to have a potential epidemiological link to LL5 despite being diagnosed in 2007 and 2008, but were found to have isolates with different MIRU-VNTR types. A summary of the epidemiological information on each case is outlined in Table 4.1, with LL1 (the landlord) suspected as the most likely source of the outbreak and potentially was a super-spreader.

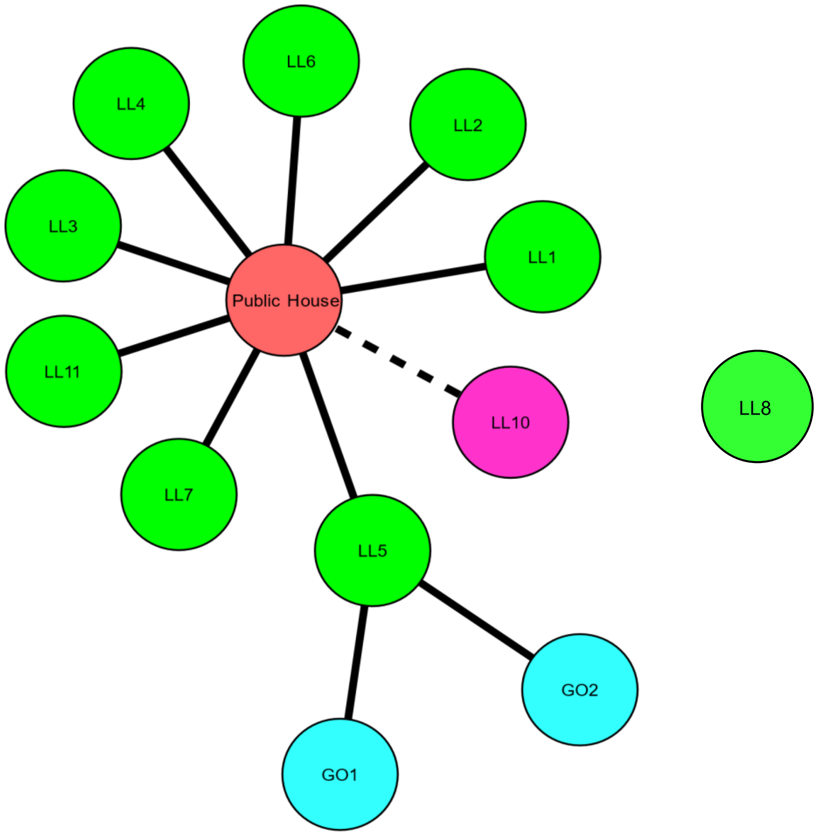


Figure 4.2: Diagram showing the epidemiological relationship between each isolate with a confirmed or potential epidemiological link to the outbreak. Isolate LL9 had a different MIRU-VNTR type and no epidemiological link to the public house or any other isolate within the outbreak is excluded from this diagram. Different colours indicate different MIRU-VNTR types.

Table 4.1: Overview of the basic epidemiology and MIRU-VNTR results on the Llwynhendy tuberculosis outbreak cases, colour coded according to MIRU-VNTR including the two isolates from the Gorseinon outbreak.

Isolate	Location	Date of Onset/ Diagnosis	Attended the public house?	Comments	MIRU-VNTR match?
LL1	Llwynhendy	Jan-10/Aug-10	Yes	Public house landlord	Yes
LL2	Llwynhendy	Aug-10	Yes	Wife of landlord	Yes
LL3	Llwynhendy	Apr-11	No	Mother-in-law of landlord	Yes
LL4	Llwynhendy	Aug-10	Yes	Neighbour and regular of the public house	Yes
LL5	Llangennech	Nov-2009/Mar-2010	Yes	Weekly regular of the public house, also connections with public houses associated with the Gorseinon outbreak	Yes
LL6	Llwynhendy	Late 2010	Yes	Frequented the public house	Yes
LL7	Llwynhendy	Late 2010	Yes	Frequented the public house	Yes
LL8	Bridgend	Nov-10	No	No epidemiological association with the outbreak	Yes
LL9	Llwynhendy	May-10	No	Recent immigrant residing close to the public house.	No
LL10	Burry Port	Feb-11	Unknown	Potentially linked to the outbreak public house	No
LL11	Llanelli	Sep-11	Yes	Frequented the public house and recently moved from Llwynhendy to Llanelli	Yes
GO8	Gorseinon	Oct-2006/ Jan-2007	No	Potentially linked to LL5	No
GO9	Gorseinon	Feb-2007/April 2008	No	Potentially linked to LL5	No

4.1.2: Chapter aims

There were several aims in this chapter. The primary aim was to resolve the epidemiology of the Llwynhendy outbreak using WGS data, cgMLST and *in silico* functional prediction software to highlight physiological features of the polymorphisms found within the WGS of the outbreak isolates. Particular attention was given to the LL10 (Burry Port case) and LL8 (Bridgend case) with regards to their inclusion and exclusion from the outbreak, which were disputed by contact tracing. The relationship between LL5 and the two epidemiologically-associated cases from the separate Gorseinon outbreak (GO8 and GO9) were also resolved. The second aim was to identify the source case of this outbreak and exploit WGS data to provide evidence for a potential super-spreader.

4.2: Methods

4.2.1: Sample collection and gathering of epidemiological information

Isolate DNA was obtained as described in section 2.1. The collection of each Llwynhendy related sample was carried out as described in section 2.1. The Epidemiological information described in the introduction was obtained from face-to-face interviews with a nurse from the original PHW contact tracing investigation team and from documents produced during the outbreak investigation.

4.2.2: DNA sequencing

DNA from each isolate was sequenced using an Illumina MiSeq platform as described in section 2.2. The quality of the sequences produced was represented by Ridom SeqSphere as the percentage of “good targets”. The higher the percentage of good targets, the better the initial sequencing and the more genes were successfully aligned to the cgMLST scheme genes.

4.2.3: Core genome MLST

The cgMLST within this analysis was based on the 2891 gene *M. tuberculosis* scheme created by the Ridom SeqSphere software and published previously (Kohl *et al*., 2014). The number of “good targets” varied with the inclusion of certain sequenced isolates with sub-optimal level of quality (sections 2.3 and 2.3.1). The isolate NPTB6, from the Neath Port Talbot outbreak was also included in this analysis, because the phylogenetic results presented in chapter 3 (figure 3.4) assigned NPTB6 to the same sub-lineage as the Llwynhendy outbreak isolates. The resulting neighbour-joining and minimum-spanning trees were produced using the Ridom SeqSphere software. The threshold for direct transmission was set at 12 allelic differences as described previously in section 2.3 (Kohl *et al*, 2014, Walker *et al.*, 2013).

4.2.4: Whole Genome Sequence SNP mapping

Traditional SNP mapping was carried out by using the conserved signature indels (CSI) phylogeny application provided by the Centre of Genomic epidemiology online server, as described in section 2.5.

4.2.5: Functional analysis

Functional analysis of each non-synonymous mutation found within each isolate was carried out using a Provean algorithm, for identifying the significance of amino acid changes within proteins, as defined in section 2.5.3.

4.3 Results

4.3.1 Initial Sequencing results

Nine of the original 13 isolates included in this analysis aligned successfully to over 90% of the cgMLST scheme targets (Table 4.2), and thus were deemed reliable for further analysis. These were isolates LL1, 2, 3, 4, 5, 8, 9, GO8 and LL11. Isolates LL6, LL7, LL10 and GO9 aligned to less than 90% of the cgMLST scheme targets. Including these isolates would have reduced the number of genes within the cgMLST scheme and thus any conclusions on relationships between the isolates would not be valid according to the parameters set in this study for the cgMLST analysis (see section 2.3.1). Therefore, cgMLST analysis was only applied to those with good quality sequences (LL1, LL2, LL3, LL4, LL5, LL8, LL9, GO8 and LL11).

Table 4.2: The percentage of good targets present in each isolate following sequencing.

Isolate	% Good Targets
GO8	98.9
GO9	67.5
LL1	99.3
LL2	92.3
LL4	97.9
LL3	97.7
LL5	92.6
LL6	8.2
LL7	18.4
LL8	99.2
LL9	98.2
LL10	58.1
LL11	90.3

4.3.1.1: cgMLST

Results of the cgMLST analysis on the 9 Llwynhendy isolates and NPTB6 are shown in Figure 4.3. Five isolates (LL1, LL2, LL3, LL4 and NPTB6) were found to share fewer than 12 allelic differences indicating that they were part of an outbreak complex. A further five isolates had >200 allelic differences with all other isolates in the analysis, thus were as not part of the same outbreak. These included 3 isolates with a matching MIRU-VNTR to the public house outbreak (LL5, LL8 and LL11), the Gorseinon isolate GO8 and the suspected un-linked case LL9. The minimum spanning tree in Figure 4.3 showed that LL1 was central to the outbreak, being the closest relative to each isolate within the dataset. With regards to the number of allelic differences between each isolate within the outbreak complex individually (Table 4.3), only the relationship between isolate LL2 and NPTB6 exceeded 12 (the threshold for direct transmission), indicating that isolates from the outbreak had diverged recently. Appendix table LL1 provides a matrix of the allelic differences between all isolates within figure 4.3.

LL10 (the case form Burry Port) was not included in the analysis because only 58.1% sequence alignment with the cgMLST had been obtained. However, for interest, when sequence data for LL10 were included it appeared within the outbreak complex, separated by 3 allelic differences (Appendix figure LL1).

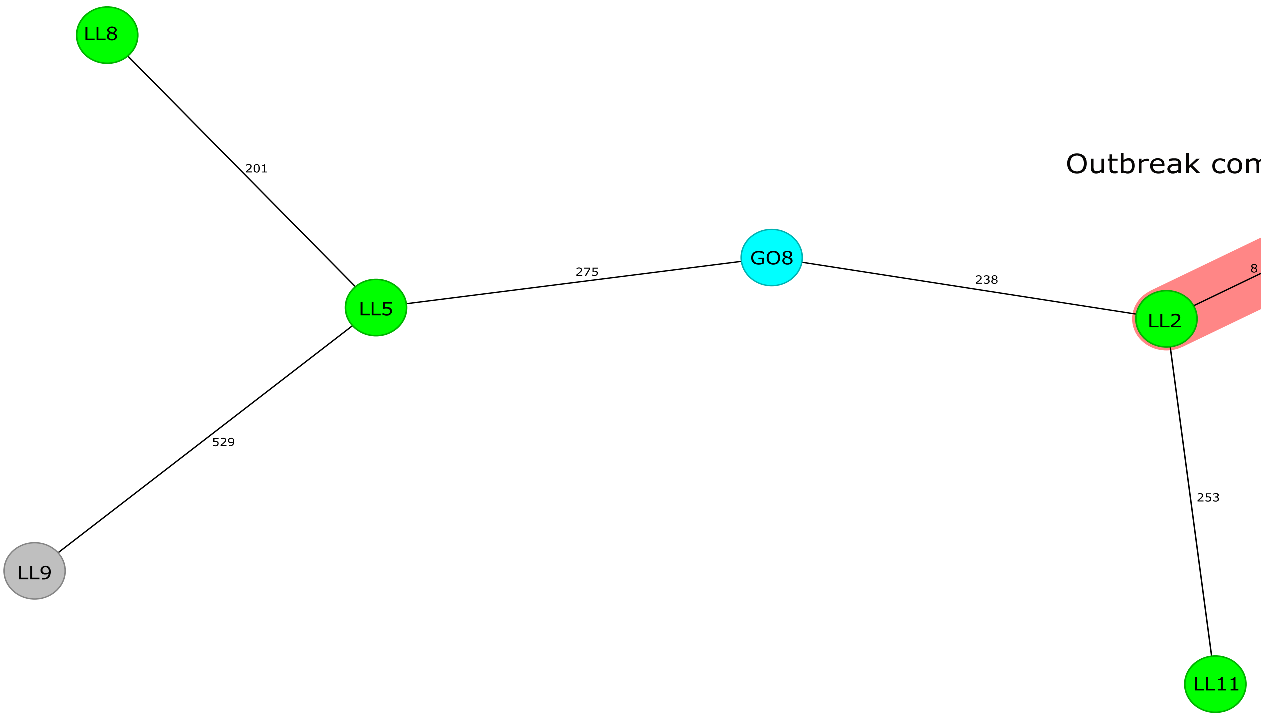


Figure 4.3: A minimum spanning tree showing the number of allelic differences across the cgMLST scheme. I their MIRU-VNTR profile. Branch distances are not to scale. Numbers alongside branches represent the num any two given isolates. The clouded region represents the isolates within the outbreak complex that share l

Table 4.3: Distance matrix showing the number of allelic differences between each of the isolates represented in the outbreak complex defined in Figure 4.3. The greater the red intensity the more different the isolates are from each other. The greater the intensity of blue in the cells the more closely related the isolates are.

Number of allelic differences					
Isolate	NPTB6	LL1	LL2	LL3	LL4
NPTB6	0	7	13	9	11
LL1	7	0	8	2	4
LL2	13	8	0	10	12
LL3	9	2	10	0	6
LL4	11	4	12	6	0

4.3.1.2: Neighbour joining analysis of the outbreak complex

A neighbour joining phylogeny was produced (Figure 4.4) based on only those five isolates that formed the outbreak complex defined in Figure 4.3. The neighbour joining tree showed LL1 to be ancestral to each isolate. Thus, in addition to being the central isolate within the outbreak, isolate LL1 also holds the most ancestral position within the phylogeny and appears to represent the source case. An early branching point separated isolates NPTB6 and LL2 from the cluster containing isolates LL1, LL3 and LL4, and suggests that an intermediate case, was the common ancestor of LL2 and NPTB6, and is likely to have been missed from in this analysis and would have held a position between the source case, LL1, and isolates NPTB6 and LL2.

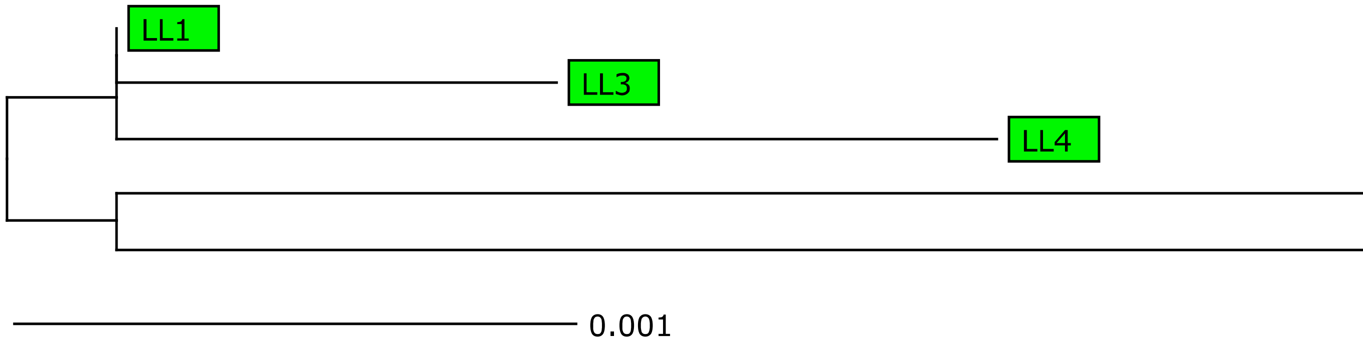


Figure 4.4: Neighbour joining tree constructed using the core genome MLST scheme. The figure contains c defined in Figure 4.3. The scale bar indicates the genetic divergence relevant to branch length measured ir 2891 genes defined in the cgMLST scheme.

4.3.2 Analysis of individual allelic differences

Through analysing each individual allelic difference, it was possible to obtain more information with regards to the pattern of transmission that occurred amongst this closely related complex. A transmission pattern was established by examining at which point certain allelic differences occurred (Table 4.4). Analysis of individual allelic differences supported the results shown in Figure 4.3, indicating that cases LL3 and LL4 were directly infected by LL1, as each carried unique allelic differences in relation to LL1 and thus could not have been infected by one another. Interestingly, an allelic difference at Rv2195 was present in LL2 and NPTB6 but not in isolates LL1, LL3 and LL4 (Table 4.4). The observation corroborates that seen in Figure 4.4, where LL2 and NPTB6 have branched off from LL1, suggesting that the transmission between isolate LL1 and the subsequently more divergent LL2 and NPTB6 isolates was missing an intermediate case that harboured the Rv2195 allele.

Table 4.4: A matrix highlighting each allelic difference between each isolate and the source isolate, LL1. Cells are coloured green where the allelic difference is unique to the isolate, and red represents an allelic difference that occurred in more than one isolate in relation to LL1.

Isolate	LL1							
LL3	Rv0932c	Rv3594						
LL4	Rv0103c	Rv0126	Rv0551c	Rv2137c				
NPTB6	Rv0133	Rv0663	Rv0959	Rv2678c	Rv3635	**Rv2195**		
LL2	Rv1069c	Rv1921c	Rv2117	Rv2153c	Rv2393	Rv2682c	Rv3705A	**Rv2195**

4.3.3: Traditional SNP mapping

Traditional SNP mapping was applied to the five outbreak complex isolates (LL1, LL2, LL3, LL4 and NPTB6) identified in Figure 4.3. The genomic distances between each given isolate in terms of SNPs across the whole genome of each outbreak associated isolate are shown in Figure 4.5 and Table 4.5. SNPs were extracted using the CSI phylogeny application (see methods sections 4.2.4 and 2.5,1 for more details). In each case, the (arbitrary) 12 SNP threshold for direct transmission was exceeded. Thus, based on this parameter no isolate can be deemed as being outbreak-related, according to this WGS SNP mapping result. However, the pattern matched that of the cgMLST results in Figure 4.3. As in the cgMLST phylogeny, isolate LL1 holds the ancestral position within the phylogeny and shows a closer relationship with LL3 and LL4. Also, as in the cgMLST phylogeny, isolates NPTB6 and LL2 are the most distantly related isolates and have a different branching point to isolates LL1, LL3 and LL4 again suggesting the presence of an intermediate case. Therefore, although the divergence between the isolates in the traditional SNP mapping results is greater, the conclusion on the pattern of relationships between the isolates is the same.

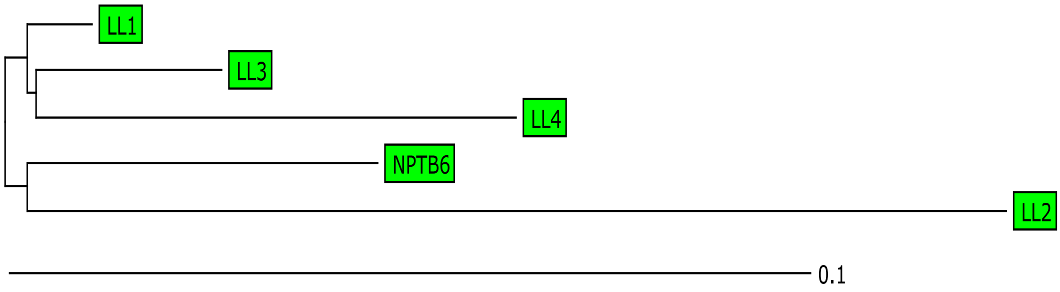


Figure 4.5: A neighbour joining tree based on the WGS data obtained from the conserved signature indels (CSI) phylogeny analysis (see methods section 4.2.4 and 2.5.1 for more details) of the outbreak complex subset. The scale bar indicates the genetic divergence relevant to branch length measured in units of nucleotide differences per site across the WGS of each isolate.

Table 4.5: The number of SNPs present between each isolate defined as being part of the outbreak complex in Figure 4.3. The greater the red intensity the more different the isolates are from each other; the greater the intensity of blue the more closely related they are.

	Number of SNPs across the Whole Genome				
Isolate	**NPTB6**	**LL1**	**LL2**	**LL3**	**LL4**
NPTB6	0	25	72	31	49
LL1	25	0	59	14	30
LL2	72	59	0	67	81
LL3	31	14	67	0	36
LL4	49	30	81	36	0

4.3.4: Functional analysis of the allelic differences within the Llwynhendy outbreak isolates.

Allelic differences were found in 20 loci and the functional effect of each in isolates LL1, LL2, LL3, LL4 and NPTB6, according to Provean functional prediction software (section 4.2.5 and 2.5.3), is summarised in Tables 4.6a and 4.6b. The closest relative to LL1 was LL3, sharing two allelic differences within loci Rv0932c and Rv3594. According to Provean analysis, the allelic difference at locus Rv0932c had no effect on the respective protein function. The allele carried by LL3 for gene Rv3594 contained mutations at six different base positions (Table 4.6b). Each mutation was non-synonymous and according to Provean analysis the effect on the protein was not deleterious. LL4 harboured 4 allelic differences relative to LL1 (Rv0126, Rv0551c, Rv2137c and Rv0103c), one of which was synonymous and three were non-synonymous, with its RV2137c allele having a mutation that caused a deleterious effect on its respective protein (hypothetical protein). The most divergent isolate from LL1 was LL2 with eight allelic differences; one was a synonymous mutation in RV2153c and the remaining 7 were non-synonymous (Rv1069c, Rv1921c, Rv2117, Rv2195, Rv2393, Rv2682c and Rv3705A). According to Provean functional prediction, the effect of each non-synonymous allelic difference in LL2 was neutral. NPTB6 had 7 allelic differences relative to LL1 (Rv0133, Rv0422c, Rv0663, Rv0959, Rv2678c, Rv2678c and Rv2195). These were non-synonymous mutations, with five (Rv0133, Rv0422c, Rv0663, Rv0959, Rv2678c and Rv2195) harbouring mutations predicted to have a deleterious effect on their protein functions. Isolates NPTB6 and LL2 both have a common allelic difference from LL1, at the Rv2195 locus. However, the mutation does not affect the function of the encoding protein.

Table 4.6a: A detailed matrix based on the allelic differences between LL1, LL2, LL3, LL5 and NPTB6. The ma each isolate as designated by the cgMLST scheme, allelic difference are synonymous or non-synonymous, S (aa) changes, and position of nucleotide (nt) changes. “X” refers to the deletions of the given nucleotide l columns refer to the Provean results: Provean score ≤-2.0 = deleterious mutation, >-2.0 = neutral mutation.

		Allele number relevant to the cgMLST scheme								
Locus	GenBank product	LL1	LL2	LL4	LL3	NPTB6	Non-synonymous /Synonymous	SNP (nt)	Position nt change	ch
Rv0103c	cation-transporter P-type ATPaseB	3	3	374	3	3	Synonymous	G>C	1350	N
Rv0126	trehalose synthase/amylase TreS	1	1	166	1	1	Non-synonymous	A>G	1780	T
Rv0133	GCN5-like N-acetyltransferase	1	1	1	1	132	Non-synonymous deletion	ACC>XXX	282-284	Th
Rv0422c	hydroxymethylpyrimidine/ phosphomethylpyrimidine kinase	1	1	1	1	104	Non-synonymous	A>G	652	T
Rv0551c	fatty-acid--CoA ligase FadD8	1	1	270	1	1	Non-synonymous	C>A	62	A
Rv0663	arylsulfatase AtsD	33	33	33	33	408	Non-synonymous	C>A	729	H
Rv0932c	phosphate ABC transporter substrate-binding lipoprotein PstS	1	1	1	151	1	Non-synonymous	ACC>GGT	333-335	I
Rv0959	hypothetical protein	1	1	1	1	354	Non-synonymous	A>C	407	D
Rv1069c	hypothetical protein	1	271	1	1	1	Non-synonymous	T>G	532	I
Rv1921c	lipoprotein LppF	1	214	1	1	1	Non-synonymous	C>A	291	S
Rv2117	hypothetical protein	1	48	1	1	1	Non-synonymous	G>A	281	I
Rv2137c	hypothetical protein	1	1	73	1	1	Non-synonymous	G>A	190	C

		Allele number relevant to the cgMLST scheme							
Locus	GenBank product	LL1	LL2	LL4	LL3	NPTB 6	Non-synonymous /Synonymous	SNP (nt)	Position nt change
Rv2153c	UDP-N-acetylglucosamine-N-acetyl muramyl-(pentapeptide) pyrophosphoryl-undecaprenol-N-acetylglucosamine transferase	17	225	17	17	17	Synonymous	C>A	306
Rv2195	ubiquinol-cytochrome C reductase rieske iron-sulfur subunit	3	1	3	3	1	Non-synonymous	G>A	40
Rv2393	Ferrochelatase	1	95	1	1	1	Non-synonymous	A>G	116
Rv2678c	uroporphyrinogen decarboxylase	1	1	1	1	167	Non-synonymous	G>A,	814
Rv2682c	1-deoxy-D-xylulose 5-phosphate synthase	1	213	1	1	1	Non-synonymous	G>T	659
Rv3594*	hypothetical protein	1	1	1	95	1			
Rv2678c	transmembrane protein	1	1	1	1	277	Non-synonymous	C>A	788
Rv3705A	proline-rich protein	1	123	1	1	1	Non-synonymous	C>A	292

*Rv3594 analysis contains 6 mutations that are presented in Table 4.6b.

Table 4.6b: Mutations observed within the Rv3594 gene of LL3 *represents a stop codon, and X represents t
acid at that position. The columns are the same as in Table 4.6a but exclude the column for each isolates' al

Nucleotide position(s)	Mutation(s)	Non-synonymous /Synonymous	Amino acid change	Amino acid position	Provean sco
812	C>G	Non synonymous	P>R,	271	-0.698
816	X>G(insertion)	Non synonymous	X>R,	272	-0.26
819	A>G	Non synonymous	*>W,	273	0.565
821	C>X(deletion)	Non synonymous	H>X,	274	0.404
826-827	TA>GG	Non synonymous	V>G	276	0.016

4.4: Discussion

This chapter described methods for resolving whether 11 *M. tuberculosis* isolates from the Llwynhendy region of South West Wales were part of the same outbreak. In addition to evaluating the results of MIRU VNTR typing against WGS, this study also aimed to evaluate the epidemiological links made by the contact tracing team. Specifically, the study used a combination of cgMLST, WGS SNP mapping, detailed evaluation of each mutation, and a physiological prediction-based method to evaluate the Llwynhendy outbreak. The application of traditional SNP mapping provided evidence supporting the conclusions of cgMLST analysis, with the resulting phylogeny and pattern of genomic distances between isolates being similar. However, the number of SNPs seen between isolates through mapping was substantially greater than that seen by cgMLST analysis. This greater amount of divergence in SNP mapping in comparison with the cgMLST scheme was documented in a previous study (Kohl *et al*., 2014), although to a lesser degree. Unlike that study, the results presented here found a level of divergence that exceeded the 12 SNP threshold, indicating that there was no direct association between any isolate in this study. Due to the strong epidemiological links, especially between isolates LL1, LL2, LL3 and LL4, together with the cgMLST conclusions that showed LL1, LL2, LL3 and LL4 to be closely related, it seems likely that the SNP mapping results seen here are not reliable. However, although the threshold provides a consistent parameter for analysing the outbreak status of isolates, it is arbitrary in its nature as it was developed elsewhere and based originally on different data and this should be considered when interpreting outbreak relations (Walker *et al*., 2013). In addition, as the actual pipeline used in the Kohl *et al* (2014) study was not disclosed publicly, a reliable comparison between the results presented in this chapter and other studies that have used in-house pipelines (Gardy *et al*., 2011, Walker *et al*., 2013, Walker *et al.*, 2015), cannot be done due to the uncertainty that the parameters used were the same. This raises the issue of the need for standardized methods for *M. tuberculosis* outbreak investigations. However, the data presented here supports previous studies that found cgMLST schemes provide a standardized method for analysis of a threshold for outbreak inclusion and exclusion which is reproducible across WGS datasets (Sheppard *et al*., 2012, Maiden *et al*, 2013). The results presented here show for the first time the application of sequencing carried out directly from *M. tuberculosis* boilate samples for outbreak investigation. Boilate extraction has previously and justifiably been criticized for its impure/lower yield of DNA, compared to other extraction methods used for sequencing (Aldous *et al*., 2005). The major advantages of using a boilate extraction for WGS are the lower cost, ease and speed (Aldous, *et al*,2005), particularly as the cost of sequencing and slow growth of *M. tuberculosis* are two key barriers in outbreak investigation. Boilate extraction using TE buffer has been documented to cost around 0.03$ a sample, one hundred times cheaper than the Infection Diagnostics, Inc (IDI) lysis tube extraction method which has been documented to yield the most and best quality DNA (Aldous *et al*., 2005). This study provides proof in principle that combined with cgMLST this may be a useful cost-effective approach where resources are tight.

Detailed evaluation of each mutation has not been described for other WGS-based investigations (Gardy, J *et al*., 2011, Walker T.M *et al.,* 2013, Kohl, T.A *et al.,* 2014, Walker, T.M *et al.,* 2015). Provean analysis was used to predict the functional effect each mutation had on genes which differed amongst the outbreak isolates (Choi and Chan, 2015). If the amino acid change is neutral and so does not cause a functional difference, then clinically the mutation is probably not relevant. Out of a total of 25 mutations across the outbreak, 23 were non-synonymous. Isolate LL2, despite carrying the most allelic differences, was physiologically identical to LL1. Allelic differences in LL4 also had no effect on its proteins. Therefore, in terms of both physiology and genomic relation, LL4 was the most similar to LL1. NPTB6 contained the greatest number of deleterious allelic differences and thus had the potential to be the most different from LL1 in terms of physiology. The mutation in the RV0133 gene of NPTB6 is predicted to have a deleterious effect on an acetyltransferase protein which is a known membrane protein (Lew *et al*., 2011). The literature reports this gene to be non-essential to the growth of the H37Rv reference strain (Sassetti *et al*., 2003), and so the mutation of Rv0133 in

NPTB6 may not confer any clinically-relevant physiological change. Similarly, the mutation in Rv0663 is also reported to be non-essential to *M. tuberculosis* growth and unrelated to virulence factors (Sassetti *et al*., 2003). NPTB6 did carry a deleterious mutation in its Rv2678c gene. The Rv2678c gene, *hem*E, is an essential gene for *in vitro* growth of the H37Rv reference strain (Sassetti *et al*., 2003). It has been identified as being a protein that can mediate drug resistance to isoniazid, ofloxacin, and ethambutol (Raman and Chandra, 2008). The protein produced by Rv2678c is a uroporphyrinogen decarboxylase and has been identified as a potential co-protein for inhibition of resistance amongst bacteria and is noted as a high confidence drug target (Raman and Chandra, 2008). Co-proteins are used by the resistome of bacteria as communication pathways between the drug target and resistance machinery. Simultaneous inhibition of the co-protein (often referred to as a co-target) and the primary protein for resistance aids the binding of the respective drug to the target binding site and allows the drug to become effective once again (Raman and Chandra, 2008). Therefore, the deleteriously mutated *hemE* gene in NPTB6 suggests NPTB6 lacks the ability to resist drugs via this protein. The mutations seen in Rv0959 in NPTB6 and Rv2137 were in hypothetical proteins, neither of which had known functions and thus no physiological information could be gleaned (TubercuList, 2017). In summary, Provean analysis has shown that each outbreak isolate was physiologically either identical or very similar.

The contact tracing team raised the issue of a possible super--spreader related to the public house being central to the Llwynhendy outbreak. Such a question cannot be answered by MIRU-VNTR typing, as stated previously (Walker *et al.*, 2013a, Walker *et al*,. 2013b). This study found that WGS supported the outbreak team's assumption that a super-spreader was present within the outbreak, and that LL1 was the likely candidate. The gene-by-gene analysis produced a phylogeny with a star-like structure (Figure 4.3), a feature previously shown to represent the presence of a super-spreader within an outbreak (Liu *et al*., 2006, Hirsh *et al*., 2004, Walker *et al*., 2013c, Kohl *et al*., 2014). The phylogeny showed cases LL2, LL4 and LL3 to originate from LL1 (the landlord) at the centre of the phylogeny. The data agreed with previous work that showed the published cgMLST scheme developed by Ridom SeqSphere can identify a potential super-spreader, as can traditional SNP mapping (Walker *et al.*, 2013 Kohl *et al*., 2014). Multiple factors are known to contribute to super-spreading such as immune suppression, late diagnosis/failure to diagnose, co-infection with another agent, airflow dynamics and delayed treatment (Stein *et al*., 2011). LL1 had onset of symptoms eight months prior to his hospital admission for treatment and this delay is likely to have contributed to his ability to infect a disproportionate number of secondary cases (Riley *et al*., 2003, Stein *et al*., 2011). More rapid case reporting has been shown to reduce the rate of secondary cases from super-spreading cases (Stein *et al*., 2011).

The phylogenetic analysis described in Chapter 3 (Figure 3.4) led to the inclusion of isolate NPTB6 in this analysis of the Llwynhendy outbreak. The cgMLST went on to suggest that the origin for NPTB6`s infection was the Landlord, case LL1. Although there was genomic evidence of a direct link between LL1 and NPTB6, the isolates were separated by a relatively larger geographical distance. At the time of the outbreak, no epidemiological evidence of links between NPTB6 and the Llwynhendy outbreak was sought, due to the geographical distance between the two outbreaks, and it having a different MIRU-VNTR type. Therefore, it would be interesting to investigate retrospectively the potential epidemiological links between NPTB6 and LL1. The findings of this analysis support previous claims that phylogenetic classification can provide data that is clinically relevant within tuberculosis outbreaks (Gutacker *et al*., 2002, Filiol *et al*., 2006, Coll F *et al.*, 2014).

4.5 Conclusion

WGS, through cgMLST analysis, was used to investigate a TB outbreak in Llwynhendy, South West Wales. Four out of 11 cases considered part of the outbreak by MIRU-VNTR were found not to be according to WGS. WGS also identified an isolate, NPTB6, as clonally related to the outbreak. This association had not been detected by MIRU-VNTR typing or epidemiological investigation and was only picked up by WGS because prior phylogenetic analysis had identified a similarity. Additionally, WGS analysis allowed the construction of a transmission chain and identification of a super-spreader. Provean analysis of the amino acid changes that occurred because of the DNA SNP mutations within the outbreak isolates highlighted the potential advantages WGS analysis provides in addition to strain typing and outbreak resolution. To summarize, 5 isolates were confirmed to be part of a direct transmission chain within the Llwynhendy outbreak in South West Wales, and the study established LL1 (the landlord) as the likely origin of the outbreak and potential super-spreader.

Chapter 5

Resolution of an outbreak of *Mycobacterium tuberculosis* in the Neath Port Talbot area of South Wales using Whole Genome Sequencing.

5.1 Introduction

On the 11th of September 2006 an outbreak of tuberculosis in the Sandfields area of Port Talbot, South Wales*,* came to the attention of Public Health Wales. The outbreak involved eight cases with cultured isolates *and* appeared to be circulating amongst individuals who frequented five local public houses within the Sandfields area, with one public house, Public House X, having connections to several cases in the outbreak. The outbreak sparked a review by Public Health Wales of tuberculosis case records in the area and continued until 2011. Over this period a further five cases were reported and molecular typing, through MIRU-VNTR profiling, was carried out on each isolate.

5.1.1 Epidemiology and Typing of Neath Port Talbot Outbreak Cases

Contrary to the assumptions of the outbreak investigation team based on contact tracing data, the isolates did not show a single MIRU-VNTR pattern, despite geographical, chronological and epidemiological connections between several of the cases. Two MIRU-VNTR profiles were identified within the isolate collection (Table 5.1), thus the outbreak investigation team assigned the isolates as either “A” or “B” strains according to their MIRU-VNTR profile (Table 5.1). In addition, isolate NPTB2, harboured a further polymorphism at the ETR C locus (Table 5.1, isolate coloured green), which was not present in other “A” or “B” isolates.

Table 5.1: A descriptive summary of each outbreak case and their MIRU-VNTR patterns, generated by the P are highlighted in blue; and NPTB isolates are in red, except for nptb2, coloured green as it differs from the

Case	Date of Diagnosis (month-year)	Onset of disease (month-year)	Attended Public House X	Strain	MIRU-VNTR profile								
					ETRA	ETRB	ETRC	ETRD	ETRE	MIRU2	MIRU10	MIRU16	MIRU20
NPTA1	May-04	May-04	Yes (landlord)	A	3	2	4	3	3	2	3	1	2
NPTA2	Apr-05	Apr-05	Unknown	A	3	2	4	3	3	2	3	1	2
NPTA3	Jun-05	Jun-05	Unknown	A	3	2	4	3	3	2	3	1	2
NPTA4	Nov-05	Nov-05	Yes	A	3	2	4	3	3	2	3	1	2
NPTA5	Jun-06	Jun-06	Yes	A	3	2	4	3	3	2	3	1	2
NPTA6	Aug-06	Aug-06	Yes	A	3	2	4	3	3	2	3	1	2
NPTA7	Feb-07	Nov-06	Yes**	A	3	2	4	3	3	2	3	1	2
NPTA8	Apr-07	Apr-07	Unknown	A	3	2	4	3	3	2	3	1	2
NPTB1	Oct-05	Oct-05	Unknown	B	3	2	4	3	3	2	3	1	2
NPTB2	Feb-06	Feb-06	Unknown	B	3	2	**3**	3	3	2	3	1	2
NPTB3	Nov-06	Mar-06	Unknown	B	3	2	4	3	3	2	3	1	2
NPTB4	Jan-07	Jan-07	No	B	3	2	4	3	3	2	3	1	2
NPTB5	Jun-09	Jun-09	No	B	3	2	4	3	3	2	3	1	2
NPTB6	Oct-11	2007-2011	Unknown	B	3	2	4	3	3	2	3	1	2

**NPTA7 did not admit to frequenting Public House X, but information acquired during contact tracing suggested t

5.1.2 Proposed transmission chain

The index case, that had pulmonary tuberculosis, was the landlord of Public House X. This case was identified in May 2004 (NPTA1) and contact tracing of close contacts and the pub's regular customers was carried out promptly. Contact tracing at the time detected no other cases. The landlord believed he had caught the infection from workers who had been visiting the public house before and at the time of his infection. However, between April 2005 and September 2006, a further 5 cases of pulmonary tuberculosis (NPTA2, NPTA3, NPTA4, NPTA5, NPTA6) were identified in persons with connections to the public house (Table 5.1). NPTA5 stated that they did not frequently visit the public house, whilst NPTA6 was a woman who was a regular at Public House X and at other public houses within the area. She also lived and socialised frequently with people in sheltered accommodation. In September 2006, Public Health Wales assembled the Neath Port Talbot (NPT) outbreak team. In early 2007, another male was diagnosed with pulmonary tuberculosis (NPTA7) within the outbreak area, with a MIRU-VNTR matching that of the NPTA strains. This case had an onset of tuberculosis-related symptoms several months prior to his diagnosis. Despite being known to other cases as being a regular at Public House X, he denied that this was the case. At this point all isolates had a matching MIRU-VNTR profile, classifying them as belonging to outbreak A (Table 5.1).

In November 2006, a case of tuberculosis (NPTB3) was diagnosed in the same area, with disease onset likely to have been several months prior to diagnosis. However, NPTB3 had a MIRU-VNTR profile that differed at two loci from the 'A' strains (Table 5.1). In early 2007, the estranged wife of NPTB3 was diagnosed with pulmonary tuberculosis (NPTB4), and with the same MIRU-VNTR type. Thus, the contact tracing team were confident that NPTB3 and NPTB4 shared a direct transmission event, and an MIRU-VNTR profile was assigned as Neath Port Talbot 'B' (NPTB). Retrospective investigations identified case NPTB2, diagnosed in February 2006, as possibly being part of the NPTB outbreak; although NPTB2 showed a single locus difference (ETRC) from the other ‘B’ genotype cases (Table 5.1). In April 2007, a further NPTA case (NPTA8) was identified in the area although there was no apparent epidemiological link between the case and any public house or other case within the area. A further ‘B’ strain case (NPTB5) was then diagnosed in June 2009 in Pontypool, 50 miles away from the outbreak area, but there were no apparent epidemiological connections between NPTB5 and the NPT outbreak despite the MIRU-VNTR match. The contact tracing team considered it unlikely that the case was directly linked to the NPT cases. Further retrospective typing identified another case from the area (NPTB1), diagnosed in May 2005. NPTB1 was then deemed the earliest isolate in the ‘B’ outbreak as the identification of NPTB1 showed that chronologically ‘A’ and ‘B’ strains had co-existed. In addition, further contact tracing also found NPTB1 to have an indirect link with Public House X, as her husband was a regular at the pub. At this stage, the contact tracing team began to question the heterogeneity of the ‘A’ and ‘B’ genotypes. The team felt it likely that evolution had occurred between cases, changing the MIRU-VNTR profile at two loci, and that probably the ‘B’ isolates were directly related to the ‘A’ ones. In October 2011, a further and final MIRU-VNTR ‘B’ type strain was identified (NPTB6). This case had been exposed twice to TB over the previous four years (2007-2011) and was a close childhood friend of the landlord of Public House X (NPTA1). By 2011, fourteen isolates had been reported to be connected to the NPT outbreak with 8 of type ‘A’ and 6 of type ‘B’. Figure 5.1 below provides a diagrammatic view of the epidemiological connections relative to public house X.

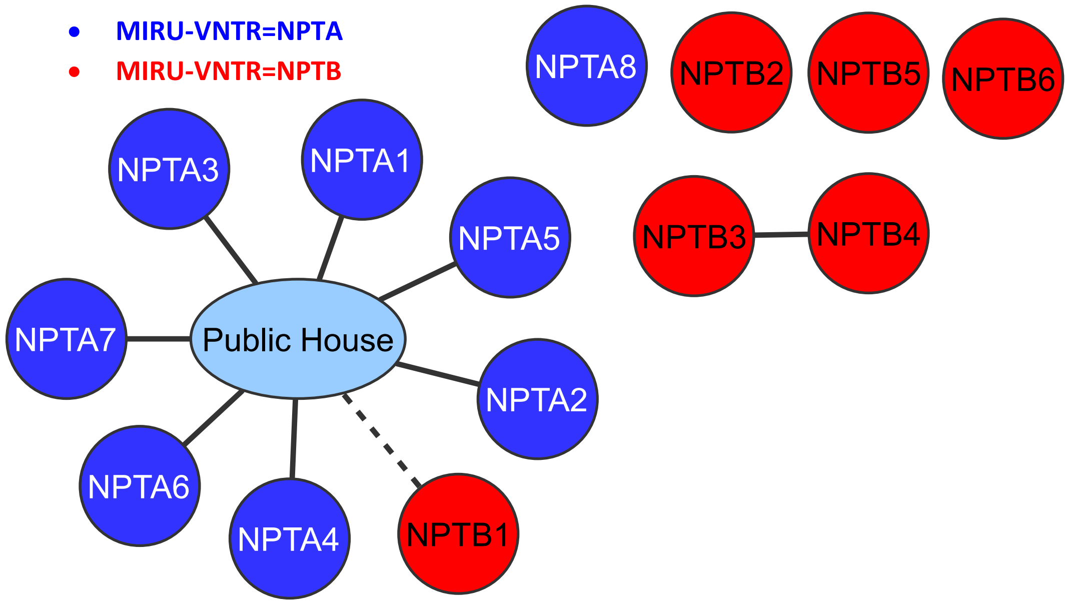


Figure 5.1: Diagram showing the epidemiological relationship between each isolate with a confirmed or indirect epidemiological link to the outbreak. Hard connecting lines represent a direct epidemiological link. The dashed connecting lines represent an indirect link. The public house labelled within the figure represents public house "X".

In addition to being sceptical about the presence of two separate outbreaks, the outbreak team also questioned NPTB5`s inclusion in the 'outbreak', as the case was geographically and epidemiologically separate from the other outbreak isolates. It was also believed that a "super spreader" was likely to be present within the outbreak, with NPTA6 or NPTA1 being the most likely candidates. The team desired further confirmation that NPTA1, the landlord, was the source of the outbreak and that the Public House X was central to the outbreak, but classical molecular typing could not answer these questions.

5.1.3: Aims

The primary aim of this study was to use WGS data to resolve the uncertainties surrounding the Neath Port Talbot *M. tuberculosis* outbreak(s) and identify the source case and any potential super spreaders within this outbreak(s). Both the established cgMLST scheme and an in-house cgMLST scheme was used to resolve whether the NPTA and NPTB outbreaks were actually one large outbreak, and whether NPTB5 was wrongly included in the outbreak by MIRU-VNTR. A novel ancestral dating method was introduced to provide evidence on the origin of outbreak isolates, whilst also elucidating the presence of micro-evolution within the outbreaks. As stated in chapter 4 and shown in chapter 3 (figure 3.4), NPTB6 clusters as the same sub lineage as the isolates from the Llwynhendy outbreak (Chapter 4). Previous cgMLST analysis in chapter 4 also showed NPTB6 to have a direct relation to isolates of the Llwynhendy outbreak (chapter 4). Thus, within this chapter, we aim to confirm the exclusion of NPTB6 from the Neath Port Talbot outbreak

5.2 Methods

5.2.1: Sample collection and epidemiological information

Isolate DNA was obtained as described in section 2.1. The collection of each Neath Port Talbot related sample was carried out as described in section 2.1. The Epidemiological information described in the introduction was obtained from face-to-face interviews with a nurse from the original PHW contact tracing investigation team and from documents produced during the outbreak investigation.

5.2.2: DNA Sequencing

A total of 14 isolates (Table 5.1) were sequenced initially, as described in section 2.2, through a Nextera XT protocol. Isolates were also processed by ancient DNA based BEST methods (Section 2.13.1 and 2.13.2).

5.2.3: Phylogenetic filtering of original dataset

The robust SNP barcode, presented in Chapter 3, Figure 3.4, provided a phylogenetic classification for each of the 14 outbreak-associated isolates included in the Neath Port Talbot outbreaks. A further three background isolates (BK1, BK2 and BK3) also clustered clonally as T family isolates and were included in further downstream analysis. 17 isolates (the 14 Neath Port Talbot isolates and the 3 T family background cases) were included in further analysis.

5.2.4: Core genome MLST (cgMLST) analysis

5.2.4.1: Established cgMLST scheme

A cgMLST scheme developed by Ridom seqSphere (Junemann *et al*., 2013) (section 2.3.1.), based on 2981 core genes was used to analyse the relationship between each of the 17 isolates selected following phylogenetic filtering. The 17 isolates include 3 background ones with phylogenetic associations in addition to the 14 previously defined Neath Port Talbot outbreak isolates.

5.2.4.2: In-house core genome MLST (cgMLST) scheme

An in-house cgMLST was developed containing all genes present across the 16 isolates (section 2.3.2) Each NPT isolate sequence was included in the "cgMLST definer" application, which used *M. tuberculosis* H37Rv as a reference genome. NPTA6 was removed from the final in-house cgMLST scheme as its inclusion significantly reduced the number of genes that could be included. The resulting in-house core genome MLST scheme consisted of 2562 loci, present in high quality across each isolate.

5.2.5: SNP analysis

5.2.5.1: Extraction of individual mutations

Using the method described in section 2.5.2, single nucleotide polymorphisms (SNPs) were extracted from 2981 core genes from the consensus sequence of outbreak 1 and outbreak 2 (defined in Figure 5.2 by the core genome MLST threshold of 12 allelic differences, see methods section 2.3.1).

5.2.5.2: Prediction Functional Genomic

The prediction of the functional effect of each mutation present within the dataset was carried out using the Provean software, described in section 2.5.3.

5.2.6: WGS SNP mapping

The published Centre of Genomic Epidemiology online server (Kaas *et al.*, 2014) was used to produce a WGS SNP mapping phylogeny for the outbreak isolates using the conserved signature indels (CSI) phylogeny pipeline (section 2.5).

5.2.7: Ancestral dating

Ancestral dating was carried out as described in section 2.4. SNPs were extracted from the consensus of the largest outbreak (outbreak 1) and the individual isolates of the smaller outbreak 2. A total of 90 SNPs was concatenated, and their ancestral dating was calculated as described in section 2.4. A cgMLST common across NPTB2, NPTB4 and the consensus of the NPTA outbreak was used, with BEAST software indicating the date of the most recent common ancestor (TMRCA) between the two outbreaks.

5.2.8: Statistical calculation of average variation between cgMLST schemes

The number of allelic differences between each isolate and the remaining isolates within the dataset was taken from both the established and in house core genome MLST results and the average number of differences each isolate had across the whole dataset was calculated using MS Excel (2013). These values were then used to determine the average amount of variations seen across the dataset for both cgMLST schemes. An unpaired T-test was used to calculate the P Value for whether the amount of variation seen in the results of the in house cgMLST scheme and Established cgMLST scheme is different. See section 2.15 for further details.

5.3 Results

5.3.1 Outbreak resolution using the cgMLST scheme

The results of the cgMLST are presented in Figure 5.2 and include those from three background isolates (BK1, BK2 and BK3) analysed as they were assigned the same sub-lineage in the phylogenetic analysis shown in Figure 3.4. The use of the established cgMLST scheme, revealed that there were technically isolates from two outbreaks present within the NPT dataset, but their composition did not correlate with the MIRU-VNTR typing defined outbreaks, NPTA and NPTB (Table 5.1). The outbreaks defined by cgMLST contained 9 isolates in outbreak 1 and 2 in outbreak 2. In outbreak 1 there were six NPTA isolates, one NPTB isolate (NPTB1) and BK2, previously thought of as an unrelated background case. NPTA3 showed 16 allelic differences from its closest relative (NPTA7) and thus according to the definition of no more than 12 allelic differences (Walker *et al*., 2013, Kohl *et al*., 2014) could not be directly linked to the outbreak. Five isolates showed no evidence of direct transmission with any other isolate within the dataset: these included three NPTB isolates (NPTB2, NPTB5 and NPTB6), and two background ones (BK1 and BK3).

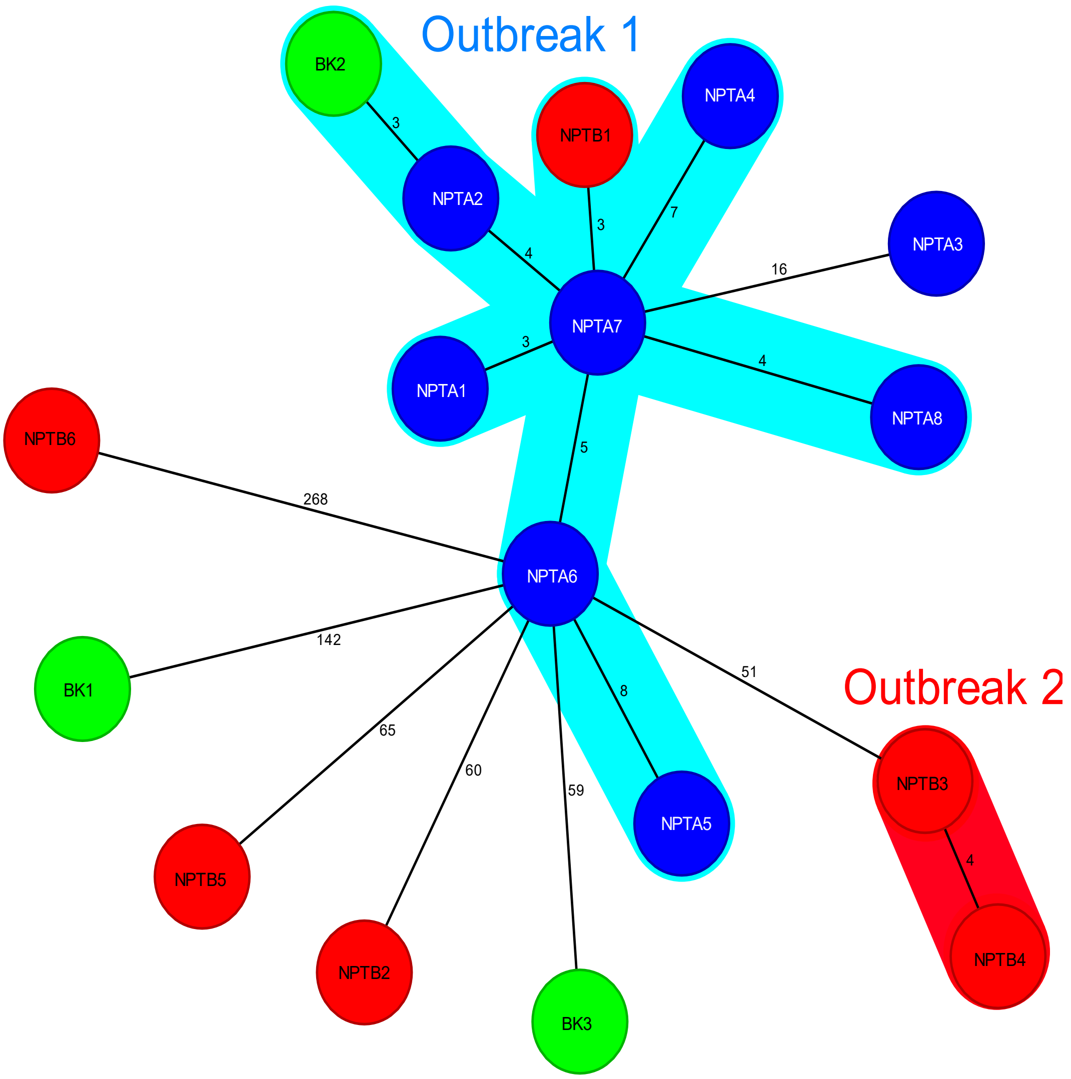


Figure 5.2: A minimum spanning tree of 17 cases constructed using Ridom SeqSphere software and based on a 2891 cgMLST scheme. Blue=NPTA, Red=NPTB, Green=Background. Isolates sharing less than 12 allelic difference are classed as direct transmission links and are thus part of a clonal outbreak and are grouped according with light blue = outbreak 1, and red=outbreak 2.

According to cgMLST, NPTA7 was the source case within this outbreak, having a direct link with seven out of the nine outbreak 1 cases, suggesting that NPTA7 was the potential super-spreader. Isolates NPTB3 and NPTB4 were similar enough to indicate direct transmission and make up outbreak 2. No other NPTB isolate showed any direct link with any other NPTB isolate, with NPTB1 in fact appearing to be directly linked with the NPTA outbreak isolates.

5.3.2 Mutation breakdown

The SNPs relevant to the consensus of allelic differences between outbreaks 1 and 2 were extracted for analysis as described in sections 2.7 and 5.2.5. The proportion of synonymous to non-synonymous mutations seen between the consensus sequence of Outbreaks 1 and 2 (Figure 5.2) showed a predominance of the latter between the two outbreaks (Table 5.3). Twenty-two of a total of 31 mutations were non-synonymous.

Table 5.2: Synonymous and non-synonymous mutations, by gene, between the consensus sequences of outbreaks 1 and 2. The dS/dN ratio is the ratio of synonymous to non-synonymous mutations seen between the two outbreaks. Only genes present across each of the isolates in outbreak 1 and 2 are compared.

Composition of Variants	
Synonymous	Non-synonymous
Rv0259c	Rv0695
Rv0272c	Rv0696
Rv0275c	Rv0913c
Rv0397A	Rv0933
Rv1858	Rv0987
Rv2141c	Rv1171
Rv2875	Rv1304
Rv3586	Rv1411c
Rv3675	Rv1482c
	Rv1551
	Rv1688
	Rv1931c
	Rv2207
	Rv2284
	Rv2536
	Rv2694c
	Rv2737c
	Rv2911
	Rv2990c
	Rv3090
	Rv3554
	Rv3802c
Total:9	Total:22

5.3.3 Functional Analysis

Of the 22 non-synonymous mutations in outbreak 1 and 2 isolates, a Provean algorithm, (sections 2.5.3 and 5.2.5.2), predicted 7 to have a deleterious effect on their respective protein (negative score), and 15 non-synonymous mutations that had no predicted effect on their given protein (positive score), thus classified as neutral mutations (Table 5.3).

Table 5.3: Functional analysis on the non-synonymous mutation genes between outbreak 1 and outbreak 2. aa change column - codon position, outbreak 1 allele and outbreak 2 allele. aa - amino acid; * - stop codon.

Functional effect of non-synonymous variants				
Gene	**Product**	**aa change**	**Provean Score**	**Effect**
Rv0695	mycofactocin system creatinine amidohydrolase family protein MftE	216,P,A	-7.567	Deleterious
Rv0696	mycofactocin biosynthesis glycosyltransferase MftF	16,G,R	2.037	Neutral
Rv0913c	Dioxygenase	376,D,A	1.215	Neutral
Rv0933	phosphate ABC transporter ATP-binding protein PstB	150, W,*	-21	Deleterious
Rv0987	adhesion component ABC transporter permease	139,G,R	-0.196	Neutral
Rv1171	hypothetical protein	124,V,I	0.505	Neutral
Rv1304	ATP synthase subunit A	222,F,L	-3.742	Deleterious
Rv1411c	lipoprotein LprG	157,R,G	-3.9	Deleterious
Rv1482c	hypothetical protein	23,R,S	2.067	Neutral
Rv1551	acyltransferase PlsB	566,S,F	-5.741	Deleterious
Rv1688	3-methyladenine DNA glycosylase	116,A,T	-1.272	Neutral
Rv1931c	transcriptional regulator	114,P,T	0.832	Neutral
Rv2207	nicotinate-nucleotide-dimethylbenzimidazol phosphoribosyltransferase	105,A,T	1.548	Neutral
Rv2284	esterase LipM	276,G,D	5.47	Neutral
Rv2536	transmembrane protein	155,E,K	-1.404	Neutral
Rv2694c	hypothetical protein	2,G,E	-0.304	Neutral
Rv2737c	recombinase A	522,S,P	-0.007	Neutral
Rv2911	penicillin-binding protein DacB2	97,*,Q	-8.19	Deleterious
Rv2990c	hypothetical protein	139,D,G	-5.611	Deleterious
Rv3090	alanine/valine-rich protein	17,G,V	1.056	Neutral
Rv3554	electron transfer protein FdxB	16,P,T	0.047	Neutral
Rv3802c	membrane protein	259,W,R	7.128	Neutral
Total:22				

Interestingly stop codons were identified in Rv2911 of outbreak 1 strains, and in Rv0933 of outbreak 2 strains. The presence of these is likely to cause truncations across the remainder of the gene, thus explaining how the mutations at Rv2911 and Rv0933 confer the most negative Provean analysis values of -8.19 and -21 respectively.

5.3.4 Outbreak resolution using an in-house core genome MLST scheme

The in-house cgMLST scheme included 2562 genes and the generated phylogeny results are presented in Figure 5.3. Isolate NPTA6 only aligned succesfully to 66 % of the potential in-house cgMLST "targets" and since its inclusion substantially reduced the number of genes that could be analysed to 1919 genes it was excluded. Again the threshold for outbreak inclusion was <12 allelic differences (Kohl *et al.*, 2014, Walker *et al.*, 2013). Compared to the results shown in Figure 5.2, it has slightly different features with regards to the relation of NPTB1 and NPTA7 to outbreak 1, but as a whole has the same phylogenic shape as seen using the published cgMLST scheme (Figure 5.2). In contrast to that result, the phylogeny has NPTA8 as the central isolate while NPTA7 appears to have no direct relationship to the outbreak cases. In addition NPTB1, which was outbreak- related in Figure 5.2, showed a higher level of divergence and exceeded the 12 allelic difference threshold for direct transmission within outbreak 1.

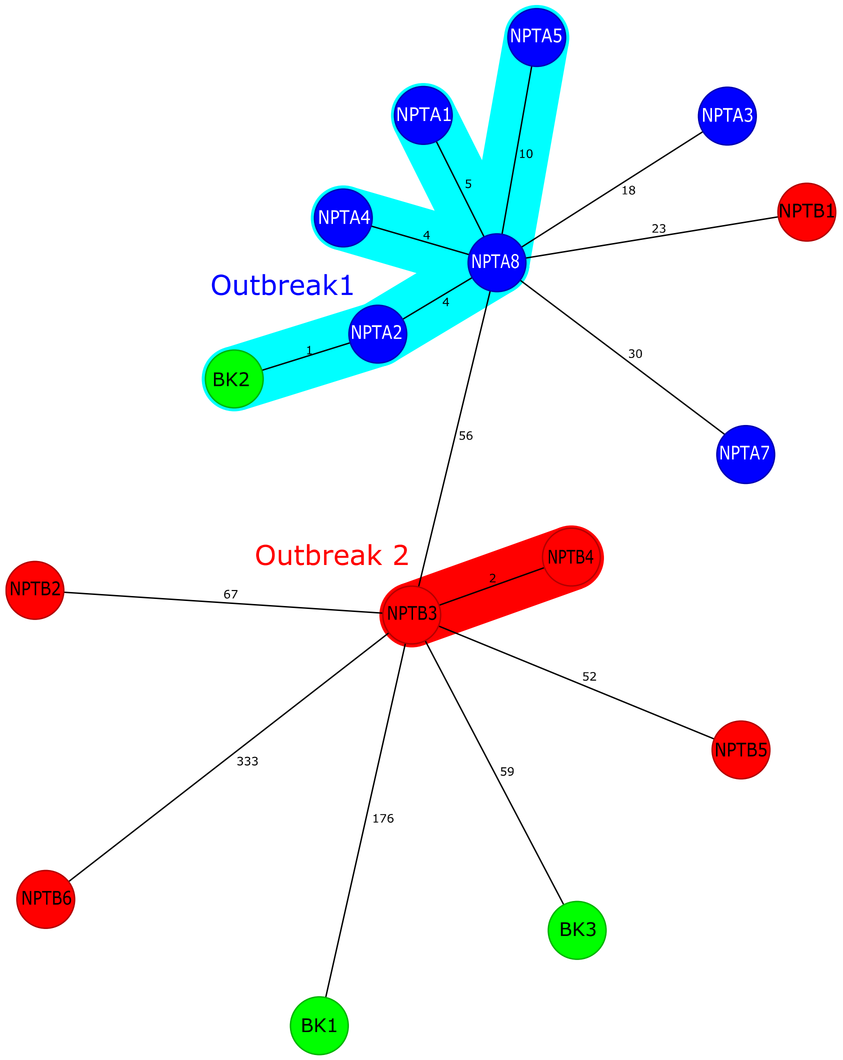
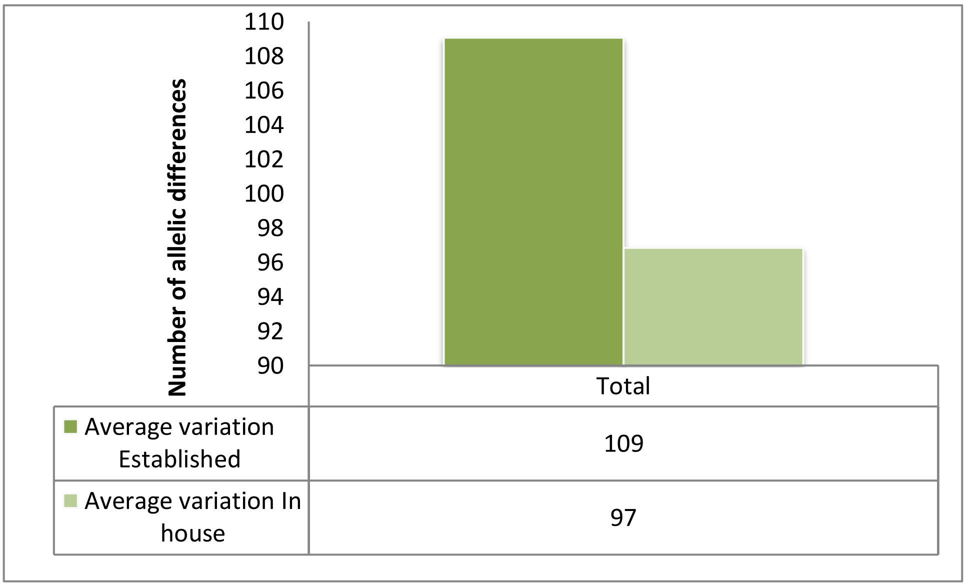


Figure 5.3: A minimum spanning tree generated by in-house cgMLST and based on 2562 genes present with high quality scores across each isolate within the dataset. Blue = NPTA ,Red = NPTB, Green = background. Numbers adjacent to branches represent the number of allelic differences between given isolates.

5.3.5 Comparison between the in-house and established cgMLST schemes

The average number of allelic differences seen between strains for the two cgMLST schemes are shown in Figure 5.4. Across the isolates included in the established and in-house cgMLST analyses respectively, the in-house cgMLST scheme showed fewer differences between strains with an average 97 allelic differences per strain compared with an average 109 per strain for the established scheme. However, the variation (amount of allelic differences) between isolates was not significantly different (P>0.05) in one cgMLST scheme over the other (see section 5.2.8). Thus despite using a different set of core genes, the resulting relationships between isolates on average are not significantly different when using either the in house or established cgMLST scheme.



	Total
🟩 Average variation Established	109
🟩 Average variation In house	97

Figure 5.4: A graph showing the average number of allelic differences across the total set of isolates.

5.3.6 Whole Genome Sequence SNP mapping

The SNP mapping shown in Figure 5.5a and Table 5.4, highlight the degree of divergence seen between the outbreak isolates when analysed using the published Centre of Genomic Epidemiology SNP mapping protocol (section 2.5 and 5.2.7). The neighbour- joining phylogeny in Figure 5.5a does not include isolate NPTB6 as it had an excesive amount of genomic variation, diverging by over 700 SNPs with every isolate in the dataset, as presented in Table 5.4. The amount of divergence within the phylogeny exceeds that seen within the cgMLST, with no isolates sharing fewer than 12 SNPs and thus could not be classified as an oubtreak complex. However, the overall structure of genomic relationships emulates that seen in the cgMLST results (Figure 5.2 and Figure 5.5b), although the inclusion of NPTB6 in the cgMLST results should be considered when analyzing Figure 5.5b. As in the cgMLST analysis, the results in Figure 5.5a showed that the NPTA and NPTB isolates cluster separately. The NPTA cluster holds an ancestral position within the phylogeny and the NPTB isolates branch off at a later point in Figure 5.5a. As in the cgMLST analysis, NPTB1 and BK2 again cluster with the NPTA isolates (Figure 5.5a and 5.5b).

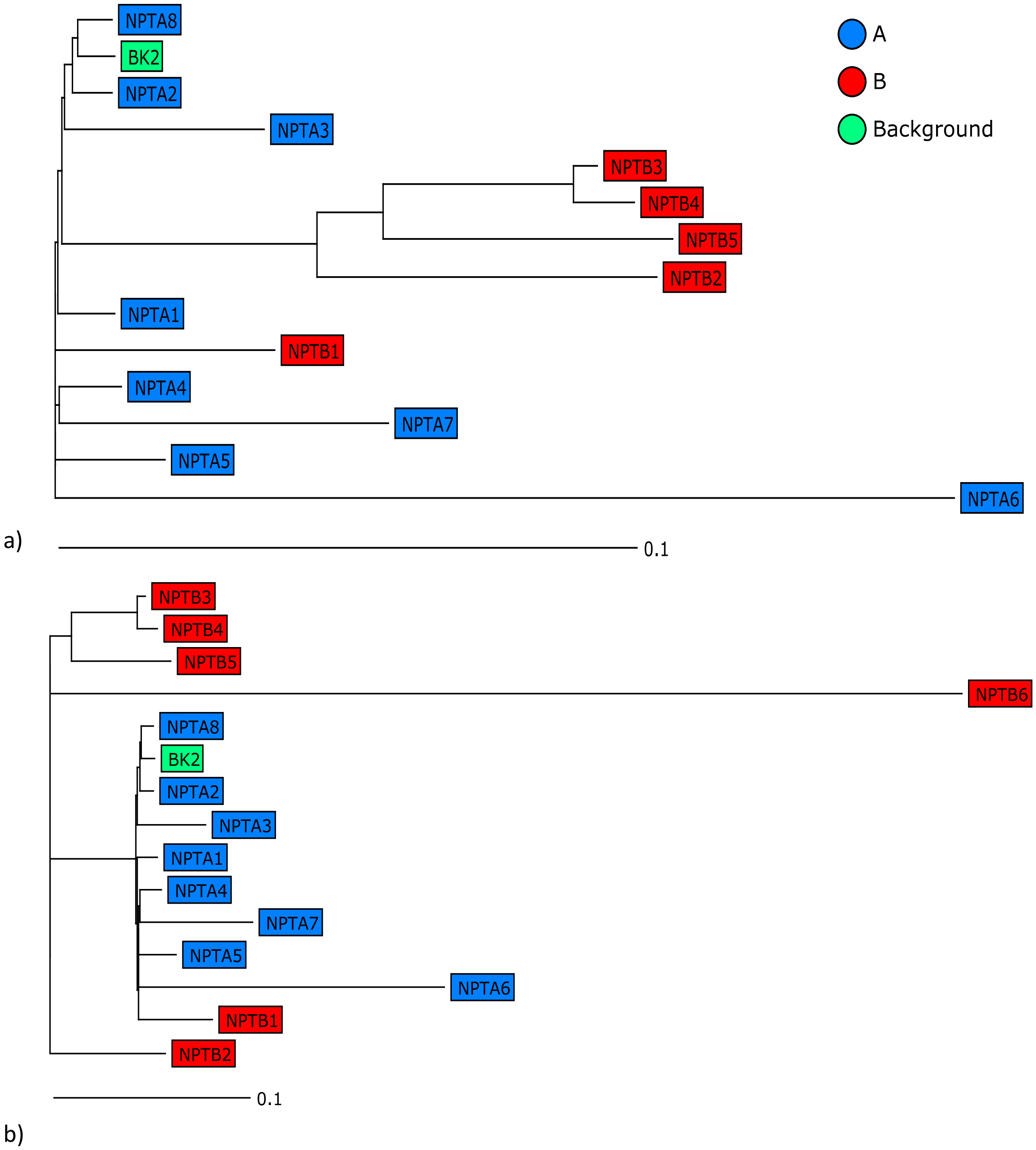


Figure 5.5: Neighbour- joining phylogenies constructed from a total of 1391 SNPs. (a) traditional WGS SNP mapping, and (b) cgMLST analysis derived from results presented in MST format in Figure 5.2. Blue = NPTA, Red = NPTB, Green = background. The scale bar for (a) represents the genomic distance measured in nucleotide differences across 1391 loci, and for (b) is measured in units of allelic differences.

Table 5.4. A correlating distance matrix to figure 5.5, produced by traditional WGS SNP mapping described i
Blue= NPTA isolates. The darker blue the cell the more closely related the isolates are. The darker red the ce

	Number of SNP`s across the Whole Genome										
Isolate	NPTA1	NPTA2	NPTA3	NPTA4	NPTA5	NPTA6	NPTA7	NPTA8	NPTB1	NPTB2	NP
NPTA1	0	28	62	30	42	230	94	21	67	161	94
NPTA2	28	0	60	28	40	230	94	21	67	155	14
NPTA3	62	60	0	66	78	266	132	59	103	191	17
NPTA4	30	28	66	0	42	232	94	29	69	161	14
NPTA5	42	40	78	42	0	242	106	43	79	169	15
NPTA6	230	230	266	232	242	0	296	229	269	361	34
NPTA7	94	94	132	94	106	296	0	95	133	223	20
NPTA8	21	21	59	29	43	229	95	0	68	158	14
NPTB1	67	67	103	69	79	269	133	68	0	196	18
NPTB2	161	155	191	161	169	361	223	158	196	0	14
NPTB3	94	141	177	147	155	347	209	144	182	146	0
NPTB4	151	149	189	157	167	356	219	146	194	160	20
NPTB5	159	161	193	167	175	365	229	158	200	168	12
NPTB6	720	720	754	726	734	922	786	717	759	727	71

The phylogeny and distance matrix (Figure 5.5 and Table 5.4 respectively) also show NPTA6 to be the most divergent isolate within the phylogeny, with over 200 SNPS with every other isolate. The closest relationship within the phylogeny was between isolates BK2 and NPTA8 with 17 SNPs. In support of the cgMLST results, WGS also showed NPTB3 and NPTB4 clustered more closely with each other (sharing 20 SNPs) than they do to any other NPTB isolate. These NPTB isolates differed from all the NPTB isolates by over 100 SNPs. Within the NPTA strain cluster, NPTA6 was the most divergent and NPTA7 the next most. NPTA7's closest relations were NPTA8, NPTA4, NPTA1 and NPTA2, with 94 SNPs. Isolate NPTB1 clustered within the NPTA isolates, rather than the other NPTB isolates, and showed a closer relation to isolates NPTA1, NPTA2, NPTA8, NPTA54 and NPTA5 than to NPTA7. Thus, the SNP mapping results support previous findings in this study that NPTB1 was more closely related to the NPTA isolates than to the NPTB isolates, with which it shared a clonal MIRU-VNTR profile. The closest relationships between the NPTA isolates are between NPTA1, NPTA2 and NPTA8 equally with 23 SNPs. NPTA4 and NPTA5 then also share fewer than 50 SNPs with NPTA1, NPTA2 and NPTA8. NPTA3 does show at least 60 SNPs with each other NPTA isolate and thus along with NPTA7 is less clonally related to the other NPTA isolates. The SNP mapping method provided a similar phylogenetic structure to that seen previously, although the level of divergence was greater than that seen with the cgMLST analysis.

5.3.7 Ancestral dating

The results so far all provided evidence for the presence of two separate outbreaks, despite also showing discrepancies in the status of NPTB5, NPTB1, NPTB2 and NPTB6. However, the original phylogenetic analysis in Figure 3.4 indicated that the NPTA and NPTB outbreak isolates were of the same sub-lineage. To investigate the possibility that both outbreaks were caused by the same circulating strain, with unreported cases causing direct transmission links, a previously defined ancestral reconstruction method was used (section 5.2.8 and section 2.4). Figure 5.6 depicts the chronological divergence that has occurred between the two outbreaks and how long ago both outbreaks shared their most recent common ancestor.

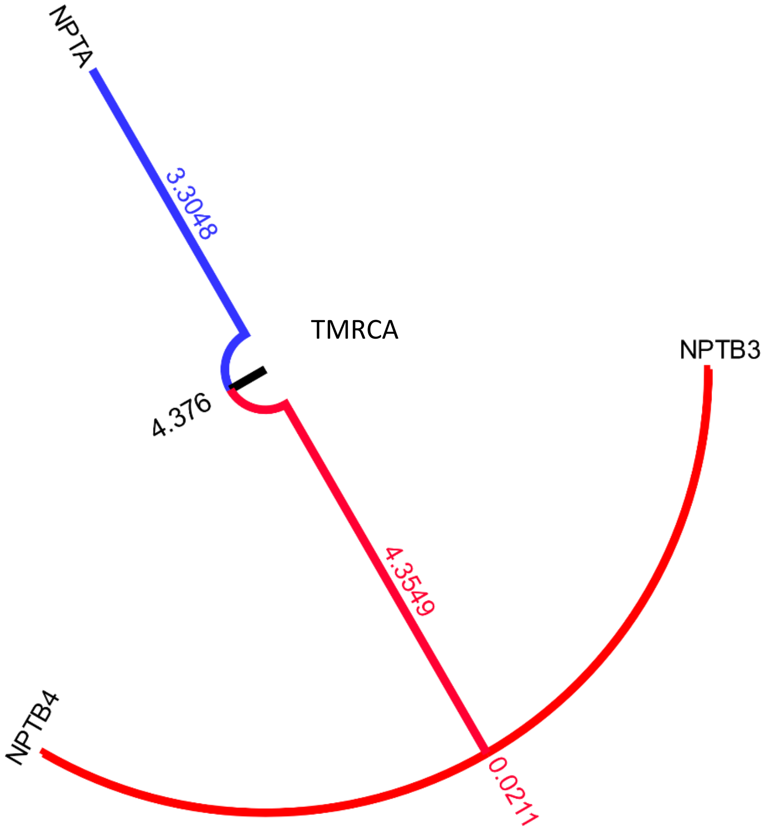


Figure 5.6: A phylogenetic tree based on the ancestral reconstruction and dating of the most recent common ancestor (TMRCA) of both NPTA and NPTB outbreaks. A consensus sequence of NPTA outbreak strains was produced and aligned to the corresponding cgMLST sequences of both NPTB3 and NPTB4. Red = divergence of NPTB isolates; Blue = divergence of NPTA outbreak isolates from TMRCA of both. Values are scaled in years.

The data indicated that it was 4.38 years since NPTA and NPTB isolates shared a common ancestor. As can be seen by the lower branch length numbers, the NPTA isolates are of a closer relation to TMRCA (3.30 years of divergence) than the NPTB ones (4.38 years of divergence). NPTB2 and NPTB4 last shared a common ancestor 0.02 years before their divergence to their current state. Thus, the two outbreaks diverged from each other around 4.38 years before their current state and NPTB2 and NPTB4 isolates share a very recent common ancestor.

5.4: Discussion

This chapter has described how through WGS of *M. tuberculosis* boilate samples it was possible to apply phylogenetics, cgMLST methods, SNP mapping and ancestral sequence dating to analyse the Neath Port Talbot outbreak. This study supports previous ones that have shown the higher resolution of WGS data provides over MIRU-VNTR typing (Walker *et al*., 2013b, Walker *et al*., 2013a, Roetzer *et al*., 2013).

One aim of this study was to evaluate whether a phylogenetic method could aid epidemiological investigation within a tuberculosis outbreak. The phylogenetic characterisation of tuberculosis isolates into respective lineages and sub-lineages has been shown previously to provide clinically relevant data in terms of vaccines and antibiotic resistance risks (Filliol *et al*., 2006, Comas *et al*., 2009, Homolka *et al*., 2012, Feuerriegel *et al*., 2014). Phylogenetics has also been used epidemiologically in other infectious diseases, such as Ebola and respiratory syncytial virus (Dudas and Rambaut, 2014, Nabeya *et al*., 2017). In this current study, phylogenetic analysis confirmed that all the outbreak isolates except NPTB6, were clustered within the same sub-lineage, the Euro-American T family. The results within this chapter further confirms NPTB6`s wrongful inclusion as being part of the Neath Port Talbot outbreak. In addition, the SNP barcode method seen in figure 3.4(chapter 3) identified three further apparently unrelated local isolates that clustered within this phylogeny; substantiating that phylogenetic characterisation may be useful in tuberculosis outbreak investigation.

A previous study highlighted how a novel gene-by-gene cgMLST scheme could provide a more standardised method for outbreak resolution over the more traditional SNP typing methods (Kohl *et al*., 2014). This method has already been applied to *Campylobacter* spp. (Sheppard *et al*., 2012)*, Salmonella* spp. (Maiden *et al*., 2013), *Legionella* spp. (Moran-Gilad *et al*., 2015) and *Neisseria meningitidis* (Agnememel *et al*., 2016) outbreaks. An advantage of gene-by-gene MLST methods is that unlike SNP mapping, they do not rely on high quality genomes, and can thus provide analysis where only partial sequence data are available (Sheppard *et al*., 2012, Maiden *et al*., 2013), as was the case in this study. Through the use of this established cgMLST scheme (Kohl *et al*., 2014), the relationship between the NPT outbreak isolates was resolved, and in accordance with the MIRU-VNTR results, two outbreaks were confirmed, although they did not correlate completely with those generated by MIRU-VNTR. The cgMLST analysis confirmed that each of the NPTA cases was directly linked, except for NPTA3. The cgMLST analysis also confirmed that the cases were directly linked to Public House X, as assumed by initial contact tracing team. However, in the cases of NPTB, cgMLST analysis found only two out of the six NPTB isolates to be directly related and found NPTB1 to cluster with the main NPTA outbreak, confirming the link between NPTB1 and the Public House X, as proposed by the contact tracing team. The close genomic relationship and contact tracing links between the NPTA isolates (the main outbreak) and NPTB1 may have resulted from microevolution, which refers in this context to the gradual evolution of a single clonal strain during an outbreak (Walker *et al*., 2013, Liu et al., 2015, Takiff and Feo., 2015), altering the MIRU-VNTR typing results, as postulated in previous studies (Walker *et al*., 2013b, Takiff and Feo., 2015).

Previous studies have highlighted that WGS can identify super spreaders and that this is indicated by a star-like phylogeny (Kohl *et al*., 2014;Walker *et al* 2013). In agreement with those studies, the cgMLST result in this one also generated a star-like phylogeny. However, the contact tracing teams assumed the super-spreaders were NPTA1 and NPTA6, but this was not shown by cgMLST, which indicated NPTA7 was the most likely source case and super-spreader. A super-spreader is validated epidemiologically as a case that has disproportionally infected more secondary cases than others infected with the same disease(Galvani *et al*., 2005, Stein *et al*, 2011). Epidemiological investigation had identified that NPTA7 was likely to have frequented Public House X (although he did not admit to it), and had identified that NPTA7 had numerous social contacts, including family members, who also frequented the pub.

One of the key potential advantages of WGS over other typing methods is that its application is not restricted to typing and epidemiological applications. The data have the potential to provide information on the physiological properties of isolates within a given outbreak. Bioinformatic tools and databases have and are being developed to predict the functional effect that variations between genomes have on the phenotype for a given sample (Bromberg and Rost, 2007, Wong and Zhang, 2014, Johnson *et al*., 2005, Thomas *et al*., 2003). One particular tool, called Provean (Choi and Chan, 2015), allows prediction of whether non-synonymous mutations have deleterious effects on a given protein; and has been previously used in studies on virulence in the tuberculosis reference strains H37Rv and H37Ra (Jena *et al*., 2014), the prediction of colistin resistance genes in *Klebsiella pneumoniae* (Olaitan *et al*., 2014), and in the analysis of host adaptation in *Salmonella enterica* (Octavia *et al*., 2017). In this current study, a total of 31 allelic differences between the outbreak 1 and outbreak 2 isolates were identified, 22 were non-synonymous and 9 were synonymous. Although the information on synonymous mutations between the outbreaks is useful in terms of understanding the natural evolution between the two outbreaks (Kimura, 1984, Hurst, 2002, Li, 2014), the properties of the non-synonymous mutations may yield more information on the physiological differences between the outbreak isolates, which may be clinically relevant.

Provean analysis on the 22 non-synonymous mutations found only ones in seven genes conferred a deleterious effect on their respective protein (Table 5.4); 5 of which were clinically relevant. RV1411c produces a TLR-2 ligand that inhibits human macrophage class II antigen processing, allowing the organism to avoid recognition by CD4+ T-cells (Gehring *et al*., 2004). Rv0695 encodes the protein MftE which is part of the myofactin system (Lew *et al*., 2011) shown to be essential for *in vitro* growth on cholesterol (Lew *et al*., 2011, Griffin *et al*., 2011), and for the ability of *M. tuberculosis* to use cholesterol as a carbon source in order to persist within animal tissue (Griffin *et al*., 2011). Recent studies have highlighted that statins are associated with reduced risk of active tuberculosis infection (Lai *et al*, 2016).

A further three mutations were found which have a possible influence on drug susceptibility. Rv0933 is a cell wall gene encoding an ABC transporter (efflux pump; Lew *et al*., 2011, Pang *et* *al*., 2013), and its hypothetical function is to enable phosphate transportation across the cell wall which is essential to the survival of *M. tuberculosis* (Betts *et al*., 2002). A previous study showed that in rifampicin mono-resistant clinical samples, a high level of Rv0933 expression was seen (Pang *et al*., 2013). Rv1304 is an ATP synthase subunit and is an essential gene for *in vitro* growth and *in vivo* in a mouse host (Lew *et al*., 2011). It is situated and expressed within the *Sig1* regulon, a sigma factor (Lee *et al*., 2012), and its loss of function has been shown to correlate with isoniazid resistance which does not hinder fitness (Lee *et al*., 2012). Lastly, a deleterious functional mutation was detected in the *Dac*B2 gene, Rv2911. The deletion of *Dac*B2 has been shown to cause a hyper-virulent phenotype, whereby the bacteria survive and grow at a significantly higher rate within macrophages than wild type isolates (Bourai *et al*., 2012). Recent *in vitro* evidence has shown the potential efficacy of a combination of a carbapenem and amoxicillin-clavulanic acid in treating tuberculosis infection, through the inhibition of DacB2 by meropenem (Diacon *et al*., 2016, Kumar *et al*., 2012). Hence, functional prediction of non-synonymous mutations may therefore be conferring potentially clinically useful information, not least in terms of therapy. Although all the isolates in this study were known to be phenotypically sensitive to standard anti-tuberculosis treatment, having been tested at PHW WCM, it is interesting to consider that a subset of isolates may have increased susceptibility to beta-lactams, and this is something that could be tested in future to investigate whether previous reports on the function of *Dac*B2 can be corroborated. A previous study highlighted how WGS data will be an important future tool for *M. tuberculosis* drug susceptibility and resistance detection (Walker *et al*., 2015). This current study showed that Provean functional analysis provided clinically relevant evidence for not only antibiotic susceptibility, but also sheds light on other clinical feature, such as the ability to metabolise cholesterol.

The in-house cgMLST scheme confirmed the presence of two separate outbreaks and results correlated with those obtained using the established cgMLST scheme previously defined, showing minor contrasts for isolates NPTB1 and NPTA7. Certain genes within the in-house cgMLST scheme, such as the *pe/ppe* genes, are known to be polymorphic. The *ppe/pe* genes encode secreted cell surface proteins involved in host immune evasion and antigen variation (McEvoy *et al*., 2012, Gutacker *et al*., 2002). It has been shown previously that non-synonymous mutations are 3.0 and 3.3 times more likely, in *pe* and *ppe* genes respectively, than in other ones (McEvoy *et al*., 2012, Gutacker *et al*., 2002). Other studies have removed such loci from their alignment analysis due to their polymorphic properties (Stucki and Gagneux., 2013) and this includes the established cgMLST scheme used here (Kohl *et al*., 2014). This may at least partly explain the discrepancies in degree of divergence seen between the two methods.

In line with previous studies, the SNP mapping produced results that reflected the pattern seen in the cgMLST analyses (Kohl *et al*., 2014). However, in contrast to previous studies (Walker *et al*., 2013b;Kohl *et al*, 2014), the number of SNPs far exceeded the previously published 12 SNP threshold for direct transmission. As stated in Chapter 4, the SNP calling pipeline used in this study (see section 5.2.6 and 2.5) is likely to be based on parameters that are not identical to the publications from which the 12 SNP threshold was derived as the exact pipeline is not presented by the original Walker *et al (*2013) publication (Walker *et al*; 2013a). Therefore the application of this threshold may not to be suitable for outbreak associations made by the WGS SNP mapping, highlighting the non-standardized nature of SNP calling pipelines (Sheppard *et al.*, 2012, Maiden *et al*., 2013). The effective use of cgMLST typing in this study provides promise in providing a standardized platform for the widespread use of WGS data for outbreak investigation.

Despite this study providing evidence for the presence of two separate outbreaks, phylogenetic analysis found isolates from both outbreaks to be of the same sub-lineage and epidemiological analysis also suggested a single, larger outbreak. Ancestral dating methods showed that both outbreaks shared a recent common ancestor of less than 5 years previously and were probably caused by a single endemic circulating T family strain, thus suggesting that isolates from both outbreaks 1 and 2 derive from the same origin, and possibly due to microevolution of a circulating T family strain. Investigation into intermediate cases that may be responsible for the divergence seen between both outbreaks, should be carried out in future work following this study. A study of a tuberculosis outbreak in Germany with the prominent "Hamburg clone" used BEAST ancestral reconstruction (Ansari *et al*, 2016) to assess the clonality of a large outbreak, spaced over a large period of time, where the interpretation of the genomic results was found to be complicated by microevolution (Roetzer *et al*., 2013a). Ancestral dating methods have also been applied to other infectious agents for identification of the evolution of a certain trait, outbreak investigation or species origin in the SARS outbreak (Jombart *et al*., 2014)*, Salmonella enterica* (Hawkey *et al*., 2013) and *Yersinia pestis* (Whittles and Didelot, 2016) respectively. Such information cannot be obtained via MIRU-VNTR typing, thus this study's results supports the ability of ancestral dating to elucidate confusions within outbreak resolution and along with phylogenetic grouping, provide convincing evidence with regards to outbreak origin, as stated in previous studies (Ansari and Didelot, 2016; Roetzer *et al*., 2013a).

5.5 Conclusion

Using WGS data, an outbreak of tuberculosis in Neath Port Talbot, in which epidemiology and molecular typing by MIRU-VNTR were inconsistent with one and other, was resolved. Although original cgMLST analysis showed evidence of two separate outbreaks, the introduction of ancestral dating indicated that the two "outbreaks" were likely to be one continuous one caused by the same endemic strain that had micro-evolved over time, with intermediate cases likely to have been missed form this investigation. The phylogenetic clonality, seen in the robust SNP barcode results in Chapter 3 provides further support to the presence of one outbreak caused by a circulating strain. The results also confirm Public House X as being central to the outbreak, and the presence of a super-spreader, with NPTA7 being the most likely super-spreader according to the established cgMLST scheme results. The results also confirm exclusion of isolate NPTB6 from the Neath Port Talbot outbreak. For future work identification, culturing and sequencing of missing intermediate cases should be carried out and followed by further genomic analysis including these missing isolates in addition to those already in this dataset.

Chapter 6

The study of a *M. tuberculosis* outbreak in Gorseinon, South Wales, using a novel ancient DNA library build and sequencing protocol.

6.1: Introduction

During 2003-2004, an outbreak of tuberculosis was identified centred on a public house in the Gorseinon area of South West Wales. Another outbreak arose in a school in the same area in 2007. In addition, in 2008, a further outbreak occurred in the Townhill area of nearby Swansea. MIRU-VNTR typing showed that both the Gorseinon pub and school outbreak isolates shared identical patterns, with the Townhill outbreak isolates differing by one locus (MIRU16) from the Gorseinon outbreak isolates (Table 6.1 below). Due to the proximity of the Townhill outbreak to the other two in both time and place, and the existence of at least one case with contacts to the Townhill as well as the Gorseinon outbreak, the investigation team suspected that all three outbreaks might be part of the same one, and that micro-evolution had caused the slight divergence between the MIRU-VNTR types. To investigate this, WGS was applied to a subset of 13 isolates from the outbreaks.

Table 6.1: Epidemiological data for each isolate of the 13 isolates included within this dataset.

Case	Date of Diagnosis	Date of symptom onset	Outbreak MIRU-VNTR	Notes
GO1	Jun-03	Apr-03	Gorseinon public house	The first Gorseinon case only frequented Public House M a GO3. Potential link with GO1,TH2 and TH3.
GO2	Nov-03	Nov-03	Gorseinon public house	Linked to 3 - 4 public houses in Gorseinon, was not in direc
GO3	Feb-04	Mar-03	Gorseinon public house	Barman at Public House M for 3 years. Potential super-spr potential Townhill link (TH2, TH1 and TH3). Direct contact
GO4	Jul-04	Feb-03	Gorseinon public house	Potential super-spreader in the Gorseinon outbreak and contact with all public house cases.
GO5	Aug-04	Jul-04/03	Gorseinon public house	Potentially infectious for a year prior to diagnosis.
GO6	May-05	Feb-05	Gorseinon public house	Gorseinon outbreak secondary case.
GO7	Oct-08	Jan-08	Gorseinon Public house	Potential link case between Gorsienon and Townhill.
GS1	Apr-08	Feb-07	Gorseinon school outbreak	Potential link between Gorseinon school and public house
TH1	Mar-08	Oct-06	Townhill	Original Townhill outbreak case and potential super-sprea
TH2	Dec-02	Jul-02	Townhill	Thought to be the main link between the Gorseinon and T
TH3	Jan-07	Oct-06	Townhill	Potential link to Gorseinon outbreak.
GO8	Oct-08	Jul-08	Unique	Contrasting MIRU-VNTR and epidemiological data, either a ba Townhill and Gorseinon outbreaks.
GO9	Feb-04	Aug-02	Unique	Closely related through epidemiology to the public house outbre

6.1.2: Background

6.1.2.1: Gorseinon public-house associated outbreak (2003-2004 and 2007)

On 10th February 2004, Public Health officials noted that the number of cases of *M. tuberculosis* reported in the Gorseinon area of South West Wales was higher than expected. In addition, all the cases were found to drink regularly at a public house (M) in Gorseinon. GO1 was the first case diagnosed (notified June 2003), with onset of disease occurring in April 2003. GO1 was a customer at Public House M and another one in the Gorseinon area, with evidence of direct contact with another case from the area (GO3). Case GO2 was notified in December 2003, with onset of symptoms one month earlier. This person was a customer at three public houses in Gorseinon, including Public House M, and had direct contact with GO3 and GO4. Case GO3, a barman in Public House M, was notified in February 2004, having symptoms going back to at least December 2003, and potentially as far back as spring 2003. He had direct contact with at least six other *M. tuberculosis* cases within the area that visited Public House M (GO1, GO2, GO4, GO5, GO6 and GO7). Case GO4 was notified in July 2004, with onset of disease in February or March 2003, was considered highly infectious, and had direct contact with each of the original three cases, especially GO3. At this point, it was postulated that GO3 and GO4 might represent potential 'super-spreaders' within the public house outbreak, as both had direct links to multiple cases. Case GO5 was diagnosed in August 2004, although the Public Health team believed that this individual had had active pulmonary TB for a year prior to diagnosis. In May 2005, GO6 was diagnosed with pulmonary tuberculosis, with symptom onset probably in February 2005. All these cases had identical MIRU-VNTR typing patterns. A summary of the MIRU-VNTR profiles for each isolate is shown in Table 6.2.

Table 6.2: A table showing the MIRU-VNTR profiles across 15 loci seen for each of the 13 isolates within this dataset.

		ETR A	ETR B	ETR C	ETR D	ETR E	MIRU2	MIRU10	MIRU16	MIRU20	MIRU23	MIRU24	MIRU26	MIRU27	MIRU39	MIRU40
Outbreak	Townhill	3	2	3	3	3	2	5	3	2	5	1	5	3	2	3
	Gorseinon Pub	3	2	3	3	3	2	5	4	2	5	1	5	3	2	3
	Gorseinon school	3	2	3	3	3	2	5	4	2	5	1	5	3	2	3
Isolates	TH1	3	2	3	3	3	2	5	3	2	5	1	5	3	2	3
	TH2	3	2	3	3	3	2	5	3	2	5	1	5	3	2	3
	TH3	3	2	3	3	3	2	5	3	2	5	1	5	3	2	3
	GS1	3	2	3	3	3	2	5	4	2	5	1	5	3	2	3
	GO1	3	2	3	3	3	2	5	4	2	5	1	5	3	2	3
	GO2	3	2	3	3	3	2	5	4	2	5	1	5	3	2	3
	GO3	3	2	3	3	3	2	5	4	2	5	1	5	3	2	3
	GO4	3	2	3	3	3	2	5	4	2	5	1	5	3	2	3
	GO5	3	2	3	3	3	2	5	4	2	5	1	5	3	2	3
	GO6	3	2	3	3	3	2	5	4	2	5	1	5	3	2	3
	GO7	3	2	3	3	3	2	5	4	2	5	1	5	3	2	3
	GO8	3	2	4	3	3	2	4	3	2	5	1	5	3	2	3
	GO9	3	2	4	3	2	2	5	3	2	5	2	5	3	2	2

The second cluster of cases in the Gorseinon area was identified in 2007. GS1 worked as a dinner lady in a Gorseinon school and was diagnosed in April 2007, with symptom onset being in February of that year. She had the same MIRU-VNTR genotype as the 2003-04 pub cases (Table 6.2) and also frequented Public House M. The identification of case GS1 led to an extension of the outbreak investigation into the local school with 67 children being screened. Seven had positive tuberculin skin tests and 4 needed prophylactic treatments. GS1 had been seen by a medical practitioner three times prior to her diagnosis and been wrongly diagnosed with pleurisy. The final case related to this public house (GO7) was identified in October 2008. This individual was closely linked to GO3 and had contact with GO4, with symptom onset in January 2008. Case GO7 worked in Public House M alongside GO3 and went on to work with GO3 at other public houses in Townhill, Swansea. GO7 was friendly with the spouse of GO3 sleeping regularly in their home and was a close friend of GO4`s daughter. The Public Health team were uncertain whether GO7 had been infected by GO3 or GO4, as all had identical MIRU-VNTR types. A diagrammatic view of the epidemiological connections associated with the public house outbreak in Gorseinon can be seen in Figure 6.1.

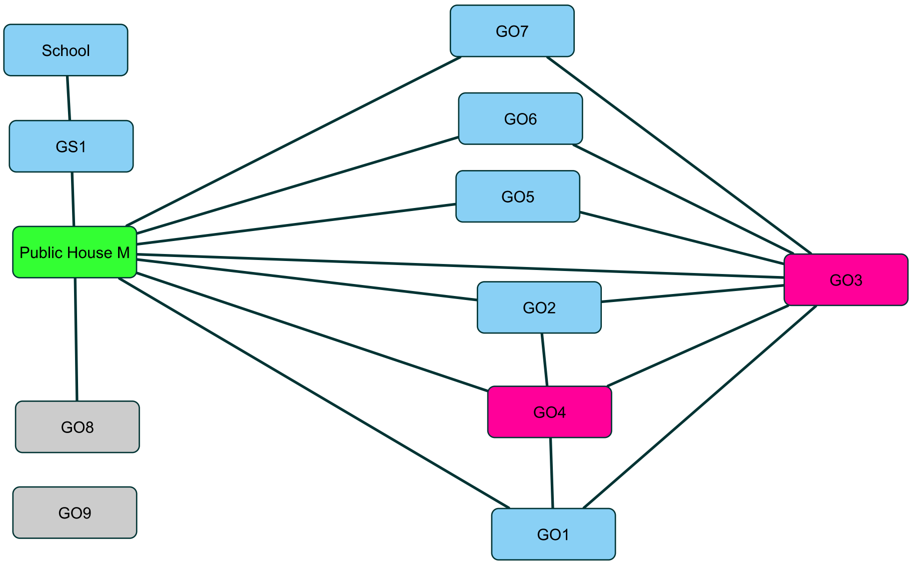


Figure 6.1: A diagrammatic view of the epidemiological connections between all isolates associated with the public house outbreak in Gorseinon. Isolates in pink, GO3 and GO4, represent potential "super spreaders" within the outbreak, and isolates in grey do not share a common MIRU-VNTR profile to the outbreak.

A further two cases (GO8 and GO9) were linked epidemiologically with the Gorseinon outbreak and included in the investigation. However, the MIRU-VNTR profiles of GO8 and GO9 did not match those of the other Gorseinon isolates or one another. Case GO8 admitted to frequenting Public House M during the outbreak but was not diagnosed until October 2008. Although GO9 was diagnosed in August 2004, the individual had onset of symptoms two years earlier in April 2002. Case GO9 was considered potentially involved in the Gorseinon public house and/or school outbreaks but appeared to have no direct contacts with other cases. Due to lack of direct epidemiological links for GO8 and GO9 and the fact that neither of the isolates shared the same MIRU-VNTR pattern as the outbreak isolates, they were not considered part of the outbreak.

6.1.2.2: Townhill outbreak (2006-2008)

In 2008, an outbreak of *M. tuberculosis* occurred in the Townhill region of Swansea. The first case, TH1, was initially admitted to hospital in March 2008 with a suspected chest infection but was found to be suffering from tuberculosis. TH1 had been severely symptomatic for a prolonged period prior to hospital admission, with coughing, weight loss and night sweats going back as far as October 2006. He or she was considered to have been highly infectious for some time. TH1 had disseminated tuberculosis with lung cavities and tuberculomas in the brain and had a history of drug and alcohol abuse. Contact tracing identified numerous contacts. Thirteen cases were found among the contacts - 8 latent and 5 with active infection. Retrospective case-finding identified another Townhill-related case (TH3), diagnosed in January 2007. Patient TH3 resided in the Gendros area of Swansea, about a mile from Townhill and 3.2 miles from Gorseinon (Figure 6.1), but contact tracing documented that TH3 visited public houses in both the Townhill and Gorseinon region during the two outbreaks. Initially no apparent direct links were identified between TH1 and TH3. However, detailed contact tracing of GO3 revealed that he regularly met and socialised with TH3 and GO1 in Public House M. With no direct link to TH1 and with close regular contact with cases of the Gorseinon outbreak, it was believed initially that TH3 was part of that outbreak.

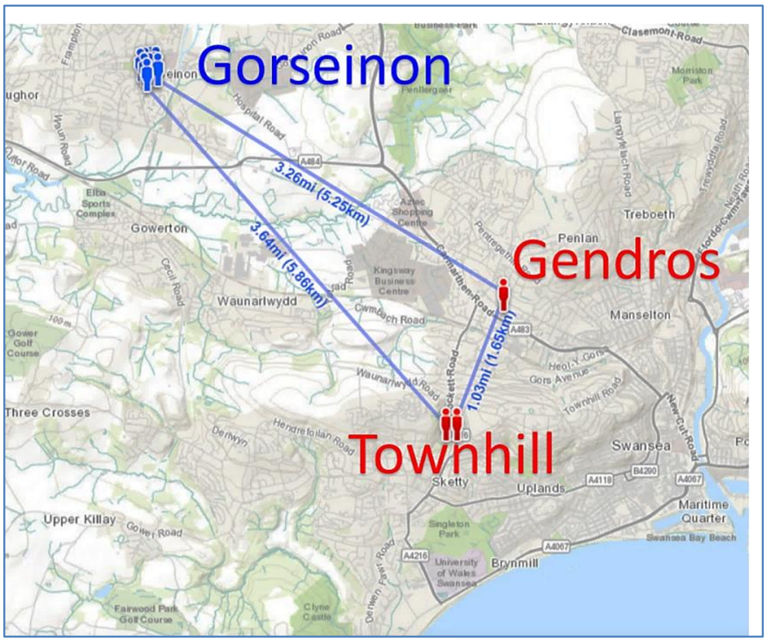


Figure 6.2: A map showing the distances between the 9 Gorseinon and 3 Townhill outbreak cases. One 'Townhill' case resided in the adjacent Gendros area. Further details on the construction of the map, are described in section 2.14.

When compared to the Gorseinion public house outbreak, MIRU-VNTR typing showed that the Townhill outbreak isolates differed by one locus (MIRU16). The MIRU-VNTR typing of the TH3 isolate was identical to that of TH1, showing the same single polymorphism at locus MIRU16. It was at this point that contact tracing began to further investigate links between the Gorseinon and Townhill outbreaks. In November 2008, further retrospective investigations found that the uncle of TH1, TH2, was diagnosed with open tuberculosis in December 2002. The onset of symptoms for TH2 was in July 2002 and thus he was potentially infectious for six months. TH2 also claimed to have been exposed to tuberculosis during time spent in prison between 1989 and 1990. Initial screening in 2002, focused on immediate family (son and wife) and a next-door neighbour, meaning TH1 was not screened during the time of TH2`s contact tracing investigation. It was found that TH2 regularly visited numerous public houses in Swansea, Townhill and Gorseinon and regularly socialised and drank at Public House M at the time of the Gorseinon outbreak. TH2 said he drank in Gorseinon with aman who lived opposite him in Townhill, who was diagnosed with tuberculosis years earlier. Furthermore, another neighbour of TH2 regularly drank, smoked marijuana and consumed other drugs with TH2 around this time, and was also diagnosed with tuberculosis, but the isolate was un-culturable. TH2 also smoked marijuana with TH1. In addition, TH2 was found to socialise heavily with people from the Gendros area where TH3 resided. Thus, the contact tracing team concluded that the Gorseinon and Townhill outbreaks must be linked, despite the one locus difference in MIRU-VNTR typing.

Retrospective contact tracing also found further epidemiological links between isolates from the Gorseinon and Townhill outbreaks. It was found that cases GO3, GO4, GO7 and GO9 all had links with the Townhill area. In particular, it was found that GO4 had an estranged wife who was associated with the Townhill outbreak. GO1 had frequented the same public houses as TH2 and thus was a potential link, and furthermore, there was evidence that the potential super-spreader GO3 had contact with both TH1 and TH2. During the Townhill outbreak, GO7 had contact with a doorman in the Townhill area and thus, despite having matching MIRU-VNTR to the Gorseinon public house outbreak and strong links to potential super-spreader GO4, GO7 was considered a potential link case between the two outbreaks. GO9 frequented public houses in the Townhill area during 2008 in addition to Public House M in Gorseinon. A diagrammatic view of the epidemiological connections related to the Townhill cases is shown in Figure 6.3. The Public Health team concluded that the outbreaks of Gorseinon and Townhill were probably linked in some way and that the small amount of divergence seen in the MIRU-VNTR profiles of both outbreaks must be due to micro-evolution of the strain, with TH2 being deemed the most likely case to link the two outbreaks.

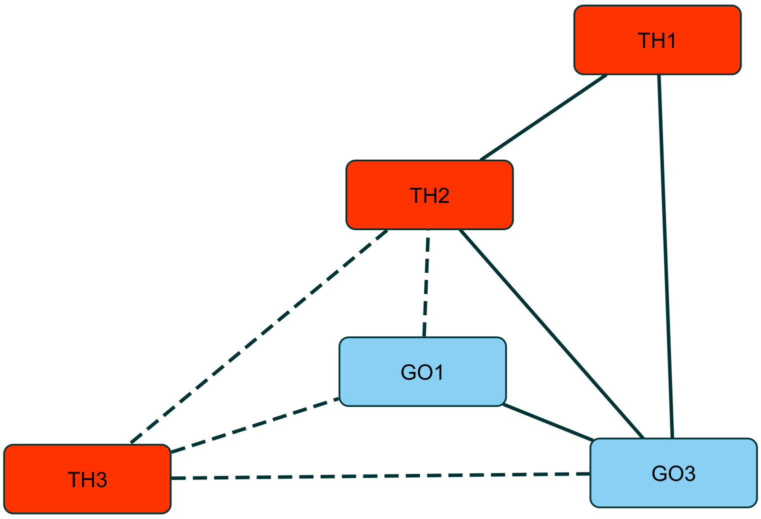


Figure 6.3: A diagrammatic view of the epidemiological connections between the Townhill isolates TH1, TH2 and TH3 and the potentially linked Gorseinon isolates GO1 and GO3. Solid lines refer to confirmed direct contact between the cases; dashed lines refer to potential, but not confirmed contact between cases.

6.1.3: Aims

There were four aims to the work presented this chapter. Firstly, WGS was carried out on the 13 isolates and the previously established gene-by-gene cgMLST typing method used to identify the source case and any super-spreaders within the outbreaks. Secondly the relationship between the Gorseinon and Townhill outbreaks was investigated using genomic data. Thirdly, the outbreaks were further analysed using novel *in silico* spoligotyping (Brudey *et al*., 2006), and their evolutionary relationship elucidated using BEAST ancestral dating (Jombart *et al*., 2014). The final aim was to support or refute the conclusions from the gene-by-gene cgMLST and assess whether this analysis provided a useful alternative to SNP mapping methods for using WGS in clinical studies.

6.1.4: DNA sequencing problems

For 9 of the isolates studied, significant difficulties were experienced in obtaining sequence data from the available DNA. The samples were provided in boilate form. Although boilate extraction does release DNA, it is crude, inconsistent and has been documented to yield DNA of lower integrity, in low quantity and of poorer quality in comparison with other extraction methods (Aldous *et al.*, 2005, Timms *et al.*, 2015).

As might be expected, poor sample quality has a negative effect on the standard Nextera XT library preparation (Tyler *et al.*, 2016), which was used for analysis of the Gorseinon samples. The Nextera XT library preparation is widely used and established within microbial genomics (Tyler *et al.*, 2016). Given the clinical importance of generating WGS data to resolve this outbreak, solutions to this problem would be very beneficial. In this regard, lessons might be learnt from the field of palaeogenomics, which routinely attempts to generate genome-scale data from samples that are highly fragmented, contaminated with non-target DNA, and often contain residual chemical impurities (Pääbo., 1989, Lindahl, 1993., Dabney *et al*., 2013,. Schubert *et al* 2014, Orlando *et al*, 2015, Carøe C *et al*, 2017). For example, recent developments within palaeogenomic methodologies have allowed WGS data to be obtained from a wide range of samples, spanning ancient humans and hominids (Rasmussen, Morten, *et al*, 2010 Green, Richard E., *et al*, 2010), mammals (Miller W.,*et al*, 2008, Orlando L., *et al*, 2013, Skoglund P,. *et al* 2015, Palkopoulou E., *et al* 2015, Mak, S.S.T., *et al* 2017), plants (Jaenicke-Despres, Viviane, *et al*, 2003., Li, Chunxiang, *et al*, 2011., Wales, N., *et al*, 2013.,) and even pathogens (Verena J. Schuenemann *et al*, 2013) – many of which contain DNA with fragment lengths of <80bp. We therefore hypothesised that the application of ancient DNA (aDNA), which refers to DNA derived from ancient specimens such as fossils that is often degraded significantly (M Stoneking, 1995), library preparation and sequencing protocols might overcome the impossibility of retrieving genomic data from the low purity and low-quality DNA of crude *M. tuberculosis* boilate samples.

The Geogenetics Gilbert group at the University of Copenhagen is renowned for research into the molecular analysis and sequencing of degraded ancient DNA (Grealy, A., Phillips *et al*, 2017., Carøe C *et al*, 2018) A European Microbiology Organisation research grant made it possible to visit the Gilbert group and use the ancient DNA sequencing facilities at the University of Copenhagen to obtain usable sequence from nine of the samples.

An ancient DNA library build method, termed the BEST library building protocol (Carøe C *et al*., 2017),was employed on the assumption that the boiled samples would contain DNA in a similarly fragmented state to that from ancient samples (Carøe C *et al*., 2017). The small fragment size of the target DNA is a feature that is sometimes incompatible with standard sequencing equipment such as the Nextera XT protocol used in the previous chapters (Nextera XT, Illumina, 2010). These standard methods use a library preparation step which targets amplicons of a certain consistent size of over 300bps (Illumina, 2015), whilst the ancient DNA method targets fragments as small as 50bps (Carøe C *et al.,* 2017). The use of a library preparation method which originates from museum based paleogenomic research has not previously been attempted in the clinical setting with regards to outbreaks of infectious disease.

6.2: Methods

6.2.1: Standard Sequencing and assembly

The 13 isolates were supplied in boilate form by PHW WCM, Cardiff, Wales, as described in section 2.1. DNA sequencing was carried out directly on these samples. Genomic DNA libraries, genome sequencing and genome assembly were carried out as described in section 2.2. The percentage of reference genome coverage was calculated through submission of isolate sequences to the CSI phylogeny web-server version 1.3 (https://cge.cbs.dtu.dk/services/CSIPhylogeny/). The percentage of core genome MLST coverage was calculated using the Ridom SeqSphere software as described in previous chapters.

6.2.2: Ancient DNA sequencing protocol

Briefly, isolates were purified using the extraction protocol described in Section 2.13.1, before being subjected to the BEST ancient DNA library building protocol and sequencing described in Section 2.13.2 (Carøe C *et al*, 2017).

6.2.3 Core Genome MLST

A previously established gene-by-gene cgMLST typing method (Kohl *et al*., 2014) was used, as described in Sections 2.3 and 2.3.1. Typing was performed successfully on 11 of the sequenced isolates (TH1, TH2, GO1, GO2, GO3, GO4, GO5, GO6, GO7 and GO9).

6.2.4: Extended gene–by-gene MLST.

In addition to the core genes used in the cgMLST, the Ridom SeqSphere also included an accessory genome scheme consisting of 755 genes. Through including the accessory genes to the core genome MLST, a gene-by-gene MLST scheme containing 3646 genes was used for the resolution of relations between the 11 remaining outbreak isolates, see section 2.3.3 for further description.

6.2.5: WGS SNP mapping

Traditional SNP mapping was carried out by using the conserved signature indels (CSI) phylogeny application provided by the Centre of Genomic epidemiology online server, as described in section 2.5.

6.2.6 In silico spoligotype

In silico spoligotyping was carried out using the Python-based SpolTyping (Xia *et al*., 2016) *in silico* software as described in Methods section 2.6.

6.2.7: Ancestral dating

This was carried out as described in section 2.4. SNPs present between the consensus sequence of 'outbreak 1' isolates and TH1, TH2 and GO1 of 'outbreak 2' were extracted using the "extract SNVs from target groups" application within the Ridom SeqSphere software. Outbreak 1 and 2 isolates were as defined by the cgMLST scheme (see results in figure 6.4a). The extraction was based on loci from the cgMLST, and missing alleles were ignored. Indels and SNPs within 10 bps of each other were removed from the analysis, thus a total of 132 SNPs were concatenated.

6.3: Results

6.3.1: Standard sequencing

Unfortunately, standard sequencing failed to provide any sequence data for 11 of the 13 isolates associated with the Gorseinon and Townhill outbreaks (Table 6.3). The heat-to-kill process may have denatured and fragmented the DNA within the samples to a level whereby downstream molecular processing was not possible using standard sequencing protocol. Sequence data could only be obtained for isolates GO8 and GO9 (Table 6.3). Further attempts to clean up and sequence isolates TH3 and GS1 resulted in the loss of those samples, leaving 11 for further analysis.

Table 6.3: Basic sequencing results for the 13 Gorseinon and Townhill outbreak isolates. Included in the table are the number of contigs produced, the largest contig followed by the total length of the reference genome that was covered by sequencing results, the percentage of the cgMLST genes present and the percentage of the reference genome covered according to the CSI phylogeny algorithm provided by the Centre for Genomic Epidemiology (Kaas *et al.*, 2014).

Standard sequencing results					
Sample ID	**#contigs**	**Largest contig**	**Total length**	**% of cgMLST**	**% of reference genome covered***
GO1	0	0	0	0	0
GO2	0	0	0	0	0
G03	0	0	0	0	0
GO4	0	0	0	0	0
GO5	0	0	0	0	0
GO6	0	0	0	0	0
GO7	0	0	0	0	0
GS1	0	0	0	0	0
TH1	0	0	0	0	0
TH2	0	0	0	0	0
TH3	0	0	0	0	0
GO8	143	228436	4353711	98.89	99.41
GO9	1514	20613	4163985	67.45	94.11

6.3.2: Ancient DNA sequencing

Ancient DNA sequencing was carried out successfully on the remaining 9 samples: TH1, TH2, GO1, GO2, GO3, GO4, GO5, GO6 and GO7 (Table 6.4), providing a total length for each isolate of over 4 370 000 bps. The sequence in each case provided data for >99% of the core genes used in the cgMLST scheme. In addition, sequence data were obtained from all isolates that covered >98% of the reference genome according to the CSI phylogeny software of the Centre for Genomic Epidemiology.

Table 6.4: Sequencing results from 9 Gorseinon and Townhill outbreak isolates that failed to sequence using the standard sequencing protocol, obtained using the ancient DNA protocol.

Ancient DNA sequencing results					
Sample ID	#Contigs	Largest contig	Total length	% of cgMLST	% of reference genome covered
TH1	797	174895	4397736	99.52	98.27
TH2	276	171547	4372794	99.34	98.18
GO1	271	174836	4379530	99.52	98.45
GO2	250	174750	4375457	99.52	98.07
GO3	1296	211047	4574259	99.52	98.26
GO4	230	257745	4379980	99.41	98.15
GO5	230	228152	4380862	99.45	98.50
GO6	248	210603	4384763	99.48	98.36
GO7	247	174721	4374335	99.52	98.50

6.3.3: *In silico* spoligotyping

Following ancient DNA sequencing, WGS was now available for a final set of 11 isolates from these 2 outbreaks, and 10 of the 11 isolates were successfully spoligotyped *in silico* (table 6.5) as described in section 6.2.6 and 2.6. SpolTyping software could not process the sequence data provided for GO9 as the quality of its sequence data did not meet that required by the software (Xia *et al*., 2016). All 10 successfully predicted isolates that could be assigned to the lineage 4 Euro-American and Haarlem H1 phylogenetic lineage and spoligotype clade respectively. Within this, three different spoligotypes were found, correlating to three separate international types. Two Gorseinon public house cases (GO2 and GO4) were of international type 742. This spoligotype differs by one locus from that of isolates GO3, GO5, GO6, GO7 and GO8, which were all international type 47. GO1 did not have the same spoligotype pattern as any other Gorseinon public house outbreak isolate but had the same spoligotype pattern and international type as Townhill outbreak cases TH1 and TH2.

Table 6.5: *In silico* spoligotyping results for each of the Gorseinon and Townhill-associated outbreak isolates, showing the predicted spoligotype (produced by SpolDB4), international spoligotype (SITVIT database), lineage assignment and clade assignment (both outputted from the TB-insight online server).

Spoligotype clade assignment				
Isolate	**Predicted Spoligotype**	**International spoligotype**	**Lineage**	**Clade**
GO3	777777774020771	47	Euro-American	H1
TH1	777777770000000	46	Euro-American	H1
TH2	777777770000000	46	Euro-American	H1
GO1	777777770000000	46	Euro-American	H1
GO2	777777770020771	**742**	Euro-American	H1
GO4	777777770020771	**742**	Euro-American	H1
GO5	777777774020771	47	Euro-American	H1
GO6	777777774020771	47	Euro-American	H1
GO7	777777774020771	47	Euro-American	H1
GO8	777777774020771	47	Euro-American	H1
GO9	110100020400000*	Unsuccessful	Unsuccessful	Unsuccessful

6.3.4: cgMLST analysis

The cgMLST-based phylogenies are shown in Figures 6.4 and showed the presence of two outbreaks within the dataset, separated by 124 allelic differences (Figure 6.4). The cgMLST analysis showed each of the Gorseinon public house outbreak cases GO2, GO3, GO4, GO5, GO6 and GO7 formed one clonal outbreak, 'outbreak 1'. Isolates GO2 and GO6 were identical, and GO3, GO4 and GO7 differed from them by only one allelic difference (Table 6.4a). The largest divergence seen within outbreak 1 was 3 allelic differences between GO5 and isolates GO2, GO4 and GO6 (Figure 6.4a). Outbreak 2 was represented by isolates TH1, TH2 and GO1. A clonal complex was also observed in outbreak 2, containing TH1 and GO1, with TH2 diverging from them both by 4 allelic differences (Figure 6.4ba).

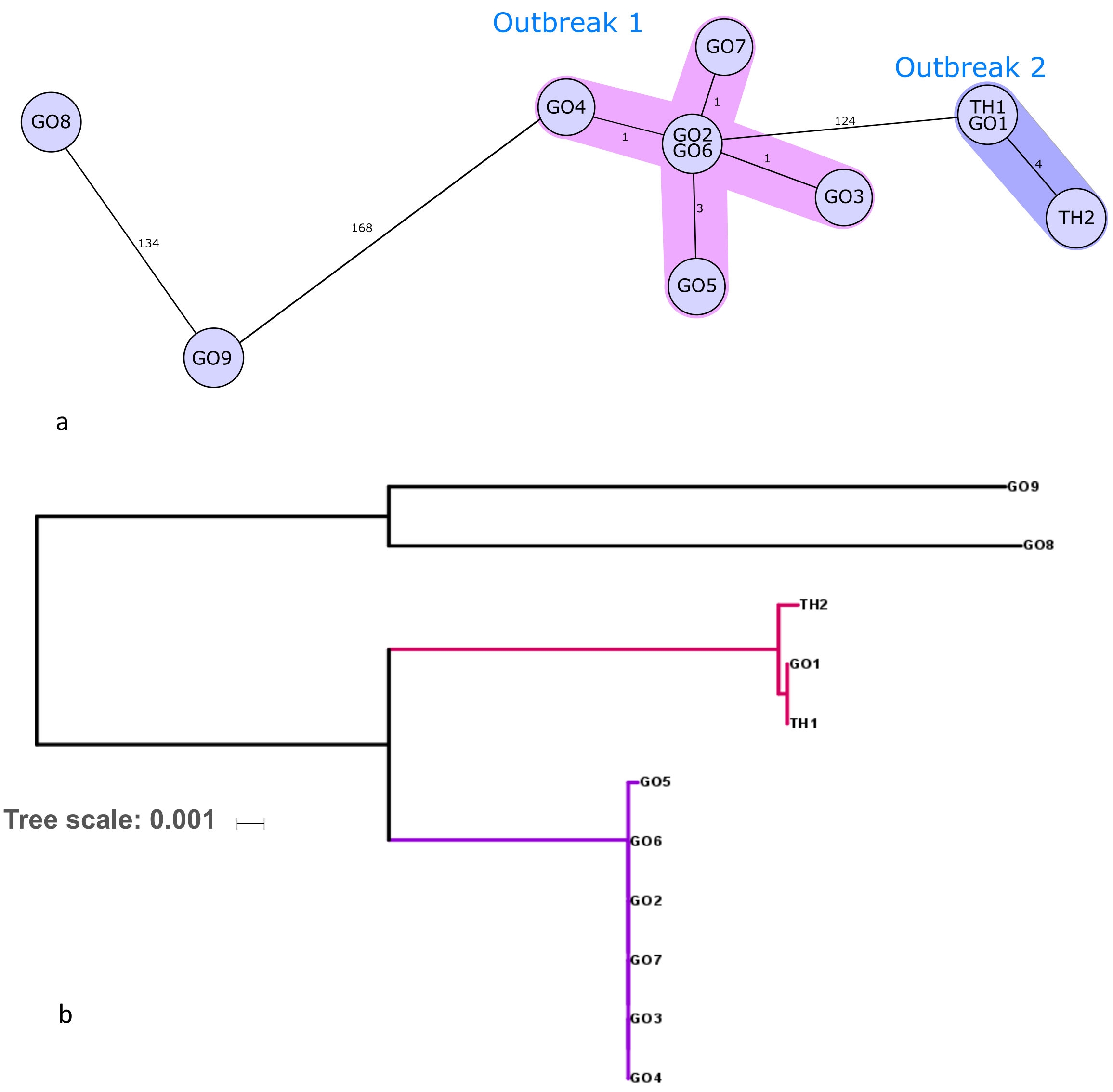


Figure 6.4: cgMLST-based phylogenies of the 11 successfully sequenced isolates. a) A minimum spanning tree containing each of the 11 sequenced isolates. Numbers represent the number of allelic differences between isolates, however the branch lengths between isolates does not represent a measure of divergence. b) A neighbour-joining phylogenetic tree based on the 2891 core genes that comprise the cgMLST. Townhill cluster is coloured pink with the Gorseinon cluster coloured purple. The scale bar highlights the genetic divergence relevant to branch length measured in units of allelic differences per site across the cgMLST scheme of the 11 isolates included in the figure.

Both GO8 and GO9 substantially exceeded the 12-allelic difference threshold for direct transmission from any of the other cases within this dataset. Therefore, according to the cgMLST, they were not outbreak-related isolates, a result consistent with the MIRU-VNTR results. The neighbour-joining phylogeny produced from the cgMLST analysis provided further information on the evolutionary relationships of the isolates. The neighbour joining tree above (Figure 6.4b), shows that the Gorseinon cluster (GO2, GO3, GO4, GO5, GO6 and GO7) harbours a more ancestral position within the tree than the Townhill cluster (GO1, TH1 and TH2) following on from the branching off of both clusters.

The shorter branch length indicated a more ancestral position for the Gorseinon public house cluster (purple) relative to cases TH1, TH2 and GO1. Interestingly, although having a short branch length, TH2 branched off before TH1 and GO1, suggesting it was ancestral to both isolates.

6.3.5: Extended gene-by-gene analysis

The introduction of 755 additional accessory genes not specified as part of the cgMLST scheme, see section 6.2.4 for details, resulted in a minimum spanning tree with the same pattern as cgMLST alone, (Figure 6.5), with the only differences being in the number of allelic differences between the two outbreaks: with 154 allelic differences for extended gene-by-gene analysis of outbreak 1 instead of 124 for cgMLST and within outbreak 2, 5 allelic differences instead of 4. The addition of 755 genes also increased the divergence seen for isolates GO8 and GO9. Outbreak resolution was not impacted by the addition of the accessory genes.

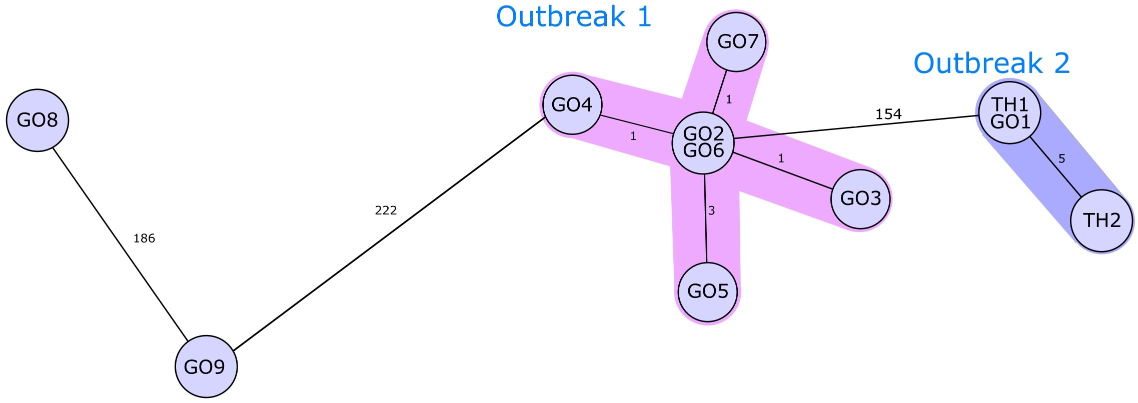


Figure 6.5: A minimum spanning tree of the 11 isolates, based on the cgMLST scheme and 755 accessory genes. Numbers represent the number of allelic differences between isolates, and the branches between isolates do not represent a measure of divergence.

6.3.5.2: Whole Genome SNP mapping

WGS SNP mapping using the Centre for Genomic Epidemiology software (see section 6.2.5 and 2.5 for more details), provided more resolution than any of the previous methods, with all isolates showing at least 4 SNP differences between one and other (Figure 6.6a; Table 6.6). The results found that GO1 was central to outbreak 2 and to be the intermediate isolate between TH1 and TH2, whilst GO3 was identified as the central case within outbreak 1. Again, GO8 and GO9 were found to be clearly divergent from the rest of the isolates. As in the previous cgMLST neighbour-joining tree (Figure 6.4b), the neighbour-joining one here (Figure 6.6b) showed outbreak 1 isolates harboured an ancestral position to TH1, TH2 and GO1, with TH2 branching off at a slightly ancestral position to isolates TH1 and GO1.

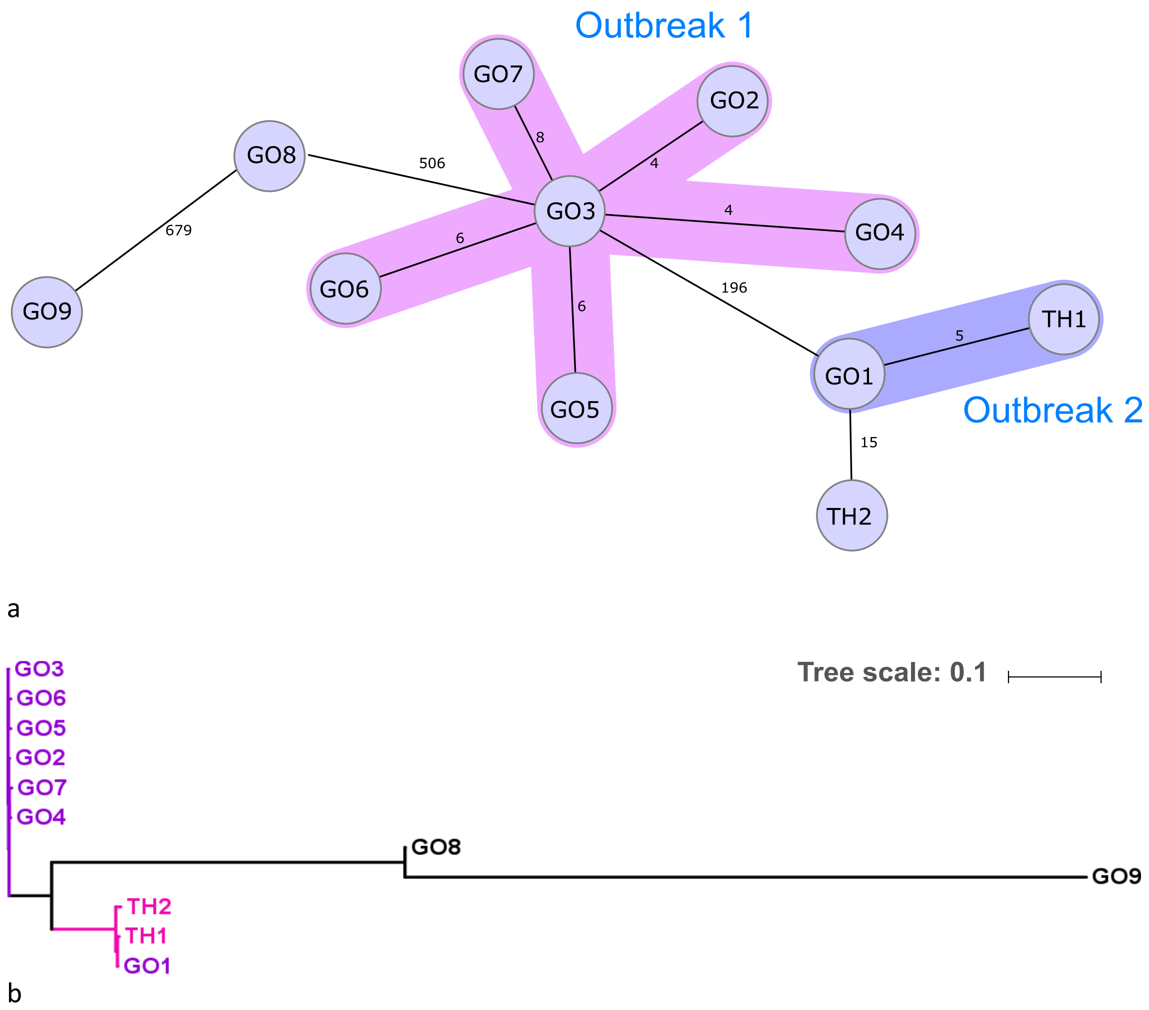


Figure 6.6: Phylogenetic analysis results from WGS SNP mapping. a) A minimum spanning tree based on a total of 1123 SNPs across 11 isolates related to the Gorseinon and Townhill outbreaks. Clusters are highlighted by colour with the threshold for cluster inclusion being <12 SNPs. The numbers seen adjacent to the branches between isolates are the number of SNPs that differ between two given isolates. b) A neighbour-joining phylogeny showing Townhill cases in pink, Gorseinon outbreak cases in purple and the non-matching G08 and GO9 in black. The scale bar (figure 6.6b) highlights the genetic divergence relevant to branch length measured in units of nucleotide differences per site across the Whole Genome sequences of the 11 isolates.

Table 6.6: A table showing the number of SNPs between each isolate found in the WGS SNP mapping phylog within the table are related to the number of SNPs between isolates. The redder the cell the more SNPs, th

	Whole Genome SNP mapping							
Isolate	TH1	GO6	GO7	TH2	GO1	GO2	GO3	GO4
TH1	0	201	199	18	5	197	197	197
GO6	201	0	10	201	200	8	6	9
GO7	199	10	0	201	198	8	8	8
TH2	18	201	201	0	15	197	197	198
GO1	5	200	198	15	0	196	196	196
GO2	197	8	8	197	196	0	4	6
GO3	197	6	8	197	196	4	0	4
GO4	197	9	8	198	196	6	4	0
GO5	197	10	12	203	196	8	6	10
GO8	538	511	511	544	537	509	507	509
GO9	823	794	794	829	822	794	792	794
min: 4 max: 829								

6.3.6: Ancestral dating

To determine the relationship between outbreaks 1 and 2, ancestral dating of the isolates was carried out (Sections 2.4 and 6.2.7). The most recent common ancestor (TMRCA) refers to the last time either outbreaks (or isolates) shared an identical sequence. A diagrammatic representation of the distance in terms of evolutionary chronology between outbreak 1 and outbreak 2 isolates as defined by cgMLST is shown in Figure 6.7. The phylogeny was rooted at 6.001, meaning that the TMRCA between outbreak 1 and 2 isolates existed 6.001 years prior to their current state. The phylogeny also showed that the TMRCA of all three outbreak 2 isolates diverged from the TMRCA of the whole phylogeny (phylogeny root) 5.755 years before diverging into the three isolates, TH1, TH2 and GO1. In addition, the results found that TH2 diverged from TH1 and GO1 0.246 years previous, with TH1 and GO1 diverging from each other more recently, around 0.0387 years prior to their current state.

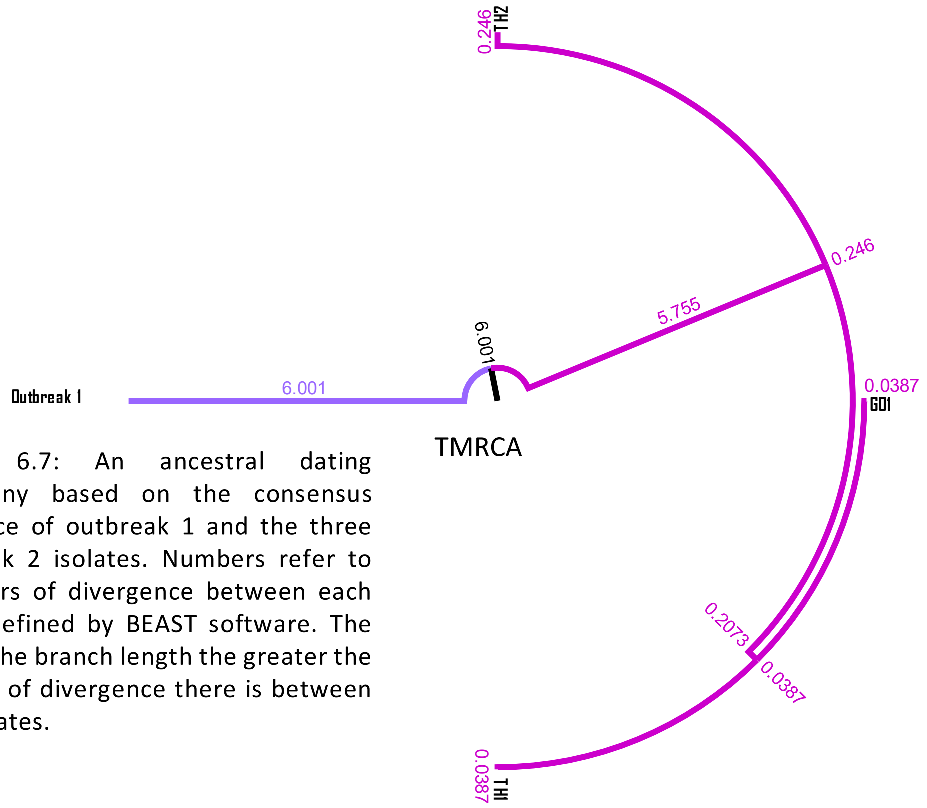


Figure 6.7: An ancestral dating phylogeny based on the consensus sequence of outbreak 1 and the three outbreak 2 isolates. Numbers refer to the years of divergence between each node, defined by BEAST software. The longer the branch length the greater the amount of divergence there is between the isolates.

6.4: Discussion

The application of an ancient DNA-based library building and sequencing method has allowed the analysis of isolates associated with the outbreaks across the Gorseinon and Townhill regions of South Wales, where traditional WGS failed. WGS analysis confirmed the presence of two outbreaks within the dataset correlating with the MIRU-VNTR results and that isolate GO1 had a clonal relationship with the Townhill isolates, in contrast with MIRU-VNTR data, supporting the assumption that both outbreaks were connected. WGS results confirmed isolates GO8 and GO9 to be unrelated to the outbreak, agreeing with the results of MIRU-VNTR. Ancestral sequence reconstruction and dating showed that the isolates of both outbreaks shared a TMRCA six years previously. The evidence presented also supported the hypothesis of micro-evolution proposed by the contact tracing team and a direct relationship between the Townhill and Gorseinon cases.

The DNA used had been extracted from boilate samples of *M. tuberculosis*, and previous studies have highlighted how boilate extractions, although cheap and easily carried out, provide DNA of the lowest quantity and quality in comparison with other extraction methods (Aldous *et al*., 2005, Tyler *et al*., 2016, Cruz., 2017). The standard Nextera XT protocol used in this current study has been shown previously to produce poor or no sequencing results when the starting sample is of poor quality (Tyler *et al*., 2016).

Ancient DNA specialised library building and sequencing methods have been developed to overcome problems of sample contamination, low DNA starting quantity and integrity. They have allowed, for example, the sequencing of Neanderthal, cave bear and mammoth fossil samples (Meyer *et al*., 2007, Dabney *et al*., 2013,. Lindahl, 1993) with the oldest successfully sequenced sample being DNA from a horse discovered in 700,000 year old permafrost (Orlando *et al*., 2013, Dabney *et al*., 2013). Following the application of the BEST ancient DNA library preparation, it was possible to sequence the DNA from 11 of the boilates (Carøe C *et al*., 2017).

Importantly, this is the first clinical application of this ancient DNA sequencing protocol, and despite not being the method of first choice, it has the potential to be extremely useful in cases where DNA sequencing is critical and available DNA quality is poor.

One of the aims was to identify the source case and potential super-spreader(s) within this dataset. Previous studies have highlighted the use of WGS data for the identification of super-spreaders within an outbreak, which cannot be identified by MIRU-VNTR or other typing methods (Walker *et al*., 2013b, Witney *et al*., 2016). Super-spreaders represent cases which have a higher propensity to cause secondary cases and represent highly infectious individuals (Lloyd-Smith *et al*., 2005, Stein., 2011, Walker *et al*., 2013b). They have been associated with outbreaks of *M. tuberculosis* within schools (Ewer *et al*., 2003) and have also been described as central to some community-based outbreaks (Walker *et al*., 2013b). Due to the evolution of *M. tuberculosis* being by descent, the presence of a super-spreader within a cluster of outbreak cases is characterised by a star-like phylogenetic structure (Walker *et al*., 2013b, Kohl *et al*., 2014). Within the Gorseinon outbreak cluster, isolates GO3 and GO4 were considered the most likely super-spreaders epidemiologically, being regulars in the public houses in the area and GO4 having been highly infectious. In addition, contact tracing identified GO5 as being potentially infectious for a year prior to diagnosis and treatment and thus GO5 was also deemed to have the potential to have caused multiple secondary cases.

However, in both the cgMLST and WGS SNP mapping results (Figures 6.3a and 6.5a), a star-like structure was present within the phylogenies. Gene-by-gene based cgMLST identified isolates GO2 and GO6 as the central cases, which differed from what the contact tracing team suspected. The use of a WGS SNP calling method also produced a star-like structure but identified GO3 as the super-spreader central to the outbreak, causing direct transmission to each of the five secondary cases within outbreak 2. As stated, GO3 was considered, along with GO4, the most likely super-spreader in the public house outbreak according to the epidemiological data. Thus, use of the WGS SNP calling method supported at least some of the epidemiological assumptions.

However, the WGS SNP calling method determined that GO4 and GO5 were not super-spreaders despite their epidemiological status as being potentially highly infectious. In addition, the WGS SNP calling results confirmed the status of GO6 and GO7 as secondary cases within the outbreak and found that GO7 was infected by GO3 and not GO4 (a question arising from contact tracing prior to typing). A previous study has highlighted the increased resolution WGS SNP calling phylogenies provide in comparison with cgMLST analysis (Kohl *et al.*, 2014). This study supported those conclusions. However this study also contrasts with the study by Kohl *et al* (2014), which found that despite providing lower resolution, cgMLST provided an adequate alternative to WGS SNP mapping in terms of matching the epidemiological conclusions on transmission events and super-spreader identification. Here, in contrast, cgMLST failed to assign a distinct super-spreader and was not in agreement with transmission events as identified by the SNP mapping and epidemiological data, in terms of both the central role GO3 played within the outbreak and the secondary role GO2 and GO6 had within the outbreak. Thus, despite showing a similar overall outbreak structure, this study found that, for the identification of a super-spreader, the gene-by-gene cgMLST method and WGS SNP calling gave conflicting results. The inferior accuracy of the cgMLST found here also contrasts with the results in other publications which support the use of gene-by-gene WGS MLST methods (such as the cgMLST used here) as an adequate alternative to SNP mapping for analysing outbreak transmission (Sheppard *et al.*, 2012, Maiden *et al.*, 2013).

Despite TH2`s epidemiological links with the Gorseinon area during the outbreak, no genomic evidence of direct transmission between this person and any of the Gorseinon public house cases was found. The TH2 isolate differed from the others in outbreak 2 by at least 15 SNPs. With TH1 being the nephew of TH2 and both being in regular contact, the epidemiological links between the two are overwhelming. A previous study based on *Clostridium difficile* nosocomial outbreaks highlighted how the transmission chain missed intermediate cases which then skewed the true genomic distances and subsequent clinical relationship between isolates within an outbreak, the study applied ancestral dating to WGS data which in turn highlighted the likely presence of such cases (Didelot *et al*., 2012). Due to the overwhelming epidemiological evidence, it is likely that intermediate cases between TH2 and TH1 must have been missed which in turn caused the excessive divergence documented by WGS SNP mapping. Unfortunately, the 2 isolates that could have been the intermediate cases could not be included in the study as isolate TH3 could not be sequenced and the other isolate was unculturable.

In-silico spoligotyping showed that each isolate belonged to the Euro-American lineage H1 Haarlem spoligotype. Diversity was seen with regards to the octal pattern and international types given to each isolate and the results supported those from the WGS data results that showed isolates TH1, TH2 and GO1 to be part of a separate clonal complex (outbreak 2). *In-silico* spoligotyping also supported the cgMLST and epidemiological conclusions which categorise TH2 as being part of a direct transmission chain with isolates TH1 and GO1. This is consistent with the hypothesis that the reason for the divergence seen in the WGS SNP calling was due to an unidentified intermediate case between TH2 and isolates TH1 and GO1.

Isolates GO2 and GO4 had a different octal pattern and international type from isolates GO3, GO5, GO6 and GO7 despite being part of the same clonal complex according to cgMLST and WGS SNP calling methods. It has been documented previously that micro-evolution of the Direct Repeat (DR) locus responsible for spoligotype designation, can mask the true epidemiological relationships between clinically-related isolates (Fang *et al*., 1998, Warren *et al*., 2002). With both cgMLST and WGS SNP calling methods in agreement that GO2 and GO4 are directly related to isolates GO3, GO5, GO6 and GO7, and with there being only one locus difference between the two octal codes for GO2 and GO4 in comparison with the other outbreak 1 isolates, it is plausible that micro-evolution of the DR locus may be responsible for the divergence seen. Isolate GO8 harboured the exact same spoligotype octal pattern and international type as isolates GO3, GO5, GO6 and GO7 despite being divergent in the WGS based result. Therefore, despite not being directly related to any of the isolates within this dataset, the correlating *in* *silico* spoligotype indicates that isolate GO8 *may have* potentially been infected by the same circulating strain that infected GO3, GO5, GO6 and GO7. Micro-evolution of the DR-locus has been shown to occur mainly within the sub-families of spoligotypes (Warren *et al*., 2002, Van Embden *et al*., 2000, Fang *et al*., 1998). In terms of providing a phylogenetic assignment above sub-family level the *in silico* spoligotyping provided reliable results. However, this integrity is not matched when spoligotyping was used to define direct evolutionary relationships between related clinical isolates. Thus it is not reliable for epidemiological conclusions when used alone (Warren *et al*., 2002).

One of the original assumptions by the contact tracing team was that the Townhill and Gorseinon public house outbreaks were caused by the same circulating strain of *M. tuberculosis* that had micro-evolved. *In silico* spoligotyping provided evidence that supported the presence of an endemic circulating strain in the Gorseinon and Townhill areas supporting the contact tracing assumptions. To further assess the possibility of a endemic circulating strain causing both outbreaks, an ancestral dating method (Wohl *et al*., 2016, Jombart *et al*., 2014) was used to estimate the date of divergence between the consensus sequence of 'outbreak 1' and isolates TH1, TH2 and GO1 of 'outbreak 2'. The use of such methods has previously aided infectious disease investigation for identification of the evolution of a certain trait, for a previous outbreak of severe acute respiratory syndrome (SARS) (Jombart *et al*., 2014), outbreak resolution for *Salmonella enterica* (Hawkey *et al.,* 2013), identification of missing cases of *M. tuberculosis* (Bjorn-Mortensen *et al., 2016)*, or species origin of *Yersinia pestis* (Whittles and Didelot, 2016). The results found TH2 to be the ancestral isolate of TH1 and GO1, which correlates to the ancestral position TH2 has within the cgMLST and WGS SNP calling phylogenies. This ancestral position of isolate TH2 was also supported by the epidemiological data that showed TH2 to be the earliest clinical isolate within outbreak 2 (although date of clinical presentation does not necessarily reflect date of infection in tuberculosis). In addition, the phylogeny also showed GO1 and TH1 to have diverged from TMRCA 0.0387 years prior to their isolation, which correlates to the close genetic relationship seen in the results. The data predicted that the transmission chain that linked 'outbreak 1' and 'outbreak 2' dated back 5.755 years prior to the isolation of case TH2. The divergence observed between the two outbreaks could be explained by the lack of intermediate cases between the estimated dates of the TMRCA of both outbreaks.

Previous studies have highlighted how ancestral dating techniques can elucidate more on confusions within outbreak resolution (Roetzer *et al*., 2013a., Ansari and Didelot, 2016). Through the use of ancestral dating, which was further supported by in silico spoligotyping and strong epidemiological data, this study concludes that the divergence of TH1 from TH2 was likely because of missing cases and thus supports previous studies which uses ancestral dating on WGS data to elucidate and help resolve outbreaks where original SNP mapping is questionable (Roetzer *et al*., 2013a., Ansari and Didelot, 2016). This result extends on from Chapter 5 of this thesis which also highlights how ancestral dating aided outbreak understanding whereby genomic divergences seem to unrealistically exceed epidemiological expectations.

6.5: Conclusion

This study has described the first use of an ancient DNA sequencing protocol for clinically relevant outbreak isolates that could not be sequenced by standard protocols. The data presented were broadly in agreement with MIRU-VNTR typing, namely that based on cgMLST and SNP mapping parameters there were two separate outbreaks that occurred separately. However, the application of *in silico* spoligotyping and ancestral reconstruction suggested that the Public Health team was correct in assuming that all isolates were related, as the origin of both outbreaks emerged only 6.001 years earlier than the final case and all isolates spoliogtyped identically as H1 Haarlem strains. The data also supported epidemiological conclusions that a super-spreader was present amongst the cases and suggests that GO3 was the most likely super-spreader, which is supported by the epidemiological data on this investigation. In contrast to work in previous chapters, WGS SNP mapping provided results that correlated with the epidemiological data, whereby cgMLST analysis provided results that could be questioned based on epidemiological data. For future work, it is the potentially missing intermediate cases that warrant further investigation.

Chapter 7

General Discussion

The work carried out for this thesis aimed to provide an in-depth analysis of *M. tuberculosis* outbreaks and phylogenetics primarily using WGS and gene-by-gene MLST analysis software developed and published by Ridom SeqSphere (Jünemann, S *et al*, 2013). Previous studies have shown that the use of a gene-by-gene MLST analysis for WGS data works well (Sheppard *et al.*, 2012, Maiden *et al.*, 2013, Kohl *et al.*, 2014, Bratcher H.B, 2014). However, the authors emphasised the need for the provision of the raw data in a uniform and standardised format. With affordable WGS allowing multiple microbial genomes to be sequenced and analysed, refining the way data are analysed and used for clinical outbreak resolution will inevitably be an issue for application in clinical practice. Previous studies, using traditional SNP mapping pipelines have provided excellent resolution of *M. tuberculosis* outbreaks, going above and beyond earlier typing methods with regards to confirming outbreak clusters (Gardy *et al*., 2011, Walker *et al*., 2013, Walker *et al*., 2015,), identifying outbreak source cases (Walker *et al*., 2018), identification of super-spreaders (Takiff and Feo, 2015), highlighting direct transmission events and more (Walker *et al*., 2013, Witney *et al*., 2016). These studies have confirmed the advantages of using WGS data for outbreak resolution and investigation. However, there are limitations as identified in the present study; the use of "in-house" SNP mapping pipelines across various published studies has meant that the widespread use of WGS in clinical practice is challenging due to the lack of a uniform and standardised format for bioinformatic analysis on large WGS datasets.

Analysis of three outbreaks in South West Wales using cgMLST allowed for adequate resolution of each one, supporting the results of a previous study that used this method for *M. tuberculosis* outbreak resolution (Kohl *et al*., 2014). However, for the Llwynhendy and Gorseinon outbreaks, certain discrepancies were found with regards to cgMLST conclusions. Previous studies have highlighted that WGS gene-by-gene MLST methods allow isolates, for which only partial sequence data were obtained (Sheppard *et al*., 2012), to be included in the downstream bioinformatic analysis (Sheppard *et al*., 2012). However, the results of the Llwynhendy outbreak analysis highlighted that where a parameter was set, such as the 12-allelic difference threshold, isolates with partial sequence data could not be included confidently, such as for LL10. The cgMLST indicated that LL10 showed little divergence from the other isolates associated with the Llwynhendy outbreak, however only 1466 genes (51%) of the 2891 cgMLST scheme could be analysed for LL10. Whether or not this divergence would exceed the threshold when all 2891 genes were present is not known. Therefore, although partial sequence data for isolates can be analysed using the cgMLST scheme, doing so cannot provide a conclusive result, only that LL10 may be worth re-sequencing (Sheppard *et al.,* 2012, Kohl *et al.*, 2014). Although the compromised cgMLST analysis provided evidence for LL10`s inclusion in the outbreak, re-sequencing and the cost that goes with it would still be needed to make any valid conclusions; thus this method would not save time or money in clinical practice. Inevitably, for WGS data to be used clinically in a widespread manner for *M. tuberculosis*, a threshold defining outbreak resolution is needed for successful standardisation. In short, the emphasis should be placed on consistent high-quality sequencing of outbreak isolates rather than focusing on being "able" to use partial sequence data, a strategy which has been suggested previously (Sheppard *et al.*, 2012, Maiden *et al.*, 2013).

In contrast, for the Gorseinon outbreak (Chapter 6), cgMLST successfully resolved conflicts between MIRU-VNTR typing and traditional epidemiological conclusions. The cgMLST clarified that two outbreaks were present; that isolate GO1 despite having the Gorseinon MIRU-VNTR type actually clustered as a Townhill outbreak isolate and confirmed suspicions that isolates GO8 and GO9 were not directly related to the either the Gorseinon or Townhill outbreaks. However, the cgMLST analysis could not define a specific case as the likely source case or super-spreader as postulated by the original contact tracing team. In fact, the results for the Gorseinon outbreak largely supported those seen by Kohl *et al*, (2014), in finding that the cgMLST method provided adequate resolution of the outbreak but was inferior to the resolution provided by traditional SNP mapping. SNP mapping did reveal the presence of GO3 as the super-spreader within the outbreak, a conclusion which was supported by the traditional epidemiological data.

Although the traditional SNP mapping method provided superior resolution for the Gorseinon outbreak, its application for the Llwynhendy and Neath Port Talbot outbreaks (Chapter 4 and 5 respectively) was inconsistent and provided results that were unlikely according to the epidemiology. In these cases, the cgMLST analysis was superior to traditional SNP mapping. Thus, within the present study of three outbreaks, there were inconsistent results from using WGS traditional SNP mapping for outbreak resolution, based on the published parameter of 12 SNPs for outbreak association defined by Walker *et al* (2013). Within this study, the published CSI phylogeny SNP mapping programme available on the Centre of Genomic Epidemiology online server, applying the standard default parameters to each isolate of each outbreak, was used. A major flaw in the application of traditional SNP mapping with regards to consistency was identified. The resolution of the outbreaks within this study was based on using a 12 SNP threshold defined in previous studies (Walker *et al.,* 2013, Kohl *et al*, 2014). However, the studies that originally defined the threshold did not define the specific computational pipeline or sequence quality parameters used to call the SNP`s from their WGS data and rather refer to their pipeline as "in-house". Therefore, as the detailed parameter`s used in those studies are unknown, the 12 SNP threshold applied for the traditional WGS SNP mapping across this thesis is not reliable, as the threshold was calculated based on an "in-house" pipeline as stated by Walker *et al* (2013) which may differ from the CSI phylogeny pipeline used in this study. The inconsistency in the computational pipelines used to call SNP`s from sequence data across different datasets is a flaw in traditional SNP mapping in terms of its widespread use in outbreak resolution.

In contrast, the 12 allelic difference threshold applied for each outbreak was derived from a published set of 2891 core genes used in a previous study (Kohl *et al*, 2014) and commercially available using the Ridom seqSphere software (Jünemann S *et al.*, 2013). Understanding the exact parameters for SNP mapping is essential for consistent outbreak resolution.

The SNP mapping analysis, carried out using the CSI phylogeny pipeline, was based on SNPs across the WGS of each outbreak isolate, thus parameters to exclude more polymorphic regions of the genome were not included in the analysis. The basis to exclude such sites was not defined by Walker *et al* (2013), which first defined the 12 SNP threshold. However, through comparison of the WGS SNP mapping with both an in-house constructed cgMLST and an established cgMLST scheme, it could be concluded that the discrepancies observed between the SNP mapping and the established cgMLST analysis were likely to be caused by the inclusion of polymorphic loci within the SNP mapping analysis. Within *M. tuberculosis*, polymorphic loci refer to hyper variable loci whereby the rate of divergence is disproportionally higher than the rest of the loci in the genome, these loci notoriously include genes of the PPE and PE gene family which accumulate non-synonymous mutations 3 times that of the rest of the genome (Gutacker *et al*., 2002, McEvoy *et al*., 2012). This is a point to consider with regards to the use of WGS gene-by-gene MLST methods as an alternative to WGS SNP mapping.

In addition, the application of the 12 SNP "cut-off" thresholds may also be invalid for our outbreaks for reasons relating to the inconsistency of the pipeline used. Previous research has emphasised that, due to the relatively slow but variable molecular clock of *M. tuberculosis* (Bryant *et al*, 2013), the use of a simple SNP threshold as a "cut-off" point of any kind may not be applicable to this organism in terms of confirming transmission events alone (Bryant *et al*, 2013), a concept supported by the work in this thesis. Instead of providing evidence of direct links between isolates for the Llwynhendy and Neath Port Talbot outbreaks, the SNP mapping results mirrored those in the original Gardy *et al* (2011) SNP mapping study where the pattern of divergence between outbreak isolates in the phylogenies produced by WGS SNP mapping aided the original epidemiological data to resolve discrepancies with MIRU-VNTR typing and identified the presence of super-spreaders (Gardy *et al*., 2011), rather than being used as a standalone method for outbreak resolution.

The outbreaks in Llwynhendy and Neath Port Talbot are more divergent genomically than the Gorseinon one in terms of the WGS SNP mapping of the isolates involved. The polymorphic regions, which were not removed from traditional SNP mapping, are likely to have exacerbated the genomic distances in the less clonal outbreaks of Llwynhendy and Neath Port Talbot. The application of the established cgMLST scheme provided results that were more standardised, reproducible and consistent across each outbreak within this study, in contrast to the results achieved by traditional SNP mapping. For clinical outbreak resolution, cgMLST results provide promise for the use of WGS data in clinical practice which is consistent, presented in a uniform format and based on a standardised set of established core genes from which outbreak association thresholds can be applied consistently using a single piece of software.

In addition to using the cgMLST method for resolution of clinical outbreaks, this was the first use, to our knowledge, of a gene-by-gene method for providing phylogenetic analysis of *M. tuberculosis* isolates. Such analysis has been carried out routinely through traditional SNP mapping for *M. tuberculosis* (Gutacker *et al* 2006, Comas *et al* 2013, Coll *et al*, 2014). The results from this study based on a set of 60 SNPs (Coll *et al*, 2014) showed a close correlation with the initial SNP bar-coding results, clearly highlighting a dominance of Euro-American lineage 4 isolates among the Welsh ones. However, cgMLST statistically could only provide confident assignments to 50 of the 59 *M. tuberculosis* isolates originally assigned into phylogenetic groups by robust SNP bar-coding. In particular, statistical analysis could not provide 100% confidence in the associations of isolates BK21 and BK25, with further sub-lineage specific genotyping needed to confirm their assignment as lineage 2 Beijing strains as previously postulated by the 60 SNP barcode (Coll *et al*, 2014). Globally, failure to confidently identify *M. tuberculosis* Beijing strains is of concern, with this sub-lineage containing strains often regarded as virulent and highly transmissible (Rad *et al* 2003, Marais *et al* 2006, Mestre *et al* 2011). Within this current study, the 60 SNP barcode analysis did provide more reliable and detailed conclusions on the phylogenetics of the Welsh isolates, giving both major and sub-lineage classifications. The 60 SNP barcode also provided a level of standardisation analysing the same 60 loci across any *M. tuberculosis* isolate for which phylogenetic assignment is needed. However, even with a standardised set of SNPs, the computational pipeline needed to extract these was based on a command line script including a combination of various, user-unfriendly command line steps which requires high level computational expertise. The computational complexity and user-unfriendly nature of such a pipeline is clinically relevant as the method would be undesirable for routine laboratory staff and would require widespread employment or training of computational expertise for its routine use across the clinical setting. The cgMLST results did not match the integrity of those made by the SNP barcode. However, gene-by-gene MLST analysis has the potential to provide phylogenetic classifications which are user-friendly, reproducible and uniform. Further research into refining the cgMLST analysis for phylogenetic classification is needed.

This study has provided insights into how WGS data can be used in various applications, other than simply for outbreak resolution. In the phylogenetic analysis alone, it was possible to classify isolates into PGGs, SNP cluster groups and extract strain-specific polymorphisms to support the lineage and sub-lineage assignments made by the application of a robust SNP barcode and the cgMLST association results. The same sequence data were used for each of the phylogenetic analyses, emphasising that WGS data allow for multiple formats to which phylogenetic classifications can be made simultaneously. This contrasts with other typing methods such as spoligotyping and MIRU-VNTR typing, where phylogenetic classification is uniformly designated and would require further separate protocols, such as PCR amplification of the *gyr*B and *kat*G loci for PGG assignment (Sreevatsan *et al.*, 1997).

Various applications of WGS data interpretation can lead to increasing levels of insight to an outbreak and the strains of *M. tuberculosis* within them. The Llwynhendy outbreak was presented initially, to provide an introduction into the application of a gene-by-gene based cgMLST method for outbreak resolution. In Llwynhendy, the epidemiological questions were simple with only the inclusion and exclusion of isolates from an outbreak surrounding the public house needing to be resolved. Therefore, the application of the cgMLST alone was satisfactory for the resolution of this outbreak. The presence of "two" separate outbreaks in Neath Port Talbot provided a more complex scenario and thus ancestral dating was introduced to provide a clearer picture. For the Neath Port Talbot outbreak, cgMLST alone did resolve discrepancies in the original typing, excluding several "B" isolates as wrongly assigned by MIRU-VNTR and including a background isolate previously considered not associated with the outbreak. However, cgMLST did confirm the presence of two outbreaks within the dataset. The use of ancestral dating provided clarity on the relationship between the "two" outbreaks identified by cgMLST, showing that, the isolates had only recently diverged. This supports the original hypothesis by traditional epidemiological conclusions that the Neath Port Talbot cases were caused by one larger outbreak. The ancestral dating indicated that further work is needed to investigate potential intermediate cases not included in the investigation which may bridge the genomic distances between isolates of the separate outbreaks.

The Gorseinon/Townhill outbreak in Chapter 6 showed complexity with the presence of "two" outbreaks from two close but distinct areas. The use of ancestral dating provided further insight into the outbreaks and highlighted that although both cgMLST and WGS SNP mapping identified two; their common ancestor was recent, indicating both were part of a larger outbreak. In addition, *in silico* spoligotyping provided more insight into the demography of the outbreak. As stated, all isolates belonged to the H1 Haarlem Euro American spoligotype family, a result that correlated with the SNP bar-coding phylogenetic assignment of the Gorseinon isolates carried out in Chapter 3. Thus, through addition of *in silico* spoligotyping it was possible to identify the presence of an endemic H1 Haarlem strain circulating in the Gorseinon and Townhill regions. Such evidence provided a more complete picture of the Gorseinon outbreak than the use of cgMLST and WGS SNP mapping alone. For resolving outbreaks of a complex nature, the work presented provides evidence that additional information through ancestral dating; *in silco* spoligotyping and SNP bar-coding is helpful prior to drawing conclusions on the outbreak.

For both the Llwynhendy (Chapter 4) and Neath Port Talbot (Chapter 5) outbreaks, functional analysis of polymorphisms was carried out. Through introducing *in-silico* functional prediction software to this study, further insights into the physiology of outbreak isolates were potentially elucidated. Within the Neath Port Talbot outbreak, Provean functional prediction identified the presence of a deleterious mutation within the *dac*B2 (Rv2195) gene. Inactivation of the DacB2 protein has been shown *in vitro* to cause susceptibility of *M. tuberculosis* to amoxicillin-clavulanic acid treatment (Kumar *et al*., 2012 Diacon *et al*., 2016). This is not to say that treatment with this drug would be indicated in this case, but it showed that, in principle, information that might be useful for treatment could be obtained. Such physiological information cannot be gleaned when typing is carried out by any other molecular typing method.

This study highlights the large amount of information that can be obtained from WGS data and its ability to resolve outbreaks in a more accurate manner than classical methods, which have been documented previously (Gardy *et al*, 2011, Walker *et al*, 2013, Walker *et al*, 2015). Prediction of functional differences caused by polymorphisms, identifying the last common ancestor of separate outbreaks, providing phylogenetic conclusions and more can all be carried out though analysis and processing of the same raw sequence data. Although outbreak resolution is greater using WGS data, it is the wider applications and extensive information obtained from the same original raw sequence data which elevates the potential of WGS for clinical purposes.

The most unique element of my work was the use of an ancient DNA library preparation method for the rescue of poor quality sample DNA essential to resolving the Gorseinon outbreak. Although ancient DNA library preparation and sequencing protocols have been used to investigate pathogens from historical outbreaks, such as the *Yersinisa pestis* plague of 541 AD (Gilbert *et al*, 2014) and the Philadelphia *Vibrio cholerae* outbreak of 1849 (Devault *et al*., 2014), no adaptation of such methods has been recorded in the resolution of a modern-day clinical tuberculosis outbreak, as far as can be established. The method used has allowed WGS data to be obtained from a wide range of samples, spanning ancient humans and hominids (Rasmussen, *et al*, 2010; Green *et al*, 2010), mammals (Miller *et al.*, 2008, Mak *et al.*, 2017) and plants (Li *et al.*, 2011; Wales *et al.*, 2013; Jaenicke-Despres *et al.*, 2003). The adaptation of the method to aid the resolution of the Gorseinon outbreak demonstrated successfully how work in another field has opened the door for the rescuing of genomic data from modern clinical samples of sub-optimal quality despite its origins being based on sequencing DNA from museum-based specimens. The cost of the ancient DNA library preparation is around £50 (Christian Caroe, personal communication), which is over double the cost of the conventional Nextera XT library preparation per sample, which stands at around £24 (Christian Caroe, personal communication). Therefore, its widespread usage would not be cost-effective. In certain circumstances, such as legal cases regarding *M. tuberculosis* infections or in investigation of characteristics important to certain outbreaks, such as drug resistance, it could be justified.

In conclusion, WGS will soon revolutionise clinical outbreak investigations. This study has shown that its applications are above and beyond any other current typing method and supports previous studies regarding the inferior resolution provided by MIRU-VNTR typing. Developments in the standardisation of WGS data without losing its accuracy, inclusion of simultaneous ancestral dating and simultaneous physiological predictions from WGS data will be able to truly exploit the WGS data provided. The cost of conventional WGS has fallen to a level where it is considered affordable for use in clinical outbreak investigations and for complex cases where poor DNA quality is a problem the ancient DNA library preparation method used in this study has potential to rescue important genomic data. Thus, over the next few years, WGS for the resolution, surveillance and treatment of *M. tuberculosis* will no doubt become widespread throughout clinical practice.

Bibliography

ABADIA, E., ZHANG, J., RITACCO, V., KREMER, K., RUIMY, R., RIGOUTS, L., GOMES, H., ELIAS, A., FAUVILLE-DUFAUX, M., STOFFELS, K., RASOLOFO-RAZANAMPARANY, V., DE VIEDMA, D., HERRANZ, M., AL-HAJOJ, S., RASTOGI, N., GARZELLI, C., TORTOLI, E., SUFFYS, P., VAN SOOLINGEN, D., REFREGIER, G. & SOLA, C. 2011. The use of microbead-based spoligotyping for Mycobacterium tuberculosis complex to evaluate the quality of the conventional method: Providing guidelines for Quality Assurance when working on membranes. BMC Infectious Diseases, 11, 110.

AGNEMEMEL, A., HONG, E., GIORGINI, D., NUÑEZ-SAMUDIO, V., DEGHMANE, A.-E. & TAHA, M.-K. 2016. Neisseria meningitidis Serogroup X in Sub-Saharan Africa. Emerging Infectious Diseases, 22, 698.

ALDOUS, W. K., POUNDER, J. I., CLOUD, J. L. & WOODS, G. L. 2005. Comparison of six methods of extracting Mycobacterium tuberculosis DNA from processed sputum for testing by quantitative real-time PCR. Journal of Clinical Microbiology, 43, 2471-2473.

AL-HAJOJ, S. A., AKKERMAN, O., PARWATI, I., AL-GAMDI, S., RAHIM, Z., VAN SOOLINGEN, D., VAN INGEN, J., SUPPLY, P. & VAN DER ZANDEN, A. G. 2010. Microevolution of Mycobacterium tuberculosis in a tuberculosis patient. Journal of Clinical Microbiology, 48, 3813-3816.

ALIX, E., GODREUIL, S. AND BLANC-POTARD, A.B., 2006. Identification of a Haarlem genotype-specific single nucleotide polymorphism in the mgtC virulence gene of Mycobacterium tuberculosis. Journal of clinical microbiology, 44(6), pp.2093-2098.

ALLAND, D., LACHER, D. W., HAZBÓN, M. H., MOTIWALA, A. S., QI, W., FLEISCHMANN, R. D. & WHITTAM, T. S. 2007. Role of large sequence polymorphisms (LSPs) in generating genomic diversity among clinical isolates of Mycobacterium tuberculosis and the utility of LSPs in phylogenetic analysis. Journal of Clinical Microbiology, 45, 39-46.

AMINIAN, M., COUVIN, D., SHABBEER, A., HADLEY, K., VANDENBERG, S., RASTOGI, N. AND BENNETT, K.P., 2014. Predicting Mycobacterium tuberculosis complex clades using knowledge-based Bayesian networks. BioMed research international, 2014.

ANDERSON, J., JARLSBERG, L.G., GRINDSDALE, J., OSMOND, D., KAWAMURA, M., HOPEWELL, P.C. AND KATO-MAEDA, M., 2013. Sublineages of lineage 4 (Euro-American) Mycobacterium tuberculosis differ in genotypic clustering. The International Journal of Tuberculosis and Lung Disease, 17(7), pp.885-891.

ANSARI, M. A. & DIDELOT, X. 2016. Bayesian Inference of the Evolution of a Phenotype Distribution on a Phylogenetic Tree. Genetics, 204, 89-98.

ARJOMANDZADEGAN, M., TITOV, L. P., SURKOVA, L. K., FARNIA, P., SHEIKHOLESLAMI, F., OWLIA, P., ESHGHINEJAD, A. & FARAZI, A. A. 2012. Determination of principal genotypic groups among susceptible, MDR and XDR clinical isolates of Mycobacterium tuberculosis in Belarus and Iran. Tuberk Toraks, 60, 153-159.

ASANTE-POKU, A., YEBOAH-MANU, D., OTCHERE, I. D., ABOAGYE, S. Y., STUCKI, D., HATTENDORF, J., BORRELL, S., FELDMANN, J., DANSO, E. & GAGNEUX, S. 2015. Mycobacterium africanum is associated with patient ethnicity in Ghana. PLoS Neglected Tropical Diseases, 9, e3370.

AUGUSTYNOWICZ-KOPEC, E., JAGIELSKI, T., KOZINSKA, M., ZABOST, A. & ZWOLSKA, Z. 2007. The significance of spoligotyping method in epidemiological investigations of tuberculosis. Pneumonol Alergol Pol, 75, 22-31.

BAKER, L., BROWN, T., MAIDEN, M. C. & DROBNIEWSKI, F. 2004. Silent nucleotide polymorphisms and a phylogeny for Mycobacterium tuberculosis. Emerging Infectious Diseases, 10, 1568.

BALASUBRAMANIAN, V., WIEGESHAUS, E. & SMITH, D. 1992. Growth characteristics of recent sputum isolates of Mycobacterium tuberculosis in guinea pigs infected by the respiratory route. Infection and Immunity, 60, 4762-4767.

BARRY, C. E., LEE, R. E., MDLULI, K., SAMPSON, A. E., SCHROEDER, B. G., SLAYDEN, R. A. & YUAN, Y. 1998. Mycolic acids: structure, biosynthesis and physiological functions. Progress in Lipid Research, 37, 143-179.

BETTS, J. C., LUKEY, P. T., ROBB, L. C., MCADAM, R. A. & DUNCAN, K. 2002. Evaluation of a nutrient starvation model of Mycobacterium tuberculosis persistence by gene and protein expression profiling. Molecular Microbiology, 43, 717-731.

BIFANI, P. J., PLIKAYTIS, B. B., KAPUR, V., STOCKBAUER, K., PAN, X., LUTFEY, M. L., MOGHAZEH, S. L., EISNER, W., DANIEL, T. M. & KAPLAN, M. H. 1996. Origin and interstate spread of a New York City multidrug-resistant Mycobacterium tuberculosis clone family. JAMA, 275, 452-457.

BISHAI, W. R., DANNENBERG, A. M., PARRISH, N., RUIZ, R., CHEN, P., ZOOK, B. C., JOHNSON, W., BOLES, J. W. & PITT, M. L. M. 1999. Virulence of Mycobacterium tuberculosisCDC1551 and H37Rv in Rabbits Evaluated by Lurie's Pulmonary Tubercle Count Method. Infection and Immunity, 67, 4931-4934.

BJORN-MORTENSEN, K., SOBORG, B., KOCH, A., LADEFOGED, K., MERKER, M., LILLEBAEK, T., ANDERSEN, A., NIEMANN, S. & KOHL, T. 2016. Tracing Mycobacterium tuberculosis transmission by whole genome sequencing in a high incidence setting: a retrospective population-based study in East Greenland. Scientific Reports, 6.

BLASER, M. J. & KIRSCHNER, D. 2007. The equilibria that allow bacterial persistence in human hosts. Nature, 449, 843.

BOURAI, N., JACOBS, W. R. & NARAYANAN, S. 2012. Deletion and over expression studies on DacB2, a putative low molecular mass penicillin binding protein from Mycobacterium tuberculosis H 37 Rv. Microbial Pathogenesis, 52, 109-116.

BRATCHER, H.B., CORTON, C., JOLLEY, K.A., PARKHILL, J. AND MAIDEN, M.C., 2014. A gene-by-gene population genomics platform: de novo assembly, annotation and genealogical analysis of 108 representative Neisseria meningitidis genomes. BMC genomics, 15(1), p.1138.

BROMBERG, Y. & ROST, B. 2007. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Research, 35, 3823-3835.

BROSCH, R., GORDON, S. V., PYM, A., EIGLMEIER, K., GARNIER, T. & COLE, S. T. 2000. Comparative genomics of the mycobacteria. International Journal of Medical Microbiology, 290, 143-152.

BROSCH R, GORDON SV, MARMIESSE M, 2002. A new evolutionary scenario for the Mycobacterium tuberculosis complex. Proceedings of the National Academy of Sciences of the United States of America. 2002;99(6):3684-3689. doi:10.1073/pnas.052548299.

BRYANT, J.M., SCHÜRCH, A.C., VAN DEUTEKOM, H., HARRIS, S.R., DE BEER, J.L., DE JAGER, V., KREMER, K., VAN HIJUM, S.A., SIEZEN, R.J., BORGDORFF, M. AND BENTLEY, S.D., 2013. Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data. BMC infectious diseases, 13(1), p.110.

BRUDEY, K., DRISCOLL, J. R., RIGOUTS, L., PRODINGER, W. M., GORI, A., AL-HAJOJ, S. A., ALLIX, C., ARISTIMUÑO, L., ARORA, J. & BAUMANIS, V. 2006a. Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiology, 6, 23.

BYNUM, H., 2012. Spitting Blood: The history of tuberculosis. Oxford University Press.
CARØE C, G. S., VINNER L, MAK SS, SINDING MH, SAMANIEGO JA *ET AL*. 2017. Single-tube library preparation for degraded DNA. Methods in Ecology and Evolution.

CAWS, M., THWAITES, G., DUNSTAN, S., HAWN, T. R., LAN, N. T. N., THUONG, N. T. T., STEPNIEWSKA, K., HUYEN, M. N. T., BANG, N. D. & LOC, T. H. 2008. The influence of host and bacterial genotype on the development of disseminated disease with Mycobacterium tuberculosis. PLoS Pathogen, 4, e1000034.

CAVE, A.J.E. AND DEMONSTRATOR, A., 1939. The evidence for the incidence of tuberculosis in ancient Egypt. British Journal of Tuberculosis, 33(3), pp.142-152.

CHIN, C.-S., SORENSON, J., HARRIS, J. B., ROBINS, W. P., CHARLES, R. C., JEAN-CHARLES, R. R., BULLARD, J., WEBSTER, D. R., KASARSKIS, A. & PELUSO, P. 2011. The origin of the Haitian cholera outbreak strain. New England Journal of Medicine, 364, 33-42.

CHOI, Y. & CHAN, A. P. 2015. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics, btv195.

CHRISTIANSON, S., WOLFE, J., ORR, P., KARLOWSKY, J., LEVETT, P. N., HORSMAN, G. B., THIBERT, L., TANG, P. & SHARMA, M. K. 2010. Evaluation of 24 locus MIRU-VNTR genotyping of Mycobacterium tuberculosis isolates in Canada. Tuberculosis (Edinb), 90, 31-8.

COLE, S. T., BROSCH, R., PARKHILL, J., GARNIER, T., CHURCHER, C., HARRIS, D., GORDON, S. V., EIGLMEIER, K., GAS, S., BARRY, C. E., TEKAIA, F., BADCOCK, K., BASHAM, D., BROWN, D., CHILLINGWORTH, T., CONNOR, R., DAVIES, R., DEVLIN, K., FELTWELL, T., GENTLES, S., HAMLIN, N., HOLROYD, S., HORNSBY, T., JAGELS, K., KROGH, A., MCLEAN, J., MOULE, S., MURPHY, L., OLIVER, K., OSBORNE, J., QUAIL, M. A., RAJANDREAM, M. A., ROGERS, J., RUTTER, S., SEEGER, K., SKELTON, J., SQUARES, R., SQUARES, S., SULSTON, J. E., TAYLOR, K., WHITEHEAD, S. & BARRELL, B. G. 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature, 393, 537-544.

COLDITZ, G.A., BREWER, T.F. AND BERKEY, C.S., 1994. Efficacy of BCG vaccine in the prevention of tuberculosis. J. Am. Med. Assoc, 271, p.698702.

COLL, F., MCNERNEY, R., GUERRA-ASSUNÇÃO, J. A., GLYNN, J. R., PERDIGÃO, J., VIVEIROS, M., PORTUGAL, I., PAIN, A., MARTIN, N. & CLARK, T. G. 2014. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. Nature Communications, 5.

COLLINS, F. M. & SMITH, M. M. 1969. A Comparative Study of the Virulence of Mycobacterium Tuberculosis Measured in Mice and Guinea Pigs 1, 2. American Review of Respiratory Disease, 100, 631-639.

COMAS, I., COSCOLLA, M., LUO, T., BORRELL, S., HOLT, K. E., KATO-MAEDA, M., PARKHILL, J., MALLA, B., BERG, S. & THWAITES, G. 2013. Out-of-Africa migration and Neolithic co expansion of Mycobacterium tuberculosis with modern humans. Nature genetics, 45, 1176-1182.

COMAS, I., HOMOLKA, S., NIEMANN, S. & GAGNEUX, S. 2009. Genotyping of genetically monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the limitations of current methodologies. PLoS One, 4, e7815.

COSCOLLA, M. & GAGNEUX, S. Consequences of genomic diversity in Mycobacterium tuberculosis. Seminars in Immunology, 2014. Elsevier, 431-444.

CRAMPIN, A. C., MWAUNGULU, J. N., MWAUNGULU, F. D., MWAFULIRWA, D. T., MUNTHALI, K., FLOYD, S., FINE, P. E. & GLYNN, J. R. 2010. Recurrent TB: relapse or reinfection? The effect of HIV in a general population cohort in Malawi. AIDS (London, England), 24, 417.

CRUZ, U. O. S. 2017. Troubleshooting DNA Separations on Agarose Gels [Online]. Available: http://bio.classes.ucsc.edu/bio20L/info/content/molbio2/molbio1/troub.htm.

CUBILLOS-RUIZ, A., SANDOVAL, A., RITACCO, V., LÓPEZ, B., ROBLEDO, J., CORREA, N., HERNANDEZ-NEUTA, I., ZAMBRANO, M. M. & DEL PORTILLO, P. 2010. Genomic Signatures of the Haarlem Lineage of Mycobacterium tuberculosis: Implications of Strain Genetic Variation in Drug and Vaccine Development. Journal of Clinical Microbiology, 48, 3614-3623.

DABNEY, J., KNAPP, M., GLOCKE, I., GANSAUGE, M.-T., WEIHMANN, A., NICKEL, B., VALDIOSERA, C., GARCÍA, N., PÄÄBO, S. & ARSUAGA, J.-L. 2013. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proceedings of the National Academy of Sciences, 110, 15758-15763.

DANIEL, T.M., 2006. The history of tuberculosis. Respiratory medicine, 100(11), pp.1862-1870

DAS, S. D., NARAYANAN, S., HARI, L., HOTI, S. L., THANGATHURAI, R. K., CHARLES, N., JAGGARAJAMMA, K. & NARAYANAN, P. R. 2005. Differentiation of highly prevalent IS6110 single-copy strains of Mycobacterium tuberculosis from a rural community in South India with an ongoing DOTS programme. Infection, Genetics and Evolution, 5, 67-77.

DAVID, S., DUARTE, E. L., LEITE, C. Q. F., RIBEIRO, J.-N., MAIO, J.-N., PAIXÃO, E., PORTUGAL, C., SANCHO, L. & DE SOUSA, J. G. 2012. Implication of the RD Rio Mycobacterium tuberculosis sublineage in multidrug resistant tuberculosis in Portugal. Infection, Genetics and Evolution, 12, 1362-1367.

DEFRA,. 2013. Department for Environment Food and Rural Affairs. Request for information: Various Bovine TB costs (2008-2013). URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/239443/DOC120913-12092013094820.pdf

DE JONG, B. C., HILL, P. C., AIKEN, A., AWINE, T., MARTIN, A., ADETIFA, I. M., JACKSON-SILLAH, D. J., FOX, A., KATHRYN, D. & GAGNEUX, S. 2008. Progression to active tuberculosis, but not transmission, varies by Mycobacterium tuberculosis lineage in The Gambia. The Journal of Infectious Diseases, 198, 1037-1043.

DE LA RUA-DOMENECH, R., 2006. Human Mycobacterium bovis infection in the United Kingdom: incidence, risks, control measures and review of the zoonotic aspects of bovine tuberculosis. Tuberculosis, 86(2), pp.77-109.

DEMAY, C., LIENS, B., BURGUIÈRE, T., HILL, V., COUVIN, D., MILLET, J., MOKROUSOV, I., SOLA, C., ZOZIO, T. AND RASTOGI, N., 2012. SITVITWEB–a publicly available international multimarker database for studying Mycobacterium tuberculosis genetic diversity and molecular epidemiology. Infection, Genetics and Evolution, 12(4), pp.755-766.

DEN BOON, S., VERVER, S., LOMBARD, C.J., BATEMAN, E.D., IRUSEN, E.M., ENARSON, D.A., BORGDORFF, M.W. AND BEYERS, N., 2008. Comparison of symptoms and treatment outcomes between actively and passively detected tuberculosis cases: the additional value of active case finding. Epidemiology & Infection, 136(10), pp.1342-1349.

DIACON, A. H., VAN DER MERWE, L., BARNARD, M., VON GROOTE-BIDLINGMAIER, F., LANGE, C., GARCÍA-BASTEIRO, A. L., SEVENE, E., BALLELL, L. & BARROS-AGUIRRE, D. 2016. β-lactams against tuberculosis—new trick for an old dog? New England Journal of Medicine, 375, 393-394.

DORMANDY, T., 1999. The white death: a history of tuberculosis.

DOU, H.Y., TSENG, F.C., LIN, C.W., CHANG, J.R., SUN, J.R., TSAI, W.S., LEE, S.Y., SU, I.J. AND LU, J.J., 2008. Molecular epidemiology and evolutionary genetics of Mycobacterium tuberculosis in Taipei. BMC infectious Diseases, 8(1), p.170.

DRUMMOND, A. J., SUCHARD, M. A., XIE, D. & RAMBAUT, A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution, 29, 1969-1973.

DUDAS, G. & RAMBAUT, A. 2014. Phylogenetic analysis of Guinea 2014 EBOV Ebolavirus outbreak. PLOS Currents Outbreaks.

ENGLAND, T. F. 2017. NFU [Online]. Available: http://www.tbfreeengland.co.uk/faqs/how-much-does-btb-cost/.

EWER, K., DEEKS, J., ALVAREZ, L., BRYANT, G., WALLER, S., ANDERSEN, P., MONK, P. & LALVANI, A. 2003. Comparison of T-cell-based assay with tuberculin skin test for diagnosis of Mycobacterium tuberculosis infection in a school tuberculosis outbreak. The Lancet, 361, 1168-1173.

European Centre for Disease Prevention and Control/WHO Regional Office for Europe, 2014. Tuberculosis surveillance and monitoring in Europe 2015.

European Centre for Disease Prevention and Control/WHO Regional Office for Europe, 2017. Tuberculosis surveillance and monitoring in Europe 2017.

FANG, Z., MORRISON, N., WATT, B., DOIG, C. & FORBES, K. 1998. IS6110 transposition and evolutionary scenario of the direct repeat locus in a group of closely related Mycobacterium tuberculosis strains. Journal of bacteriology, 180, 2102-2109.

FEIL, E. J. & SPRATT, B. G. 2001. Recombination and the population structures of bacterial pathogens. Annual Reviews in Microbiology, 55, 561-590.

FENNER, L., GAGNEUX, S., HELBLING, P., BATTEGAY, M., RIEDER, H. L., PFYFFER, G. E., ZWAHLEN, M., FURRER, H., SIEGRIST, H. H. & FEHR, J. 2012. Mycobacterium tuberculosis transmission in a country with low tuberculosis incidence: role of immigration and HIV infection. Journal of clinical microbiology, 50, 388-395.

FERRO, B. E., NIETO, L. M., ROZO, J. C., FORERO, L. & SOOLINGEN, D. V. 2011. Multidrug-resistant Mycobacterium tuberculosis, Southwestern Colombia.

FEUERRIEGEL, S., KÖSER, C. U. & NIEMANN, S. 2014. Phylogenetic polymorphisms in antibiotic resistance genes of the Mycobacterium tuberculosis complex. Journal of Antimicrobial Chemotherapy, 69, 1205-1210.

FILLIOL, I., MOTIWALA, A. S., CAVATORE, M., QI, W., HAZBÓN, M. H., BOBADILLA DEL VALLE, M., FYFE, J., GARCÍA-GARCÍA, L., RASTOGI, N., SOLA, C., ZOZIO, T., GUERRERO, M. I., LEÓN, C. I., CRABTREE, J., ANGIUOLI, S., EISENACH, K. D., DURMAZ, R., JOLOBA, M. L., RENDÓN, A., SIFUENTES-OSORNIO, J., PONCE DE LEÓN, A., CAVE, M. D., FLEISCHMANN, R., WHITTAM, T. S. & ALLAND, D. 2006b. 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. Journal of Bacteriology, 188, 759-772.

FIRDESSA, R., BERG, S., HAILU, E., SCHELLING, E., GUMI, B., ERENSO, G., GADISA, E., KIROS, T., HABTAMU, M. & HUSSEIN, J. 2013. Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia. Emerg Infect Dis, 19, 460-463.

FITZGIBBON, M., GIBBONS, N., ROYCROFT, E., JACKSON, S., O'DONNELL, J., O'FLANAGAN, D. & ROGERS, T. 2013. A snapshot of genetic lineages of Mycobacterium tuberculosis in Ireland over a two-year period, 2010 and 2011. Euro surveillance: bulletin Européen sur les maladies transmissibles, European Communicable Disease Bulletin.

FLYNN, J. L. & CHAN, J. 2001. Tuberculosis: latency and reactivation. Infection and Immunity, 69, 4195-4201.

FORD, C. B., SHAH, R. R., MAEDA, M. K., GAGNEUX, S., MURRAY, M. B., COHEN, T., JOHNSTON, J. C., GARDY, J., LIPSITCH, M. & FORTUNE, S. M. 2013. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. Nature Genetics, 45, 784-790.

FRITH, J. 2014. History of tuberculosis. Part 1-phthisis, consumption and the white plague. Journal of Military and Veterans Health, 22, 29.

FROTHINGHAM, R. & MEEKER-O'CONNELL, W. A. 1998a. Genetic diversity in the Mycobacterium tuberculosis complex based on variable numbers of tandem DNA repeats. Microbiology, 144, 1189-1196.

GAGNEUX, S. & SMALL, P. M. 2007. Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. The Lancet Infectious Diseases, 7, 328-337.

GAGNEUX, S. 2012. Host–pathogen coevolution in human tuberculosis. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 367, 850-859.

GAGNEUX, S., DERIEMER, K., VAN, T., KATO-MAEDA, M., DE JONG, B. C., NARAYANAN, S., NICOL, M., NIEMANN, S., KREMER, K. & GUTIERREZ, M. C. 2006. Variable host–pathogen compatibility in Mycobacterium tuberculosis. Proceedings of the National Academy of Sciences of the United States of America, 103, 2869-2873.

GALVANI, A.P. AND MAY, R.M., 2005. Epidemiology: dimensions of superspreading. *Nature*, *438*(7066), p.293.

GARDY, J., JOHNSTON, J., HO SUI, S., COOK, V., SHAH, L., BRODKIN, E., REMPEL, S., MOORE, R., ZHAO, Y., HOLT, R., VARHOL, R., BIROL, I., LEM, M., SHARMA, M., ELWOOD, K., JONES, S., BRINKMAN, F., BRUNHAM, R. & TANG, P. 2011. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med, 364, 730 - 739.

GEHRING, A. J., DOBOS, K. M., BELISLE, J. T., HARDING, C. V. & BOOM, W. H. 2004. Mycobacterium tuberculosis LprG (Rv1411c): a novel TLR-2 ligand that inhibits human macrophage class II MHC antigen processing. The Journal of Immunology, 173, 2660-2668.

GHEBREMICHAEL, S., GROENHEIT, R., PENNHAG, A., KOIVULA, T., ANDERSSON, E., BRUCHFELD, J., HOFFNER, S., ROMANUS, V. & KÄLLENIUS, G. 2010. Drug resistant Mycobacterium tuberculosis of the Beijing genotype does not spread in Sweden. PLoS One, 5, e10893.

GLICKMAN, M. S. & JACOBS, W. R. 2001. Microbial pathogenesis of Mycobacterium tuberculosis: dawn of a discipline. Cell, 104, 477-485.

GRANGE, J. 2001. Mycobacterium bovis infection in human beings. Tuberculosis, 81, 71-77.

GREENWOOD, D. 2012. Medical Microbiology, With STUDENTCONSULT online access, 18: Medical Microbiology, Elsevier Health Sciences.

GREEN, RICHARD E., *ET AL*. "A draft sequence of the Neandertal genome." science 328.5979 (2010): 710-722.

GRIFFIN, J. E., GAWRONSKI, J. D., DEJESUS, M. A., IOERGER, T. R., AKERLEY, B. J. & SASSETTI, C. M. 2011. High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. PLoS Pathog, 7, e1002251.

GRIMES, C. Z., TEETER, L. D., HWANG, L.-Y. & GRAVISS, E. A. 2009. Epidemiologic characterization of culture positive Mycobacterium tuberculosis patients by katG-gyrA principal genetic grouping. The Journal of Molecular Diagnostics, 11, 472-481.

GUTACKER, M. M., MATHEMA, B., SOINI, H., SHASHKINA, E., KREISWIRTH, B. N., GRAVISS, E. A. & MUSSER, J. M. 2006a. Single-nucleotide polymorphism–based population genetic analysis of Mycobacterium tuberculosis strains from 4 geographic sites. Journal of Infectious Diseases, 193, 121-128.

GUTACKER, M. M., SMOOT, J. C., MIGLIACCIO, C. A. L., RICKLEFS, S. M., HUA, S., COUSINS, D. V., GRAVISS, E. A., SHASHKINA, E., KREISWIRTH, B. N. & MUSSER, J. M. 2002. Genome-wide analysis of synonymous single nucleotide polymorphisms in Mycobacterium tuberculosis complex organisms: resolution of genetic relationships among closely related microbial strains. Genetics, 162, 1533-1543.

GUTIERREZ, M. C., AHMED, N., WILLERY, E., NARAYANAN, S., HASNAIN, S. E., CHAUHAN, D. S., KATOCH, V. M., VINCENT, V., LOCHT, C. & SUPPLY, P. 2006. Predominance of ancestral lineages of Mycobacterium tuberculosis in India. Emerging Infectious Diseases, 12, 1367.

HALL, T. A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series, 1999. 95-98.

Hawkey J, Edwards DJ, Dimovski K, et al. Evidence of microevolution of Salmonella Typhimurium during a series of egg-associated outbreaks linked to a single chicken farm. BMC Genomics. 2013;14:800. doi:10.1186/1471-2164-14-800.

HEALTH, N. I. O. 2010. National Library of Medicine–Medical Subject Headings. Available from. Accessed May, 15, 2010.

HERMANS, P., VAN SOOLINGEN, D. & VAN EMBDEN, J. 1992. Characterization of a major polymorphic tandem repeat in Mycobacterium tuberculosis and its potential use in the epidemiology of Mycobacterium kansasii and Mycobacterium gordonae. Journal of Bacteriology, 174, 4157-4165.

HERSHBERG, R., LIPATOV, M., SMALL, P. M., SHEFFER, H., NIEMANN, S., HOMOLKA, S., ROACH, J. C., KREMER, K., PETROV, D. A. & FELDMAN, M. W. 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., GERNAEY, A.M., GALILI, E., ESHED, V., GREENBLATT, C.L., LEMMA, E. AND BAR-GAL, G.K., 2008. Detection and molecular characterization of 9000-year-old Mycobacterium tuberculosis from a Neolithic settlement in the Eastern Mediterranean. PloS one, 3(10), p.e3426.

H HERZOG, B., 1998. History of tuberculosis. *Respiration*, *65*(1), pp.5-15.

HESSELING, A.C., RABIE, H., MARAIS, B.J., MANDERS, M., LIPS, M., SCHAAF, H.S., GIE, R.P., COTTON, M.F., VAN HELDEN, P.D., WARREN, R.M. AND BEYERS, N., 2006. Bacille Calmette-Guérin vaccine—induced disease in HIV-infected and HIV-uninfected children. *Clinical infectious diseases*, *42*(4), pp.548-558.
HIRSH, A. E., TSOLAKI, A. G., DERIEMER, K., FELDMAN, M. W. & SMALL, P. M. 2004. Stable association between strains of Mycobacterium tuberculosis and their human host populations. Proceedings of the National Academy of Sciences of the United States of America, 101, 4871-4876.

HOMOLKA, S., PROJAHN, M., FEUERRIEGEL, S., UBBEN, T., DIEL, R., NUBEL, U. & NIEMANN, S. 2012. High resolution discrimination of clinical Mycobacterium tuberculosis complex strains based on single nucleotide polymorphisms. PLoS One, 7, e39855.

HUARD, R. C., CHITALE, S., LEUNG, M., LAZZARINI, L. C. O., ZHU, H., SHASHKINA, E., HO, J. L. (2003). The Mycobacterium tuberculosis Complex-Restricted Gene cfp32 Encodes an Expressed Protein That Is Detectable in Tuberculosis Patients and Is Positively Correlated with Pulmonary Interleukin-10 . Infection and Immunity, 71(12), 6871–6883. http://doi.org/10.1128/IAI.71.12.6871-6883.2003

HURST, L. D. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. TRENDS in Genetics, 18, 486-487.

ILLUMINA. 2015. Nextera XT Library Prep: Tips and Troubleshooting [Online]. Available: https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_nextera/nextera-xt/nextera-xt-troubleshooting-guide.pdf.

JACKSON, C., MOSTOWY, J. H., STAGG, H. R., ABUBAKAR, I., ANDREWS, N. & YATES, T. A. 2016. Working conditions and tuberculosis mortality in England and Wales, 1890–1912: a retrospective analysis of routinely collected data. BMC Infectious Diseases, 16, 215.

JAENICKE-DESPRES, VIVIANE, *et al*. "Early allelic selection in maize as revealed by ancient DNA." Science 302.5648 (2003): 1206-1208.

JENA, L., KASHIKAR, S., KUMAR, S. & HARINATH, B. 2014. Effect of single amino acid mutations on function of Mycobacterium tuberculosis H37RV and H37RA by computational approaches. The Indian Journal of Tuberculosis, 61, 200-206.

JEONG, Y. J., LEE, K. S., KOH, W.-J., HAN, J., KIM, T. S. & KWON, O. J. 2004. Nontuberculous mycobacterial pulmonary infection in immunocompetent patients: comparison of thin-section CT and histopathologic findings. Radiology, 231, 880-886.

JOHNSON, M. M., HOUCK, J. & CHEN, C. 2005. Screening for deleterious nonsynonymous single-nucleotide polymorphisms in genes involved in steroid hormone metabolism and response. Cancer Epidemiology and Prevention Biomarkers, 14, 1326-1329.

JOLLEY, K. A. & MAIDEN, M. C. 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics, 11, 595.

JOLLEY, K. A., HILL, D. M., BRATCHER, H. B., HARRISON, O. B., FEAVERS, I. M., PARKHILL, J. & MAIDEN, M. C. 2012. Resolution of a meningococcal disease outbreak from whole-genome sequence data with rapid Web-based analysis methods. Journal of Clinical Microbiology, 50, 3046-3053.

JOMBART, T., CORI, A., DIDELOT, X., CAUCHEMEZ, S., FRASER, C. & FERGUSON, N. 2014. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLoS Comput Biol, 10, e1003457.

JONSSON, J., HOFFNER, S., BERGGREN, I., BRUCHFELD, J., GHEBREMICHAEL, S., PENNHAG, A. & GROENHEIT, R. 2014. Comparison between RFLP and MIRU-VNTR genotyping of Mycobacterium tuberculosis strains isolated in Stockholm 2009 to 2011. Plos One, 9, e95159.

JUNEMANN, S., SEDLAZECK, F. J., PRIOR, K., ALBERSMEIER, A., JOHN, U., KALINOWSKI, J., MELLMANN, A., GOESMANN, A., VON HAESELER, A., STOYE, J. & HARMSEN, D. 2013. Updating benchtop sequencing performance comparison. Nat Biotechnol, 31, 294-6.

KAAS, R. S., LEEKITCHAROENPHON, P., AARESTRUP, F. M. & LUND, O. 2014. Solving the problem of comparing whole bacterial genomes across different sequencing platforms. PLoS One, 9, e104984.

KAMERBEEK, J., SCHOULS, L., KOLK, A., VAN AGTERVELD, M., VAN SOOLINGEN, D., KUIJPER, S., BUNSCHOTEN, A., MOLHUIZEN, H., SHAW, R. & GOYAL, M. 1997. Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. Journal of Clinical Microbiology, 35, 907-914.

KENDALL, W.B., 1915. Artificial Pneumothorax in the Treatment of Tuberculosis. Canadian Medical Association Journal, 5(3), p.206.

KHANIPOUR, S., EBRAHIMZADEH, N., MASOUMI, M., SAKHAEI, F., ALINEZHAD, F., SAFARPOUR, E., FATEH, A., NEMATOLLAHI, A. N., TASBITI, A. H. & ZOLFAGHARI, M. R. 2016. Haarlem 3 is the predominant genotype family in multidrug-resistant and extensively drug-resistant Mycobacterium tuberculosis in the capital of Iran: A 5-year survey. Journal of Global Antimicrobial Resistance, 5, 7-10.

KIMURA, M. 1984. The neutral theory of molecular evolution, Cambridge University Press.
KING, H. C., KHERA-BUTLER, T., JAMES, P., OAKLEY, B. B., ERENSO, G., ASEFFA, A., KNIGHT, R., WELLINGTON, E. M. & COURTENAY, O. 2017. Environmental reservoirs of pathogenic mycobacteria across the Ethiopian biogeographical landscape. PloS One, 12, e0173811.

KOHL, T. A., DIEL, R., HARMSEN, D., ROTHGÄNGER, J., WALTER, K. M., MERKER, M., WENIGER, T. & NIEMANN, S. 2014. Whole-genome-based Mycobacterium tuberculosis surveillance: a standardized, portable, and expandable approach. Journal of Clinical Microbiology, 52, 2479-2486.

KÖSER, C.U., HOLDEN, M.T., ELLINGTON, M.J., CARTWRIGHT, E.J., BROWN, N.M., OGILVY-STUART, A.L., HSU, L.Y., CHEWAPREECHA, C., CROUCHER, N.J., HARRIS, S.R. AND SANDERS, M., 2012. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. New England Journal of Medicine, 366(24), pp.2267-2275.

KUMAR, P., ARORA, K., LLOYD, J. R., LEE, I. Y., NAIR, V., FISCHER, E., BOSHOFF, H. I. & BARRY, C. E. 2012. Meropenem inhibits D, D-carboxypeptidase activity in Mycobacterium tuberculosis. Molecular microbiology, 86, 367-381.

LARKIN, M. A., BLACKSHIELDS, G., BROWN, N., CHENNA, R., MCGETTIGAN, P. A., MCWILLIAM, H., VALENTIN, F., WALLACE, I. M., WILM, A. & LOPEZ, R. 2007. Clustal W and Clustal X version 2.0. Bioinformatics, 23, 2947-2948.

LARROUY-MAUMUS, G., BISWAS, T., HUNT, D. M., KELLY, G., TSODIKOV, O. V. & DE CARVALHO, L. P. S. 2013. Discovery of a glycerol 3-phosphate phosphatase reveals glycerophospholipid polar head recycling in Mycobacterium tuberculosis. Proceedings of the National Academy of Sciences, 110, 11320-11325.

LE, V. T. M. & DIEP, B. A. 2013. Selected insights from application of whole genome sequencing for outbreak investigations. Current Opinion in Critical Care, 19, 432.

LEE, J.-H., AMMERMAN, N. C., NOLAN, S., GEIMAN, D. E., LUN, S., GUO, H. & BISHAI, W. R. 2012. Isoniazid resistance without a loss of fitness in Mycobacterium tuberculosis. Nature communications, 3, 753.

LETUNIC, I. AND BORK, P., 2006. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics, 23(1), pp.127-128.

LEVITT, I. 2003. TB, Glasgow and the Mass Radiography Campaign in the Nineteen Fifties: A Democratic Health Service in Action, University of Glasgow Centre for the History of Medicine.

LEW, J., KAPOPOULOU, A., JONES, L. & COLE, S. 2011. TubercuList-10years after. Tuberculosis (Edinb), 91, 1 - 7.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. AND DURBIN, R., 2009. The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), pp.2078-2079.

LI, H. AND DURBIN, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 25(14), pp.1754-1760.

LI, Z. 2014. Detecting the correlated mutations based on selection pressure with Core Mut. URL : https://bioc.ism.ac.jp/packages/3.0/bioc/vignettes/CorMut/inst/doc/CorMut.pdf.

Li, Chunxiang, *et al*. "Ancient DNA analysis of desiccated wheat grains excavated from a Bronze Age cemetery in Xinjiang." Journal of Archaeological Science 38.1 (2011): 115-119.

LINDAHL, T. 1993.Instability and decay of the primary structure of DNA. Nature, 362, 709-715.

LIU, X., GUTACKER, M. M., MUSSER, J. M. & FU, Y.-X. 2006. Evidence for recombination in Mycobacterium tuberculosis. Journal of Bacteriology, 188, 8169-8177.

LIU, Q., VIA, L.E., LUO, T., LIANG, L., LIU, X., WU, S., SHEN, Q., WEI, W., RUAN, X., YUAN, X. AND ZHANG, G., 2015. Within patient microevolution of Mycobacterium tuberculosis correlates with heterogeneous responses to treatment. Scientific reports, 5, p.17507.

LLOYD-SMITH, J. O., SCHREIBER, S. J., KOPP, P. E. & GETZ, W. M. 2005. Superspreading and the effect of individual variation on disease emergence. Nature, 438, 355-359.

MAIDEN, M. C., VAN RENSBURG, M. J. J., BRAY, J. E., EARLE, S. G., FORD, S. A., JOLLEY, K. A. & MCCARTHY, N. D. 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. Nature Reviews Microbiology, 11, 728-736.

MAIDEN, M.C., VAN RENSBURG, M.J.J., BRAY, J.E., EARLE, S.G., FORD, S.A., JOLLEY, K.A. AND MCCARTHY, N.D., 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. Nature Reviews Microbiology, 11(10), p.728.

MAK, S.S.T., GOPALAKRISHNAN, S., CARØE, C., GENG, C., LIU, S., SINDING, M.H.S., KUDERNA, L.F., ZHANG, W., FU, S., VIEIRA, F.G. AND GERMONPRÉ, M., 2017. Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. GigaScience, 6(8), pp.1-13.

MANCA, C., TSENOVA, L., BARRY, C. E., BERGTOLD, A., FREEMAN, S., HASLETT, P. A., MUSSER, J. M., FREEDMAN, V. H. & KAPLAN, G. 1999. Mycobacterium tuberculosis CDC1551 induces a more vigorous host response in vivo and in vitro, but is not more virulent than other clinical isolates. The Journal of Immunology, 162, 6740-6746.

MARAIS, B. J., VICTOR, T. C., HESSELING, A. C., BARNARD, M., JORDAAN, A., BRITTLE, W., REUTER, H., BEYERS, N., VAN HELDEN, P. D. & WARREN, R. M. 2006. Beijing and Haarlem genotypes are overrepresented among children with drug-resistant tuberculosis in the Western Cape Province of South Africa. Journal of Clinical Microbiology, 44, 3539-3543.

MARCH, F., COLL, P., COSTA, R., RODRÍGUEZ, P., MORENO, C., GARRIGÓ, M. & PRATS, G. 1996. Usefulness of DR, PGRS, and spoligotyping in the typing of Mycobacterium tuberculosis. Comparison with IS6110. Enfermedades Infecciosas y Microbiologia Clinica, 14, 160-166.

MARDASSI, H., NAMOUCHI, A., HALTITI, R., ZARROUK, M., MHENNI, B., KARBOUL, A., KHABOUCHI, N., VAN PITTIUS, N. C. G., STREICHER, E. M. & RAUZIER, J. 2005. Tuberculosis due to resistant Haarlem strain, Tunisia. Tuberculosis.

MAZARS, E., LESJEAN, S., BANULS, A.L., GILBERT, M., VINCENT, V., GICQUEL, B., TIBAYRENC, M., LOCHT, C. AND SUPPLY, P., 2001. High-resolution minisatellite-based typing as a portable approach to global analysis of Mycobacterium tuberculosis molecular epidemiology. Proceedings of the national academy of Sciences, 98(4), pp.1901-1906.

MCEVOY, C. R., CLOETE, R., MÜLLER, B., SCHÜRCH, A. C., VAN HELDEN, P. D., GAGNEUX, S., WARREN, R. M. & VAN PITTIUS, N. C. G. 2012. Comparative analysis of Mycobacterium tuberculosis pe and ppe genes reveals high sequence variation and an apparent absence of selective constraints. PloS One, 7, e30593.

MELLMANN, A., HARMSEN, D., CUMMINGS, C.A., ZENTZ, E.B., LEOPOLD, S.R., RICO, A., PRIOR, K., SZCZEPANOWSKI, R., JI, Y., ZHANG, W. AND MCLAUGHLIN, S.F., 2011. Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104: H4 outbreak by rapid next generation sequencing technology. PloS one, 6(7), p.e22751.

MÉRIC, G., YAHARA, K., MAGEIROS, L., PASCOE, B., MAIDEN, M. C., JOLLEY, K. A. & SHEPPARD, S. K. 2014. A reference pan-genome approach to comparative bacterial genomics: identification of novel epidemiological markers in pathogenic Campylobacter. PloS One, 9, e92798.

MESTRE, O., LUO, T., DOS VULTOS, T., KREMER, K., MURRAY, A., NAMOUCHI, A., JACKSON, C., RAUZIER, J., BIFANI, P. & WARREN, R. 2011. Phylogeny of Mycobacterium tuberculosis Beijing strains constructed from polymorphisms in genes involved in DNA replication, recombination and repair. PLoS One, 6, e16020.

MEYER, M., BRIGGS, A. W., MARICIC, T., HÖBER, B., HÖFFNER, B., KRAUSE, J., WEIHMANN, A., PÄÄBO, S. & HOFREITER, M. 2007. From micrograms to picograms: quantitative PCR reduces the material demands of high-throughput sequencing. Nucleic Acids Research, 36, e5-e5.

MICHAEL, P. 2008. Public Health in Wales (1800-2000).

MILBURN, H., ASHMAN, N., DAVIES, P., DOFFMAN, S., DROBNIEWSKI, F., KHOO, S., ORMEROD, P., OSTERMANN, M. & SNELSON, C. 2010. Guidelines for the prevention and management of Mycobacterium tuberculosis infection and disease in adult patients with chronic kidney disease. Thorax, 65, 559-570.

MILLER, W., DRAUTZ, D.I., RATAN, A., PUSEY, B., QI, J., LESK, A.M., TOMSHO, L.P., PACKARD, M.D., ZHAO, F., SHER, A. AND TIKHONOV, A., 2008. Sequencing the nuclear genome of the extinct woolly mammoth. Nature, 456(7220), p.387.

MILLÁN-LOU, M. I., LÓPEZ-CALLEJA, A. I., COLMENAREJO, C., LEZCANO, M. A., VITORIA, M. A., DEL PORTILLO, P., OTAL, I., MARTÍN, C. & SAMPER, S. 2013. Global study of IS6110 in a successful Mycobacterium tuberculosis strain: clues for deciphering its behaviour and for its rapid detection. Journal of Clinical Microbiology, 51, 3631-3637.

MITCHISON, D., BHATIA, A., RADHAKRISHNA, S., SELKON, J., SUBBAIAH, T. & WALLACE, J. 1961. The virulence in the guinea-pig of tubercle bacilli isolated before treatment from South Indian patients with pulmonary tuberculosis: 1. Homogeneity of the investigation and a critique of the virulence test. Bulletin of the World Health Organization, 25, 285.

MITCHISON, D., WALLACE, J., BHATIA, A., SELKON, J., SUBBAIAH, T. & LANCASTER, M. 1960. A comparison of the virulence in guinea-pigs of South Indian and British tubercle bacilli. Tubercle, 41, 1-22.

MORAN-GILAD, J., PRIOR, K., YAKUNIN, E., HARRISON, T., UNDERWOOD, A., LAZAROVITCH, T., VALINSKY, L., LUCK, C., KRUX, F. & AGMON, V. 2015. Design and application of a core genome multilocus sequence typing scheme for investigation of Legionnaires' disease incidents. Euro Surveillance, 20, 21087.

MOSTOWY, S., COUSINS, D., BRINKMAN, J., ARANAZ, A. & BEHR, M. A. 2002. Genomic deletions suggest a phylogeny for the Mycobacterium tuberculosis complex. The Journal of Infectious Diseases, 186, 74-80.

MÜLLER, R., ROBERTS, C. A. & BROWN, T. A. Genotyping of ancient Mycobacterium tuberculosis strains reveals historic genetic diversity. Proc. R. Soc. B, 2014. The Royal Society, 20133236.

MUSSER, J. M., AMIN, A. & RAMASWAMY, S. 2000. Negligible genetic diversity of Mycobacterium tuberculosis host immune system protein targets: evidence of limited selective pressure. Genetics, 155, 7-16.

NABEYA, D., KINJO, T., PARROTT, G. L., UEHARA, A., MOTOOKA, D., NAKAMURA, S., NAHAR, S., NAKACHI, S., NAKAMATSU, M. & MAESHIRO, S. 2017. The clinical and phylogenetic investigation for a nosocomial outbreak of respiratory syncytial virus infection in an adult hemato-oncology unit. Journal of Medical Virology.

NGUYEN, V. 1997. Diagnosis and treatment of disease caused by nontuberculous mycobacteria. This official statement of the American Thoracic Society was approved by the

Board of Directors, March 1997. Medical Section of the American Lung Association. Am J Respir Crit Care Med, 156, S1-S25.

NICOL, M. P., SOLA, C., FEBRUARY, B., RASTOGI, N., STEYN, L. & WILKINSON, R. J. 2005. Distribution of strain families of Mycobacterium tuberculosis causing pulmonary and extrapulmonary disease in hospitalized children in Cape Town, South Africa. Journal of Clinical Microbiology, 43, 5779-5781.

NIEMANN, S., KUBICA, T., BANGE, F., ADJEI, O., BROWNE, E., CHINBUAH, M., DIEL, R., GYAPONG, J., HORSTMANN, R. & JOLOBA, M. 2004. The species Mycobacterium africanum in the light of new molecular markers. Journal of clinical microbiology, 42, 3958-3962.

NORTH, R. J., RYAN, L., LACOURCE, R., MOGUES, T. & GOODRICH, M. E. 1999. Growth rate of mycobacteria in mice as an unreliable indicator of mycobacterial virulence. Infection and Immunity, 67, 5483-5485.

OCTAVIA, S., WANG, Q., TANAKA, M. M., SINTCHENKO, V. & LAN, R. 2017. Genomic heterogeneity of Salmonella enterica serovar Typhimurium bacteriuria from chronic infection. Infection, Genetics and Evolution, 51, 17-20.

OJO, O. O., SHEEHAN, S., CORCORAN, D. G., NIKOLAYEVSKY, V., BROWN, T., O'SULLIVAN, M., O'SULLIVAN, K., GORDON, S. V., DROBNIEWSKI, F. & PRENTICE, M. B. 2010. Molecular epidemiology of Mycobacterium tuberculosis clinical isolates in Southwest Ireland. Infection, Genetics and Evolution, 10, 1110-1116.

OLAITAN, A. O., DIENE, S. M., KEMPF, M., BERRAZEG, M., BAKOUR, S., GUPTA, S. K., THONGMALAYVONG, B., AKKHAVONG, K., SOMPHAVONG, S. & PABORIBOUNE, P. 2014. Worldwide emergence of colistin resistance in Klebsiella pneumoniae from healthy humans and patients in Lao PDR, Thailand, Israel, Nigeria and France owing to inactivation of the PhoP/PhoQ regulator mgrB: an epidemiological and molecular study. International Journal of Antimicrobial Agents, 44, 500-507.

ORGANIZATION, W. H. 2017. Guidelines for the treatment of drug-susceptible tuberculosis and patient care, 2017 Update.

ORLANDO, L., GINOLHAC, A., ZHANG, G., FROESE, D., ALBRECHTSEN, A., STILLER, M., SCHUBERT, M., CAPPELLINI, E., PETERSEN, B. & MOLTKE, I. 2013. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature, 499, 74-78.

PÄÄBO, S., HIGUCHI, R.G. AND WILSON, A.C., 1989. Ancient DNA and the polymerase chain reaction: the emerging field of molecular archaeology (Minireview). The Journal of biological chemistry, 264(17), pp.9709-9712.

PANG, Y., LU, J., WANG, Y., SONG, Y., WANG, S. & ZHAO, Y. 2013. Study of the rifampin monoresistance mechanism in Mycobacterium tuberculosis. Antimicrobial Agents and Chemotherapy, 57, 893-900.

PARISH, T. & BROWN, A. C. 2009. Mycobacterium: genomics and molecular biology, Horizon Scientific Press.

PARK, Y.-K., BAI, G.-H. & KIM, S.-J. 2000. Restriction Fragment Length Polymorphism Analysis of Mycobacterium tuberculosis Isolated from Countries in the Western Pacific Region. Journal of Clinical Microbiology, 38, 191-197.

PÉREZ-LAGO, L., COMAS, I., NAVARRO, Y., GONZÁLEZ-CANDELAS, F., HERRANZ, M., BOUZA, E. & GARCÍA-DE-VIEDMA, D. 2013. Whole genome sequencing analysis of intrapatient microevolution in Mycobacterium tuberculosis: potential impact on the inference of tuberculosis transmission. The Journal of Infectious Diseases, 209, 98-108.

PÉREZ-LAGO, L., HERRANZ, M., LIROLA, M. M., BOUZA, E. & DE VIEDMA, D. G. 2011. Characterization of microevolution events in Mycobacterium tuberculosis strains involved in recent transmission clusters. Journal of Clinical Microbiology, 49, 3771-3776.

PUBLIC HEALTH ENGLAND.2013. Tuberculosis mortality and mortality rate, England and Wales, 1913-2013. In: STATISTICS, O. F. N. (ed.).

PUBLIC HEALTH ENGLAND. 2014. Tuberculosis in the UK 2014 report. In: ENGLAND, P. H. (ed.).

PUBLIC HEALTH ENGLAND. 2016. Tuberculosis in England: 2016 report. In: ENGLAND, P. H. (ed.).

PUBLIC HEALTH WALES. 2014. Tuberculosis in Wales. Annual Report 2014: Data to the end of 2013.

PUBLIC HEALTH WALES. 2016. Tuberculosis in Wales. Annual Report 2016: Data to the end of 2015.

RAMAN, K. & CHANDRA, N. 2008. Mycobacterium tuberculosis interactome analysis unravels potential pathways to drug resistance. BMC Microbiology, 8, 234.

RAMAZANZADEH, R., ROSHANI, D., SHAKIB, P. & ROUHI, S. 2015. Prevalence and occurrence rate of Mycobacterium tuberculosis Haarlem family multi-drug resistant in the worldwide population: A systematic review and meta-analysis. Journal of research in medical sciences: The Official Journal of Isfahan University of Medical Sciences, 20, 78.

RASMUSSEN, MORTEN,.2010. "Ancient human genome sequence of an extinct Palaeo-Eskimo." Nature 463.7282 (2010): 757-762.

RASIGADE, J.-P., BARBIER, M., DUMITRESCU, O., PICHAT, C., CARRET, G., RONNAUX-BARON, A.-S., BLASQUEZ, G., GODIN-BENHAIM, C., BOISSET, S. & CARRICAJO, A. 2017. Strain-specific estimation of epidemic success provides insights into the transmission dynamics of tuberculosis. Scientific Reports, 7, 45326.

REED, M. B., PICHLER, V. K., MCINTOSH, F., MATTIA, A., FALLOW, A., MASALA, S., DOMENECH, P., ZWERLING, A., THIBERT, L. & MENZIES, D. 2009. Major Mycobacterium tuberculosis lineages associate with patient country of origin. Journal of Clinical Microbiology, 47, 1119-1128.

REILING, N., HOMOLKA, S., WALTER, K., BRANDENBURG, J., NIWINSKI, L., ERNST, M., HERZMANN, C., LANGE, C., DIEL, R. & EHLERS, S. 2013. Clade-specific virulence patterns of Mycobacterium tuberculosis complex strains in human primary macrophages and aerogenically infected mice. MBio, 4, e00250-13.

RIE, A. V., PAGE-SHIPP, L., SCOTT, L., SANNE, I. & STEVENS, W. 2010. Xpert® MTB/RIF for point-of-care diagnosis of TB in high-HIV burden, resource-limited countries: hype or hope? Expert Review of Molecular Diagnostics, 10, 937-946.

RILEY, S., FRASER, C., DONNELLY, C. A., GHANI, A. C., ABU-RADDAD, L. J., HEDLEY, A. J., LEUNG, G. M., HO, L.-M., LAM, T.-H. & THACH, T. Q. 2003. Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. Science, 300, 1961-1966.

RINDI, L., LARI, N. & GARZELLI, C. 2012. Large Sequence Polymorphisms of the Euro-American lineage of Mycobacterium tuberculosis: a phylogenetic reconstruction and evidence for convergent evolution in the DR locus. Infection, Genetics and Evolution, 12, 1551-1557.

RODRIGUES, L.C., MANGTANI, P. AND ABUBAKAR, I., 2011. How does the level of BCG vaccine protection against tuberculosis fall over time?. BMJ (Clinical research ed), 343, p.d5974.

ROETZER, A., DIEL, R., KOHL, T. A., RÜCKERT, C., NÜBEL, U., BLOM, J., WIRTH, T., JAENICKE, S., SCHUBACK, S. & RÜSCH-GERDES, S. 2013. Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. PLoS Med, 10, e1001387.

ROHDE, H., QIN, J., CUI, Y., LI, D., LOMAN, N. J., HENTSCHKE, M., CHEN, W., PU, F., PENG, Y. & LI, J. 2011. Open-source genomic analysis of Shiga-toxin–producing E. coli O104: H4. New England Journal of Medicine, 365, 718-724.

RUDDY, M. C., DAVIES, A. P., YATES, M. D., YATES, S., BALASEGARAM, S., DRABU, Y., PATEL, B., LOZEWICZ, S., SEN, S., BAHL, M., JAMES, E., LIPMAN, M., DUCKWORTH, G., WATSON, J. M., PIPER, M., DROBNIEWSKI, F. A. & MAGUIRE, H. 2004. Outbreak of isoniazid resistant tuberculosis in north London. Thorax, 59, 279-285.

RYAN, K. J. & RAY, C. G. 2004. Mycobacteria. Sherris Medical Microbiology: An Introduction to Infectious Diseases. 4th Edition, McGraw-Hill, New York, 439.

SAMBROOK, J. AND RUSSELL, D.W., 2006. Purification of nucleic acids by extraction with phenol: chloroform. Cold Spring Harbor Protocols, 2006(1), pp.pdb-prot4455.

SARKAR, R., LENDERS, L., WILKINSON, K. A., WILKINSON, R. J. & NICOL, M. P. 2012. Modern lineages of Mycobacterium tuberculosis exhibit lineage-specific patterns of growth and cytokine induction in human monocyte-derived macrophages. PloS One, 7, e43170.

SASSETTI, C. M., BOYD, D. H. & RUBIN, E. J. 2003. Genes required for mycobacterial growth defined by high density mutagenesis. Molecular Microbiology, 48, 77-84.

SCHUITEMAKER, A. G. 1968. bacteriphages lysing mycobacteria.

SCHUBERT, M., JÓNSSON, H., CHANG, D., DER SARKISSIAN, C., ERMINI, L., GINOLHAC, A., ALBRECHTSEN, A., DUPANLOUP, I., FOUCAL, A., PETERSEN, B. AND FUMAGALLI, M., 2014. Prehistoric genomes reveal the genetic foundation and cost of horse domestication. Proceedings of the National Academy of Sciences, 111(52), pp.E5661-E5669.

SCHÜRCH, A. C. & VAN SOOLINGEN, D. 2012. DNA fingerprinting of Mycobacterium tuberculosis: From phage typing to whole-genome sequencing. Infection, Genetics and Evolution, 12, 602-609.

SHABBEER, A., COWAN, L.S., OZCAGLAR, C., RASTOGI, N., VANDENBERG, S.L., YENER, B. AND BENNETT, K.P., 2012. TB-Lineage: an online tool for classification and analysis of strains of Mycobacterium tuberculosis complex. Infection, Genetics and Evolution, 12(4), pp.789-797.

SHARMA, A., BLOSS, E., HEILIG, C. M. & CLICK, E. S. 2016. Tuberculosis Caused by Mycobacterium africanum, United States, 2004–2013. Emerging Infectious Diseases, 22, 396-403.

SHEPPARD, S. K., JOLLEY, K. A. & MAIDEN, M. C. 2012. A gene-by-gene approach to bacterial population genomics: whole genome MLST of Campylobacter. Genes, 3, 261-277.

SEGEN, J.C., 1992. The dictionary of modern medicine. CRC Press.

SINGH, B. 1964. The guinea pig virulence of Indian tubercle bacilli. American Review of Respiratory Disease, 89, 1-11.

SMALL, P. M., MCCLENNY, N., SINGH, S., SCHOOLNIK, G., TOMPKINS, L. & MICKELSEN, P. 1993. Molecular strain typing of Mycobacterium tuberculosis to confirm cross-contamination in the mycobacteriology laboratory and modification of procedures to minimize occurrence of false-positive cultures. Journal of Clinical Microbiology, 31, 1677-1682.

SOINI, H., PAN, X., AMIN, A., GRAVISS, E. A., SIDDIQUI, A. & MUSSER, J. M. 2000. Characterization of Mycobacterium tuberculosis Isolates from Patients in Houston, Texas, by Spoligotyping. Journal of Clinical Microbiology, 38, 669-676.

SPAHR, U. & SCHAFROTH, K. 2001. Fate of Mycobacterium avium subsp. paratuberculosis in Swiss hard and semihard cheese manufactured from raw milk. Applied and Environmental Microbiology, 67, 4199-4205.

SREEVATSAN, S., PAN, X., STOCKBAUER, K., CONNELL, N., KREISWIRTH, B., WHITTAM, T. & MUSSER, J. 1997a. Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination. Proc Natl Acad Sci U S A, 94, 9869 - 9874.

Statistics bulletin, 2013: Migration Statistics Wales 2011. Statistics for Wale.: https://gov.wales/statistics-and-research/migration-statistics/?lang=en.

STEIN, R. A. 2011. Super-spreaders in infectious diseases. International Journal of Infectious Diseases, 15, e510-e513.

STONEKING, M., 1995. Ancient DNA: how do you know when you have it and what can you do with it? American Journal of Human Genetics, 57(6), p.1259.

STORY, A., ALDRIDGE, R., ABUBAKAR, I., STAGG, H., LIPMAN, M., WATSON, J. & HAYWARD, A. 2012. Active case finding for pulmonary tuberculosis using mobile digital chest radiography: an observational study. The International Journal of Tuberculosis and Lung Disease, 16, 1461-1467.

STUCKI, D. & GAGNEUX, S. 2013. Single nucleotide polymorphisms in Mycobacterium tuberculosis and the need for a curated database. Tuberculosis, 93, 30-39.

STUCKI, D., BRITES, D., JELJELI, L., COSCOLLA, M., LIU, Q., TRAUNER, A., FENNER, L., RUTAIHWA, L., BORRELL, S. & LUO, T. 2016. Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. Nature Genetics.

SUPPLY, P., ALLIX, C., LESJEAN, S., CARDOSO-OELEMANN, M., RÜSCH-GERDES, S., WILLERY, E., SAVINE, E., DE HAAS, P., VAN DEUTEKOM, H., RORING, S., BIFANI, P., KUREPINA, N., KREISWIRTH, B., SOLA, C., RASTOGI, N., VATIN, V., GUTIERREZ, M. C., FAUVILLE, M., NIEMANN, S., SKUCE, R., KREMER, K., LOCHT, C. & VAN SOOLINGEN, D. 2006. Proposal for Standardization of Optimized Mycobacterial Interspersed Repetitive Unit-Variable-Number Tandem Repeat Typing of Mycobacterium tuberculosis. Journal of Clinical Microbiology, 44, 4498-4510.

SUPPLY, P., MAGDALENA, J., HIMPENS, S. & LOCHT, C. 1997. Identification of novel intergenic repetitive units in a mycobacterial two-component system operon. Molecular Microbiology, 26, 991-1003.

SUPPLY, P., WARREN, R. M., BAÑULS, A. L., LESJEAN, S., VAN DER SPUY, G. D., LEWIS, L. A., TIBAYRENC, M., VAN HELDEN, P. D. & LOCHT, C. 2003. Linkage disequilibrium between minisatellite loci supports clonal evolution of Mycobacterium tuberculosis in a high tuberculosis incidence area. Molecular Microbiology, 47, 529-538.

TAKIFF, H. E. & FEO, O. 2015. Clinical value of whole-genome sequencing of Mycobacterium tuberculosis. The Lancet Infectious Diseases, 15, 1077-1090.

TB FREE ENGLAND: http://www.tbfreeengland.co.uk/faqs/how-much-does-btb-cost/
TESSEMA, B., BEER, J., MERKER, M., EMMRICH, F., SACK, U., RODLOFF, A. C. & NIEMANN, S. 2013. Molecular epidemiology and transmission dynamics of Mycobacterium tuberculosis in Northwest Ethiopia: new phylogenetic lineages found in Northwest Ethiopia. BMC Infectious Diseases, 13, 1.

THIERRY, D., BRISSON-NOËL, A., VINCENT-LEVY-FREBAULT, V., NGUYEN, S., GUESDON, J.L. AND GICQUEL, B., 1990. Characterization of a Mycobacterium tuberculosis insertion sequence, IS6110, and its application in diagnosis. Journal of clinical microbiology, 28(12), pp.2668-2673.

THOMAS E HERCHLINE, M. J. K. A. C. E. M. S. B., MD 2017. Tuberculosis (TB) Treatment & Management. Medscape.

THOMAS, P. D., CAMPBELL, M. J., KEJARIWAL, A., MI, H., KARLAK, B., DAVERMAN, R., DIEMER, K., MURUGANUJAN, A. & NARECHANIA, A. 2003. PANTHER: a library of protein families and subfamilies indexed by function. Genome Research, 13, 2129-2141.

THOMPSON, J. D., GIBSON, T. & HIGGINS, D. G. 2002. Multiple sequence alignment using ClustalW and ClustalX. Current Protocols in Bioinformatics, 2.3. 1-2.3. 22.

THWAITES, G., CAWS, M., CHAU, T. T. H., D'SA, A., LAN, N. T. N., HUYEN, M. N. T., GAGNEUX, S., ANH, P. T. H., THO, D. Q. & TOROK, E. 2008. Relationship between Mycobacterium tuberculosis genotype and the clinical phenotype of pulmonary and meningeal tuberculosis. Journal of Clinical Microbiology, 46, 1363-1368.

TRAUNER, A., BORRELL, S., REITHER, K. AND GAGNEUX, S., 2014. Evolution of drug resistance in tuberculosis: recent progress and implications for diagnosis and therapy. Drugs, 74(10), pp.1063-1072.

TUBERCULIST 2017. Mycobacterium tuberculosis H37Rv: TubercuList.

TYLER, A. D., CHRISTIANSON, S., KNOX, N. C., MABON, P., WOLFE, J., VAN DOMSELAAR, G., GRAHAM, M. R. & SHARMA, M. K. 2016. Comparison of sample preparation methods used for the next-generation sequencing of Mycobacterium tuberculosis. PloS One, 11, e0148676.

USLAN, D. Z., KOWALSKI, T. J., WENGENACK, N. L., VIRK, A. & WILSON, J. W. 2006. Skin and soft tissue infections due to rapidly growing mycobacteria: comparison of clinical features, treatment, and susceptibility. Archives of Dermatology, 142, 1287-1292.

VAN EMBDEN, J., CAVE, M., CRAWFORD, J., DALE, J., EISENACH, K., GICQUEL, B., HERMANS, P., MARTIN, C., MCADAM, R., SHINNICK, T. & SMALL, P. 1993. Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol, 31, 406 - 409.

VAN EMBDEN, J., VAN GORKOM, T., KREMER, K., JANSEN, R., VAN DER ZEIJST, B. & SCHOULS, L. 2000. Genetic variation and evolutionary origin of the direct repeat locus of Mycobacterium tuberculosis complex bacteria. Journal of Bacteriology, 182, 2393-2401.

VAN INGEN, J., RAHIM, Z., MULDER, A., BOEREE, M. J., SIMEONE, R., BROSCH, R. & VAN SOOLINGEN, D. 2012. Characterization of Mycobacterium orygis as *M. tuberculosis* complex subspecies. Emerging Infectious Diseases, 18, 653.

VAN SOOLINGEN, D. 2001. Molecular epidemiology of tuberculosis and other mycobacterial infections: main methodologies and achievements. Journal of Internal Medicine, 249, 1-26.

VAN SOOLINGEN, D., HOOGENBOEZEM, T., DE HAAS, P. E., HERMANS, P. W., KOEDAM, M. A., TEPPEMA, K. S., BRENNAN, P. J., BESRA, G. S., PORTAELS, F. & TOP, J. 1997. A novel pathogenic taxon of the Mycobacterium tuberculosis complex, Canetti: characterization of an exceptional isolate from Africa. International Journal of Systematic and Evolutionary Microbiology, 47, 1236-1245.

VOTINTSEVA, A.A., BRADLEY, P., PANKHURST, L., DEL OJO ELIAS, C., LOOSE, M., NILGIRIWALA, K., CHATTERJEE, A., SMITH, E.G., SANDERSON, N., WALKER, T.M. AND MORGAN, M.R., 2017. Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples. Journal of Clinical Microbiology, 55(5), pp.1285-1298.

WALES, N., ALLABY, R., WILLERSLEV, E. AND GILBERT, M.T.P., 2013. PALEOBOTANY| Ancient Plant DNA.

WALKER, T. M., IP, C. L. C., HARRELL, R. H., EVANS, J. T., KAPATAI, G., DEDICOAT, M. J., EYRE, D. W., WILSON, D. J., HAWKEY, P. M., CROOK, D. W., PARKHILL, J., HARRIS, D., WALKER, A. S., BOWDEN, R., MONK, P., SMITH, E. G. & PETO, T. E. A. 2013. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. The Lancet Infectious Diseases, 13, 137-146.

WALKER, T. M., KOHL, T. A., OMAR, S. V., HEDGE, J., ELIAS, C. D. O., BRADLEY, P., IQBAL, Z., FEUERRIEGEL, S., NIEHAUS, K. E. & WILSON, D. J. 2015. Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. The Lancet Infectious Diseases, 15, 1193-1202.

WALKER, T., MONK, P., GRACE SMITH, E. & PETO, T. 2013a. Contact investigations for outbreaks of Mycobacterium tuberculosis: advances through whole genome sequencing. Clin Microbiol Infect.

WARREN, R., STREICHER, E., SAMPSON, S., VAN DER SPUY, G., RICHARDSON, M., NGUYEN, D., BEHR, M., VICTOR, T. & VAN HELDEN, P. 2002. Microevolution of the direct repeat region of Mycobacterium tuberculosis: implications for interpretation of spoligotyping data. Journal of Clinical Microbiology, 40, 4457-4465.

WATSON, J., ABUBAKAR, I., STORY, A., WELFARE, R., WHITE, P., GARNETT, G., MUGFORD, M., GARRETT, J. & S HAYWARD, A. 2007. Mobile targeted digital chest radiography in the control of tuberculosis among hard to reach groups.URL: http://discovery.ucl.ac.uk/56609/. UCL Discovery.

WENIGER, T., KRAWCZYK, J., SUPPLY, P., NIEMANN, S. & HARMSEN, D. 2010. MIRU-VNTRplus: a web tool for polyphasic genotyping of Mycobacterium tuberculosis complex bacteria. Nucleic Acids Res, 38, W326-31.

WHITTLES, L. K. & DIDELOT, X. Epidemiological analysis of the Eyam plague outbreak of 1665–1666. Proc. R. Soc. B, 2016. The Royal Society, 20160618.

WHO 2015. World Health Organization Global tuberculosis report 2015. In: ORGANISATION, W. H. (ed.).

WHO 2016. World Health Organization Global Tuberculosis report 2016:Executive summary. In: PREVENTION, T. S. (ed.).

WHO 2017. World Health Organization Global Tuberculosis Report 2017: Executive summary. In: PREVENTION.

WHO, G. 2010. Guidelines for treatment of tuberculosis. WHO Press Geneva, Switzerland.

WILLIAMS, A., JAMES, B. W., BACON, J., HATCH, K. A., HATCH, G. J., HALL, G. A. & MARSH, P. D. 2005. An assay to compare the infectivity of Mycobacterium tuberculosis isolates based on aerosol infection of guinea pigs and assessment of bacteriology. Tuberculosis, 85, 177-184.

WILSON, G.S., 1943. The pasteurization of milk. *British Medical Journal*, *1*(4286), p.261.
WIRTH, T., HILDEBRAND, F., ALLIX-BÉGUEC, C., WÖLBELING, F., KUBICA, T., KREMER, K., VAN SOOLINGEN, D., RÜSCH-GERDES, S., LOCHT, C. & BRISSE, S. 2008a. Origin, spread and demography of the Mycobacterium tuberculosis complex. PLoS Pathog, 4, e1000160.

WITNEY, A. A., COSGROVE, C. A., ARNOLD, A., HINDS, J., STOKER, N. G. & BUTCHER, P. D. 2016. Clinical use of whole genome sequencing for Mycobacterium tuberculosis. BMC Medicine, 14, 1.

WOHL, S., SCHAFFNER, S. F. & SABETI, P. C. 2016. Genomic Analysis of Viral Outbreaks. Annual Review of Virology, 3, 173-195.

WONG, K.-C. & ZHANG, Z. 2014. SNPdryad: predicting deleterious non-synonymous human SNPs using only orthologous protein sequences. Bioinformatics, 30, 1112-1119.

WORLD HEALTH ORGANISATION: JIM YONG KIM, A. S., ARACHU CASTRO, CHRIS VANDE, PAUL FARMER. 2017. Tuberculosis control [Online]. Available: http://www.who.int/trade/distance_learning/gpgh/gpgh3/en/index7.html.

WORLD ORGANISATION OF ANIMAL HEALTH, F. A. A. O. O. T. U. N., International Union Against Tuberculosis and Lung Disease 2016. zoonotic tb. In: world health organisation, w. o. o. a. h., food and agriculture organisation of the united nations, international union against tuberculosis and lung disease (ed.).

XIA, E., TEO, Y.-Y. & ONG, R. T.-H. 2016. SpoTyping: fast and accurate in silico Mycobacterium spoligotyping from sequence reads. Genome Medicine, 8, 19.

YEBOAH-MANU, D., ASARE, P., ASANTE-POKU, A., OTCHERE, I.D., OSEI-WUSU, S., DANSO, E., FORSON, A., KORAM, K.A. AND GAGNEUX, S., 2016. Spatio-temporal distribution of Mycobacterium tuberculosis complex strains in Ghana. PloS one, 11(8), p.e0161892.

YEBOAH-MANU, D., ASANTE-POKU, A., BODMER, T., STUCKI, D., KORAM, K., BONSU, F., PLUSCHKE, G. AND GAGNEUX, S., 2011. Genotypic diversity and drug susceptibility patterns among M. tuberculosis complex isolates from South-Western Ghana. PloS one, 6(7), p.e21906.

YIMER, S. A., NORHEIM, G., NAMOUCHI, A., ZEGEYE, E. D., KINANDER, W., TØNJUM, T., BEKELE, S., MANNSÅKER, T., BJUNE, G. & ASEFFA, A. 2015. Mycobacterium tuberculosis lineage 7 strains are associated with prolonged patient delay in seeking treatment for pulmonary tuberculosis in amhara region, Ethiopia. Journal of Clinical Microbiology, 53, 1301-1309.

YOUNG, R. A., MEHRA, V. & SWEETSER, D. 1985. Leprosy parasite Mycobacterium leprae. Nature, 316, 65K.

ZACHARIAH, R., SPIELMANN, M.P., HARRIES, A.D., GOMANI, P., GRAHAM, S.M., BAKALI, E. AND HUMBLET, P., 2003. Passive versus active tuberculosis case finding and isoniazid preventive therapy among household contacts in a rural district of Malawi. The international journal of tuberculosis and lung disease, 7(11), pp.1033-1039.

Appendix

Figure PH1**:** The figure below shows an un-rooted phylogeny of all isolates originally included in the cgMLST analys phylogeny presented in figure 3.6, chapter 3. Clades are colour coordinated to the lineages present within them wi coloured in red.

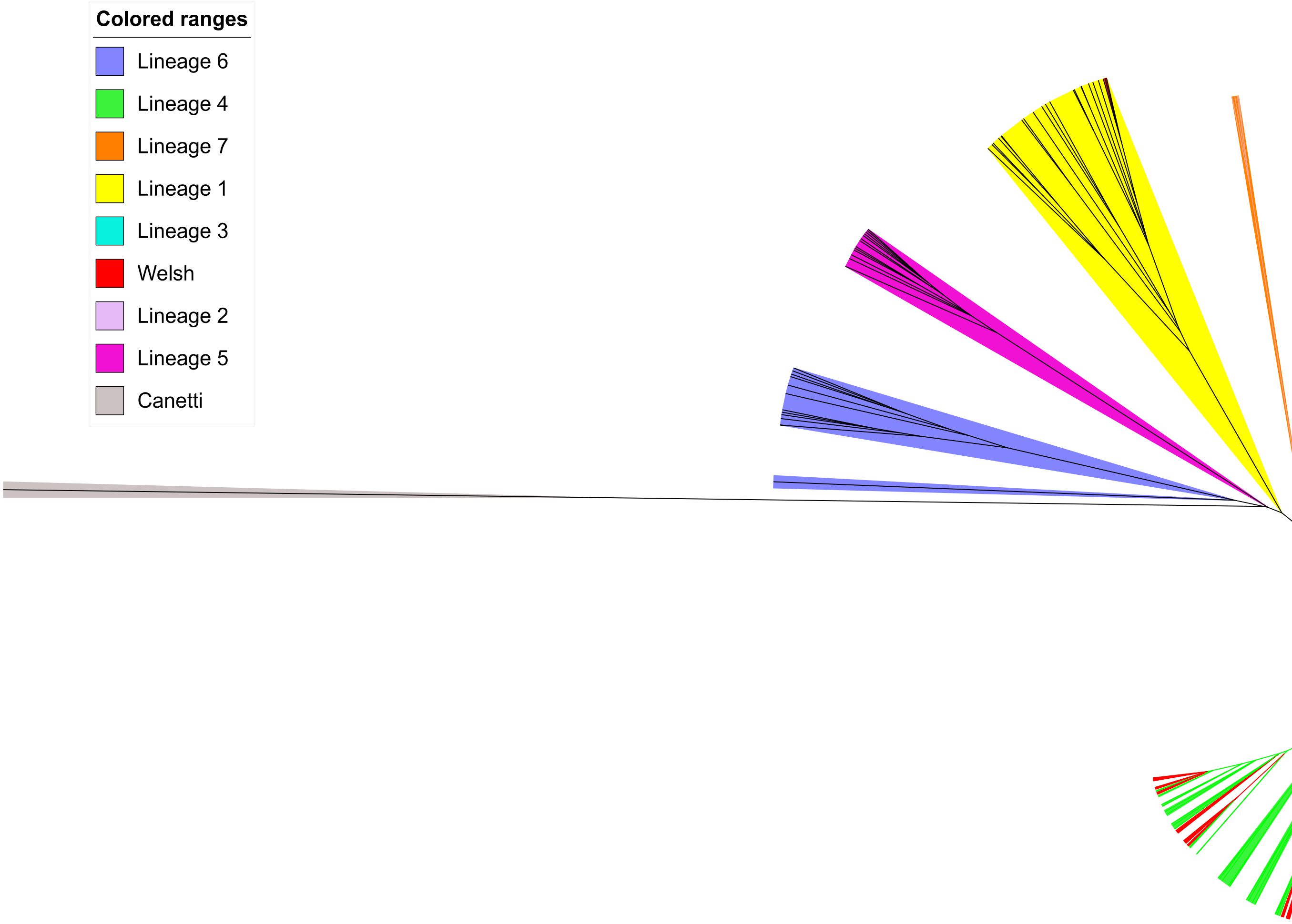


Table PH1: Below is a conclusion table showing the phylogenetic classification`s across all methods described in cha grouping refers to both the Principal Genetic Group and SNP cluster group assignments. The further most right colu outbreak or background status. The bottom section highlights those isolates whereby phylogenetic classification fo completed due to lack of sequence data, “X” = no phylogenetic assignment possible.

Isolate	**Classical grouping**		**cgMLST association**	**SNP barcode**	
BK22	1	SCG 1	Indo-oceanic(Baku *et al*.)	Indo-oceanic(Baku *et al*.)	Indo-oceanic(EAI3;EAI5
LL9	1	SCG 2	Beijing(2)	Beijing(2)	Beijing
BK25	1	SCG 2	Beijing(2)	Beijing(2)	Beijing
BK21	1	SCG 2/3a	Beijing(2)*	Beijing(2)	Beijing
LL5	2	SCG 3b	Euro-american(4)	Euro-american(4)	Haarlem(T1;H1)
LL8	2	SCG 3b	Euro-american(4)	Euro-american(4)	Haarlem(T1;H1)
BK23	2	SCG 3b	Euro-american(4)	Euro-american(4)	Haarlem(T1;H1)
BK10	2	SCG 3b	Euro-american(4)	Euro-american(4)	Haarlem(T1;H1)
BK11	2	SCG 3b	Euro-american(4)	Euro-american(4)	Haarlem(T1;H1)
TH1	2	SCG 3b	Euro-american(4)	Euro-american(4)	Haarlem(T1;H1)
TH2	2	SCG 3b	Euro-american(4)	Euro-american(4)	Haarlem(T1;H1)
GO1	2	SCG 3b	Euro-american(4)	Euro-american(4)	Haarlem(T1;H1)
GO2	2	SCG 3b	Euro-american(4)	Euro-american(4)	Haarlem(T1;H1)
GO3	2	SCG 3b	Euro-american(4)	Euro-american(4)	Haarlem(T1;H1)
GO4	2	SCG 3b	Euro-american(4)	Euro-american(4)	Haarlem(T1;H1)
GO5	2	SCG 3b	Euro-american(4)	Euro-american(4)	Haarlem(T1;H1)
GO6	2	SCG 3b	Euro-american(4)	Euro-american(4)	Haarlem(T1;H1)
GO7	2	SCG 3b	Euro-american(4)	Euro-american(4)	Haarlem(T1;H1)

GO8	2	SCG 3c	Euro-american(4)	Euro-american(4)	X family(X2)
LL11	2	SCG 3c	Euro-american(4)	Euro-american(4)	X family(Baku *et al.*)
BK14	2	SCG 3c	Euro-american(4)	Euro-american(4)	X family (Baku *et al.*)
BK15	2	SCG 3c	Euro-american(4)	Euro-american(4)	X family(Baku *et al.*)
BK12	2	SCG 3c	Euro-american(4)	Euro-american(4)	X family(X2)
BK13	2	SCG 3c	Euro-american(4)	Euro-american(4)	X family(X2)
NPTB6	2	SCG 4	Euro-american(4)	Euro-american(4)	X family(X1;X3)
LL1	2	SCG 4	Euro-american(4)	Euro-american(4)	X family(X1;X3)
LL3	2	SCG 4	Euro-american(4)	Euro-american(4)	X family(X1;X3)
LL4	2	SCG 4	Euro-american(4)	Euro-american(4)	X family(X1;X3)
BK18	2	SCG 4	Euro-american(4)	Euro-american(4)	X family(X1;X3)
BK16	2	SCG 4	Euro-american(4)	Euro-american(4)	X family(X1;X3)
BK17	2	SCG 4	Euro-american(4)	Euro-american(4)	X family(X1;X3)
BK8	2	SCG 5	Euro-american(4)	Euro-american(4)	Latin American Mediterra
BK20	2	SCG 5	Euro-american(4)	Euro-american(4)	Latin American Mediterra
BK9	2	SCG 5	Euro-american(4)	Euro-american(4)	Latin American Mediterra
LL10	2	SCG 6	Euro-american(4)*	Euro-american(4)	X family(X1;X3)
BK19	3	SCG 6	Euro-american(4)*	Euro-american(4)	T family
NPTB5	3	SCG 6a	Euro-american(4)	Euro-american(4)	T family
NPTA8	3	SCG 6a	Euro-american(4)	Euro-american(4)	T family
NPTA6	3	SCG 6a	Euro-american(4)*	Euro-american(4)	T family
NPTA7	3	SCG 6a	Euro-american(4)	Euro-american(4)	T family
NPTA4	3	SCG 6a	Euro-american(4)	Euro-american(4)	T family
NPTA5	3	SCG 6a	Euro-american(4)	Euro-american(4)	T family
NPTB4	3	SCG 6a	Euro-american(4)	Euro-american(4)	T family

NPTB1	3	SCG 6a	Euro-american(4)	Euro-american(4)	T family
NPTA2	3	SCG 6a	Euro-american(4)	Euro-american(4)	T family
NPTA3	3	SCG 6a	Euro-american(4)	Euro-american(4)	T family
NPTA1	3	SCG 6a	Euro-american(4)	Euro-american(4)	T family
BK1	3	SCG 6a	Euro-american(4)	Euro-american(4)	T family
BK2	3	SCG 6a	Euro-american(4)	Euro-american(4)	T family
BK3	3	SCG 6a	Euro-american(4)	Euro-american(4)	T family
NPTB3	3	SCG 6a	Euro-american(4)	Euro-american(4)	T family
NPTB2	3	SCG 6a	Euro-american(4)	Euro-american(4)	T family
BK4	3	SCG 6b	Euro-american(4)	Euro-american(4)	H37Rv-like
BK5	3	SCG 6b	Euro-american(4)	Euro-american(4)	H37Rv-like
BK6	3	SCG 6b	Euro-american(4)*	Euro-american(4)	H37Rv-like
BK7	3	SCG 6b	Euro-american(4)	Euro-american(4)	H37Rv-like
H37Rv	3	SCG 6b	Euro-american(4)	Euro-american(4)	H37Rv

With missing data						
Isolate	**PGG**	**SCG**	**Lineage**	**Lineage**	**Sublineage**	**Outbreak/Background**
BK24	X	SCG 3b	Euro-american(4)	Euro-american(4)	T family	Background
BK26	X	SCG 3b/c	X	X	X	Background
GO9	X	SCG 3c	Euro-american(4)*	Euro-american(4)	X family(X2)	Outbreak isolate
BK27	X	SCG 3c	X	X	X	Background
LL2	X	SCG 4	Euro-american(4)	Euro-american(4)	X family(X1;X3)	Outbreak isolate
LL7	X	SCG 4	X	X	X	Outbreak isolate
NPTB7	X	SCG 6	X	X	X	Outbreak isolate
BK28	X	SCG 6	X	X	X	Background
BK29	X	SCG 6	X	X	X	Background

*represents isolates that could not be confidently assigned to a lineage by cgMLST association.

Table PH2a and PH2b (below): Tables showing the full SNP pattern of 48 loci identified in the Mestre *et al* (2011) st For both tables PH2a and PH2b, the first line indicates the gene and the second line indicates the position on th identified in relation to *M. tuberculosis* H37Rv strain for Beijing genotype classification. Due to the large amou presented with table PH2a providing the pattern across first 23 positions and PH2b providing the pattern Polymorphisms that characterize and allow the discrimination of the 26 sequence types defined previously, those wi in red. Also included in the table are the Welsh Beijing isolates (in red) LL9, BK21, BK25 and the control isolates (in Beijing welsh strain) and the reference strain H37Rv.

PH2a	*ligD*				*radA*			*recF*		*recX*			*dnaQ*				*recR*	
Isolate	485	1030	1038	1738	456	557	827	734	807	23	175	458	227	263	483	631	130	2
LL9	A	C	C	C	A	T	T	C	G	C	C	G	A	T	C	T	T	C
BK21	A	C	C	C	A	T	T	C	G	C	G	G	A	T	C	T	T	C
BK25	A	C	C	C	A	T	T	C	G	C	C	G	A	T	C	T	T	C
H	A	C	C	C	A	T	T	C	G	C	G	G	G	T	C	T	G	C
BK4	A	T	C	C	G	T	T	T	G	C	G	G	G	T	C	G	G	C
H37Rv	A	T	C	C	G	T	T	T	G	C	G	G	G	T	C	G	G	C
BmyC1	A	C	C	C	A	T	T	C	G	C	G	G	A	T	C	T	G	C
BmyC2	A	C	C	T	A	T	T	C	G	C	G	G	A	T	C	T	G	C
BmyC3	C	C	C	T	A	T	T	C	G	C	G	G	A	T	C	T	G	C
BmyC4	A	C	C	C	A	T	T	C	G	C	G	G	A	T	C	T	T	C
BmyC5	A	C	T	C	A	T	T	C	G	C	G	G	A	T	C	T	T	C
BmyC6	A	C	C	C	A	T	T	C	G	C	G	G	A	T	C	T	T	C
BmyC7	A	C	C	C	A	T	T	C	G	C	G	G	A	T	C	T	T	C
BmyC8	A	C	C	C	A	T	T	C	G	T	G	G	A	T	C	T	T	C
BmyC9	A	C	C	C	A	T	T	C	G	C	C	G	A	T	C	T	T	C
BmyC10	A	C	C	C	A	T	T	C	G	C	C	G	A	T	C	T	T	C
BmyC11	A	C	C	C	A	T	T	C	G	C	C	G	A	T	C	T	T	T
BmyC12	A	C	C	C	A	T	T	C	T	C	C	G	A	T	C	T	T	C
BmyC13	A	C	C	C	A	T	T	C	G	C	C	G	A	T	C	T	T	C
BmyC14	A	C	C	C	A	T	T	C	G	C	C	G	A	T	T	T	T	C
BmyC15	A	C	C	C	A	T	T	C	G	C	C	G	A	T	C	T	T	C

BmyC16	**A**	**C**	**C**	**C**	**A**	**T**	**T**	**C**	**G**	**C**	**C**	**G**	**A**	**T**	**C**	**T**	**T**	**C**
BmyC17	**A**	**C**	**C**	**C**	**A**	**T**	**T**	**C**	**G**	**C**	**C**	**G**	**A**	**T**	**C**	**T**	**T**	**C**
BmyC18	**A**	**C**	**C**	**C**	**A**	**T**	**T**	**C**	**G**	**C**	**C**	**G**	**A**	**T**	**C**	**T**	**T**	**C**
BmyC19	**A**	**C**	**C**	**C**	**A**	**T**	**T**	**C**	**G**	**C**	**C**	**G**	**A**	**T**	**C**	**T**	**T**	**C**
BmyC20	**A**	**C**	**C**	**C**	**A**	**T**	**C**	**C**	**G**	**C**	**C**	**A**	**A**	**T**	**C**	**T**	**T**	**C**
BmyC21	**A**	**C**	**C**	**C**	**A**	**T**	**C**	**C**	**G**	**C**	**C**	**A**	**A**	**T**	**C**	**T**	**T**	**C**
BmyC22	**A**	**C**	**C**	**C**	**A**	**T**	**T**	**C**	**G**	**C**	**C**	**G**	**A**	**T**	**C**	**T**	**T**	**C**
BmyC23	**A**	**C**	**C**	**C**	**A**	**C**	**T**	**C**	**G**	**C**	**C**	**G**	**A**	**T**	**C**	**T**	**T**	**C**
BmyC2C	**A**	**C**	**C**	**C**	**A**	**T**	**T**	**C**	**G**	**C**	**C**	**G**	**A**	**T**	**C**	**T**	**T**	**C**
BmyC25	**A**	**C**	**C**	**C**	**A**	**T**	**T**	**C**	**G**	**C**	**G**	**G**	**A**	**T**	**C**	**T**	**T**	**C**
BmyC26	**A**	**C**	**C**	**C**	**A**	**T**	**T**	**C**	**G**	**C**	**C**	**G**	**A**	**T**	**C**	**T**	**T**	**C**

PH2b	*ruvB*	*ligB*		*recD*			*tagA*		*uvrD1*	*dnaZX*		*nei*	*nth*			*alkA*		*ligC*	
Isolate	843	230	271	360	416	831	537	385	1384	274	291	229	5	101	365	31	34	630	938
LL9	G	T	T	C	T	G	C	G	G	C	T	A	G	A	T	G	G	C	A
BK21	G	T	T	C	T	G	C	G	G	C	T	A	G	A	T	G	G	C	A
BK25	G	T	T	C	T	G	C	G	G	C	T	A	G	A	T	G	G	C	A
H	G	T	T	A	T	G	C	G	G	C	C	A	C	A	T	G	A	G	A
BK4	G	T	C	A	T	G	C	G	G	C	C	A	C	A	T	G	A	G	G
H37Rv	A	T	C	A	G	G	C	G	G	C	C	A	C	A	T	G	A	G	G
BmyC1	G	T	T	C	G	G	C	G	G	C	T	C	G	A	T	G	G	C	A
BmyC2	G	T	T	C	G	G	C	G	G	C	T	C	G	A	T	G	G	C	A
BmyC3	G	T	T	C	G	G	C	G	G	C	T	C	G	A	T	G	G	C	A
BmyC4	G	T	T	C	G	G	C	G	G	C	T	C	G	A	T	G	G	C	A
BmyC5	G	T	T	C	G	G	C	G	G	C	T	C	G	A	T	G	G	C	A
BmyC6	G	T	T	C	G	G	C	G	G	C	T	C	G	A	T	G	G	C	A
BmyC7	G	T	T	C	G	G	C	G	G	C	T	C	G	A	T	G	G	C	A
BmyC8	G	T	T	C	G	G	C	G	G	C	T	C	G	A	T	G	G	C	A
BmyC9	G	T	T	C	G	G	C	G	G	C	T	C	G	A	T	G	G	C	A
BmyC10	G	T	T	C	G	G	C	G	G	C	T	C	G	A	T	G	G	C	A
BmyC11	G	T	T	C	G	G	C	G	G	C	T	C	G	A	T	G	G	C	A
BmyC12	G	T	T	C	G	G	C	G	G	C	T	C	G	A	T	G	G	C	A
BmyC13	G	T	T	C	G	G	C	G	A	C	T	C	G	A	T	G	G	C	A
BmyC14	G	C	T	C	G	G	C	G	A	C	T	C	G	A	T	G	G	C	A
BmyC15	G	T	T	C	G	G	C	G	G	C	T	C	G	A	G	G	G	C	A
BmyC16	G	T	T	C	G	G	C	G	G	T	T	C	G	A	T	G	G	C	A
BmyC17	G	T	T	C	G	G	C	G	G	C	T	C	G	C	T	G	G	C	A

BmyC18	G	T	T	C	G	G	C	G	G	C	T	C	G	A	T	A	G	C	A
BmyC19	G	T	T	C	G	G	C	A	G	C	T	C	G	A	T	G	G	C	A
BmyC20	G	T	T	C	G	G	C	G	G	C	T	C	G	A	T	G	G	C	A
BmyC21	G	T	T	C	T	G	C	G	G	C	T	C	G	A	T	G	G	C	A
BmyC22	G	T	T	C	G	A	C	G	G	C	T	C	G	A	T	G	G	C	A
BmyC23	G	T	T	C	G	G	C	G	G	C	T	C	G	A	T	G	G	C	A
BmyC2C	G	T	T	C	G	G	T	G	G	C	T	C	G	A	T	G	G	C	A
BmyC25	G	T	T	C	G	G	C	G	G	C	T	C	G	A	T	G	G	C	A
BmyC26	G	T	T	C	G	G	C	G	G	C	T	C	G	A	T	G	G	C	A

Table PH3a: A distance matrix based on the number of polymorphisms across the 48 defined in the Mestre *et al* (2011) stud
number of polymorphisms there are between any two given isolates. The bluer the cell the closer the lesser number of polymor

Isolate	LL9	BK21	BK25	H	BK4	H37Rv	BmyC1	BmyC2	BmyC3	BmyC4	BmyC5	BmyC6	BmyC7	BmyC8	BmyC9	BmyC10	BmyC11	BmyC12	BmyC13	BmyC14	BmyC15	BmyC16	BmyC17	BmyC18	BmyC19	BmyC20
LL9		4	0	13	19	21	7	8	9	6	8	5	7	6	5	2	3	3	3	5	3	3	3	3	4	4
BK21	4		4	11	17	19	5	6	7	4	6	3	3	4	5	6	7	7	7	9	7	7	7	7	8	8
BK25	0	4		13	19	21	7	8	9	6	8	5	7	6	5	2	3	3	3	5	3	3	3	3	4	4
H	13	11	13		6	8	10	11	12	11	13	12	14	13	14	15	16	16	16	18	16	16	16	16	17	17
BK4	19	17	19	6		2	16	17	18	17	19	18	20	19	20	21	22	22	22	24	22	22	22	22	23	23
H37Rv	21	19	21	8	2		16	17	18	17	19	18	20	19	20	21	22	22	22	24	22	22	22	22	23	23
BmyC1	7	5	7	10	16	16		1	2	1	3	2	4	3	4	5	6	6	6	8	6	6	6	6	7	7
BmyC2	8	6	8	11	17	17	1		1	2	4	3	5	4	5	6	7	7	7	9	7	7	7	7	8	8
BmyC3	9	7	9	12	18	18	2	1		3	5	4	6	5	6	7	8	8	8	10	8	8	8	8	9	9
BmyC4	6	4	6	11	17	17	1	2	3		2	1	3	2	3	4	5	5	5	7	5	5	5	5	6	6
BmyC5	8	6	8	13	19	19	3	4	5	2		3	5	4	5	6	7	7	7	9	7	7	7	7	8	8
BmyC6	5	3	5	12	18	18	2	3	4	1	3		2	1	2	3	4	4	4	6	4	4	4	4	5	5
BmyC7	7	3	7	14	20	20	4	5	6	3	5	2		3	4	5	6	6	6	8	6	6	6	6	7	7
BmyC8	6	4	6	13	19	19	3	4	5	2	4	1	3		3	4	5	5	5	7	5	5	5	5	6	6
BmyC9	5	5	5	14	20	20	4	5	6	3	5	2	4	3		3	4	4	4	6	4	4	4	4	5	5
BmyC10	2	6	2	15	21	21	5	6	7	4	6	3	5	4	3		1	1	1	3	1	1	1	1	2	2
BmyC11	3	7	3	16	22	22	6	7	8	5	7	4	6	5	4	1		2	2	4	2	2	2	2	3	3
BmyC12	3	7	3	16	22	22	6	7	8	5	7	4	6	5	4	1	2		2	4	2	2	2	2	3	3
BmyC13	3	7	3	16	22	22	6	7	8	5	7	4	6	5	4	1	2	2		2	2	2	2	2	3	3
BmyC14	5	9	5	18	24	24	8	9	10	7	9	6	8	7	6	3	4	4	2		4	4	4	4	5	5
BmyC15	3	7	3	16	22	22	6	7	8	5	7	4	6	5	4	1	2	2	2	4		2	2	2	3	3
BmyC16	3	7	3	16	22	22	6	7	8	5	7	4	6	5	4	1	2	2	2	4	2		2	2	3	3
BmyC17	3	7	3	16	22	22	6	7	8	5	7	4	6	5	4	1	2	2	2	4	2	2		2	3	3
BmyC18	3	7	3	16	22	22	6	7	8	5	7	4	6	5	4	1	2	2	2	4	2	2	2		3	3
BmyC19	4	8	4	17	23	23	7	8	9	6	8	5	7	6	5	2	3	3	3	5	3	3	3	3		4
BmyC20	4	8	4	17	23	23	7	8	9	6	8	5	7	6	5	2	3	3	3	5	3	3	3	3	4	
BmyC21	3	7	3	16	22	24	8	9	10	7	9	6	8	7	6	3	4	4	4	6	4	4	4	4	5	1
BmyC22	3	7	3	16	22	22	6	7	8	5	7	4	6	5	4	1	2	2	2	4	2	2	2	2	3	3
BmyC23	3	7	3	16	22	22	6	7	8	5	7	4	6	5	4	1	2	2	2	4	2	2	2	2	3	3
BmyC24	3	7	3	16	22	22	6	7	8	5	7	4	6	5	4	1	2	2	2	4	2	2	2	2	3	3
BmyC25	6	2	6	13	19	19	3	4	5	2	4	1	1	2	3	4	5	5	5	7	5	5	5	5	6	6
BmyC26	4	4	4	13	19	19	3	4	5	2	4	1	3	2	1	2	3	3	3	5	3	3	3	3	4	4

Table PH3b: A subset of the distance matrix seen above, highlighting the number of SNP`s, across the 48 previously defined loci, between only those isolates most closely related to the Welsh isolates LL9, BK21 and BK25.

Table PH3b			
Isolates	LL9	BK21	BK25
Closest relative	BmyC10	BmyC25	BmyC10
Number of SNP`s	2	2	2

The figure below highlights the genomic relationship between LL10 and the outbreak associated isolates defined in the original cgMLST analysis within chapter 4, figure 4.3. The minimum spanning tree was constructed on only the 1446 core genes that could be successfully aligned between LL10 and the cgMLST scheme defined in method section 2.3.1. The figure below shows that LL10 shares 3 allelic differences with both LL1 and LL4, equal to the divergence seen with LL3, more than the divergence seen with LL2 and less than the divergence seen between NPTB6 and isolates LL1 and LL4. Although not analysed using the full compliment, the analysis here across 1466 core genes shows LL10 to have less than the 12 allelic differences with its closet relative and thus supports original epidemiological data which postulated that LL10 may be part of the Llwynhendy outbreak. This result suggests that further analysis of LL10 should be carried out to ascertain whether the close genomic relationship is replicated when the full complement of the cgMLST scheme is analysed.

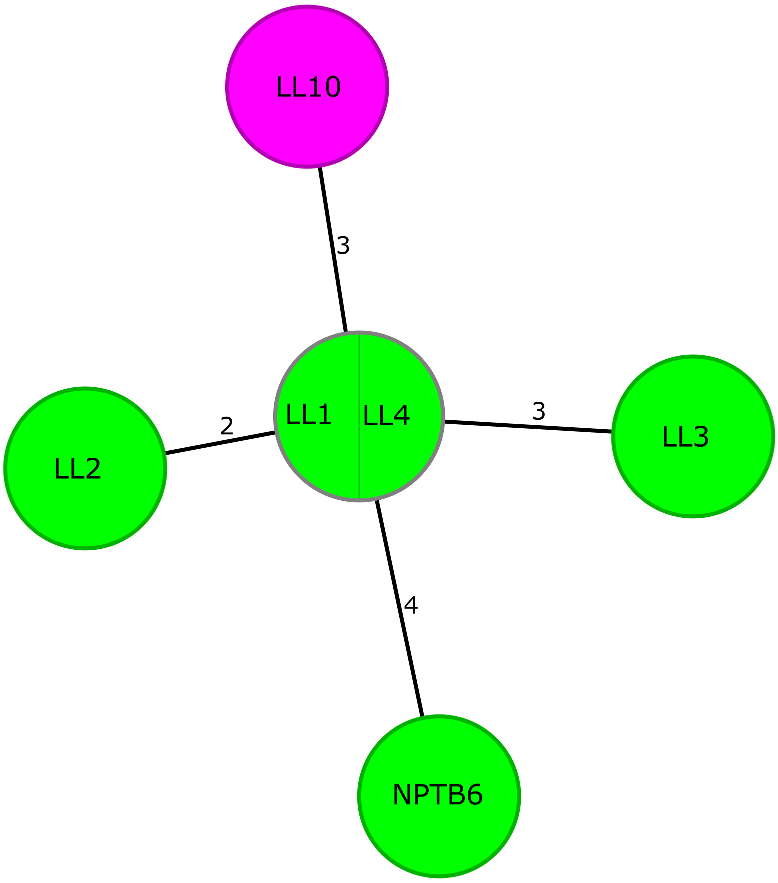


Figure LL1: A minimum spanning tree based on 1466 core genes successfully aligned to the cgMLST scheme. Those in the original cgMLST analysis in section 4, figure 4.3, are coloured green with LL10 coloured pink.

Table LL1: Distance matrix showing the number of allelic differences between all isolates defined in Figure 4.3. The
figure and includes isolates that exceeded the 12 allelic differences threshold and thus were not outbreak associate
more different the isolates are from each other. The greater the intensity of blue in the cells the more closely related

Isolates	GO8	NPTB6	LL1	LL2	LL3	LL4	LL5	LL8
GO8	0	246	245	238	243	248	275	
NPTB6	246	0	7	13	10	12	294	
LL1	245	7	0	8	2	4	296	
LL2	238	13	8	0	11	13	285	
LL3	243	10	2	11	0	6	296	
LL4	248	12	4	13	6	0	296	
LL5	275	294	296	285	296	296	0	
LL8	307	328	327	311	325	330	201	
LL9	573	570	585	554	579	581	529	
LL11	266	268	266	253	267	267	297	

Table PH4: A table showing the % probability that each isolate is associated with a given lineage.

Isolate	Lineage 4	Lineage 3	Lineage 2	Lineage 1
NPTB5	100%	0%	0%	0%
NPTA8	100%	0%	0%	0%
NPTA6	9%	2%	0%	89%
NPTA7	100%	0%	0%	0%
NPTA4	100%	0%	0%	0%
NPTA5	100%	0%	0%	0%
GO8	100%	0%	0%	0%
GO9	1%	3%	0%	96%
NPTB4	100%	0%	0%	0%
NPTB1	100%	0%	0%	0%
NPTA2	100%	0%	0%	0%
NPTA3	100%	0%	0%	0%
NPTA1	100%	0%	0%	0%
NPTB6	100%	0%	0%	0%
LL1	100%	0%	0%	0%
LL2	100%	0%	0%	0%
LL3	100%	0%	0%	0%
LL4	100%	0%	0%	0%
LL5	100%	0%	0%	0%
LL8	100%	0%	0%	0%
LL9	0%	0%	100%	0%
LL10	0%	23%	0%	77%
LL11	100%	0%	0%	0%
BK4	100%	0%	0%	0%
BK8	100%	0%	0%	0%
BK14	100%	0%	0%	0%
BK5	100%	0%	0%	0%
BK20	100%	0%	0%	0%
BK18	100%	0%	0%	0%
BK10	100%	0%	0%	0%
BK6	0%	10%	0%	90%
BK15	100%	0%	0%	0%
BK9	100%	0%	0%	0%
BK19	0%	3%	0%	97%
BK1	100%	0%	0%	0%
BK12	100%	0%	0%	0%
BK21	75%	2%	16%	7%
BK7	100%	0%	0%	0%
BK13	100%	0%	0%	0%
BK16	100%	0%	0%	0%
BK2	100%	0%	0%	0%

BK11	100%	0%	0%	0%
BK22	0%	0%	0%	100%
BK3	100%	0%	0%	0%
BK25	9%	24%	5%	63%
BK17	100%	0%	0%	0%
TH1	100%	0%	0%	0%
GO6	100%	0%	0%	0%
GO7	100%	0%	0%	0%
NPTB3	100%	0%	0%	0%
NPTB2	100%	0%	0%	0%
TH2	100%	0%	0%	0%
GO1	100%	0%	0%	0%
GO2	100%	0%	0%	0%
GO3	100%	0%	0%	0%
GO4	100%	0%	0%	0%
GO5	100%	0%	0%	0%