How can we use whole genome sequencing and mathematical modelling to understand tuberculosis transmission and inform our public health practices?

Hollie-Ann Hatherell

A dissertation submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

of

University College London

Department of CoMPLEX

University College London

10 October 2019

Declaration

I, Hollie-Ann Hatherell, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Tuberculosis (TB) remains a public health problem in cities in high-income, low-incidence countries, such as London, where it disproportionately affects particular population groups and, as such, more effective intervention strategies are needed. With whole genome sequencing (WGS) data being increasingly used for TB epidemiology, I investigated how WGS data alongside statistical inference and mathematical modelling can improve our understanding of transmission in these population groups. By reviewing the literature on WGS in TB epidemiology studies, I concluded that whilst genomic data can improve our understanding of TB transmission, including epidemiological data alongside is helpful for mitigating uninformative genomic data or strengthening conclusions. I then employed a statistical inference method on sequencing data from a Canadian outbreak and used the inferred transmission network to determine that the outbreak had ended, demonstrating the use of genomic epidemiology in public health. As we must analyse genomic data using bioinformatics and sometimes phylogenetic methods before we can interpret it for epidemiological purposes, I undertook bioinformatics analysis of 415 genomes from a London TB outbreak and attempted to create a timed-phylogenetic tree that could be used for genomic epidemiology inferences. However, the data proved difficult to interpret resulting in a tree with little confidence, potentially due to little variation amongst the sequences. Finally, I constructed a novel mathematical transmission model to recapitulate the London outbreak and investigate public health interventions to conclude that despite loss-to-follow-up being considered an important factor amongst the cohort anecdotally, focusing interventions on reducing loss-to-follow-up or increasing re-engagement does not significantly reduce the number of outbreak cases. Finding infectious cases early achieves the most impact. In conclusion, combining epidemiological and sequencing with novel quantitative analysis using statistical inference and transmission modelling, provides useful insight into the spread of TB in urban outbreaks and illustrates the limitations of new approaches and data.

Impact statement

The research presented here may prove useful within public health studies of infectious disease. In my systematic review, I explored why and where there may be limitations to the use of whole genome sequencing for investigating TB transmission. This will prove a useful resource for academic research to help in designing and evaluating studies of TB using whole genome sequencing, a now popular approach. I explore two new methods, a statistical inference method and mathematical model, for studying TB transmission for public health purposes. The statistical inference method seeks to infer the infection dates of TB cases, an often hidden aspect of TB disease, and is improved in this thesis. The finding of infection dates allows for the studied TB outbreak to be declared over, which is important for TB public health teams worldwide as it allows them to stop unnecessarily using resources for the outbreak. The mathematical model developed in the final chapter of this thesis seeks to model a (previously unmodelled) large ongoing TB outbreak and determine which of several interventions may prove the most effective. The results of the model could be applied in future epidemics and the model may be used to study other outbreaks in similar settings.

Acknowledgements

It goes without saying that this work has only been made possible thanks to the love and support of many people. Firstly, I would like to thank my supervising panel: Dr Charlotte Jackson and Dr Helen Stagg, you have believed in me even when I had no belief in myself. I greatly appreciate your attention to the little things. Thank you both for your kindness. To Professor Caroline Colijn, I have been lucky to have you to guide me whenever I tied myself up in modelling knots and am honoured to have collaborated with you. Finally, Professor Ibrahim Abubakar, I am eternally grateful for your ability to look at the facts and come up with a plan of action. You never gave up on me even when I wholly believe others would have. Thank you and thank you all for the opportunity.

Next to all those at CIDE who helped make the office a fun and enjoyable place to work. Especially, (Dr!) Joanne Winter, who began this journey with me, I am glad we could be systematic review buddies. A big thank you to Maeve Lalor, Charlotte Anderson, Jennifer Davidson, Helen Maguire, Neil MacDonald, Ted Cohen, Xavier Didelot, Jennifer Gardy and more who have collaborated with me and advised me in the various undertakings of this PhD.

To all those in the CoMPLEX family, particularly the 2013 intake, I am especially thankful for the laughs and learnings you have provided over the years. It would not have been as fun without you all and I could not have chosen a nicer bunch of people to embark on this journey with. Just a shame we never did the write-up retreat in the Scottish Highlands.

Before coming to UCL, I was lucky enough to meet people who have nurtured and guided me onto this path, mostly importantly, Dr Jon Pitchford who first gave me the infectious disease modelling bug.

Of course, I have to thank my friends and family. Thank you to the York group who have been such great friends, in particular, Ruby – thank you for always proof reading for me, including this thesis! To my family, my mum, Sarah, and brothers, Ashley, Max and Charlie, without you constantly asking me how the PhD was coming along I am sure I never would have got to this point. But in all seriousness, thank you all for your love and support for the past 28 years. And to my nan, Yvonne, who sadly passed before I could get to this point, I hope you would be proud.

Last but not least, my husband Joe. When I embarked on this journey five years ago, I had just dragged you away from your family to live in a tiny, expensive London flat as my boyfriend. You took a massive step outside of your comfort zone to let me follow my dream and I am grateful every day. Since then we have moved multiple times, got married, bought a house and adopted a cat, Indy (whose snuggles have also made this thesis possible). I am so glad that I got to do this with you and amazed that you have put up with my ramblings about genomes and disease transmission for so long. You are a true hero.

Contents

Index of tables

Index of Figures

Abbreviations

BIC	Bayesian Information Criterion
BEAST	Bayesian evolutionary analysis sampling trees
DNA	Deoxyribonucleic acid
GTR	Generalised time reversible
HCV	Hepatitis C virus
HIV	Human Immunodeficiency virus
LFU	Loss to follow up
LTBI	Latent tuberculosis infection
MCMC	Markov chain Monte Carlo
MDR	Multidrug resistant
ML	Maximum likelihood
MIRU-VNTR	Mycobacterial interspersed repetitive units – variable number tandem repeats
MRCA	Most recent common ancestor
MRSA	Methicillin-resistant *Staphylococcus aureus*
M. tb.	*Mycobacterium tuberculosis*
PHE	Public Health England
PRCC	Partial Rank Correlation Coefficient
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RFLP	Restriction fragment length polymorphism
SARS	Severe Acute Respiratory Syndrome
SIR	Susceptible Infected Removed
SNP	Single nucleotide polymorphism
SRF	Social risk factor
STROME-ID	Strengthening the reporting of molecular epidemiology for infectious diseases
TB	Tuberculosis
WGS	Whole genome sequencing
WHO	World Health Organisation

Publications and Presentations

Two peer-reviewed journal articles have been published from the work done as part of this PhD:

> H.-A. Hatherell, C. Colijn, H. R. Stagg, C. Jackson, J. R. Winter and I. Abubakar, "Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review," *BMC Medicine,* vol. 14, no. 1, p. 21, 2016. **(Chapter 2)**
>
> H.-A. Hatherell, X. Didelot, S. Pollock, P. Tang, A. Crisan, J. C. Johnston, C. Colijn and J. L. Gardy, "Declaring a tuberculosis outbreak over with genomic epidemiology," *Microbial Genomics,* vol. 2, no. 5, 2016. **(Chapter 4)**

One journal article building on work from this thesis is currently under review:

> Y. Xu, H. Topliffe*, J. Stimson, H. Stagg, I. Abubakar and C. Colijn, "Transmission analysis of a large TB outbreak in London: mathematical modelling study using genomic data," *bioRxiv,* 2019. **(Chapter 5)**
>
> *Topliffe is my married name.

One abstract was accepted for poster presentations:

> H.-A. Hatherell, J. Gardy, X. Didelot, C. Colijn, "What can we learn about tuberculosis transmission from phylogenetics: a renewal equation model," 2015. Epidemics[5] Fifth International Conference on Infectious Disease Dynamics, Florida, USA. **(Chapter 4)**

1 INTRODUCTION

Tuberculosis (TB) is often considered a disease of the poor or of the past, it is thus often assumed that high-income, developed countries are free from the burden of this ancient infectious disease. However, with over 5,000 TB cases in the UK in 2017 [1], this is far from the case. Although there has been a year-on-year decrease in the number of cases in the country since 2011, some boroughs in London have rates comparable to high-incidence countries: more than 40 cases per 100,000 in Brent and Newham [2]. Additionally, TB disproportionately affects particular population groups; the proportion of TB cases with at least one social risk factor, classed as current alcohol misuse that would impact on the patient's ability to take treatment, current or history of drug misuse, homelessness and/or imprisonment, has increased from 9.4% in 2014 to 12.6% in 2017 [1]. This would suggest we need a different approach for tackling the TB epidemic in these small pockets of communities, and perhaps a better understanding of what is driving TB outbreaks in low-incidence countries would help find a solution.

With the advent of whole genome sequencing (WGS), now is the perfect time to use the ever-increasing amount of genomic data to uncover more about transmission patterns and determine if it can provide better insights into the best way to tackle the TB epidemic. Namely, we need to establish how best to direct our current control efforts. In this thesis I investigate questions around TB control using whole genome sequencing, statistical inference, and mathematical modelling, using data collected from real-world TB outbreaks in two low incidence TB countries, the UK and Canada.

1.1 BIOLOGY OF TB

The bacterium *Mycobacterium tuberculosis (M. tb.)* is the causative agent of most human TB disease. Transmitted through the air via droplet nuclei expelled whilst coughing, singing, or sneezing [3], it primarily affects the lungs (pulmonary disease), and this is when it is in its most communicable form. For a proportion of cases, the bacteria will affect the body outside of the lungs (extra-pulmonary disease), for example the spine [4].

Once infected with the bacterium, the body's immune system fights the invading bacteria. Some individuals are able to eradicate the infection entirely; in others the immune system cannot clear the infection but manages to contain the bacteria within granulomas, collections of immune cells and *M. tb.* (which may be dead) [5]. If successfully contained, they can remain latently infected until the end of their lives and never experience active TB disease. For a small proportion of cases, often quoted as 5-10% although recently debated [6, 7, 8], active TB disease will develop when the immune system is no longer able to contain the bacteria, although this may be weeks or decades after infection [6]. Those with impaired immune systems, e.g. HIV-positive individuals, have a higher risk of developing active TB

[9]. The mechanisms behind eradication, containment, long latency periods and the breakdown of immune response are an area of interest for research as being able to enhance the body's innate immune response to TB could provide a viable method for control [10].

When individuals progress to active TB disease shortly after infection, defined to be within two (or three years), they are classed as having primary TB disease. We consider those progressing after a long period of latency (more than two years) to have reactivated, secondary, or post-primary TB. This may be as a result of endogenous reactivation, where the immune system breaks down releasing the bacterium from the granuloma, or exogenous, where the individual is re-infected [11, 12].

There are multiple methods that can help diagnose a patient with pulmonary TB, the most common are via sputum culture, sputum smear, or chest x-ray and these are ideally used in conjunction. Sputum culture and sputum smear require a sputum sample, produced via coughing up phlegm, which is then examined for the presence of *M. tb.*, only then is the case considered bacteriologically confirmed. It is possible for the smear and culture to produce contrasting results [13, 14], and even both to come back negative for the presence of *M.tb.* but the individual still classified as having TB via another means [15]. Thus, we often describe TB cases as having sputum/culture negative/positive TB, especially as this factor can determine certain features of the TB disease for that individual, e.g. the infectiousness of the patient [16, 17]. Radiological findings of lesions on the lungs can also suggest infection with TB [18]. Once diagnosed, the typical line of treatment for TB as recommended by the National Institute for Health and Care Excellence (NICE) is a six-month course of antibiotics, isoniazid and rifampicin supplemented by pyrazinamide and ethambutol in the first two months [19]. If the disease is bacteriologically confirmed we can test the sputum sample for drug resistance, which may result in an alternative treatment course, should the bacterium be resistant to one of the drugs used for treatment.

Once an individual diagnosed with bacteriologically confirmed TB has finished a prescribed course of treatment, we consider them to have been officially cured of (drug-sensitive) TB if they are culture negative/smear negative in the last month of treatment and on at least one previous occasion [20].

1.2 GLOBAL TB

Despite the existence of anti-tuberculous drugs [21], a vaccine [22], and multiple, albeit imperfect, methods for detection and diagnosis [23], TB remains a devastating problem globally with an estimated 10.4 million people contracting TB and 1.7 million dying from the disease in 2016 [24]. In particular, low income countries, such as Liberia with an estimated TB incidence rate of 308 per 100,000 in 2016 [24], are affected to a greater extent than high income countries. These figures have earned TB the title of the world's deadliest infectious

disease, clearly highlighting that there is a serious need to address the epidemic. The ability of *M. tb.* to develop drug resistance [25], remain latent for decades [26], and supposedly establish an infection after brief exposure [27], have all contributed to its success in evading eradication.

Numerous initiatives around ending the TB epidemic exist worldwide, such as the World Health Organization's (WHO) End TB strategy [28], which outlines a goal to dramatically reduce suffering and death due to TB by 2035, similar to the United Nations' Sustainable Development Goals (Goal 3: Good health and well-being) . Others include the Zero TB initiative [29] and the Stop TB partnership [30]. These initiatives often centre on increasing the rates of diagnosis and treatment, improving distribution of drug, vaccine and technological resources (especially to low-income countries), and finding better drug regimens, better diagnostics and a new vaccine. The major focus of these initiatives is to tackle the TB epidemic in low-income countries with high burdens.

1.3 TB IN LOW INCIDENCE COUNTRIES

Whilst developing countries shoulder the largest proportion of the TB burden, high-income developed countries, such as the UK, still experience pockets of high incidence, especially in large cities [31]. These high incidence rates are mostly thought to be driven by a high proportion of migrants from high-incidence countries, immuno-supressed, e.g. HIV-positive, individuals, and socially deprived individuals living in crowded and poor sanitary conditions, as large, urban cities have a high number of these populations [32, 33]. As a result, one key area of control for the TB epidemic in low-incidence countries centres on effective allocation and use of resources (medication, technology, etc.) for targeting of these hard-to-reach individuals, as opposed to how to afford and obtain the resources foremost, as is the case in many developing countries [34].

Public Health England (PHE), an executive agency of the Department of Health and Social Care responsible for safeguarding the public health of the country, has determined its own TB strategy for tackling the epidemic, the Collaborative TB strategy for England 2015-2020 [35]. Their aims are to find cases (and start them on treatment) earlier, increase the number of cases finishing their course of treatment, reduce the amount of drug-resistant TB and tackle TB in underserved populations (i.e. the hard-to-reach).

We define hard-to-reach individuals as people who do not engage with the standard healthcare system and are difficult to treat [36]. Within a metropolitan city, this may be because they have a social risk factor (SRF), such as they take drugs, abuse alcohol, are frequently in prison or have no fixed abode, amongst other reasons such as having a disability or speaking a different language. As a consequence, in London, numerous monitoring indicators are worse in these groups than in those without a SRF: the most

deprived 10% of the population have a TB incidence rate more than 7 times higher than the least deprived 10%, treatment outcomes are worse (6.5% of those with at least one SRF are lost to follow-up at final outcome vs. 3.3% without an SRF), and time to diagnosis is longer (33.9% of people with pulmonary TB and a SRF who experienced a delay from symptom onset to treatment start of more than four months versus 31.6% of those without an SRF) [1]. These individuals therefore often act as an untreated source of infection and potentially have large contact networks (i.e. come into contact with or share a common setting with a large number of individuals), due to exposure to locales with large mixing populations e.g. prisons and homeless shelters. As a result, extra interventions besides the standard are necessary to control the epidemic in this group. Standard interventions would consist of efficacious treatment and passive case finding, which involves only screening cases (testing) for TB when individuals present at a healthcare setting. Extra interventions focus on finding and diagnosing people in these groups quickly, using active case finding or contact tracing [37, 38], which involves finding more cases through population screening and screening contacts of infected cases; encouraging them to take treatment and come to follow up appointments, using peer support and cash incentives [39, 40]; and tracing those who become lost to follow-up [41, 37].

Clearly, even though these efficacious interventions exist, they are only effective and cost-effective if put into practise correctly. A key priority for public health systems is to use the minimal amount of intervention for the maximum amount of gain. In order to maximise the effectiveness and cost-effectiveness of interventions, it is crucial to determine where their usage would be most beneficial; this may entail targeting a certain geographical location, a certain locale (e.g. public houses or churches) or a certain population group. We can determine these targets through a good understanding of the biological, environmental and social components of TB transmission.

1.4 TB TRANSMISSION DYNAMICS

In order to examine TB transmission dynamics (the who, where, when of TB transmission), we need to do two things:

1. Identify transmission – who has infected whom?
2. Identify patterns by relating data on who infected whom to personal information about the individuals

If we can find patterns, then we can try to identify interventions that might exploit these patterns and break the chain of transmission. How we can undertake each of these tasks is detailed below.

1.4.1 IDENTIFYING TB TRANSMISSION

Identifying when transmission has occurred is difficult because we cannot see TB macroscopically during transmission and therefore cannot definitively say that person A infected person B at time X. Thus, we need to infer the possibility of such transmission events using evidence based on the infected individuals (locations visited, length of time infected etc.) and their infecting strains of bacteria.

Traditional epidemiology investigates transmission using evidence based on the individuals; examining their demographics, their social lives and their environments and relating these factors to the risk of disease [42]. This may involve building a social network of contacts (contact network) for the infected individuals, i.e. a "map" of individuals who know those who are infected or have some kind of social setting in common with them. These contacts are said to have an "epidemiological link", thus providing a potential for transmission. A lack of epidemiological link between infected individuals would then preclude the possibility of them having infected each other. This kind of information can be time-intensive to collect as it involves fielding questionnaires and sometimes deep investigation in order to identify an opportunity where transmission may have occurred, as even brief contact has been found to yield the potential for transmission. In addition, it is extremely difficult to obtain a complete dataset as some individuals, those who are hard-to-reach in particular (see Section 1.3), will not be available or willing to answer the questionnaire, potentially due to illicit activity. To compound these difficulties, as it may be decades between infection and active disease [43], the likelihood of remembering contacts after so many years is diminished.

Molecular epidemiology, however, is concerned with the bacteria sampled from the infected individuals and tries to determine the possibility of transmission using knowledge of the bacterial genomics. The practise uses genotyping data to determine whether bacterial samples are genetically similar and therefore may have shared a recent ancestor, this in turn suggests that the two individuals the samples were taken from are linked in recent transmission [44, 45]. There are a number of difficulties with molecular epidemiology. Firstly, it requires a sample, which may not always be obtainable, for example, sputum is the primary sample type used for diagnosing TB, but some patients are unable to cough up sputum. Secondly, there are numerous types of genotyping methods and they each have their own advantages and disadvantages for transmission analysis as explored in the following section.

In Chapter 4, I employ molecular epidemiological techniques on outbreak WGS data to explore transmission patterns.

1.4.1.1 METHODS FOR GENOTYPING TB

Genotyping involves examining the genetic material of an organism. *M. tb.* has a DNA genome of size roughly 4.4 million base pairs (bp), where variation is mostly generated through insertions of small DNA fragments, deletions of small DNA fragments or single base mutations (called single nucleotide polymorphisms (SNPs)) as opposed to horizontal gene

transfer and recombination. As a result, *Mycobacteria* are very monomorphic and highly clonal, with 99.9% of the DNA sequence of sub-species being identical. An additional consequence is that the genome is considered very stable, and slowly changing when compared to other bacteria [46], with documented mutation rates from 0.3-0.5 SNP per genome per year [45, 47, 48].

M.tb. has co-evolved with man over the past 40,000 years, with the oldest known case of molecularly confirmed (found mycobacterial DNA) TB disease to have been estimated at 9,000 years ago [49]. As man migrated from the Horn of Africa, TB migrated also and over time the species split into multiple lineages [50]. As a result, different lineages are correlated with geography; often resulting in lineages being named after the locations in which they are endemic, such as the Euro-American and East Asian lineages [51, 52]. However, as global travel has become commonplace, the boundaries around this have become more blurred.

Three core genotyping methods have historically been used for *M. tb*: Restriction fragment length polymorphism (RFLP), mycobacterial interspersed repetitive units – variable number tandem repeats (MIRU-VNTR), and spoligotyping (see Table 1.1). They operate by examining certain sections of the mycobacterial genome that form distinct patterns, for example the presence of insertion elements (*IS6110* RFLP), and then representing these patterns in a form, such as a string of numbers (representing the number of repeats present at certain loci), which can be used to compare two genomes.

However, the granularity of these methods and the overall homogeneity of *M. tb.* means that the methods cannot always identify differences between strains, leading to the conclusion that potentially unrelated strains are closely related. As a result, transmission between the hosts is assumed and transmission is often over-estimated [53]. This is because if the genetic marker being studied has not evolved enough over time (in accordance with the timescale of transmission) then genotypes from samples not linked in transmission will look identical or similar, making it difficult to conclusively rule-out the possibility of transmission having taken place [54]. WGS has been considered a solution to this problem.

Strain typing method	What it is?	Advantages	Disadvantages
Insertion sequence 6110 (IS6110) Restriction fragment length polymorphism (RFLP)	Determines the position and number of copies of IS6110. Digested DNA is run on a gel and the copies produce a banding pattern	More discriminatory than spoligotyping or MIRU-VNTR	Requires weeks of culturing. Not very reproducible
Spacer oligonucleotide typing (spoligotyping)	Looks for presence of spacer sequences in a particular region of the genome. Produces a binary code that is converted to its octal equivalent	Reproducible. Requires only small amount of sample. Quick	Not very discriminatory
12/15/24 Mycobacterial interspersed repetitive units – variable number tandem repeats (MIRU-VNTR)	Number of tandem repeats (repeats adjacent to each other) at 12-, 15- or 24-loci in the genome. Produces a numerical code	Reproducible. Requires small amount of sample. Quick. More discriminatory than spoligotyping	Less discriminatory than IS6110 RFLP
Whole genome sequencing (WGS)	The nucleotide sequence for the entire length of the genome. Samples are compared to a reference strain and single nucleotide polymorphisms (SNPs) determined	Most discriminatory of all the methods	Analysis requires high skillset and specialist software. More expensive than other methods

Table 1.1 Summary of different genotyping methods used for Mycobacterium tuberculosis strains and their advantages and disadvantages

1.4.1.1.1 WHOLE GENOME SEQUENCING

Whereas other methods look only at certain areas of the genome that may exhibit some genomic signal, WGS examines the genome in its entirety at the very finest level of detail - the building blocks of the entire genome - therefore any and all differences between two genomes can be compared, base by base. At the time of beginning this thesis (October 2014), the increased availability of WGS was advancing the field of molecular epidemiology [55] by permitting discrimination between bacterial (and viral) strains that are indistinguishable using other methods and therefore providing a superior method for revealing transmission networks.

Since the advent of WGS into the field of molecular epidemiology, there have been high expectations for the insights that it can bring to the understanding of TB [56]. The ability to inspect the raw sequence data means that all genomic variation can be identified, giving WGS a clear advantage over molecular genotyping methods, which have struggled to distinguish between closely related *M. tb.* strains (due to its slow molecular clock) [43]. The increased discrimination is expected to provide better answers to questions of genomic relatedness and, by extension, the transmission of strains [57].

WGS usually involves first culturing *M. tb.* from a patient sample, e.g. sputum, and then extracting the DNA (although this can be done culture-free [58]). The DNA is then broken up into short fragments of known length, which could vary from 50 to 500 bp [59]. Polymerase chain reaction (PCR) is then used to amplify the DNA fragments and make many copies, which are then sequenced using high-throughput next-generation sequencing technologies from companies such as Roche and Illumina [60]. The result after sequencing is millions of short sequences of As, Ts, Gs and Cs called “reads”. These then need to be reconstructed to recover the original genome sequence using bioinformatics.

Despite having several advantages over molecular genotyping, WGS is still hindered by the homogeneity of the *M. tb.* genome [61], as even though it is possible to examine the genome at the finest scale, if there is no variation between genomes then we are still unable to make certain inferences about transmission. As well as this, ambiguity around the best approach to bioinformatics analysis and interpretation of diversity in the genome has made it difficult to compare studies and produced conflicting results. As a result, WGS has not brought the enlightenment that was envisaged for public health. In Chapter 3, this thesis addresses some of these issues by reviewing some of the methods of WGS analysis and what results we can realistically draw from such analyses for public health interventions.

1.4.1.1.1.1 PHYLOGENETIC TREES

WGS data can be used for numerous types of investigation; one can look for mutations that confer antibiotic resistance, the number of SNP differences between a group of genomes or identify species and/or subtype of an isolate. It can also be used to build a phylogenetic tree,

a pictorial way of representing the relationships between multiple whole genome sequences which in the past few years, has increasingly been used to investigate transmission. One method for doing this is employed in Chapter 4 of this thesis.

Phylogenetic trees can show how different organisms are related through either shared physical characteristics or genetics. The anatomy of a tree consists of branches and tips: each tip represents an organism and the tips are at the end of branches. Two branches are joined together at a node, which represent their most recent common ancestor (MRCA). Some trees also contain a "root", meaning a unique node where the earliest splitting occurs; this denotes the MRCA of all the samples in the phylogenetic tree, if no such root is present then the tree is considered unrooted. The most basic interpretation of a phylogenetic tree is that the more recent a common ancestor between two organisms is, the more closely related they are [62].

Phylogenetic trees can be constructed using several different methods and metrics [63]. For example, the simpler variety involves calculating the (genomic) difference between all the organisms using a certain measure (e.g. SNPs) and constructing a distance matrix. An algorithm then seeks to build a tree which minimises the distances between all the organisms, for example neighbour-joining. These do not use an evolutionary model, i.e. a substitution model that describes how the bases in the sequences change over time and is typically stated as rates of change between any two given bases, say from an A to a G. These can vary from all being identical to each being different. More complicated methods, such as maximum likelihood (ML) trees, do include evolutionary models and work by assessing the probability of the sequence data given all possible phylogenetic trees and a substitution model, and then choosing the tree that maximises this probability. Finally, there are Bayesian methods for building trees, which also assume an evolutionary model but seek to construct a tree after sampling.

Trees can be broadly split into two types: timed versus untimed. Typically, the scale of a phylogenetic tree is based on genomic distance (untimed), meaning that the length of a branch is proportional to genomic distance. However, timed trees are on a timescale and do so by relating genomic distance to evolutionary time via a molecular clock i.e. the idea that genomic variation accumulates over time in some consistent way.

1.4.2 TB TRANSMISSION PATTERNS

There are multiple levels at which to examine an outbreak of disease. Firstly, we could think about the individual level: the mode of transmission, the host characteristics, the patient's contacts. Then there is the outbreak level: where we can take all the aggregate information about the individuals linked in transmission and determine if there are any overarching patterns that correlate with other data such as location or drug resistance profile. As an example, on an individual level we would want to monitor a patient's treatment to ensure that

they are taking their treatment course correctly. A result of that may be identifying that that individual is frequently missing doses or become lost to follow up. If we gather that information on all patients, we would get a good idea of the proportion of patients that do not take their treatment correctly and therefore are at risk of continuing to be infectious, a useful input into a model of the outbreak. It may also be found that the lost to follow up individuals are all from the same area or are a similar age; information like this could help refine the implementation into the model and potentially identify a way of targeting interventions in this population.

Once we can identify transmission through traditional and molecular epidemiology methods as mentioned in Section 1.4.1, we can build transmission networks. Transmission networks are a way of presenting infectious disease cases alongside who infected whom, usually depicted with the cases as nodes and transmission represented as arrows from one node to another. It is then through identifying transmission events and building transmission networks that then patterns can be identified by examining the additional information available for the cases. These patterns can then be used for targeting interventions, e.g. a pattern of TB occurring between patients at hospitals would suggest nosocomial transmission is common in TB, thus a potentially effective intervention would be isolation for TB patients at hospitals [64].

1.5 MODELLING APPROACHES

Once transmission patterns have been identified they can be used to help determine the most effective interventions. This can be done speculatively for example, identifying that a large proportion of cases in an outbreak are currently in prison and implementing an intervention in prisons, e.g. screening. But ultimately it is preferable if the effect of an intervention can be quantified in a rigorous way in order to determine whether it is cost- or time-effective. Due to constraints on money and time that would be needed for experimental studies to test interventions as well as any potential ethically issues, this is often done by constructing a mathematical model [65]. A mathematical model will be built that incorporates some or all of the transmission patterns and demographic patterns identified in the outbreak population and then different interventions can be implemented to examine where and how they may produce the largest positive impact on incidence and/or mortality. This is a common use of modelling in public health, which has been employed for many infectious diseases [66].

Of course, the development of a mathematical model involves a plethora of decisions around model type, structure, parameter choices, and more. For example, a common discrepancy in mathematical models of TB transmission is the implementation of latent TB infection; some models include only one latent compartment, others include two

consecutively whereas some include two but in parallel [67]. All these assumptions will have an impact on the outcome and interpretations.

The most suitable choice of model would depend on the type of patterns that need to be incorporated as well as the level of complexity desired. Some examples of how transmission patterns potentially determined from transmission data could influence the design of a mathematical model of infectious disease are listed in Table 1.2.

Transmission pattern	Example of model implementation
Number of secondary infections per person is larger amongst those who are non-adherent	Separate the infectious population into adherent and non-adherent, with a higher transmission rate for non-adherent
Infection mostly occurs between individuals of a similar age	Stratify the population by age and set inter- and intra-strata transmission rates, with intra-strata rates higher than inter-strata
Time between being infected and infecting others is short	Set the rate of progression from latent to active disease to be large, corresponding to a shorter latent period

Table 1.2 Examples of transmission patterns and their possible implementations in a mathematical model

The most common type of mathematical model for infectious diseases is the compartmental model, where a set of individuals who share some characteristic related to the disease (for example they are in a certain stage of disease) are grouped together and there is a flow of individuals from one compartment to another as their characteristic changes, e.g. disease progresses. The most basic of these models is the Susceptible-Infected (SI) model, where individuals are either infected with the disease or not; this can then be extended to the common Susceptible-Infected-Removed (SIR) model or even further to include more disease stages (e.g. latency, vaccinated or immunity) or separate populations in parallel [68, 69, 70].

The compartmental model is a population-based, deterministic model. As a contrast to this, there are models which can track individuals in the population (individual-based) and models which are stochastic i.e. include random effects to better model real-life where stochasticity is inherent and can have a big effect. The choice of which type of model is best for the study will rest on keeping the complexity low whilst being able to effectively study the outputs of interest.

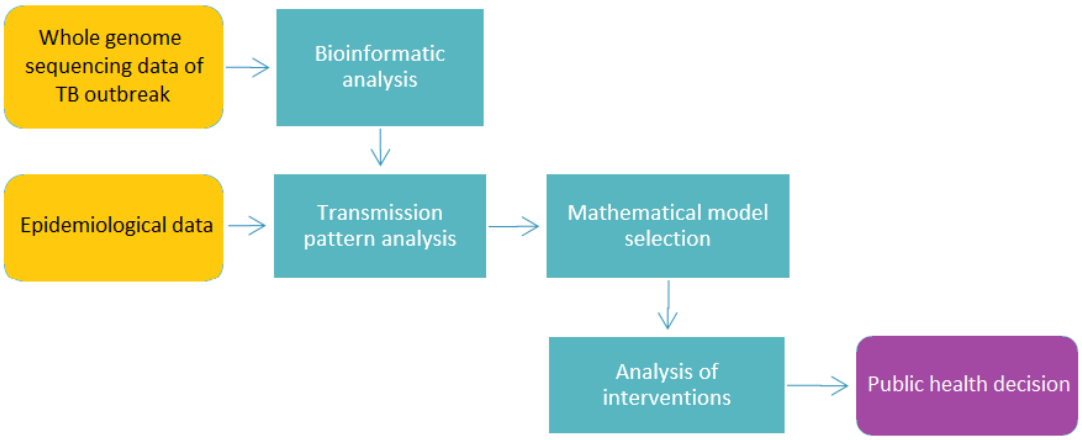

Figure 1.1 An example workflow of how whole genome sequencing data can be used to inform public health decisions around interventions for a tuberculosis outbreak

Data (yellow boxes) feeds into the analysis steps (blue boxes) and produces the final output (purple box)

1.6 SUMMARY

With all the aforementioned techniques in mind (bioinformatics analysis, transmission analysis, mathematical modelling), it is possible to identify a cohesive workflow that draws on all the methods and can be used to better understand TB transmission and use WGS and epidemiological data to inform public health decisions (Figure 1.1). This workflow requires numerous skillsets, numerous different programming software, multiple datasets, and experience with different datatypes. Although these are demanding requirements it has many advantages; it incorporates complementary datasets, it can be adapted for use with different infectious diseases and includes steps in order to assess quality, sensitivity and uncertainty.

During my thesis, I explore each aspect of this workflow.

2 AIMS AND OBJECTIVES

2.1 AIM

The overall aim of this thesis was to examine the use of WGS, statistical methods and mathematical modelling for understanding TB transmission and evaluate how these methods can be used for public health applications.

2.2 OBJECTIVES

Objective 1: To systematically review the literature for uses of WGS data to understand TB transmission, namely, to answer whether we can use WGS to:

- distinguish relapse from re-infection
- confirm or rule out transmission events
- determine direction of transmission

and therefore, identify transmission patterns.

Objective 2: To infer a transmission network with infection timings for a TB outbreak in British Columbia, Canada from WGS data using a novel Bayesian inference method and interpret the findings for public health purposes.

Objective 3: To analyse WGS data from a drug resistant TB outbreak in London, UK using the same Bayesian inference method to reveal information on transmission patterns useful for guiding the public health response.

Objective 4: To build a mathematical model of TB transmission for the same outbreak in London and investigate a number of public health interventions.

2.3 OUTLINE OF THE THESIS

The outline of the remaining chapters of this thesis are as follows:

Chapter 3: Literature review of WGS studies

This chapter contains a systematic review of the literature for different methods used to interpret whole-genome sequencing data for TB transmission and comments on the advantages and limitations of each method for investigating certain characteristics of outbreaks, for example the patterns of drug resistance or determining re-infection from relapse.

Chapter 4: Bayesian inference method on Canadian TB outbreak

This chapter introduces the Bayesian inference method used to infer outbreak transmission dynamics from whole genome sequences taken from a TB outbreak in Canada.

Chapter 5: Phylogenetic analysis of WGS data from London TB outbreak

This chapter presents the methods for a bioinformatic and phylogenetic analysis of whole genome sequencing data of a large drug-resistant TB outbreak.

Chapter 6: Mathematical modelling and analysis of TB outbreak interventions

This chapter contains an explicit description of a mathematical model used to describe an outbreak of TB in London and present the effects of different interventions.

Chapter 7: Discussion and conclusion

This chapter discusses the results of all the chapters and the potential for future work and presents the conclusions from all aspects of the thesis.

3 REVIEW OF APPROACHES USED TO INTERPRET WHOLE GENOME SEQUENCING DATA FOR TUBERCULOSIS TRANSMISSION (OBJECTIVE 1)

Since the advent of cheaper and faster genome sequencing, the use of sequencing data in epidemiological investigations of infectious diseases has increased dramatically [71]. PHE now routinely uses WGS to investigate clusters of potential TB transmission [72] and the use of more traditional typing methods has all but been phased out. Whilst more genomic information is always better than less, WGS is only useful if we can understand the data and what it means for transmission, drug resistance, etc. For sequencing data, this relies on having confidence in our bioinformatic and phylogenetic analyses and being aware of the consequences of our assumptions, and therefore requires training and retooling for those wanting to explore transmission with these data. In this chapter, I aim to understand how different bioinformatic and phylogenetic approaches have been used to infer aspects of transmission dynamics that are useful for public health purposes and outline the limitations of these methods.

To do this, a systematic review of the literature was performed. This involved comprehensive searches of multiple electronic databases for studies that presented methods used to interpret WGS data for investigating TB transmission. Two authors independently selected studies to be included and extracted data. Due to considerable methodological heterogeneity between studies, we present summary data with accompanying narrative synthesis rather than pooled statistical analyses.

This chapter is based on a paper by Hatherell *et al.* [73].

3.1 BACKGROUND

WGS has been considered an advance in the field of molecular epidemiology; due to its high discriminatory power it can distinguish pathogen strains when other typing methods are unable [55]. This makes it superior for public health purposes where outbreak surveillance (tracking a strain through a population) and outbreak source identification are key goals [74]. The increased discriminatory power also allows for better resolution of transmission events and their direction, particularly in the case of pathogens that have very little genetic diversity [75]. Although WGS can provide these greater insights, it is important to be able to compare the results of different WGS studies in order to understand patterns amongst different outbreaks. However, the variation in methods for producing, analysing and interpreting WGS, in the field of TB epidemiology [57, 76] and infectious disease in general [55, 77, 78], makes direct comparison between outbreaks difficult. Here I discuss the different methods

for examining outbreaks and advantages and disadvantages of methods that may help guide the decision to a more standardised method of WGS analysis and interpretation.

To date, TB molecular epidemiology using WGS has focussed on four aspects [57, 78]: identifying chains of transmission; differentiating between relapse and re-infection; measuring within-host diversity and its impact on transmission; identifying primary versus secondary (acquired) drug resistance. These all play an important role in understanding and tackling TB outbreaks in the following ways:

- Identifying chains of transmission means determining who infected whom within an outbreak so that the outbreak progression through the population can be seen. The importance of delineating transmission chains for helping to identify larger patterns of transmission were briefly discussed in Section 1.4.2.
- Differentiating between re-infection and relapse involves being able to determine the nature of a second case of TB disease and distinguish a new infection separate from the original case from a re-activation of the original uncleared case. Both have important distinct consequences for public health: re-infection implies ongoing transmission, requiring public health action, and suggests a lack of immunity to the newly infecting strain or high intensity of exposure [79, 80]. Relapse suggests initially inadequate treatment.
- Measuring within-host diversity refers to being able to detect whether there are multiple genomic variants within one host and determine the nature of it, i.e. via microevolution or mixed infection. If within-host diversity is not fully captured, transmission might be inappropriately ruled out. For example, if an individual co-infected with two dissimilar strains transmits one of these to a contact, and different strains are then sampled from the two patients, these cases would not be identified as linked [81]. As a result, it is important we can identify and measure it correctly using WGS.
- Distinguishing between primary and secondary drug resistance. Primary drug resistance is defined as an individual who has drug resistant TB because the TB strain that is transmitted to them is already drug resistant. Contrarily, secondary or acquired drug resistance is when an individual has drug resistant TB because the initially drug-sensitive strain they became infected with develops drug resistance during the course of their infection. This can have consequences for TB interventions as primary drug resistance requires tighter controls to stop transmission, whereas secondary drug resistance necessitates improvements around treatment.

Each one of these topic areas requires an awareness of the methodological choices available and their limitations, which should underpin the choice of analytical approach. For example, if the study is investigating relapse versus re-infection in a population that harbours

a TB strain that has a mutator phenotype it would be unwise to use a threshold determined from a strain that does not have a mutator phenotype because the strain is likely to evolve more quickly than the one used to set the threshold and this will likely rule out possible relapses.

In this review I describe the methods used to analyse WGS data, their limitations and implications for clinical application. Although I focus on the use of WGS for investigating TB, many of the same methods have been employed in the study of other infectious diseases (e.g. Severe Acute Respiratory Syndrome (SARS), Coronavirus [82], methicillin-resistant *Staphylococcus aureus* (MRSA) [83] and *Clostridium difficile* (*C. diff.*) [84]).

3.2 METHODS

The study was conducted, where relevant, in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement [85].

3.2.1 SEARCH STRATEGY AND STUDY SELECTION

Multiple databases were searched using the combination of variants of the key terms "genome sequencing", "tuberculosis" and "transmission" with no date or language restrictions (see Section 9.1.1 – Appendix 1). For completeness, the references of included articles were also checked for any relevant missing articles. Articles were double screened by me and a second reviewer and included if they performed WGS data analysis to investigate the transmission of *M. tb,* according to any of the four topics prioritised for this review. Disagreements were resolved by a third party. Reviews, opinion pieces, studies in non-human subjects and of other *Mycobacteria* were excluded. The studies which had full texts written in a language other than English were excluded based on their abstracts which were written in English.

3.2.2 DATA EXTRACTION

Data from each study were extracted by myself and a second party independently into a pre-designed spreadsheet that included participant characteristics, the protocol for bioinformatics analysis, and the definition of mixed infections, in line with STROME-ID guidelines [86] (see Section 9.1.2 – Appendix 1). Discrepancies between the reviewers were discussed until consensus was reached.

3.2.3 DATA SYNTHESIS AND QUALITY ASSESSMENT

Meta-analysis was considered not feasible due to the heterogeneity in the methods and the results of the included studies; thus a narrative synthesis of the main findings is presented.

Criteria from STROME-ID and Newcastle-Ottawa [87] were adapted (see Section 9.1.3 – Appendix 1) to evaluate the molecular and classical epidemiological aspects of study quality as either 'adequate', 'inadequate' or 'unknown'. I performed the quality assessment and a second party independently confirmed 10% of the results. Discrepancies between the reviewers were discussed until consensus was reached.

3.2.4 DEFINITIONS

Base call: The base or allele (A, G, C, or T) assigned to a genomic position.

Heterozygous (mixed) base call: When multiple bases are called to one genomic position**.**

Variant: In the context of one genome, it refers to a genomic position which has a different base call to the reference strain. In the context of multiple genomic sequences/subpopulations, the variant refers to the genomic sequence/subpopulation which has at least one SNP compared to another.

Strain: A genetic variant of an organism.

Sample: A specimen retrieved from the patient, typically sputum. This could potentially contain a mixture of strains.

Isolate: An organism cultured from a patient sample.

Cluster: A group of two or more cases presumed to be linked in transmission due to sharing a genotype.

3.2.5 PROTOCOL AND REGISTRATION

This review was registered on PROSPERO (CRD42014015633).

3.3 RESULTS

694 papers were found through using my search strategy (Figure 3.1). After de-duplication, 358 articles remained, with 25 (reporting on 25 studies) meeting our inclusion criteria with 97% inter-reviewer agreement (see Appendix 1). The main reason for exclusion of a paper was that the study was not looking at transmission; rather they were looking at TB cases regardless of whether they were part of an outbreak. The studies can be separated dependent on which aspect of transmission they investigated, with some studies looking at multiple aspects: the possibility of transmission regardless of direction (12 studies) [45, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97] [48]; the direction of transmission (9 studies) [44, 89, 90, 91, 92, 98, 99, 100, 101]; the nature of TB recurrences (4 studies) [56, 79, 92, 102]; within-host strain diversity in the context of transmission (7 studies) [48, 79, 88, 90, 92, 102, 103];

and the emergence of drug resistance (6 studies) [97, 104, 105, 106, 107, 108]. A wide range of populations (ages, ethnicities, co-morbidities), countries with varying TB burdens, and differing dominant *M. tb.* lineages were represented amongst the studies.

3.3.1 QUALITY OF STUDIES

In order to assess the quality of the included studies, two reviewers independently reviewed ten quality criteria and marked the studies as 'adequate', 'inadequate' or 'unknown' (Appendix 1). Inter-reviewer agreement for the quality standards of the studies was 86%.

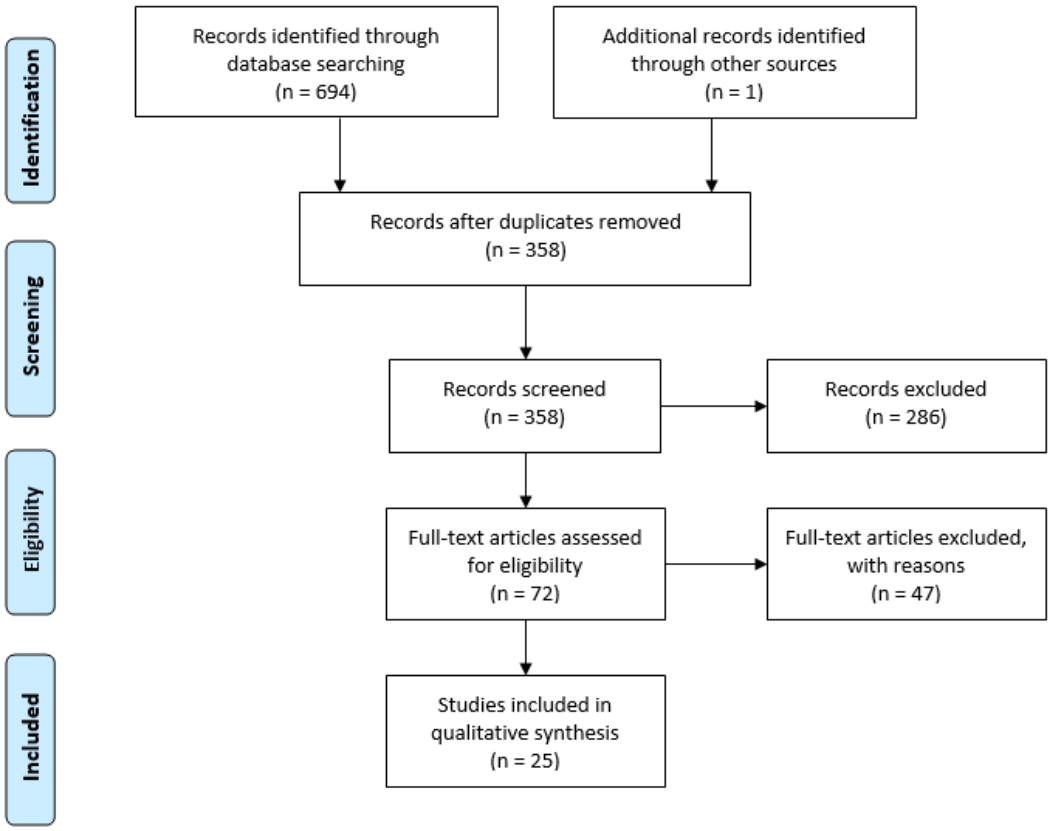

Figure 3.1 PRISMA flowchart showing the study selection process

Only a single study was assessed to have an inadequate case definition [108], this was due to the cluster being defined as samples that shared the same spoligotype and spoligotyping has been shown to have limited discriminatory power [109]. Ascertainment bias was identified in one study [44] as they only assessed the full genomes of three out of 104 outbreak samples to find eight SNPs. They then examined only these sites in the remainder of the samples to obtain a SNP type, potentially excluding variation in the rest of the genome.

The quality assessment also looked at how studies defined mixed infections. Our assessment considered using heterozygous base calls to define mixed infections to be 'adequate', given that there is no consensus on how to define mixed infections and we wanted to document and comment on how the different studies interpreted heterozygous base calls for microevolution and mixed infections. Additionally, the culturing of samples i.e. the practise of multiple passaging of the sample to obtain a culture, was not considered inadequate for defining mixed infections, as it is not possible to confirm if the process affected the presence of multiple strains without comparison WGS data from the non-passaged culture. 64% of studies did not report finding any mixed infections. Measuring or minimising cross-contamination was only documented by seven studies (28%). The comparison of WGS and epidemiological data was mixed between studies; with 20% commenting on epidemiological data but failing to compare the number of SNPs separating epidemiologically linked patients.

It has been shown that including low quality studies in meta-analyses can influence the outcome [110], however, as a meta-analysis was not performed, no studies were excluded on the basis of low quality. The aim of this review was to weigh the advantages and disadvantages of all approaches that had been used in published studies, and it was felt that excluding some due to low quality would potentially bias the outcome of the methods used and not present the whole picture.

3.3.2 CONFIRMATION OF TRANSMISSION

Studies attempting to infer transmission amongst cases in an outbreak (irrespective of direction) using WGS data used a number of different methods and interpretations [45, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97] [48]. Chiefly they combined information on SNPs as a measure of genetic distance, epidemiological data, and/or phylogeny. These methods are described in turn.

SNP thresholds were used by nine studies to confirm transmission [48, 89, 91, 92, 93, 94, 96, 97, 98] (Table 3.1). In a study which sequenced all TB cultures obtained from cases between 1995 and 2011 held at the Midlands, South Yorkshire and Humberside PHE laboratory (390 sequences), Walker *et al.* [48] investigated SNP differences within community and household clusters in the UK. By identifying epidemiological links between cases which had samples and then cross-referencing with the number of SNPs between the linked samples, they concluded that a five SNP threshold can be used to exclude transmission, as no epidemiologically linked pairs of samples exceeded this level of difference. An epidemiological link was defined as individuals having shared a space at the same time as each other or a third party.

Clark *et al.* [97], Kato-Maeda *et al.* [98], and Roetzer *et al.* [91] have similarly defined thresholds using epidemiologically linked or genotypically clustered cases. In their study to

examine the nature of drug resistant TB using 51 TB samples from 41 treatment-experienced TB patients, Clark *et al.* declared that a proportion of isolates had an excess of 50 SNPs variation and thus were not implicated in transmission and only focussed on the clusters that had a variation of less than 50 SNPs. However, the reasoning behind the use of a 50 SNP threshold is not transparent and it is not clear how the variation is defined, e.g. whether they all isolates are more than 50 SNPs different to any other in the cluster, or if the maximum number of SNPs between any two isolates is 50. Kato-Maeda *et al.* determined epidemiological links between nine TB cases and, on comparing the sequences amongst them, noted that there was a range 0-2 SNPs between all linked cases. Roetzer *et al.* in a study of 86 isolates from a long-term outbreak in Germany, found that a maximum of three SNPs was found amongst isolates from confirmed transmission chains and determined this to pose a SNP threshold for transmission. Luo *et al.* [89], Witney *et al.* [94] and Guerra-Assunção *et al.* [92] and Walker *et al.* [96] employed existing SNP thresholds to define transmission clusters.

An alternative approach by Lee *et al.* [93], employed in a remote Canadian village experiencing an outbreak of 50 TB cases, determined the variation between improbable transmission pairs first (those residing in different villages without an epidemiological link) and, as no pair had less than two SNPs difference, used 0-1 SNPs between sequences to define a cluster.

Journal article	How was threshold defined	Cut-off	Sampling fraction	Lineages
Bryant *et al*. [102]	Own data	≤6 SNPs relapse (same strain), >1306 re-infection (different)	47 sequenced out of 50 chosen	4 major lineages
Clark *et al*. [97]	Unknown	<50 SNPs defined a cluster	-	CAS, LAM, EAI, T1, T2, Beijing, X1
Guerra-Assunção *et al*. [79]	Own data	≤10 relapse, >100 re-infection	60 out of 139 WGS confirmed recurrences	4 major lineages
Guerra-Assunção *et al*. [92]	Own data (transmission), Guerra-Assunção *et al.* [79] (relapse)	≤10 SNPs confirmed transmission, ≤10 SNPs defined a relapse	1687 out of 2332 had WGS	4 major lineages
Kato-Maeda *et al.* [98]	Own data	0-2 SNPs per transmission event	-	-
Lee *et al.* [93]	Own data	0-1 SNPs confirmed transmission	631 “improbable” transmission pairs between outbreak cases and cases in other villages	Outbreak samples were Euro-American lineage
Luo *et al*. [89]	Walker *et al.* [48]			
Roetzer *et al*. [91]	Own data	3 SNPs confirmed transmission	31 out of 2301 (for the threshold). Equivalent to 8 transmission chains of	Haarlem lineage

			2-7 patients	
Walker *et al*. [48]	Own data	≤5 SNPs cluster, >12 SNPs no transmission	303 out of 609 (for the threshold)	All 5 major lineages
Walker *et al*. [96] (transmission)	Walker *et al*. [48]			
Walker *et al*. [96] (relapse versus re-infection)	Own data	SNP differences of 475, 1032, and 1096 between the original and relapse strain suggested that the patient had been secondarily infected with a different strain rather than within-host evolution.	Pulmonary vs extra pulmonary pairs from 49 patients and 110 longitudinal samples from 30 patients	All 5 major lineages
Witney *et al.* [94]	Walker *et al.* [48]			

Table 3.1 Use of single nucleotide polymorphism thresholds within papers reviewed

Mutation rates, which display the rate of genetic change of an organism, were also used to assess whether transmission was likely given the time between samples or how long ago sampling occurred (assuming that the mutation rate is constant over time) [92, 95]. In a study that included whole genome sequences from 1687 samples from patients in Malawi diagnosed between 1995 and 2010, Guerra-Assunção *et al.* [92] examined two things: construction of transmission networks amongst their clusters and analysis of risk factors for transmission. They defined their clusters using a threshold cut-off of 10 SNPs, determined by the distribution of SNP distances they found amongst all possible pairs of samples in the dataset. For their analysis of risk factors of transmission, cases with a SNP difference of 6-10 SNPs were classified as having 'uncertain' transmission links unless the mutation rate between them was less than 0·003 SNPs/day, determined by calculating the number of SNPs between two samples that shared the same RFLP type per number of days between the dates of their disease onset.

Other studies used the presence of certain insertions or deletions to help form subclusters and then assumed that no transmission could occur between patients that had samples in different subclusters [90, 91, 93]. Similarly, Gardy *et al*. [88] studied an outbreak in British Columbia, Canada amongst individuals predominantly with SRFs to determine the source case and explore the potential causes of the outbreak. To do so, they assembled a phylogenetic tree of their samples, which revealed two lineages. They then precluded transmission between patients with samples of different lineages in order to reduce the number of transmission pathways. In another study, transmission events between epidemiologically linked cases were excluded when the samples involved were not adjacent on the phylogenetic tree [45].

3.3.3 DIRECTION

As mentioned previously WGS can reveal variation between samples that are identical by other typing methods (such MIRU-VNTR) [111]. This higher resolution has been used by studies to help infer the direction of transmission between cases, i.e. distinguish between the infector and the infectee, using only WGS data. Approaches proposed include SNP accumulation, Bayesian statistical inference, and determining transmission networks alongside epidemiological data.

In a study of a TB outbreak in the Netherlands, Schürch *et al*. [44] examined transmission direction using the accumulation of SNPs between three sample sequences from epidemiologically linked patients, one sample from three separate patients. The method assumes that, over time, a strain will acquire new SNPs and retain existing ones, and direction of transmission is to the case with the additional SNPs. This approach has since been applied by others, combined with patients' TB histories and contact tracing data (Table 3.2) [90, 98, 101] to make it more robust, as this information can help exclude certain

transmission pathways. The other studies employing this method generally examined only a small number of samples and found small numbers of SNPs (eight SNPs amongst three samples [44], seven amongst nine [98], and two amongst 12 [101]). As this method requires manual examination of all the SNP sites in all of the sample genomes included in the study it is time and labour intensive and as such is only suitable for small studies with small SNP numbers.

Journal article	How was direction of transmission determined
Didelot *et al*. [99]	Epidemiological data and WGS used in a Bayesian inference framework to construct a transmission tree
Gardy *et al.* [88]	Social network analysis and contact tracing posed putative transmission, timing of infection and smear status was used to narrow down possible direction and WGS to remove transmission events involving cases with different lineages
Kato-Maeda *et al.* [98]	Contact tracing and accumulation of SNPs
Luo *et al.* [89]	Epidemiological links and timing of infection and symptoms helped propose direction of transmission between samples in the same WGS-based cluster. Transmission of mutant alleles from case with mixed base calls
Mehaffy *et al.* [90]	Genomic and epidemiological information (i.e. SNP pattern, contact information, year of diagnosis and infectiousness based on smear and chest x-ray results).
Pérez-Lago *et al.* [103]	In one case direction was proposed by the transmission of mutant alleles from a case with mixed base calls
Roetzer *et al.* [91]	Contact tracing revealed transmission chains and accumulation of variation is mentioned, although not clear if this resolved the order of the chain
Schürch *et al.* [44]	Accumulation of SNPs
Smit *et al.* [101]	Accumulation of SNPs and period of infectiousness

Table 3.2 Methods used by studies to determine direction of transmission

Statistical frameworks provide an altogether different approach by integrating WGS data with epidemiological information to estimate the probabilities of hypothesised transmission chains, rather than strictly define transmission events [82, 112]. A Bayesian framework of this kind was employed by Didelot *et al.* to infer transmission events and their direction from a phylogenetic tree, whilst taking into account within-host diversity [99], for a TB outbreak of 33 cases in British Columbia, Canada [88]. The study found that genomic data alone could produce transmission trees which capture the known epidemiology: the most likely source case and several key transmission clusters were correctly inferred by the method. Such an approach allows for identification of transmission events that a direct analysis from epidemiological and sequence data might not, as it assesses the probability of all possibilities, instead of immediately excluding some through clear cut thresholds etc. However, the fact that the Canadian outbreak was heavily sampled is a key assumption of the method which may make it impractical for other outbreaks where there is known to be many unsampled cases.

Alternatively, studies have used networks to visualise transmission using genomic data without including epidemiological data [100, 91, 89, 103]. These networks are generally created using algorithms that are run on the SNP distances between all the samples. Some examples are minimum spanning, neighbour joining or median joining networks. Three studies also created transmission networks but did include epidemiological data alongside the genomic data: Walker *et al.* [48, 96] used their own algorithm to create a similar network, which involved choosing the epidemiological links between cases that had the smallest SNP distance or shortest time between sampling; Schürch *et al.* [44] used temporal and contact tracing data to assign an index to SNP clusters and resolve a transmission network.

Inferring direction of transmission was rendered impossible between some samples in at least 18 of the 25 studies where some proportion of samples were identical, either because there was no diversity, or because they were unable to capture it due to stringent variant calling thresholds or stringent sampling/culturing methods.

3.3.4 RECURRENCES

Relapses and re-infections represent the two ways in which an individual can experience a recurrent episode of TB. The ability to distinguish the nature of recurrent TB disease can have important repercussions for public health and treatment and requires being able to compare the genomic data from the samples for the first and recurrent episode. In an ideal scenario, a relapse would be easily distinguishable from a re-infection as a relapse would be an infection with the same strain whereas a re-infection would be with a strain distinct from the first episode. However, it is possible to be re-infected with a genetically identical strain, especially in a highly endemic TB area, and these re-infections would likely be misinterpreted as relapses.

Several studies have investigated this area and have used SNP differences to attempt to distinguish between relapse and reinfection. Analyses of data from the REMox trial [102, 113] and the Karonga Prevention Study [79], found a bimodal distribution of pairwise SNP differences between longitudinal samples: 0-6 versus >1306 SNPs [102] and 0-8 versus >100 SNPs [79]. Given the very clear-cut bimodal results, both studies decided that the individuals with the smaller number of SNPs between samples were as a result of relapse, and those with a large number were a result of re-infection. Both found SNP distances larger than 1000 when they recovered different TB lineages from the two episodes, clearly representing re-infection. Influenced by this work, Guerra-Assunção *et al.* [92] used these results to classify recurrent cases of TB in their Malawian cohort, defining them as relapses if they differed by less than 10 SNPs from the initial strain. By comparison, in another study, Schürch *et al.* [44] classified a recurrent case as re-infection because the recurrent strain differed by one SNP from the initial infecting strain. This was determined because another individual in the outbreak, with whom there was an epidemiological link, had the same strain as the one determined to be the cause of the re-infection.

3.3.5 WITHIN-HOST DIVERSITY

Within-host diversity can arise via exposure to a mixed infection (a single infection event with multiple distinct strains), repeated infection events with distinct strains (i.e. superinfection) or microevolution (within-host evolution). Given that we are likely to just have one sample from one point in time, it is unlikely that we would be able to distinguish between superinfection and mixed infection, however using WGS, studies have attempted to distinguish between infection with multiple distinct strains and the microevolution of one infecting strain.

There were three approaches identified amongst the WGS studies for defining multiple co-infecting TB strains. Heterozygous base calls were one method some used to indicate the presence of two (or more) strains, by assuming each of the base calls was related to a different strain [48, 79, 89, 92, 94, 97, 98, 100, 102, 103]. However, the definition of a heterozygous base call has varied amongst the studies (Table 3.3), with some studies applying strict criteria on the minimum proportion of reads that must support the variant or the minimum number of heterozygous base calls there must be across the genome. As well, whether the presence of heterozygous base calls indicates microevolution or a mixed infection has been a source of discrepancy.

Another method used to identify patients with multiple infecting TB strains was through comparing multiple cross-sectional or longitudinal samples. Walker *et al.* [48] examined multiple cross-sectional or longitudinal samples of 79 patients and identified 3 patients as having samples differing by ≥400 SNPs with the remaining 76 having samples with 11 SNPs or less. These results led them to conclude that those with a small SNP distance between multiple samples were as a result of microevolution, whereas those with a large SNP

distance (>400) were classed as a mixed infection. If we compare this to the findings of Schürch *et al.* [44] where a single SNP difference between two samples from the same patient was classified as a re-infection, we may consider that by Walker *et al.*'s definitions such a small difference would be considered the result of microevolution.

The final method used to identify the presence of multiple co-infecting TB strains was the use of phylogenetic analysis. By constructing multiple phylogenetic trees with their samples, Gardy *et al.* [88] noticed that some of their sequences appeared to move between lineages within different tree constructions, which they believed signified that the sample had two different genomic signals corresponding to two different strains. This method was also employed by three other studies included in this review [90, 79, 102].

Journal article	Interpretation of heterozygous base calls	Definition of heterozygous base calls
Bryant *et al.* [102]	Mixed infection	Heterozygous base calls were identified if each allele was supported by at least 5% of reads (minimum read depth of four). The sites also had to satisfy quality and coverage conditions. Heterozygous sites were excluded if they were within 200 base pairs of other heterozygous sites to reduce the possibility of a mapping error. Samples were classified as a mixed infection in they had more than 80 heterozygous base calls
Guerra-Assunção *et al* [79]	Mixed infection	A position was classified as heterozygous if >1 allele accounted for ≥30% of the reads (and there were >30 reads). Samples were classified as a mixed infection if they had more than 140 heterozygous base calls
Kato-Maeda *et al.* [98]	Mixed infection	A heterozygous base call was reported when 38% of reads supported the variant, the remainder supported reference
Luo *et al* [89]	Microevolution	A position was classified as heterozygous if the variant ("less frequent allele") was present in at least five reads of high-quality and the overall coverage was at least 10
Pérez-Lago *et al.* [103]	Mixed infection	Heterozygous base calls were distinguished by the presence of a variant nucleotide ("less frequent nucleotide") in at least five reads
Walker *et al.* [48]	Microevolution	The presence of heterozygous base calls was seen as suggestive of "sub-populations" i.e. microevolution. No formal definition of heterozygous base calls was presented

Table 3.3 How included studies interpreted heterozygous base calls for classifying diversity

The studies were assessed for whether they accounted for diversity when investigating transmission. Through collecting multiple cross-sectional and longitudinal patient samples, Pérez-Lago and colleagues [103] were able to refine their transmission network. They did this by first building within-host networks for each patient from the multiple samples which depicted how the bacterium had evolved within the patients. Then when linking patients in transmission they were able to compare each of the multiple samples (instead of just one supposedly representative sample) to see if any reasonably suggested transmission. In a similar, but more simplified version, Walker *et al.* [48] refined their transmission networks by comparing multiple samples per patient and only including the sample that minimised the SNP distances in transmission chains.

Within their cohort, Kato-Maeda *et al.* [98] identified one patient sample containing a 'mixed population' of TB strains. On investigation, the two TB strains within the sample seemed to be present in two other earlier diagnosed patients' samples. Given the timing restrictions, the authors concluded that the most likely explanation was that one of the patients who had been diagnosed earlier was infected by both strains (one not being present in their sample) and had transmitted the disease to both of the other patients. By investigating the diversity of multiple TB strains within patients they were able to better resolve the picture of transmission.

Heterozygous base calls were also used by studies to help elucidate transmission events [89, 103]. Where one sequence has a heterozygous base call in one position and another has the same or similar sequence but either the reference or alternative allele is fixed at the heterozygous base, this can suggest transmission between them and even indicate direction of transmission. One explanation is that there may have been transmission from a source patient with the reference allele followed by microevolution in the infected case giving rise to an alternative allele or microevolution in the source patient giving rise to an alternative allele followed by transmission where the alternative allele becomes fixed.

Another method for estimating the probability of transmission between patients is to use knowledge of the within-host mutation rate to assess the likelihood of the amount of genomic differences found between their samples given the sampling dates. A study by Colangeli *et al.* [114] found that TB mutates less in its latent state than during active disease. Considering this finding, Mehaffy *et al.* [90] concluded that the presence of zero SNPs between multiple clusters of patients infected over the course of more than five years was an indication that the patients must have reactivated TB i.e. a long latency period followed their infection by one common source during which there was no mutation. The limitation to this method, however, is the debate around the true value of the mutation rate of TB and how this may vary over time in one patient versus during transmission chains: using longitudinal patient data, Walker *et al.* [48] found the within-host mutation rate to be lower than mutation rate during household outbreaks (0.3 vs 0.6 SNPs/genome/year respectively; conversely, Guerra-Assunção *et al.* [92] found that the within-host mutation rate was higher than

between linked pairs in their transmission networks (0.45 vs 0.26 SNPs/genome/year respectively).

3.3.6 DRUG RESISTANCE

The primary purpose of investigating the emergence of drug resistance is to determine its nature i.e. is a resistant strain being transmitted (primary resistance), requiring better transmission control; or is resistance arising separately within individuals (secondary or acquired resistance), suggesting poor treatment adherence, inadequate dosing, or individual variability in drug metabolism. Six studies included in this review investigated drug resistance using WGS.

One method employed by two studies involved constructing phylogenetic trees and examining the presence of drug-resistant strains within a clade [105, 106]. The conclusion of transmission of a drug-resistant strain was determined only if all samples in a clade had the same resistance-conferring mutation (i.e. the resistance was gained by an ancestor of the clade); otherwise drug resistance was considered to have been as a result of secondary resistance. Patterns of drug-resistance-conferring mutations were also used to investigate likely transmission patterns: in one clade, isoniazid and rifampicin resistance was common amongst all samples (suggesting transmission of a MDR strain) but resistance to fluoroquinolones was not, suggesting that the acquisition of fluoroquinolone resistance was secondary within those samples [105].

In contrast, Clark *et al.* [97] allowed for the conclusion of transmission amongst a proportion of samples within a clade, even when not all samples within a clade had the same resistance mutation. For example, in one clade of three samples, two had the same resistance mutations for isoniazid and rifampicin and were therefore considered a transmission pair, whilst the third was not considered to be in the same chain of transmission as the other two. Similarly, Casali *et al.* [104], in their study of 1,000 samples from Russia applied a principle of assuming acquired resistance when only one case in a phylogenetic clade had a certain resistance-conferring mutation.

Phylogenetic trees were not used for the examination of drug resistance in the other two studies, instead resistance-conferring mutations alone were examined. In a study of eight samples from an outbreak in Haiti, six of the eight samples were available for WGS. These six samples had the same *rpoB* and *katG* resistance mutations for rifampicin and isoniazid resistance and thus Ocheretina and colleagues [108] concluded that the outbreak represented primary resistance. Although no mention was made of how common these resistance mutations were in the setting. In a similar approach, Regmi *et al.* [107] investigated an MDR-TB outbreak in Thailand, however only four of the 54 samples were examined for resistance mutations.

3.4 DISCUSSION

3.4.1 MAIN FINDINGS: IMPLICATIONS OF ANALYTIC APPROACHES ON WGS INFERENCES

This review has identified a range of analytical approaches for interpreting WGS data to understand different aspects of TB transmission. Chief amongst them have been SNP thresholds, calculating mutation rates and phylogenetic methods. The impacts of these methods are now discussed.

A common approach amongst WGS studies to define transmission, distinguish relapse and re-infection, and distinguish microevolution and mixed infections was to use SNP thresholds. Whilst SNP thresholds are an attractive method due to their extreme simplicity, this is also one of their drawbacks: they cannot be applied to all situations. Many studies vary in their determination of a SNP, due to factors such as the definition of a 'quality' read and the number of amplification steps [99, 45, 96, 103, 112], and as such different genomics pipelines may uncover a different number of SNPs even within the same cohort. This can then dramatically affect the chosen number for a SNP cut-off. This may also partially explain the conflicting results in the number of SNP differences found between linked cases: three studies found epidemiological links between cases with larger than 12 SNP differences [45, 96, 103], the number considered by Walker *et al.* to be the cut-off for transmission.

A preferable alternative would be to consider the mutation rate and the time between the sampling dates to decide whether the number of SNP differences could have plausibly occurred during such time and therefore make transmission probable. This would then allow for transmission events where a large number of SNPs has accumulated over a long time period when they may otherwise be erroneously excluded by a strict threshold. One such alternative is an approximation to the pairwise distribution of genetic distances [115]. Despite good agreement for the mutation rate of *M. tb* across epidemiological studies [91, 45, 92, 48], it would be important to consider the presence of factors which may affect the mutation rate, such as drug resistance [45, 90].

A more rigorous method for including the mutation rate in a probabilistic framework is through the use of Bayesian inference [99]. An additional advantage of this method is that it can infer unsampled cases and also include epidemiological data, such as smear positivity and individuals' locations, which can help strengthen inferences of transmission.

One method used to determine the direction of transmission was examining the process of accumulating SNPs across patient genomes. This method is likely more suitable when there are few SNPs between samples, ideally one. A need for such small numbers of SNPs therefore highlights the need to minimise sequencing errors, as their presence, if misinterpreted as SNPs, can have a big impact on this type of inference [116]. On the other

hand, limited genomic variation i.e. no SNPs between samples due to the slow *M.tb* mutation rate [101, 90, 117], would also hinder this method. In the case of determining direction, if transmission between a pair has already been identified, information on timing of exposure and infectiousness for contacts is likely to be the best aide [101, 89].

Methods used for distinguishing between primary and secondary drug resistance have mostly relied on generating phylogenetic trees and examining whether or not clades shared the same resistance-conferring mutation [105, 106]. A limitation with this method lies in the fact that phylogenetic trees are not equivalent to transmission trees and therefore cases linked in transmission may not always be paired together in the tree. In addition, should resistance arise in the middle of a sequential transmission chain and many of the susceptible ancestral cases are sampled along with the resistant descendant cases, these may indeed still cluster together in the phylogenetic tree. However, in the case of some of these approaches, primary resistance for the resistant descendant would not be acknowledged. A perhaps more effective approach would be to first identify a transmission network amongst the cases, from there the data on drug resistance can be examined to identify if the drug resistance seen in a case is present in the one transmitting the infection and therefore whether it is being transmitted versus acquired.

Within-host diversity was investigated through the placing of samples in the phylogenetic tree and, primarily, through the examination of heterozygous base calls. The use of heterozygous base calls to distinguish microevolution and mixed infections represents the most discriminatory method, however, similarly to SNP thresholds, the studies used a variety of thresholds for the number of calls used to categorise mixed infections and often these were arbitrarily defined. One key issue with using a phylogenetic tree to identify mixed infections lies in the possibility of co-infection with strains of the same lineage [79]; this would not produce the effect of samples switching between lineages of the phylogenetic tree and therefore such types of mixed infections would go unnoticed [88].

Ideally, more diversity needs to be captured through sampling, as often only one sputum sample is available and this may not capture the pockets of different strains of TB that may be residing in the lungs [118, 119]. In this case, multiple samples from different timepoints could be a potential solution [76, 120]. Once samples are taken, the process of subculturing and selecting single colonies for sequencing may affect the amount of diversity found [121, 122, 123], although methods exist for sequencing from primary culture plates [124]. Methods for identifying more diversity at the sequencing level involve deep sequencing and examining minor variants [125].

3.4.2 STRENGTHS AND LIMITATIONS

A strength of the review is the use of a comprehensive search strategy in multiple databases without language restrictions, meaning that a wide range of studies were found with a

smaller risk of bias. Double screening was employed and is considered a strength as it helps ensure that the included studies are reflective of what was required for the review [126]. This is augmented by the fact that there was good inter-reviewer agreement. As well, the identified studies were generally good quality (Table 9.9).

A limitation is that due to the nature of the investigation, the studies included are all observational and thus a more rigorous analysis of the data, such as meta-analysis, is not tractable.

3.5 CONCLUSION

This review has served to summarise the approaches that have been employed to analyse WGS for investigating TB as well as highlight the limitations and complications of these approaches.

Several conclusions can be drawn from this review: firstly, SNP thresholds have a wide range of applications, but their application to confirming transmission should be considered in light of local TB incidence, strain diversity, the time between the samples, potential hitchhiking, homoplasy and more. Consideration of factors that affect mutation rates is essential. Secondly, epidemiological data and clinical history remain critical to outbreak investigations, especially when sequence data lacks variation. Finally, knowing how diversity arises will help resolve transmission. Better characterisation of microevolution and mixed infection will require better sampling, deeper sequencing and investigation of the within-host mutation rate.

4 BAYESIAN INFERENCE USED TO UNCOVER INFECTION DATES FOR PUBLIC HEALTH PURPOSES (OBJECTIVE 2)

In this chapter, using a dataset from a TB outbreak in British Columbia, Canada, an existing Bayesian inference method used to infer transmission networks from epidemiological and genomic data is adapted to improve its use for studying TB, namely being able to infer infection dates. From the resulting transmission network, the aim was to identify how long ago the last transmission event occurred within this outbreak and to use that information to determine whether the outbreak could be declared over. The following chapter is based on a paper by Hatherell *et al* [127].

4.1 INTRODUCTION

As described in Chapter 3, being able to infer when infection occurred for a TB case is an extremely difficult task given the long and variable latency period [26]. However, such knowledge is critical for deciding whether new cases are a result of recent infection or reactivation; it is generally considered that an individual who presents with active TB within two years of being infected is as a result of recent infection, any longer and the case is considered a result of reactivation [128]. This differentiation has a knock-on effect for public health control of an outbreak: if there are cases arising due to recent infection it is important to look for cases of active TB in the community and perform contact tracing to ensure we can find the source of transmission and forestall the outbreak by quickly preventing any already latently infected individuals from progressing to active disease. If cases are being identified as a result of reactivation then interventions may focus more on improvement of treatment or identifying further latently infected individuals, for example through community-wide LTBI screening [129]. In addition, public health practitioners consider an outbreak ‘over’ if the last transmission event was more than two years ago, i.e. no new cases have been identified for two years.

One way to identify when infection may have occurred is to identify a possible infector and determine when their infectious period was. Potential infectors would be contacts (friends, family, colleagues) that have been diagnosed with TB, however there are numerous issues that may occur when trying to identify the potential timing of the transmission event in this way: individuals may not identify all their possible contacts or contacts diagnosed with TB may not remember when their symptoms started (start of infectious period). In addition, there is the possibility that multiple contacts may have been diagnosed with TB and there are multiple different transmission pathways, especially if the individuals all live in a tight-knit

community and there is a high-incidence of disease, meaning we would not be able to definitively identify the correct infector.

As a potential solution to the problem of multiple possible infectors, genomic data provides extra discriminatory information that helps reduce the possible number of transmission pathways and thus narrow down the possibility of who infected whom and therefore when infection occurred. Examining an outbreak of Foot and Mouth Disease in the UK in 2001, Cottam *et al* [130] were one of the first to publish a method that used genomic data to determine an infectious disease transmission network, by creating a phylogeny and then determining the ML of each possible tree in relation to the known data of infection dates and infectious periods for the infected farms. Using the same outbreak, Ypma *et al* [112] and Morelli *et al* [131] then built on this method. However, these early studies were not required to infer infection date as Foot and Mouth disease has short, well-defined infectious and latent periods and thus the date of infection can be narrowed down to a range of a few days merely from knowing the date of symptom onset and subtracting the latent period. Subsequently, Jombart *et al* [82] used a Bayesian method to infer infection dates for 2003 SARS outbreak in Singapore, however it does not consider within-host diversity, thereby making it unsuitable for studying TB.

In 2013, Didelot *et al* [99] extended the above by developing a Bayesian statistical inference framework, TransPhylo, which takes into account within-host diversity and thus can be used for TB. The framework accounts for the fact that the bacterium present in each host can evolve from the one that seeded the infection and diverge into multiple distinct strains. Each strain can generate onward infections resulting in divergent strain profiles in subsequent cases. This diversity is important to account for if the method is to be used for TB as, even though it is generally believed that *M. tb.* evolves at a slow rate, there has been evidence of large genomic diversity [132], which may be going undetected with current sampling methods. Didelot *et al.*'s method was applied to the British Columbia data concerning an outbreak amongst 52 individuals presented by Gardy *et al.* [88] to demonstrate the uses of this approach versus the original analysis, which relied more heavily on a social network analysis to construct a transmission network. The re-analysis of the data looked to demonstrate whether the method captured known transmission events and identified the source case in order to test its validity.

The premise behind the method is to be able to draw information about a transmission network from a phylogenetic tree. Unfortunately, phylogenies cannot be translated directly into transmission networks, as even if two sequences are paired on a phylogenetic tree and we therefore assume they are linked in transmission, there is no way to determine the direction of transmission, a key aspect of transmission networks. Additionally, we expect that internal branching events represent ancestors of the samples at the tips, however in an outbreak where infectors and their infectees are sampled all cases will appear as tips of the phylogenetic tree despite some being ancestors of others and so internal branching events

are also associated with sampled hosts [133]. In order to change a phylogenetic tree into a transmission network we need to know where on the tree transmission happened and between whom. To do this we need the phylogenetic tree to be on a real-time scale, i.e. lengths of branches correspond to length of time, not on a genetic scale. These timed trees can be created in BEAST [134] using information on sampling time and the molecular clock of the bacterium. The latter demonstrates the timescale for evolution in the bacterium.

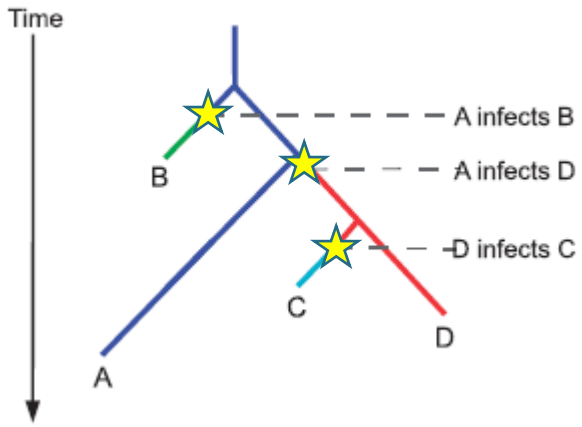

Figure 4.1 Phylogenetic tree with transmission network colouring to demonstrate statistical inference output (Adapted from Didelot *et al*. [99])
The root of the tree is coloured dark blue which corresponds to patient A (the dark blue tip). The dark blue colour changes to green and red at the yellow stars, denoting a transmission event. This represents A infecting B and D. Red then changes to light blue, denoting D infecting C. Time is increasing down the tree: the first colour change/yellow star is when A infects B, thus this happens earlier than the other transmission events. Similarly, the green tip terminates earlier than the others, denoting B being sampled earlier than the others

To depict the idea of a transmission network on a phylogenetic tree, Didelot *et al* used a colouring approach (Figure 4.1). Each tip (end of branch) belongs to a host and is a unique colour; a star denotes a transmission event where the first colour infects the second colour, going forward in time. The colour of the source case can be traced back to the root of the tree as the lineage was present in that host in the beginning. Because the tree is timed and the time of the tip denotes when the case was sampled and presumably treated and made non-infectious, a host's colour cannot appear in the tree past their sampling date.

An issue with the TransPhylo method was the assumption that the generation time (the time between being infected and infecting others) and sampling time (the time between being infected and sampled) were exponentially distributed, a consequence of using a common compartmental transmission model comprising of individuals in SIR disease stages to describe the underlying infectious disease dynamics [135]. An exponentially distributed generation time then suggests that the time from being infected with TB to becoming infectious is also exponentially distributed, which is widely considered not to be true for TB;

SIR models are considered a poor fit for TB modelling [136]. The effect of this assumption could be that the infection times inferred by TransPhylo are incorrect. As this was the key output of the model for being able to declare the outbreak over, a more suitable method to mitigate this was then explored within this work.

In order to test the changes, we revisited the Kelowna outbreak data analysed by Didelot *et al*. to identify any new insights the modified inference framework might bring with regards to the inferred infection dates of the cases and as a result answer the question of when the method postulates the date of the last transmission event, in order to assist public health measures in British Columbia, Canada. Given the new changes made, we also wanted to investigate the role of prior choice and phylogenetic uncertainty in the inference of trees and whether they affected certain characteristics of the outbreak, such as posterior generation time and sampling distributions, and the timing of the last transmission event.

The chapter is structured as follows: in Section 4.2, the outbreak is briefly introduced along with some description of the available data. Section 4.3 includes the basics of the original method and the mathematics of the extension to include the branching model. Section 4.4 presents both the analysis of the newly inferred infection dates of the cases and the impact of prior sampling and generation time distribution and phylogenetic uncertainty. Finally, section 4.5 contains a discussion of how the results can be used in a public health context to combat an outbreak and avenues for further research.

4.2 OUTBREAK AND DATA

In 2006, an outbreak of drug-sensitive tuberculosis was discovered in British Columbia, Canada [137]. The outbreak, uncovered using MIRU-VNTR typing, was relatively small, with just 41 cases in three years [88]. However, with such a low rate of background TB (the outbreak cases increased the annual incidence rate for the region by a factor of more than 10) even such a small outbreak was a big concern for the Canadian territory.

A social network questionnaire was carried out by the local health authorities to help elucidate a transmission network, aid case finding and identify who may have been the source case. The results of the questionnaire revealed that a large proportion of the outbreak population were without a fixed abode, using crack cocaine and abusing alcohol.

There was no available data on uptake of treatment or treatment delay amongst the cases; however, 85% of the first 41 cases were recorded as cured or finishing treatment.

Overall, there were 48 out of 52 cases that had samples available for sequencing (four were diagnosed out of the province or via post-mortem). The last reported case believed to be part of the outbreak occurred in July 2013. The outbreak curve can be seen in Figure 4.2

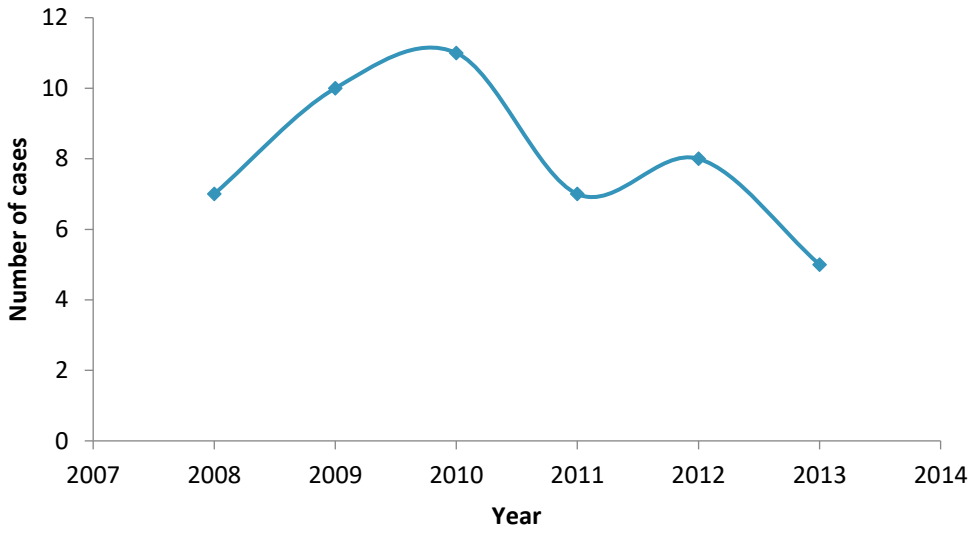

Figure 4.2 Incidence curve for Canadian outbreak

The samples were sequenced using the Illumina MiSeq platform, and the short read data for the 48 genomes is accessible via the European Nucleotide Archive; accession number: PRJEB12764 (http://www.ebi.ac.uk/ena/data/view/PRJEB12764).

4.3 METHODS

The method used for the Bayesian inference is an adaptation of the one presented by Didelot *et al*. [99]. Here a branching process is used to model the epidemic process as opposed to a stochastic, continuous time Markov chain version of the general SIR epidemic model. This circumvents the assumption of an exponentially distributed generation time by allowing the generation time distribution and the sampling distribution to be specified as a prior in the inference ensuring that the potential error introduced by the exponential distribution assumption is avoided.

To mitigate this, a branching model was used to model the number of secondary infections caused by an infected individual as Poisson-distributed with the mean connected to the generation time and the individual's infectious period.

In this section, I describe the original Bayesian method as well as the alterations made, the methods used to examine the output of the Bayesian inference, namely the timing of the last transmission event, and how we examined the effect of the generation time and sampling distribution priors on the Bayesian output.

4.3.1.1 BRANCHING PROCESS

Branching processes, also known as the Galton-Watson process, are a type of stochastic process that are normally used to describe some kind of reproduction, i.e. there are

generations of individuals and each individual in a generation produces some random number of individuals in the next generation according to a fixed probability distribution (the offspring distribution). They can generally be depicted as a tree (Figure 4.3).

In TransPhylo, the population being described by the branching process is that of individuals infected with TB, where the individuals in generation *n+1* are those that have been infected by the individuals of generation *n.*

Using a branching process allows:

- Calculation of the estimated size of generation *n,* according to the mean number of offspring each individual in a generation has, μ.
- Specification of an offspring distribution.

4.3.1.2 MODEL DESCRIPTION

The inference works by running a Monte Carlo Markov chain (MCMC) to sample a space of transmission trees. For each transmission tree proposed, the algorithm evaluates its likelihood (as specified later), uses a Metropolis-Hastings accept/reject step to determine if the current state updates or not and then proposes a new tree and so on and so forth. The chain is run until it is considered to have converged. The final set of sampled trees is then used for inference analysis e.g. finding the most sampled tree (i.e. the one with the highest likelihood), credible intervals and summary statistics. As mentioned earlier, the transmission tree is considered as a "colouring" of the phylogenetic tree on which the inference is being performed and the new proposed transmission tree is a change in the colouring, i.e. a change in the placement of the transmission events on the tree.

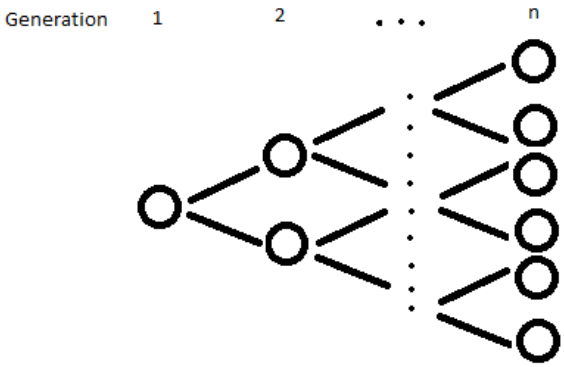

Figure 4.3 Branching process where the average number of offspring, μ, is 2.

The probability which is evaluated for each proposed transmission tree is the posterior, which is the probability of the transmission tree given the data we have i.e. the phylogenetic tree. This can be formulated using Bayes' theorem i.e. the probability of a transmission tree T given the phylogeny G is expressed as

$$P(T, \varepsilon, N_e g|G) \propto P(G|T, \varepsilon, N_e g)P(T, \varepsilon, N_e g) \quad (4.1)$$

where ε represents the parameters of the epidemiological model used to describe the outbreak and $N_e g$ represents the within-host parameter, which defines how much the bacterium is mutating, where N_e is the effective population size and g is generation time.

What is therefore required in order to calculate the posterior is the likelihood (the probability of the phylogeny given a transmission tree $P(G|T, \varepsilon, N_e g)$), and the priors for the transmission tree, epidemiological parameters, and within-host diversity parameter ($P(T, \varepsilon, N_e g)$).

Equation 4.1 can then be decomposed into:

$$P(T, \varepsilon, Neg|G) \propto P(G|T, Neg)P(T|\varepsilon)\pi(\varepsilon, Neg) \quad (4.2)$$

as G is conditional on ε through T and where π represents the priors for ε and Neg. The likelihood, $P(G|T, N_e g)$, is unchanged from its original description [99].

The second term on the right-hand side of Equation 4.2 represents the probability of the transmission tree given the branching model. We specify this by considering the probability of i having N_i infections (Poisson offspring distribution), the probability of i being sampled at t^i_{samp} (sampling distribution), and the probability of i infecting each offspring j at t^j_{inf} (generation time distribution).

By assuming that the secondary infections of an individual occur as Poisson process, we can describe the expected number of secondary infections, N_i, seeded by individual i as

$$N_i = \int_0^{t^i_{samp} - t^i_{inf}} R_o \, f_g(\tau) d\tau \quad (4.3)$$

$$= \frac{R_0}{\Gamma\left(k_g\right)} \gamma\left(k_g, \frac{t^i_{samp} - t^i_{inf}}{\theta_g}\right)$$

This is dependent on how long the individual is infectious for and their infectivity which is related to the generation time f_g. Equation 4.3 then is the mean of the offspring distribution f_o [138].

We then consider the probability of T to be

$$P(T|\varepsilon) = \prod_{i=1}^{n} f_o(N_i) f_s(t^i_{samp} - t^i_{inf}) \prod_{j=1}^{N_i} \frac{f_g\left(t^j_{inf} - t^i_{inf}\right)}{F_g\left(t^i_{samp} - t^i_{inf}\right)} \quad (4.4)$$

where f_s represents the probability density function for the sampling distribution and F_g is the cumulative distribution function for the generation time distribution. The sampling distribution

was also set to be Gamma. Equation 4.3 describes the probability for each individual $i = 1, \dots, n$ in the tree of having N_i offspring in total, specifically infecting offspring j at t_{inf}^{j} and being sampled at t_{samp}^{i} conditional on when they were infected and sampled. Thus whereas before the transmission tree prior depended on the transmission parameters of the SIR model (α and β, infection and removal rate respectively), it now depends on the parameters that determine the sampling and generation time distributions, namely k_g, θ_g, k_s and θ_s; the shapes and scales of the Gamma distributions, as well as R_0.

A Gamma distribution was chosen for the generation time distribution as it is positively skewed and more centred around the mean than an exponential distribution but also it has a long tail which therefore allows for cases of long latency and reactivation. In their attempt to characterise the incubation period or latency period of TB (defined as the time between infection and active disease), Borgdorff *et al* [139] found that there was a skewed distribution with a median of 1.26 years, and range of 0-12.8 years. Although the generation time is not the same as the incubation/latency period, the generation time will follow a similar curve as it will have to include the incubation period but also a lag between becoming infectious and infecting someone, which is assumed to be constant.

The full description of the terms is:

$$f_o(N_i) = \frac{1}{N_i!}\left(\frac{R_0}{\Gamma(k_g)}\gamma\left(k_g, \frac{t_{samp}^{i} - t_{inf}^{i}}{\theta_g}\right)\right)^{N_i} e^{-\frac{R_0}{\Gamma(k_g)}\gamma\left(k_g, \frac{t_{samp}^{i} - t_{inf}^{i}}{\theta_g}\right)} \quad (4.5)$$

$$f_s\left(t_{samp}^{i} - t_{inf}^{i}\right) = \frac{1}{\Gamma(k_s)\theta_s^{k_s}}\left(t_{samp}^{i} - t_{inf}^{i}\right)^{k_s - 1} e^{-\frac{t_{samp}^{i} - t_{inf}^{i}}{\theta_s}} \quad (4.6)$$

$$\frac{f_g\left(t_{samp}^{j} - t_{inf}^{i}\right)}{F_g\left(t_{samp}^{i} - t_{inf}^{i}\right)} = \frac{\frac{1}{\theta_g^{k_g}}\left(t_{samp}^{j} - t_{inf}^{i}\right)^{kg-1} e^{-\frac{t_{samp}^{j} - t_{inf}^{i}}{\theta_g}}}{\gamma\left(k_g, \frac{t_{samp}^{i} - t_{inf}^{i}}{\theta_g}\right)} \quad (4.7)$$

4.3.2 BIOINFORMATICS

The bioinformatics pipeline used to analyse the raw genomic data and produce the FASTA files for constructing phylogenetic trees is described in [127].

4.3.3 SETTINGS

4.3.3.1 BEAST

Phylogenetic trees were produced in BEAST, initially without constraining the molecular clock rate. The resulting trees were showing a root around 1997, which would have suggested that the index case was infected then. This was considered too early given that the inferred index case was likely infected between June and October 2007 when he was visiting Vancouver, as the circulating TB genomes from the Vancouver area had the same MIRU-VNTR type as the Kelowna outbreak strain and was likely the ancestor [88]. Once the fixed clock rate was constrained to roughly $1e^{7}$ per site per year the root was around 2005 (2003.5-2007, 95% confidence interval), which seemed a better fit.

Trees were then produced using a fixed clock model set to $1e^{-7}$ per site per year (≈ 0.44 mutations per genome per year), roughly the same posterior mean produced in BEAST by Didelot *et al*, and constant population size model [140]. We considered 100 trees randomly sampled from the latter half of a chain of 10,000,000 sampled trees, in order to allow for burn-in.

4.3.3.2 TRANSPHYLO

Fixed parameters (i.e. parameters that are not sampled by the MCMC) for the model within the MCMC inference are: $R_0 = 1.5, k_s = 1, \theta_s = 2, k_g = 2, \theta_g = 1$.

The code for the modified version entails one modified program (probTTree.m) found at https://github.com/holliehatherell/Thesis/ModifiedTransphylo with the remainder of the code available at https://github.com/xavierdidelot/TransPhyloMatlab (substituting the modified version of probTTree).

The MCMC was run using the maximum clade credibility (MCC) tree from the posterior phylogenetic trees from BEAST. The MCC tree is one tree picked from the posterior sample, chosen because it has the highest product of all its clade probabilities, where each clade (see Figure 4.4) in each tree is given a probability based on what proportion of the posterior trees contain the same clade (the more times the clade appears, the higher the score). Thus, it is a good point estimate in terms of topology for the posterior set of trees. The MCMC was run for 250,000 iterations, which was thinned so only every 250th tree from the latter half of the posterior was used for analysis, resulting in 501 transmission trees.

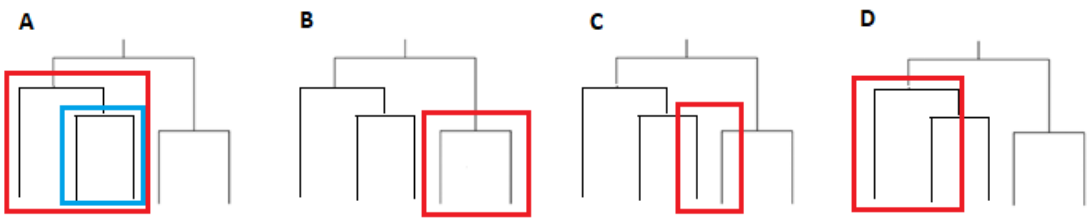

Figure 4.4 Graphical presentation of a clade within a phylogenetic tree, defined as a group of branches containing an ancestor and all its descendants

The blue box within the red box on tree A highlights a clade within a clade. The red box on tree B shows a clade, whereas those highlighted in trees C and D are not.

4.3.4 EXAMINING THE TRANSMISSION TREES FROM A PUBLIC HEALTH VIEW

In the tree object produced by TransPhylo, there is one column of infection dates, another column of sampled dates and a third column of the index of the infector. For example, consider a phylogenetic tree that describes case 1 infected case 3 who in turn infected case 2, this would be described in a tree object as in Table 4.1.

Infection date	Sample date	Infector
$\boldsymbol{inf_1}$	$samp_1$	
$\boldsymbol{inf_2}$	$samp_2$	3
$\boldsymbol{inf_3}$	$samp_3$	1

Table 4.1 Example of a tree object in TransPhylo displayed as a table.
The first row is for case 1; as they are the index case in the outbreak their infector is left blank. Case 2 was infected by case 3 so 3 is listed in the infector column for row 2 and case 3 was infected by case 1 so 1 is recorded in the infector column for row 3.

The date of the last transmission event, i.e. the latest infection date, was recorded as the latest date from column 1 of the tree object from the MCMC consensus transmission tree, a tree produced by TransPhylo which only includes transmission pairs above a certain probability. To determine whether phylogeny had any effect on the timing of the last transmission event, the date was recorded in a vector from over 501 transmission trees sampled from the MCC tree.

4.3.5 EXAMINING THE EFFECT OF PRIORS

In order to examine the effect of the prior generation time and sampling distributions on the posterior generation time and sampling distributions, a sensitivity analysis was carried out using a one-at-a-time method on the shape and scale of the Gamma distributions. To do this, the shape and the scale parameter were varied separately and then the MCMC was re-run on the MCC tree and for each infector-infectee pair inferred in the transmission trees the time was recorded between their respective infection dates as well as the time between each case's infection and sample dates.

The effect of phylogenetic uncertainty on the time of the last transmission event was also examined by running TransPhylo on a posterior sample of 100 BEAST trees. The timing of the last transmission event was recorded as the timing of the latest infection date in the MCC transmission tree, as mentioned in Section 4.3.4.

4.4 RESULTS

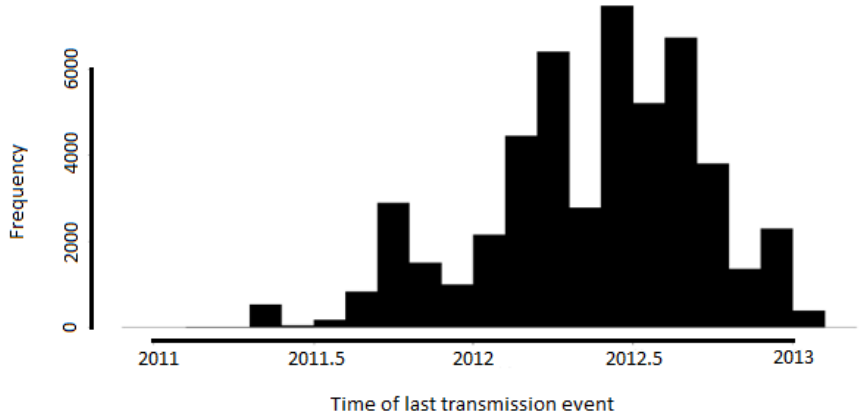

Figure 4.5 Date of last transmission event taken from 50100 transmission trees (501 posterior trees for each of 100 different phylogenetic trees)

4.4.1 OUTBREAK ANALYSIS

Examining the infection time for each of the 48 sequenced cases over 50100 transmission trees (501 transmission trees for each of 100 phylogenetic trees) revealed that the last person-to-person transmission most likely occurred in late July or early August 2012 (Figure 4.5), with 85% of the trees having a last transmission event in 2012. The point estimate from the MCMC consensus tree (run on the MCC tree) was 2012.581, i.e. 0.581 of the year into 2012, roughly 30 weeks (≈ 0.581 x 52 weeks).

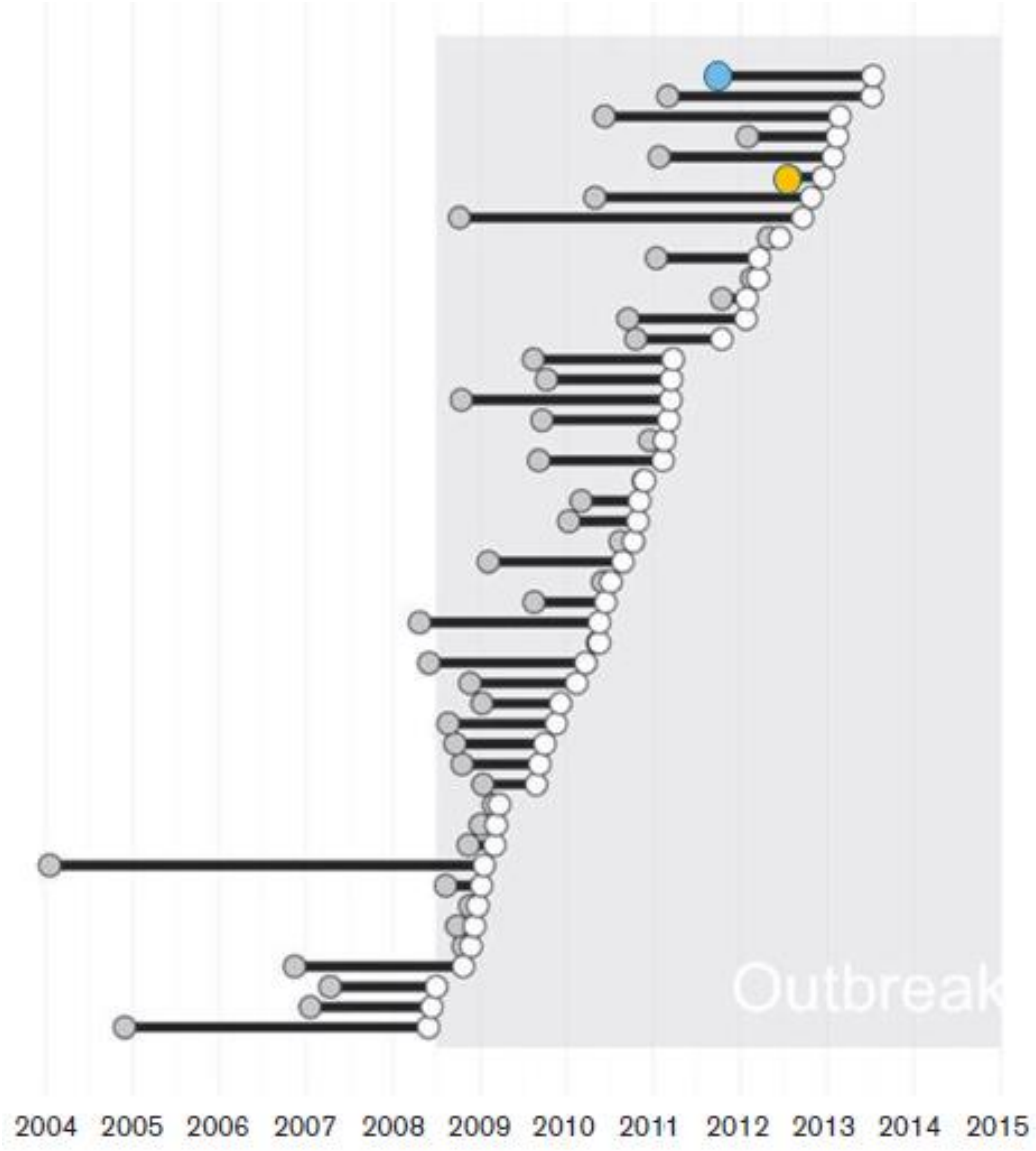

Figure 4.6 Inferred infection and sampling times for individuals in the Canadian outbreak

Adapted from Hatherell *et al*. [146]. Cases from bottom to top are in order of diagnosis date (white circle), earliest to latest. Infection date inferred from TransPhylo depicted as a grey circle. The last transmission event (latest infection date) is denoted by a yellow circle. The infection date of the most recently diagnosed case is denoted by a blue circle

Seven cases were identified as having been diagnosed after the last transmission event with

a range of late 2012 to mid-2013 (Figure 4.6). They are assumed to be instances of delayed progression as most of the transmission events leading to these cases occurred in 2010–2011, meaning their latency period was longer than 2 years.

4.4.2 MATHEMATICAL ANALYSIS OF GENERATION AND SAMPLING TIME DISTRIBUTION

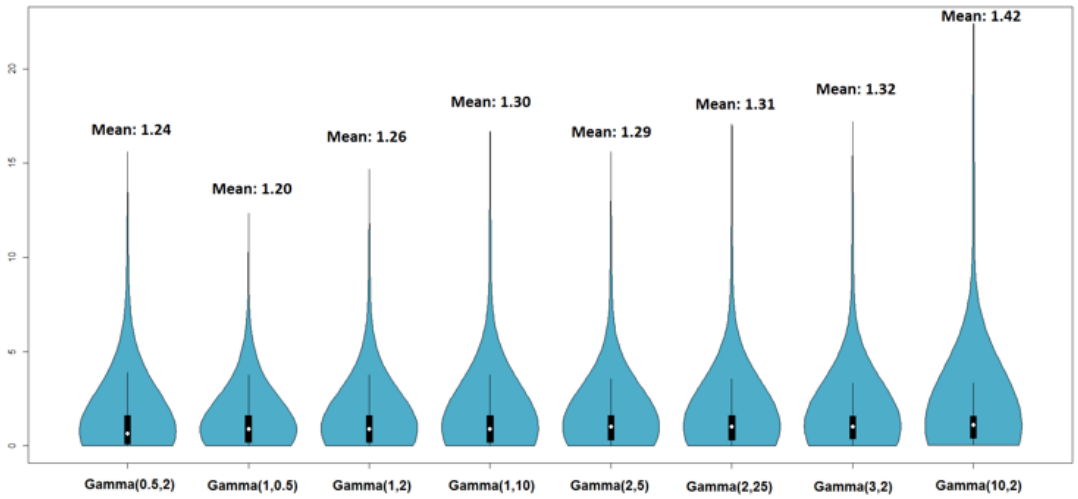

Figure 4.7 Violin plots showing posterior sampling distributions

Violin plots showing the posterior sampling distributions taken from the posterior transmission trees (time between infection and sampling for all cases) for different prior sampling distributions. The boxplots shown within the violin plots represent the prior distributions and the mean posterior sampling time is shown above each violin plot

As the generation time and sampling distribution are new parameters to the model, it is important to assess the robustness of the model posterior to the prior values. This was done by choosing different prior distributions (different scales and shape values) and plotting the resulting posterior distribution.

Figure 4.7 demonstrates the robustness of the sampling distribution to the prior chosen; over a wide range of shapes and scales for the prior sampling distribution, the posterior distribution remains right skewed and the scale is not much changed. The mean of the posterior sampling distribution is determined to be roughly 1.3 years with a maximum sampling time roughly between 17-19 years.

The generation time distribution is robust within a range of values but as the scale increases the mean generation time increases (Figure 4.8). However, the distribution remains positively skewed throughout.

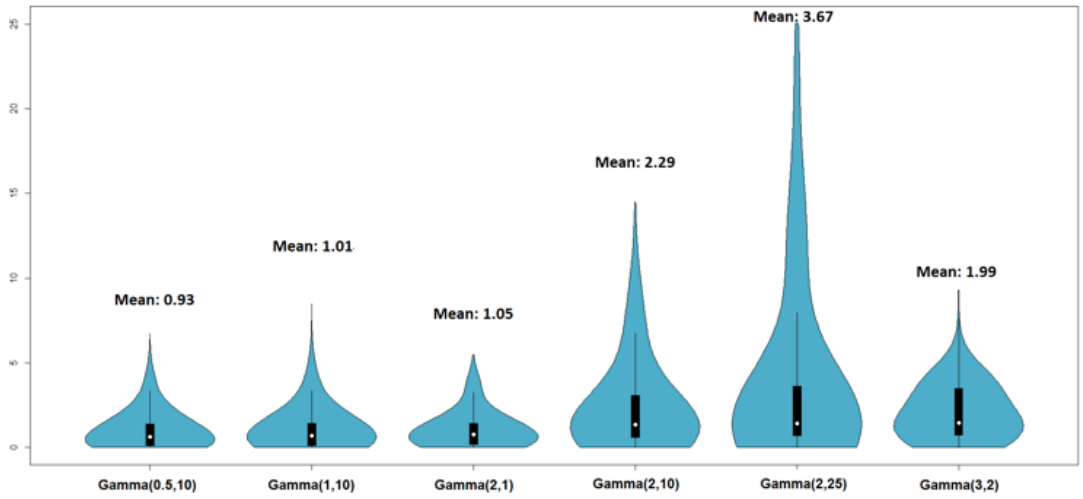

Figure 4.8 Violin plots of posterior generation time distributions
Posterior generation time distribution for different generation time prior distributions (shown as boxplots within the violin plots) and mean posterior generation time shown above each violin plot.

4.5 INFERENCES AND DISCUSSION

4.5.1 KEY FINDINGS

In this work, I aimed to reanalyse WGS data from a TB outbreak with an improved statistical inference method and interpret the timings of infection from a public health perspective. The improvement involved implementing an alternative model for transmission (i.e. a branching model versus SIR model), which then allowed for a more realistic infectious period. As shown through analysing multiple trees, even with phylogenetic uncertainty the timing of the last transmission event is robust, which makes this method useful for answering public health questions related to infection timing. The robustness comes from a robust phylogeny and the fact that generation times are constrained by the structure of the trees, namely, by tree height and branch length. However, this can lead to issues as seen when an extreme prior is used, as in order to compensate for a prior with a short mean generation time, the source case is forced to have been infected far in the past and thus has a very long generation time and skews the generation time distribution.

Whilst the phylogenetic tree is generated using WGS data alone, epidemiological data played an important part in determining whether the resulting tree ‘fit’ with the information known about the outbreak and the cases.

Although I have not presented data on the subject, intuitively, generation time will also have an impact on sampling time as sampling time must always be longer than generation time

due to the assumption that an individual cannot infect after they are sampled. This premise assumes that treatment is started once an individual is sampled and treatment is effective at reducing infectiousness immediately. Although there was no information on the uptake of treatment or treatment delay for this outbreak, given a high proportion (85%) of cases were classified as having cured, it is not an unreasonable assumption. Thus, by setting a prior generation time distribution with a large scale the mean posterior generation time is increased and, in turn, the sampling time. These effects must be considered when setting prior distributions.

A secondary output from the analysis of the sampling times was that the mean sampling time was roughly 1.3 years after infection. In general, this means that the cases were developing active disease quickly after infection, and more often as a result of recent infection as opposed to reactivation. This is also reflected by the generation times as the distribution remains positively skewed throughout, showing that most cases were fast progressors (progression from latent TB infection to active disease in less than 2 years).

4.5.2 STRENGTHS

Given that the method seeks to uncover information for which there is no ground truth, i.e. the truth is unknown, the method can only be effectively tested through simulation. The original method has been shown to produce accurate results when tested on a simulated outbreak [99].

A strength of this work was the choice of outbreak; as the outbreak was small and within a small territory the previous epidemiological investigation was able to get extremely detailed knowledge of the involved individuals. Working closely with the public health team in charge of the epidemiological investigation meant the results of the statistical analysis could be compared with the epidemiological knowledge and this helped to refine the statistical analysis.

A high sampling coverage (believed to be 48/52) is an advantage as the version of the method used in this work does not currently infer missing samples.

4.5.3 LIMITATIONS

A limitation of the method is that we cannot validate the timing of the last transmission event, the sampling times or generation times with any data as it is not possible to determine the infection date accurately with any other method. This is a fundamental limitation in any outbreak reconstruction task, as the ground truth is not known, hence the need for this analysis. The best that can be done in these circumstances is a comparison of reconstruction with a simulated outbreak as was done with the original version of TransPhylo.

This version of the method still assumes that every case involved in the outbreak has been sampled, which is known to be false given that there were four known cases for which there was no genomic data. Missing cases could affect the inference as the generation time is highly likely to be wrong if the incorrect infector is chosen. For example, in some cases, with the infector of a case missing, the infector's infector could be imputed as the infector of any they have infected. Thus, the observed generation time would, in fact, be the addition of two generation times (Figure 4.9). In this scenario, in order to capture the most accurate reconstruction (i.e. keep the approximate transmission pathways, meaning the infector of a missing case is inferred as the infector of any cases the missing case infected), TransPhylo would have to accept potentially quite long generation times, which is unlikely due to the choice of a Gamma distribution for the generation time. Since this work, an updated version of TransPhylo has been developed which can determine the possibility of unsampled cases [141].

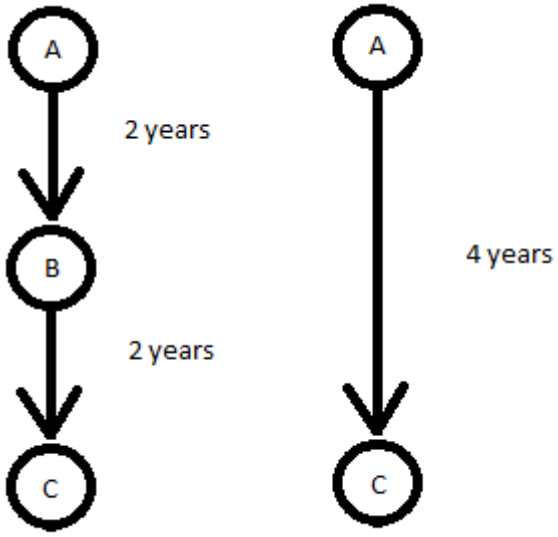

Figure 4.9 Diagram depicting effect of incomplete sampling on reconstructed transmission chains
On the left is the true transmission chain with the generation times denoted next to the arrows. On the right is the transmission chain as determined when case B is not sampled, thus the generation time is assumed to be 4 years

An additional limitation relates to the sampling of the outbreak. Cases may exist, with either latent infection or active disease, who were infected after 2012 (given that some of the known cases in the outbreak were still infectious up until 2014) but have not been diagnosed. The presence of such cases would nullify the hypothesis that the outbreak is over. This is a difficulty with any outbreak of tuberculosis; unless it can be confirmed that there are no cases of latent infection in the community, there is still a chance that an individual can reactivate in the future and cause a further wave of cases. Despite this, the

priority for public health control efforts is to stop sustained transmission, as later cases arising through occasional reactivation can be contained more easily. Once ongoing, recent transmission has been determined to have ceased then there can be more focus on detecting and treating latently infected individuals in order to prevent a recurrence.

4.5.4 RECOMMENDATIONS AND CONTEXT

The use of the method for other TB outbreaks would depend on the sampling proportion of the outbreak; this needs to be high in order to use the method as in its current state it does not account for unsampled cases. Unsampled cases may be cases that we do not know exist or cases we know about but have no available sample/sequencing data. In their method for inferring transmission networks from genomic data, Jombart *et al*. [82] included a way of dealing with unsampled cases by introducing a term in the likelihood for the probability of observing the number of genetic differences seen between i and j's genomes given that they are separated by k_i generations. A similar adaptation has since been included in the TransPhylo methodology [142] and although the outbreak was considered well sampled it could be beneficial to apply the method and confirm this assumption.

Although it was not included here, it is possible to incorporate epidemiological data into the method in order to take advantage of any further information that may help resolve the transmission network and the infection dates (see Appendix 2), as was employed by Ayabina *et al* [143]. Such information could include sputum smear results, where cases with smear-positive disease are more likely to transmit infection versus those with smear-negative disease [16, 17]. This could be included in the form of a penalty for transmission networks that propose transmission from cases with smear-negative disease. This should, in theory, help reduce any uncertainty in transmission events and their timings, however this information itself introduces its own uncertainty as people in general are often unreliable in their assessment of how long they have been unwell for. This may be compounded by issues such as drug abuse or mental health concerns.

New methods for inferring transmission networks from genomic data are still being researched. For example, Klinkenberg *et al*. [144] and Hall *et al.* [145] have developed methods that can simultaneously infer the phylogeny and transmission tree rather than just inferring the transmission tree from the phylogeny. The benefit of this being that phylogenetic uncertainty can be accounted for much more easily as opposed to having to assume that the phylogeny is absolutely true and simply accounting for the uncertainty through sensitivity analysis. Klinkenberg *et al.*'s method was tested on TB, MRSA, Foot and Mouth disease and influenza outbreak data, whereas Hall *et al.*'s method was only used for influenza. However, the method by Hall *et al.* used a susceptible-exposed-infected-recovered (SEIR) model for the epidemic process, meaning that it would still be suitable for TB.

de Maio *et al*. [146] have developed a Bayesian method that is useful for determining transmission when there may be little to no diversity between consensus sequences. This could play an important role for TB outbreaks where there have been examples of few SNP differences between outbreak genomes [61].

To generate a set of phylogenetic trees which are compatible with the known epidemiology of the outbreak, the clock rate was fixed to a single value. However, given that a Bayesian approach was used, this information could have also been specified as an informative prior on the clock rate within BEAST, which then also would have allowed for some variation.

4.6 CONCLUSION

In this work, a TB outbreak from Canada was analysed to see if the inferred infection dates from the analysis could reveal something about the current state of the outbreak, i.e. was it still ongoing? Before doing this analysis, the inference method contained an assumption, exponentially distributed generation and sampling times, considered inappropriate for TB. Given this assumption would affect the infection dates, it was therefore deemed necessary to adapt the method to improve its use for TB. This entailed changing the SIR model, which inherently contains the assumption, to a branching model, which allowed a greater freedom over the choice of distribution for the generation and sampling times. In doing this, it was necessary to assess the effect of the new choices of generation and sampling times through sensitivity analyses. The sensitivity analyses showed that as long as the choices for the prior generation time and sampling distributions were sensible, in accordance with what is known, then the posteriors were largely robust.

In summary, it was possible to produce a set of infection times for the outbreak cases and analyse how long ago the most recently sampled cases had been infected, which led to concluding the outbreak was over. This has therefore been proven to be valuable for public health investigators to help them decide on real-time public health responses.

5 PHYLOGENETIC ANALYSIS OF LONDON TB OUTBREAK GENOMES (OBJECTIVE 3)

5.1 INTRODUCTION

As WGS becomes a more enticing option for understanding infectious disease outbreaks especially in real-time, it becomes important to examine sequencing data from a variety of real-world outbreaks. The more data we have, the more we can learn about how best to interpret it, and then use those learnings in real-time to analyse current data and help control any current outbreaks. Following on from my analysis of the outbreak in British Columbia (Chapter 4), my subsequent objective focussed on the largest recorded isoniazid-resistant TB outbreak in Europe [147], an outbreak that has been largely centred in London and has been ongoing since 1995.

In this chapter, I aimed to create a timed phylogenetic tree from the raw WGS data for the London outbreak, from which statistical inference techniques as per Chapter 4 could be employed to reveal the transmission dynamics. Firstly, I describe the sequencing data that was available for analysis. Then I describe the bioinformatic analysis pipeline used, including software. Finally, I explain the phylogenetic analysis used on the resulting genomic alignment.

5.1.1 OUTBREAK

The London isoniazid-resistant TB outbreak is a cluster of over 500 TB cases that have been primarily found in London over the past two decades and are resistant to the first-line anti-TB drug isoniazid. A number of papers have been published on this cluster over the years [147, 148, 149, 150] that have documented the evolution of the outbreak over time.

This outbreak was first identified when three patients from a north London hospital were diagnosed with TB phenotypically resistant to isoniazid within a small timeframe in 2000 [149]. Genotyping found them to have indistinguishable RFLP patterns and retrospective investigation of isoniazid monoresistant samples in the north London hospital and the surrounding hospitals found earlier cases with the same RFLP type, with an index case diagnosed in 1995. From 2000 onwards, all isoniazid monoresistant cases were genotyped using rapid epidemiological typing (RAPET), which is a polymerase chain reaction (PCR)-based alternative to RFLP, and/or RFLP typing to identify if they were part of the outbreak.

In 2006, MIRU-VNTR typing was introduced; 24-loci MIRU-VNTR typing demonstrated the following outbreak strain, with one untypeable locus: 424332431515321236423-52. Searching the MIRU-VNTR*plus* database [151] for the VNTR showed that it was extremely

similar (4 loci different) to a MIRU-VNTR strain-type of the Cameroon lineage that circulates in Nigeria, where the first outbreak case originated.

As the outbreak progressed, MDR-TB was occasionally seen and was assumed to be emerging due to the large number of patients adhering poorly to treatment [147]. There were concerns that the MDR branch of the strain was being transmitted as some cases were infectious and not on treatment [152].

Over the course of the outbreak a number of characteristics emerged; firstly, it has remained surprisingly contained geographically. The majority of cases have been found in north and central London, with a smaller number from south London and the occasional case outside of London [153]. The other notable characteristic of the cases is that a large proportion has one or more risk factors determining them as 'hard-to-reach'. These risk factors are prison history, history of problem drug use, alcoholism or homelessness. A case control study found that the proportion of cases with these characteristics was higher than would be expected compared with other cases diagnosed in London during the time of the outbreak [147].

Extended contact tracing was employed throughout the early stages of the outbreak and questionnaires concerning the contacts and social hangouts were regularly distributed to clinics and returned by nurses. For these cases there is a clear picture of the treatment history and lifestyles for many of these cases; later cases are less well documented.

Despite a small decrease around 2008, the number of cases notified each year remains constant, which highlights the need for a new approach to outbreak control in order to bring it to an end (Figure 5.1). This motivates use of genomic epidemiology and phylogenetic inference to determine what fresh insights might be available.

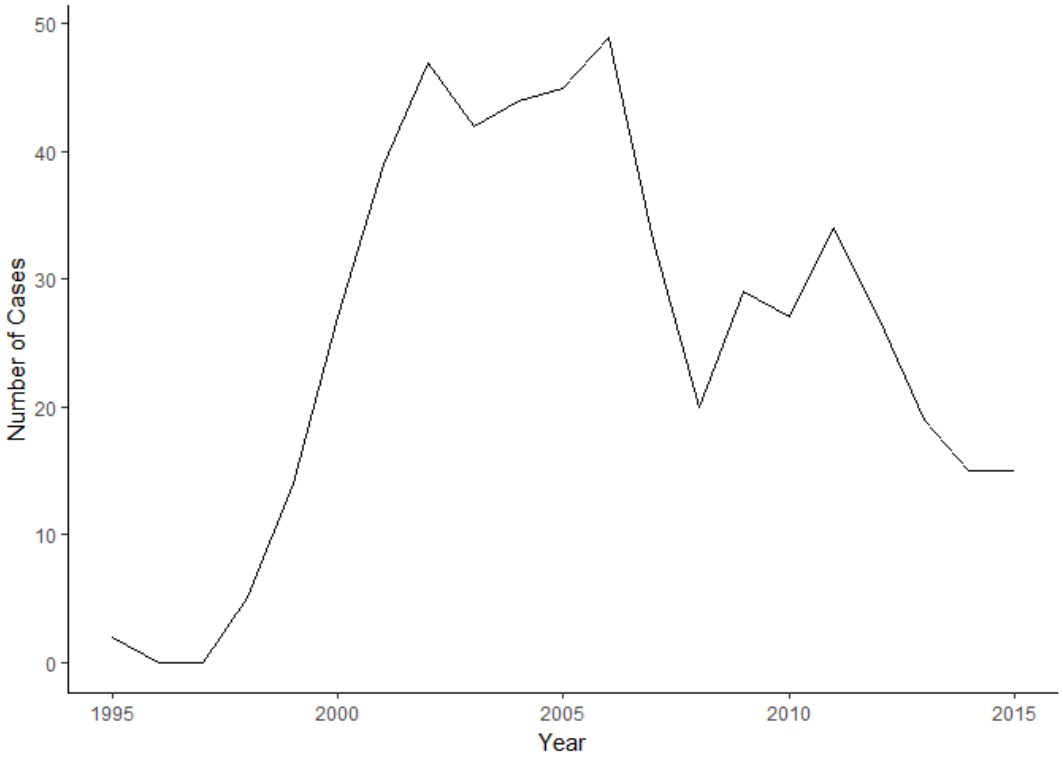

Figure 5.1 Outbreak curve for London outbreak

In 2015, the TB samples from outbreak cases kept by the National Mycobacterium Reference Laboratory (NMRL) were whole genome sequenced by the Sanger institute for a study by Casali *et al*. [61] to see what the WGS data could reveal that might give some insight into the transmission dynamics for the outbreak. As a result, the genomic data are freely available on Genbank [154]. Casali *et al*. used the WGS data to construct a minimum spanning tree and compare the SNP differences between cases known to have been epidemiologically linked. After considering the approaches laid out in Chapter 3, we sought to use the genomic data to examine transmission in more rigorous terms using a statistical approach.

5.1.2 BIOINFORMATICS

Producing SNP data firstly involves turning raw sequencing reads, which are short sequences (usually 100bp long) in no particular order, into a fully assembled genomic sequence. Genetic changes in that sequence such as single nucleotide polymorphisms (SNPs) and insertions and deletions (indels), can then be identified when compared with a reference sequence. This process often requires numerous steps using different software that can handle multiple types of files; the basic steps can be seen in Figure 5.2. There are software programs such as Genome Analysis ToolKit (GATK) [155, 156] that attempt to

handle all steps within the same package as well as software programs specialised to each step, which can be used in succession.

Before beginning the construction of the sequences and calling variants, the quality of the reads must be checked so that reads with poor quality can be discarded and minimise the number of errors. FastQC (v0.11.4) [157] assesses the reads using 12 criteria, such as per base sequence quality and per sequence GC content, and classes the files as either passed, warning or failed. The definitions of these criteria and the meaning of the outcomes are listed in the FastQC manual pages.

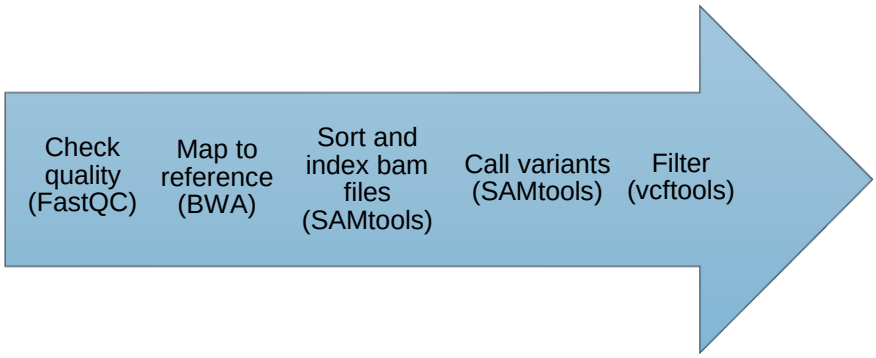

Figure 5.2 Step-by-step bioinformatic analysis process with software used for each step given in brackets

After determining the quality of the raw data files, a consensus sequence can be created from the reads. This can be done in two ways: mapping to a reference genome or *de novo* assembly.

De novo assembly entails building a genome from short reads without a guide in the form of a reference genome to help determine how the reads should be put together. It is time and computationally intensive as every read must be compared with every other read in order to determine where there may be overlaps. When overlaps between two reads are found, the two reads can be combined together and are then called a contig. Once a few reads have been joined to make long contigs, these then act as a scaffold for the remaining reads.

In comparison, the process of mapping to a reference uses a specific genome as a guide and places reads in a certain location according to how well it matches to the reference.

Mapping to a reference would generally be preferred to *de novo* assembly if there is a well-characterised reference genome as it is less computationally intensive. This is the case for *M. tb.* and the monomorphic nature of the bacterium means that there are very small amounts of variation across the lineages.

In order to undertake mapping to a reference, a reference must first be chosen; for *M. tb.* bioinformatics analysis the H37Rv strain is the most commonly used. Once a reference is

chosen a mapping software is used to align the reads from a sequenced isolate along the reference genome with some probability defined as the mapping quality and alignment quality (Table 5.2). After this has been done for all the reads, ideally every single base of the reference genome should have at least one read mapped to it i.e. a read depth of at least one. However, there is also a possibility that reads can be mapped to more than one location and some reads may not be possible to map at all. In order to then construct the consensus sequence for the isolate each base must be examined in all the reads and a 'consensus base' is chosen based on the largest proportion of reads and quality etc. For example, in Figure 5.3, at position 4, reading from left to right, the reference genome has a G, however five of the six reads mapped to the reference have an A at that position. As a result, an A is chosen for the consensus sequence at that position with a probability of 0.83.

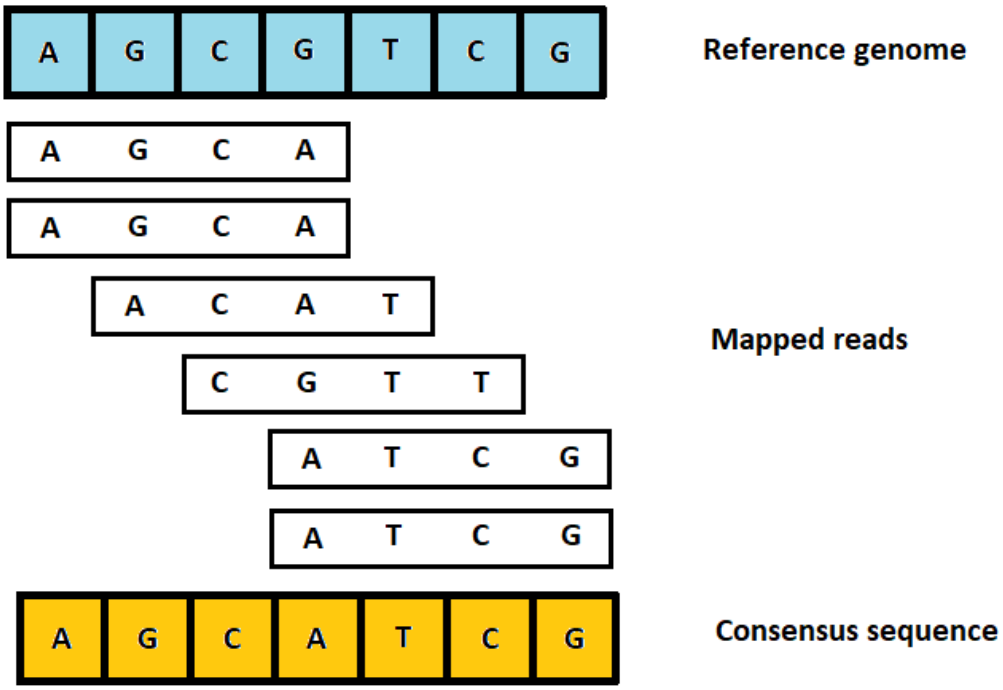

Figure 5.3 Depiction of the process of mapping to a reference genome

The aligned sequence alignment map (SAM) files are then converted to binary alignment map (BAM) files and sorted and indexed in SAMtools (v.1.2) to allow the program to find regions of the genome quickly. Using the mpileup command, it is possible to set a threshold for mapping and base quality and any bases failing this criterion will not be counted towards the determination of a genotype. The choice of these thresholds has varied across the WGS TB literature (Table 5.1) demonstrating the dearth of knowledge around the best approach for such analyses. To get a joint prior probability for ML estimations at each site [158], variants were jointly called according to a minimum mapping quality threshold and minimum base quality threshold set by the user. This uses base alignment quality (BAQ), which measures the probability of a read base being wrongly aligned and assigns it a score [159].

Study	Software	WGS quality thresholds
Bryant *et al* [45]	SAMtools and bcftools	Base quality of 50, mapping quality 30
Casali *et al* [61]	SAMtools	No base quality mentioned, mapping quality 45
Clark *et al* [97]	Not mentioned	Phred score of 30
Didelot *et al* [99]	SAMtools	Variant quality of 222, genotype quality 99, no pre-filtering
Guerra-Assunção *et al* [79]	Trimmomatic (pre-mapping filtering), SAMtools	Removed 'low quality reads' with base quality <Q27, no SAMtools filters.
Gardy *et al* [88]	Not mentioned	No mention
Guerra-Assunção *et al* [92]	Trimmomatic (pre-mapping filtering), SAMtools	Removed 'low quality reads' with base quality <Q27, no SAMtools filters.
Lee *et al* [93]	Not mentioned	Phred score of 50, no pre-filtering
Luo *et al* [89]	Not mentioned	No thresholds mentioned
Pérez-Lago *et al* [103]	SAMtools	Minimum coverage 10, minimum mapping quality 20
Regmi *et al* [107]	Not mentioned	Phred score of quality 20 post-filtering
Smit *et al* [101]	Not mentioned	Mapping quality 45
Stucki *et al* [100]	Not mentioned	Post-calling Phred score of 20
Walker *et al* [48]	Not mentioned	No thresholds mentioned
Witney *et al* [94]	SAMtools	Mapping quality score of 30, site quality score of 30

Table 5.1 A list of TB studies that have undertaken bioinformatic analysis of genomic data alongside the software used and any quality filters used

An R script, named vcfProcess, (https://www.ucl.ac.uk/computational-biology-group/research-interests/software/vcfprocess/) can be used to the filter the variants in the

Variant Call Format (VCF) file and to convert the SNPs into a FASTA file format that can be read into most phylogenetic software, such as RAxML and BEAST. Variants can be filtered according to thresholds around read depth or coverage or the Phred-scaled quality score (the probability of there being a variant at that site, not the sample level genotype confidence) or the location of the variant. Another statistic used by studies to assess the quality of base calling and mapping. Read depth or coverage is defined as the number of reads that cover each position along the genome. Intuitively, the fewer reads there are covering a position, the less confidence we can have in the base calling as there is an increased likelihood of errors.

When performing bioinformatics analyses, it is important to check the quality of the data and the alignment. This can be done at various stages of the process and different software and pipelines will measure quality with alternative language and thresholds. A brief list and explanation of some of the quality metrics that can be done are compiled in Table 5.2.

Quality measure	Definition
Base quality	This assesses the potential that the base has been reported incorrectly by the sequencer, assigned by the sequencer
Mapping quality	This assesses the potential that the read has been mapped incorrectly by the aligner. This is decreased if the read has been mapped to more than one position
Alignment quality	This assesses the similarity between the mapped read and the reference, assigned by the aligner
Variant quality	This assesses the probability that there is some kind of variation at that site
Genotype quality	This assesses the probability that the call assigned at that site for that sample is true

Table 5.2 Definitions of different quality metrics employed in bioinformatics analysis. All quality is reported as a Phred score

5.1.3 PHYLOGENETICS

Phylogenetic analysis is the process of building a tree that shows the relationships between the strains for which SNP alignments have been generated through a bioinformatics pipeline. Like bioinformatics analysis, this process also requires a number of choices and steps. There are multiple different types of phylogenetic tree as mentioned in Section 1.4.1.1.1.1, each useful for a specific purpose and as such a decision must be made about

which type of tree is most suitable dependent on the use of the tree. In this case, a timed tree was desired for use in further downstream analyses (i.e. TransPhylo transmission analyses used in Chapter 4) where timing (of transmission) is the desired goal.

In order to build reliable timed trees, there must be a clock-like signal in the sequence data, meaning that a molecular clock is present. This can be determined. by performing a linear regression analysis on the root to tip distance for each sample in the tree and its sampling date. This requires a phylogeny and the sampling dates for each genome in the tree. An R^2 value close to 1 suggests a strong clock-like signal in the data. When building the ML tree for this analysis, the main setting choice when performing this type of analysis is the choice of a substitution model. This can be set by the user based on knowledge of the data/species or chosen using a program such as jModelTest [161, 162] or ModelFinder [163], which compares different substitution models using model selection methods, such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). Bootstrapping, which involves producing a tree from a random resampling of the input sequences, helps to determine some confidence in the tree, whereby bootstrap values determine the percentage of trees a node is found in. Values closer to 100 represent greater confidence in the tree.

Once a clock-like signal has been determined, Bayesian tree sampling software, such as BEAST (Bayesian evolutionary analysis sampling trees [134]), can be used to produce a timed-tree. The data required are SNPs and the date of collection for each sample. There are a number of parameters that need specifying in BEAST, namely the type of molecular clock, a tree prior and substitution model. The type of molecular clock will determine whether the molecular clock rate (the constant rate of genetic change in a bacterium, usually measured in mutations per unit time) is strict or relaxed; strict meaning that the rate is the same across lineages versus relaxed which means it can vary across lineages [164]. A tree prior reflects how the effective population size, N_e, (of the bacterium) is believed to change over time, e.g. does it remain constant or grow exponentially etc. [134].

To decide on all of these settings, BEAST allows the user to perform a path sampling analysis. BEAUti 1.8.1 was used to generate an extensible markup language (XML) file from the sequence alignment for use in BEAST 1.8.1. It is within BEAUti that the settings for BEAST are chosen, such as the tree prior, substitution model and clock model. As recommended by the BEAST developers, initially, a simple substitution model (HKY), a coalescent tree prior and relaxed clock model was chosen to first see if the MCMC converged before choosing more complex models. Although a relaxed clock model is not recommended for intra-species data for several reasons including that large variation is not expected between lineages within a species [165], it can be used to check how much variation among rates is implied by the data and therefore determine the use of a relaxed or strict molecular clock model. To then determine the best choices for the tree prior, clock model and substitution model a comparison between these choices was made using marginal likelihood estimators (MLE) [166] and Bayes factors (BF). Bayes factors are a

statistic calculated by the ratio of MLEs from different models and then used to reject or accept one of the models according to a set of thresholds. In BEAST, this involves selecting a path sampling (PS)/stepping-stone sampling (SS) analysis, where samples are collected not only from the posterior distribution, but also a series of power posteriors [167]. The result of these analyses is a log MLE value for the model [168, 169] which can be compared between models using Kass and Raftery [170]'s statistic of 2log(BF), where BF is equal to the value of one model's MLE value minus another model's MLE value.

BEAST produces three files: one contains trees sampled by the MCMC, ones contains a log of all the parameter values over the chain, and one contains a summary of the chain. The sampled trees are the main result, but in order to produce a single tree, either one tree from the sample must be chosen or a summary tree created. A summary tree is a single point estimate tree that tries to represent the entire posterior sample of trees. There are multiple types of summary tree produced in different ways: for example, a MCC tree as described in Section 4.3.3.2 [171]. A summary tree can be produced from the trees file using TreeAnnotator. TreeAnnotator can also include the posterior probabilities of the nodes in the summary tree, as well as the highest posterior density (HPD) limits of the node heights.

The parameter log and summary files can be used to run diagnostics on the MCMC: log files are used to check the traces for the parameters (i.e. the sampled values across the chain) and the summary file contains the acceptance probabilities for the parameters, both of which determine how well the chain is mixing. The Tracer program is a tool for analysing and assessing MCMC chains from BEAST using the log files [172]. In Tracer, one can examine the trace plots for each parameter, i.e. the sampled value for each iteration in the chain. It is possible to import multiple chains and view the sampling for parameters across multiple chains simultaneously, which makes assessing convergence easier. Tracer also displays effective sample size (ESS) values for each parameter. ESS values are equal to the number of effectively independent draws from the posterior distribution to which the Markov chain is equivalent [165]. The higher the number, the less correlated the samples are and the larger the number of independent draws that are available. Larger ESS values are desired as it is only if we have independent samples, that we can apply the law of large numbers and conclude that the estimated mean of our samples will converge on the true mean (the one we are attempting to sample from) [173]. A value of 200 or more is generally considered a 'good' ESS value, although the larger the better.

5.2 DATA

Sequencing data for 415 genomes from the outbreak were downloaded as raw Illumina reads in FASTQ format from the European Nucleotide Archive (ENA) (accession number ERP003508). The meta-data for the samples were provided by the NMRL and included the reference laboratory number, the date of sample collection, patients' initials, specimen type

and drug sensitivity patterns for isoniazid, rifampicin, ethambutol, pyrazinamide and streptomycin.

Of these 415 sequences, the meta-data revealed that 17 outbreak cases had two samples sequenced, 2 cases had 3 samples sequenced and one patient sample had 27 colonies from a single culture sequenced. In the cases where multiple sequences belonged to the same patient (excluding the one with 27 sequences) the samples were retrieved at different times with a range of 8 to 1469 days between the earliest and the latest.

5.3 METHODS

5.3.1 BIOINFORMATICS

By running FastQC on all the forward and reverse read files and marking a pass as 1, warning as 0 and a fail as -1, the quality of these files can be represented in a matrix. Of the samples within this study, every file passed on the basic statistics, per sequence quality scores, per base N content, sequence length distribution, sequence duplication levels, and adapter content measures. The files overall did worse on per base sequence content and kmer content. It is unsurprising that these samples failed on the per base sequence content criteria as *M. tb* is well known to have a GC-rich genome, meaning that the difference between the number of G/Cs and A/Ts is larger than 10%, which this criteria checks.

BWA (v0.7.12) [174] was used as the mapping software using the mem command. This then produces sequence alignment map (SAM) files.

Variants were called using a minimum mapping quality threshold of 45 and minimum base quality threshold of 30; these are Phred scores (Q-scores) and thus correspond to an error of 1 in 55000 and 1 in 1000, respectively.

vcfProcess was used to the filter the variants in the Variant Call Format (VCF) file. The variants were filtered out if depth was lower than 5 reads or if the Phred-scaled quality score (the probability of there being a variant at that site, not the sample level genotype confidence) was below 20, or if they were in the proline-glutamate (PE), proline-glutamate polymorphic guanine-cytosine (GC)-rich sequences (PE-PGRS) or proline-proline-glutamate (PPE) genes as defined in TubercuList [175]. It is common practice to remove variants in these genes as they are high GC content regions [160] that make mapping reads difficult and consequently complicate the task of distinguishing true SNPs from mapping errors.

The script used to map reads and call variants is available at GitHub (http://github.com/holliehatherell/Thesis).

5.3.2 PHYLOGENETICS

In order to determine the substitution model to build a ML tree, the IQ-TREE ModelFinder program was run on the FASTA file of only the variant sites (command can be found in Appendix 3). The Bayesian Information Criterion (BIC) (a criterion for choosing between different statistical models) was used to determine the best substitution model.

A ML tree was then created in IQ-TREE using the best model as determined by ModelFinder. Bootstrapping was performed with 1000 iterations to assess the validity of the resulting tree. Full details and commands are listed in Section 9.3.1 – Appendix 3.

The program TempEst [176] was then used to determine whether there is clock-like signal in the data from the tree produced by IQ-TREE. The selection 'best-fit' root position was chosen, which roots the tree such that the sum of squared residuals from the regression line is minimised (i.e. maximises the linearity), as the tree from IQ-TREE is arbitrarily rooted. A positive correlation coefficient value, R^2, was taken as the sign of a clock-like signal in the data.

BEAST was then run on the sequence alignment. As only variant sites (SNPs) were used, an ascertainment bias correction had to be used by specifying how many invariant sites there were amongst the sequences. The values specified were the base counts for the TB reference strain H37Rv: A: 758565; C: 1449985; T: 758379; G: 1444603 [177].

Initially, simple settings for the substitution model, tree prior, clock model and rate of heterogeneity were chosen (i.e. an HKY substitution model, a coalescent tree prior and relaxed clock model) and the convergence of the model was checked via the traces of the parameters and ESS values. Then, path sampling was performed alongside the MCMC in order to compare the suitability of different choices for the settings. Firstly, different substitution models were compared, then clock models, then rates of heterogeneity, then tree priors. In each case, an MCMC chain was run for 100,000,000 iterations and then a path sampling analysis was run for 100,000,000 iterations also.

Once the path sampling results were analysed, the optimal settings were determined, and these were used to produce the final tree. The final tree was determined by running three separate MCMC chains with the chosen optimal model for 100,000,000 iterations. Once these could be assessed for convergence and mixing by checking ESS values and traces and were deemed suitable, the resulting trees from the chains were thinned by resampling at a frequency of 30,000, removing a burn-in of 40,000,000 trees and combined in LogCombiner. TreeAnnotator was then used to produce an MCC tree with posterior probabilities from the combined and thinned log. Trees have been visualised using FigTree.

5.4 RESULTS

5.4.1 BIOINFORMATICS

Manual inspection of the concatenated SNPs in Aliview showed that seven of the 415 samples had large numbers of variants and heterozygous calls in comparison to the other samples and were determined to be mixed samples or a different strain i.e. not part of the outbreak. Four more samples failed sequencing, i.e. were largely returning missing data across the sequence due to extremely low coverage. One further sequence was removed due to it being a 'repeat submission', leaving 403 sequences.

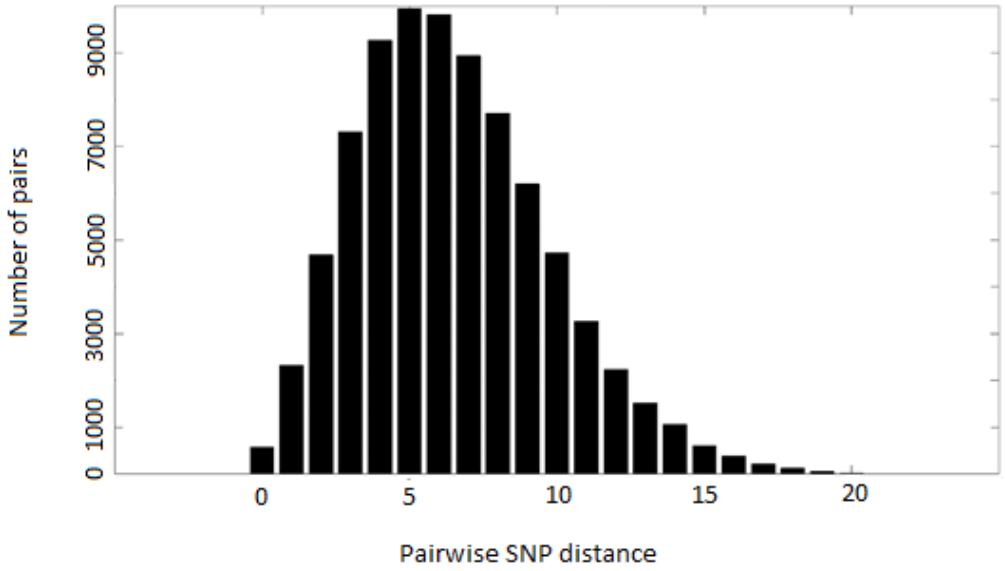

Figure 5.4 A histogram showing the distribution of the number of SNPs between every possible pair of sequences in the outbreak data

There were 980 SNPs present between at least one sample and the reference. Of these, 261 sites varied amongst the samples i.e. 719 SNPs were present in all the samples but not the reference. At 157 sites there existed a heterozygous variant in at least one of the sequences, meaning that there were calls to two variants at the site, these were given the character 'N', which denotes 'any base' as per the IUPAC code [178]. Figure 5.4 shows the pairwise distance between all of the samples. 116 samples were identical to at least one other, corresponding to 13 pairs, six groups of three, one group of four, one group of five, one group of 12, one group of 11, one group of 14 and one group of 26. The most common distance was five SNPs between any two pairs with a maximum of 20 SNPs.

There were 275 indels between samples and reference. These were excluded for phylogenetic analyses.

5.4.2 PHYLOGENETICS

5.4.2.1 IQ-TREE RESULTS

After running IQ-TREE on the sequences, the substitution model TMVe + ASC was determined to be the best fitting for the data according to the BIC. This substitution model is the transversion model, where AG=CT and there are equal base frequencies. The full set of parameters is listed in Section 9.3.1.

All models that included ASC were ranked above those without ASC.

IQ-TREE then produces a phylogenetic tree based on the results of the substitution model fitting (Figure 5.5).

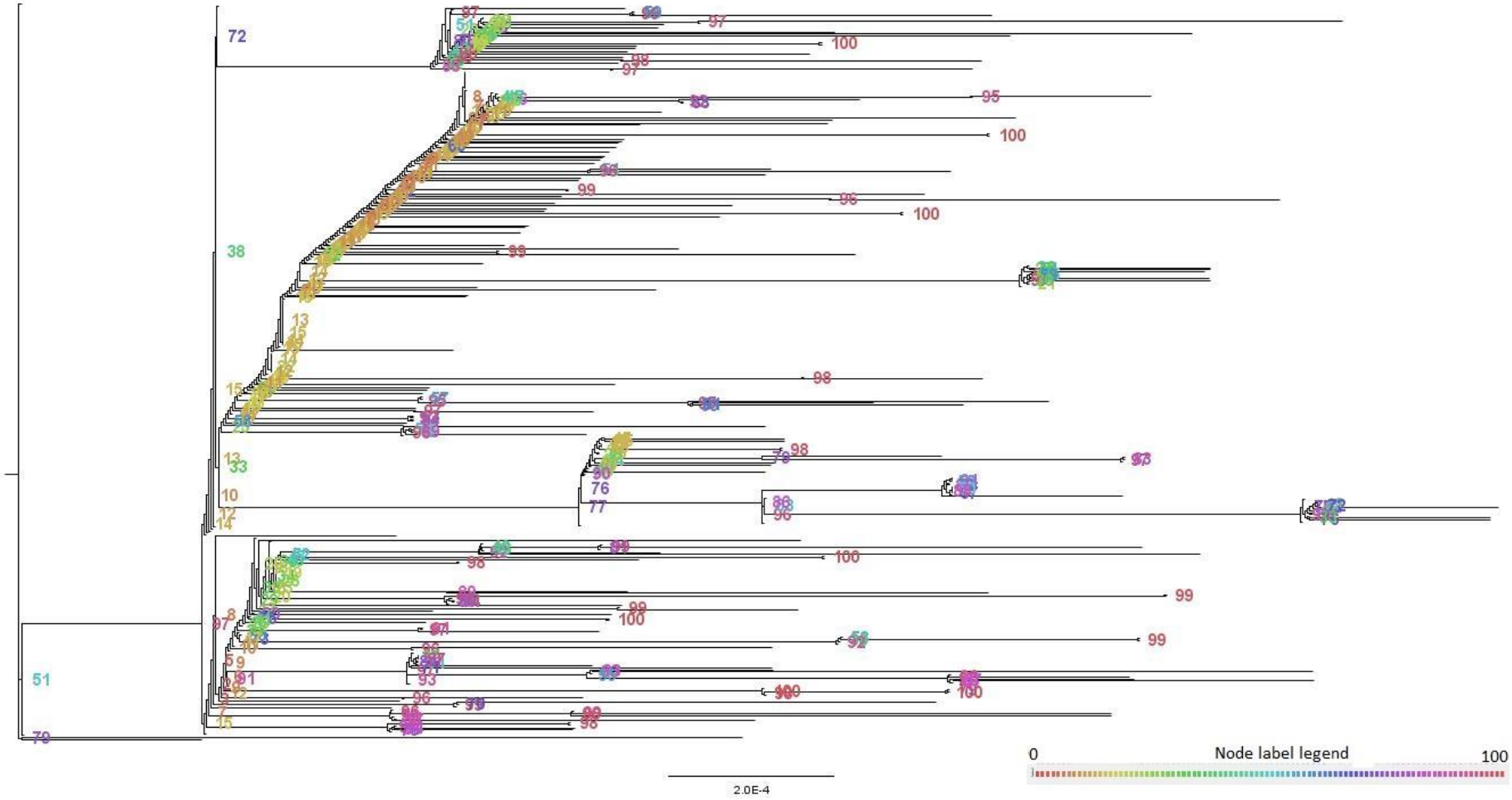

Figure 5.5 Consensus phylogenetic tree produced using IQ-TREE with a TMVe+ASC substitution model from 1000 bootstrapping replicates

The bootstrap value of each node represents the percentage of replicates that contain that node. The values are coloured according to the scale displayed in the legend

5.4.2.2 TEMPEST RESULTS

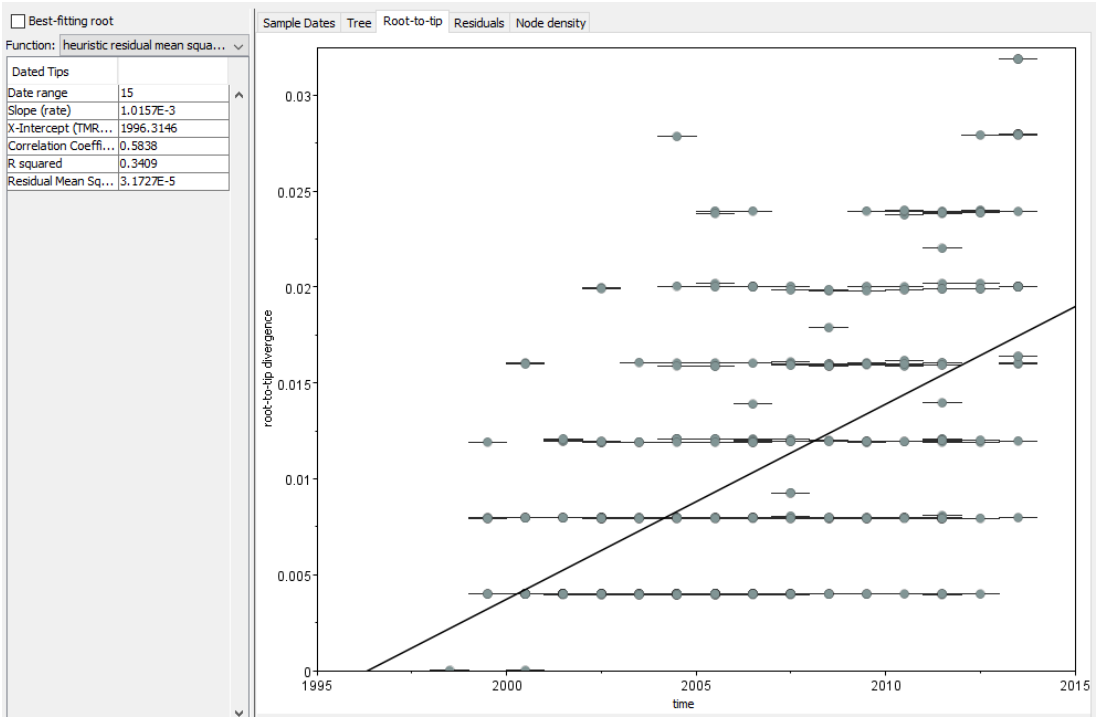

Figure 5.6 Results from TempEst for the phylogenetic tree produced by IQ-TREE.

Figure 5.6 depicts the plot of sampling date versus root-to-tip divergence in the tree without selecting a best-fitting root. The results of the TempEst analysis show a positive correlation between the time the samples were collected and the root-to-tip distance in the ML tree which implies that the data is potentially 'heterochronous' i.e. the range of time the data has been collected over is long enough to measure a reasonable amount of diversity and therefore it is possible to perform molecular clock analyses.

5.4.2.3 BEAST RESULTS

Before starting a BEAST analysis to find the best settings for the data, it is recommended by the developers that the data be first run with very simple choices (a HKY substitution model +G+I, relaxed clock and constant coalescent tree prior) to see if convergence can be achieved.

5.4.2.3.1 INITIAL SIMPLE MODEL CONVERGENCE RESULTS

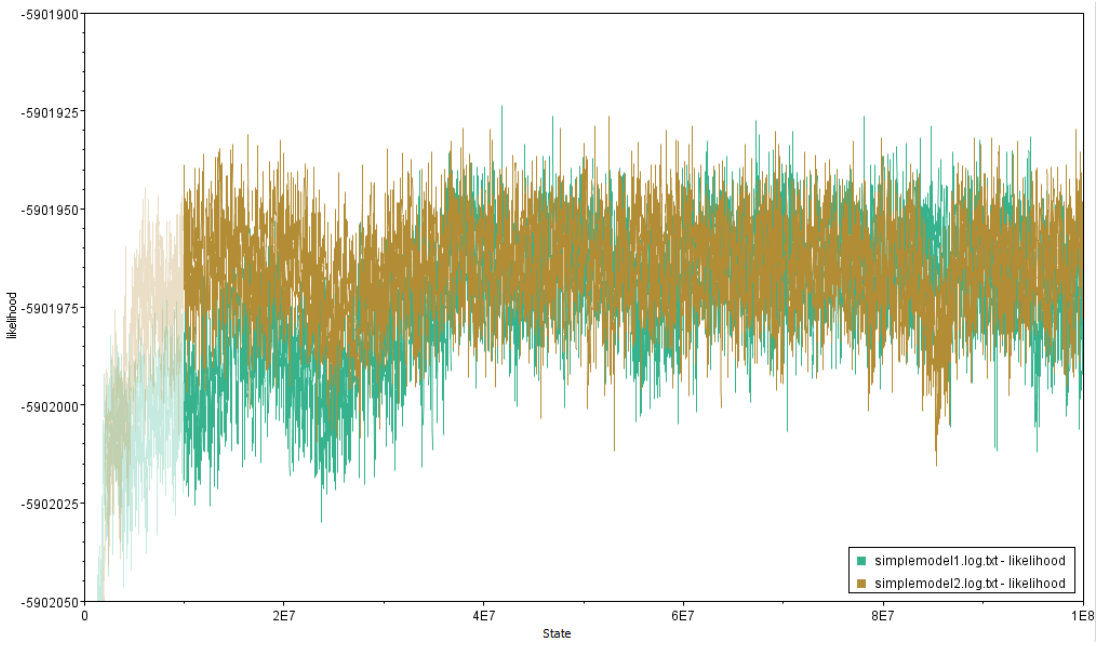

Figure 5.7 Trace of likelihood

After running the simple model for 100,000,000 states, the resulting log file can be imported into Tracer and assessed for convergence. The recommended way of assessing convergence is looking for the high ESS values (>200) for the model parameters and by looking at the traces for the parameters, i.e. the sampled values at each state. Figure 5.7 shows the trace for the likelihood and demonstrates that, after removing the burn-in (10% of number of total sampled states), the sampled values are fluctuating but are tightly around a mean. All ESS values are larger than 200 for this run (see Appendix 3). These two results together confirm that there is convergence and therefore these data should be suitable for BEAST analysis.

5.4.2.3.2 BAYES FACTOR ANALYSIS

Bayes factor analyses were then performed to compare models with different tree priors, substitution models, rates of heterogeneity, and clock models. The results were as follows.

5.4.2.3.2.1 SUBSTITUTION MODEL

BEAST allows for a generalised time-reversible (GTR) [179], HKY [180], or TN93 [181] substitution models. To compare between these three models, BEAST was run three times each with a different clock model with the rest of the model choices kept identical: no rate of heterogeneity, an exponential relaxed clock and a constant coalescent tree prior.

Substitution model	Path sampling log MLE
GTR	-5903372.370

TN93	-5903176.708
HKY	-5903156.489

Table 5.3 Log marginal likelihood estimator values for different substitution models

To compare between the models we must compute the value proposed by Kass and Raftery [170], i.e. $2\log\mathrm{MLE}_1 - 2\log\mathrm{MLE}_2$, where MLE_i is the marginal likelihood estimator for model i. They then proposed a set of ranges for when model 1 may be preferred to model 2.

Using the results of Table 5.3, if we compare the HKY model with the GTR model, we find

$$2\log\mathrm{MLE}_1 - 2\log\mathrm{MLE}_2 = 2 \times -5903156.489 - 2 \times -5903372.370 \approx 432$$

A result of 432 is greater than 10, which is the lower threshold for which Kass and Raftery propose should be used as very strong evidence for model 1 (HKY).

Comparing the TN93 model with the HKY model then, we find

$$2\log\mathrm{MLE}_1 - 2\log\mathrm{MLE}_2 = 2 \times -5903156.489 - 2 \times -5903176.708 \approx 40$$

Again, this is still larger than 10 and therefore the HKY model is considered the most fitting for the data.

5.4.2.3.2.2 CLOCK MODEL

The next step was the comparison of a strict clock, exponential relaxed clock and lognormal relaxed clock. Here, all other settings remained consistent: HKY substitution model with Gamma and Invariant sites, and a constant coalescent tree prior.

Clock model	Log MLE for PS
Strict	-5902809.593
Lognormal relaxed	-5902798.270
Exponential relaxed	-5902791.317

Table 5.4 Log marginal likelihood estimator values for different clock models

To compare the models, we must compute the difference between the log MLE values (Table 5.4) as before. First, we compare the exponential relaxed model with the lognormal relaxed:

$$2 \times -5902791.317 - 2 \times -5902798.270 \approx 13.906$$

As this value is larger than 10, we would consider that there is strong evidence for the exponential relaxed model being preferable to the lognormal relaxed.

Next, the exponential relaxed model and the strict model:

$$2 \times -5902791.317 - 2 \times -5902809.593 = 36.552$$

Again, the exponential relaxed model is preferable to the strict clock model.

5.4.2.3.2.3 RATE OF HETEROGENEITY

With the HKY model chosen as the substitution model it was then important to decide on the rate of heterogeneity that should be included. BEAST provides four options: no heterogeneity, Gamma only, Invariant sites only or both Gamma and Invariant sites. The model settings between the BEAST runs remained constant: HKY substitution model, an exponential relaxed clock model and a constant coalescent tree prior.

Rate heterogeneity	Log MLE for PS
None	-5903156.489
Gamma + Invariant sites	-5902791.317
Gamma only	-3281.630
Invariant sites	-3274.557

Table 5.5 Log marginal likelihood estimator values for different rate of heterogeneity

Comparing as before using the values in Table 5.5, invariant sites only and no rate heterogeneity:

$$2 \times -3274.557 - 2 \times -5903156.489 \approx 11799764$$

A result of larger than 10, indicates the invariant sites only is preferable to no rate heterogeneity.

Comparing invariant sites only and to both Gamma and invariant sites:

$$2 \times -3274.557 - 2 \times -5902791.317 \approx 11799033.52$$

As before, the invariant sites only model is also preferable to a model that includes both a Gamma rate of heterogeneity and invariant sites.

Finally, we compare the invariant sites only model to a Gamma only model:

$$2 \times -3274.557 - 2 \times -3281.630 \approx 14.146$$

We can also conclude that the invariant sites only model is better suited than the Gamma only model.

5.4.2.3.2.4 TREE PRIOR

Lastly, there is the choice of tree prior. BEAST provides many choices for this model so in order to reduce computation time, only a few representative models were chosen for comparison. There are two primary categories of tree prior: coalescent and non-coalescent. For the non-coalescent prior, the birth-death serially sampled prior [182, 183] was chosen. The coalescent priors are then divided into parametric and non-parametric. From the non-parametric models, the Bayesian SkyGrid prior [184, 183] was chosen (default settings of 50 parameters and 140 time at last point). Finally, as there are a wide number of parametric coalescent priors a second (exponential coalescent with doubling time [185, 183]) was chosen to compare to the constant tree prior [140, 183]. The other settings were kept consistent as: exponential relaxed clock model, HKY substitution model with invariant sites.

Tree prior	Log MLE for PS
Constant coalescent	-3274.557
Exponential coalescent	-5902717.382
Birth-Death serially sampled	Failed
Bayesian SkyGrid	-5902693.505

Table 5.6 Log marginal likelihood estimator values for different tree priors

Comparing the constant coalescent with the exponential coalescent using th values in Table 5.6, we find

$$2 \times -3274.557 - 2 \times -5902717.382 \approx 11798885.65$$

This indicates that the constant coalescent tree prior is more favourable than the exponential coalescent.

Next, we compare the constant coalescent prior with Bayesian SkyGrid prior:

$$2 \times -3274.557 - 2 \times -5902693.505 \approx 11798837.896$$

Again, the constant coalescent prior is preferable.

The Birth-Death tree prior, when used to generate a BEAST run, failed and returned an error stating that the initial tree state has a zero probability.

5.4.2.3.2.4.1 CHOSEN MODEL

As a result of the Bayes factor model comparison, the final model was an HKY with invariant sites only substitution model, a relaxed exponential molecular clock, and a constant

coalescent tree prior. This model was then used to generate the final tree by running three chains and combining the results.

5.4.2.3.2.4.2 MIXING

To assess mixing, we can look at the acceptance rates and traces for the parameters. The acceptance rates for each chain are given in Table 9.13. The majority of parameters (8/14) have an acceptance rate within a good range. The parameters: swapOperator(branchRates.categories), uniformInteger(branchRates.categories), Narrow Exchange(treeModel), Uniform(nodeHeights(treeModel)) have high acceptance rates. Whereas, Wide Exchange (treeModel) and WilsonBalding(treeModel) both have low acceptance rates.

One set of traces for a parameter from all three chains is displayed in Figure 5.11. The lines 'jump' frequently and there is no suggestion of 'sticking', where the chain repeatedly samples the same value. This is suggestive of a well-mixing chain and is repeated across the rest of the parameters (see Figure 9.1 - Figure 9.17).

5.4.2.3.2.4.3 CONVERGENCE

The ESS values for each of the three MCMC chains are listed in Table 9.14. In chains 1 and 2, there are a number of parameters, prominently likelihood and prior, that have very low ESS values. However, in chain 3, all ESS values exceed 200, considered to be a good ESS value.

Assessing the traces for the parameters across the chains, convergence should be seen as all chains would roughly end up sampling in a similar space towards the end of the chain. In Figure 5.11, although the chains start sampling from different spaces (the lines are spaced out), towards the 20,000,000th iteration of the chains, they appear to overlap suggesting they are converging. This is repeated across all of the parameters (see Figure 9.1 - Figure 9.17).

In order to improve the sampling, the three chains were resampled at a lower frequency of 300000 and a burn in of 40000000 states was removed from each of the log files and they were then combined. The resulting log file contained 541 samples. The resulting ESS values are listed in Table 5.7.

Operator (parameter)	Combined chain	Operator (parameter)	Combined chain
joint	310	frequencies3	446
prior	245	frequencies4	541
llikelihood	254	pInv	492

treeModel.rootHeight	541	uced.mean	339
age(root)	541	meanRate	335
treelength	316	coefficientOfVariation	541
constant.popSize	253	covariance	541
kappa	541	treeLikelihood	254
frequencies1	541	branchRates	-
frequencies2	541	coalescent	245

Table 5.7 ESS values for the parameters of the combined chain

These are all higher than 200, the accepted threshold for a well-converged chain.

5.4.2.3.2.4.4 RESULTING TREE

The corresponding tree files were also resampled and combined identically to the log files. The combined tree file was then used to produce the MCC tree.

The resulting tree from the combined trees file is shown in Figure 5.8. The root node and some of the nodes at the tips show high posterior support, however the majority of the central nodes show very low posterior support (0.0017). There are also a number of negative branch lengths, which is indicative of the clades occurring at an extremely low frequency within the posterior sample [186].

If the combined tree file is imported into DensiTree it is possible to see whether there is much uncertainty in the posterior sample, which would explain some of the low posterior support. As can be seen in Figure 5.9, there is a large amount of uncertainty, expressed by the ‘messiness’ of the lines. Where there is less uncertainty, the lines separate, for example, the inset in Figure 5.9, is a zoomed in section of the posterior sample of trees; here the clades are clearer and more obvious, which correlates with the higher posterior support values seen in the MCC tree, as demonstrated by the four clades highlighted in the posterior sample and MCC tree in Figure 5.10.

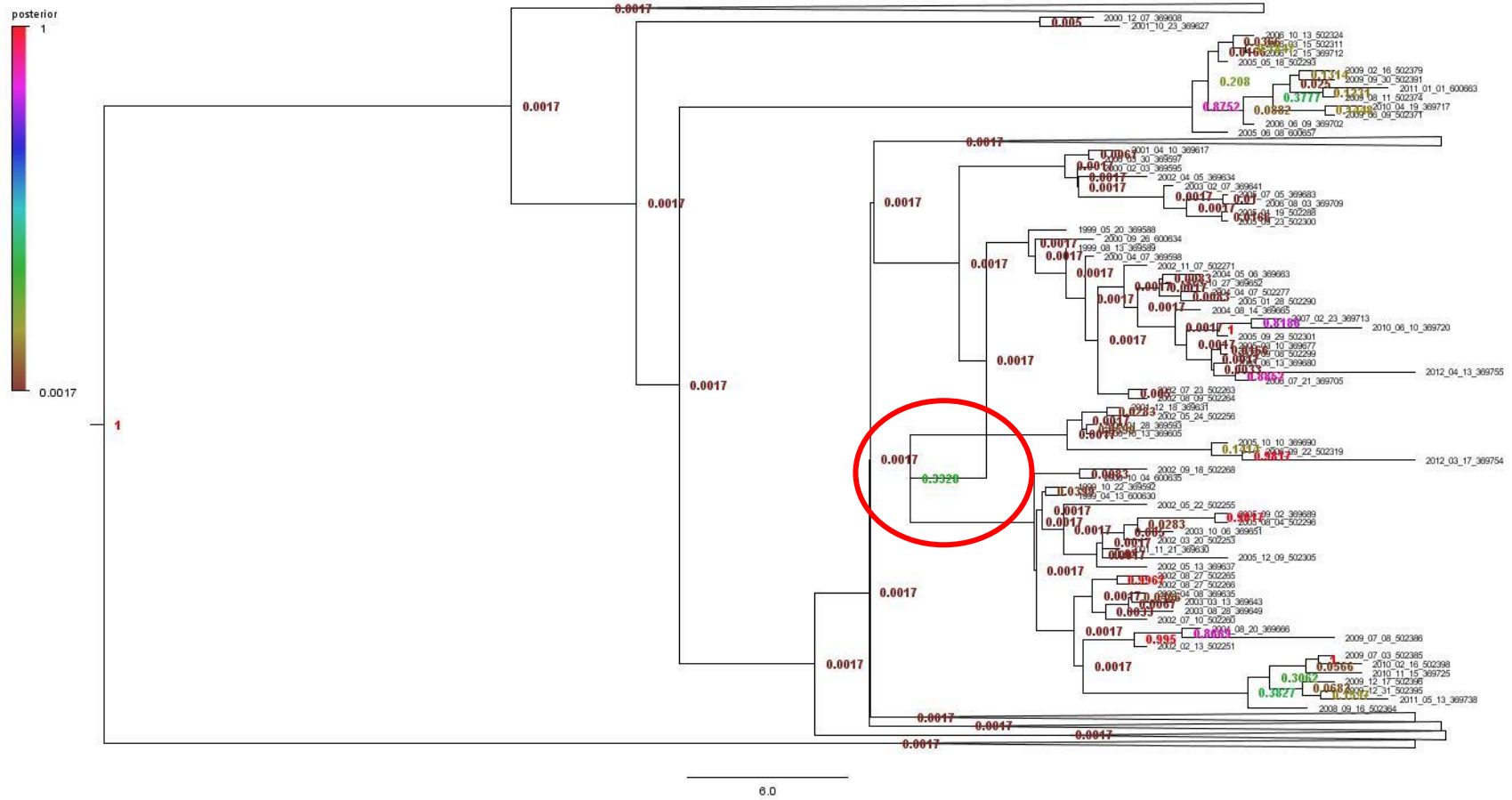

Figure 5.8 Maximum clade credibility tree from BEAST with some clades collapsed.

The posterior support for each node is shown as a node label, coloured according to the legend. A node with negative branch length is circled in red.

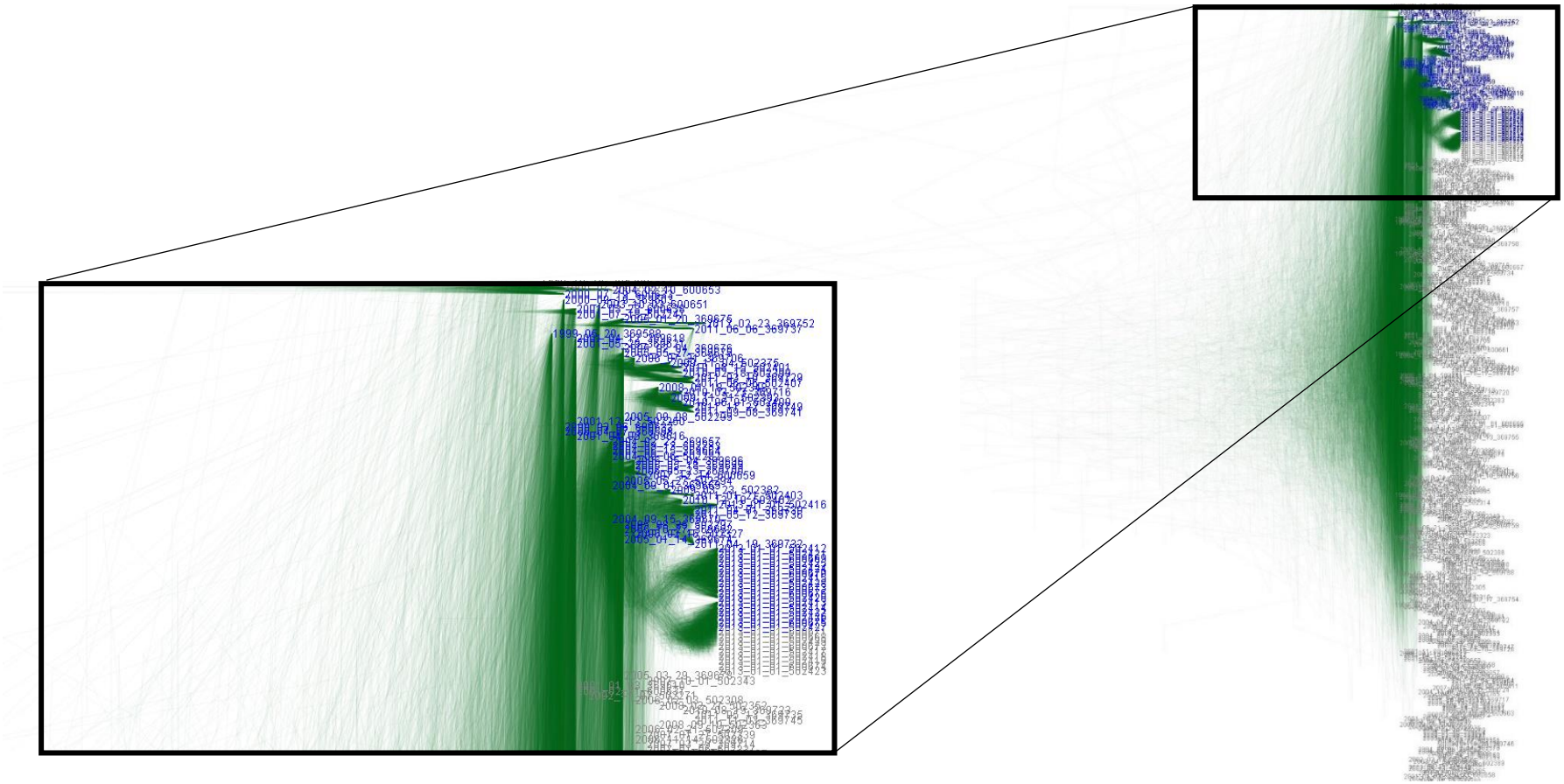

Figure 5.9 Posterior sample of BEAST trees displayed in DensiTree, where each sampled tree is one set of green lines. (Inset) The top right-hand corner is zoomed in to show clades with more certainty

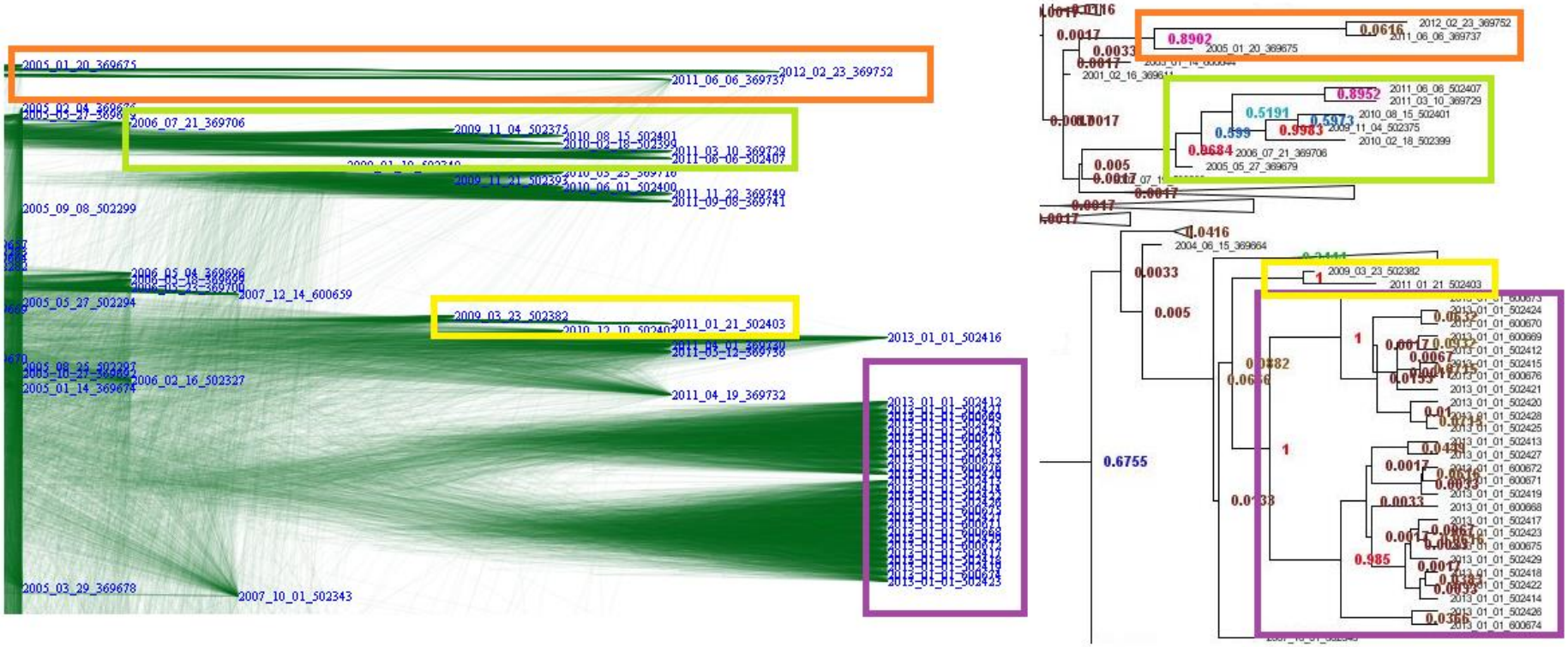

Figure 5.10 Comparison of clades from the posterior sample of trees and maximum clade credibility tree.

Left-hand figure is the posterior sample of trees, right-hand figure is the maximum clade credibility tree. Four clades are highlighted by coloured boxes in the posterior sample and correspondingly in the maximum clade credibility tree.

5.5 DISCUSSION

In this chapter, I analysed TB outbreak genomes using bioinformatic and phylogenetic techniques in order to produce a timed phylogenetic tree for use within TransPhylo.

5.5.1 KEY FINDINGS

After mapping the raw sequencing files to a genome and identifying SNPs from the assembled sequences, I attempted to construct a ML phylogenetic tree however with little success. The ML tree had extremely small bootstrap values, some as low as 5. This indicates little confidence in the tree, which can be an indication of not enough informative sites and too many taxa. Despite this, the data were determined to have a clock-like signal and a phylogenetic analysis in BEAST was performed.

Despite high ESS values and seemingly well-mixed and converged MCMC chains, when a summary tree was produced, it appeared again to have overall low posterior support for the nodes in the tree. Additionally, there were a number of branches with negative length at central nodes, which is suggestive of low sampling frequency, i.e. the clade is not well represented amongst the posterior sample. The presence of negative branch lengths in the MCC tree render it inappropriate for analyses such as TransPhylo as it would negatively impact the timings inference.

On examination of a posterior sample of trees using DensiTree, it was evident that there was a large amount of uncertainty demonstrated by a lack of clarity and definition in the overlap. When there is agreement between the sample of trees (implying more certainty), the structure of the tree is clearer.

As TransPhylo only infers transmission trees from one phylogenetic tree it is crucial that the tree be highly reliable and well-resolved, else results such as times of infection, which were key in the analysis of the Canadian outbreak (Chapter 4), will be also be highly unreliable.

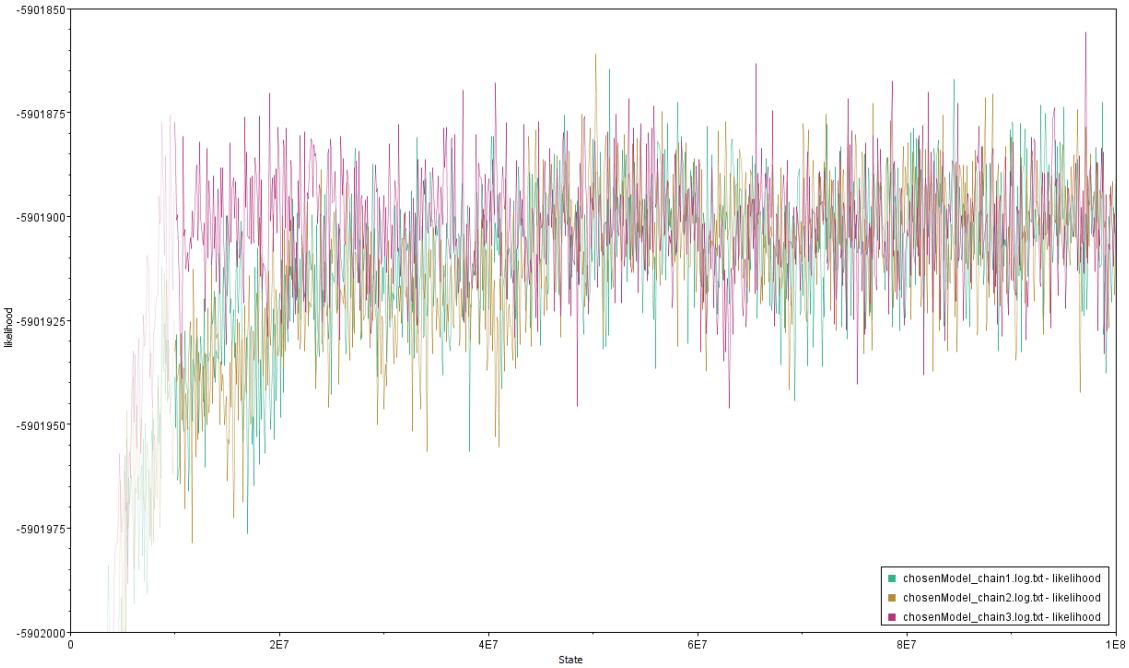

Figure 5.11 Traces for the 'likelihood' for all three MCMC chains

The consequence of the unresolved phylogenetic analysis reveals that there is either an issue with the analysis techniques used or that the data is not conducive to such phylogenetic analyses potentially due to a lack of divergence amongst the samples. Unfortunately, any inferences made with TransPhylo from the MCC tree produced would likely be highly uncertain.

If we compare the findings in this chapter with those found by Casali *et al* [61], we see that we found more diversity amongst the samples (largest group of identical sequences: 96 versus 26, maximum SNP distance: 9 versus 20) but because Casali *et al.*'s phylogenetic analysis was limited to producing a minimum-spanning tree which merely graphs the number of SNPs between samples and does not require any consideration of diversity over time and evolutionary rates etc. it is not a fair comparison.

Further phylogenetic analysis has since been attempted on these sequences [141] but using a sample set of 50 phylogenetic trees from the BEAST posterior to run TransPhylo on instead of a singular tree, in order to reduce some uncertainty from the phylogenetic tree. The authors make no comment on the reliability of the trees chosen; however, they are successful in determining 16 well-supported transmission pairs amongst the cohort, and an additional 9 pairs which involve an unsampled infector.

5.5.2 POTENTIAL CAUSES/SOLUTIONS

There are a number of potential causes of the failure to produce a high confidence phylogenetic tree with the outbreak genomes. As highlighted by the number of identical genomes in the sample and only an average of 5 SNPs difference, the diversity amongst the outbreak genomes is very low. There are potential methods that could be used to find more diversity in the sequences. Potential areas of diversity that were not included in this analysis were the PE, PPE, and PE-PGRS genes, due to the increased potential for mapping errors given these are highly repetitive areas of the genome. Therefore, a possible solution to the lack of diversity could be using *de novo* assembly for the bioinformatics analysis which attempts to create sequences without the use of a reference genome and therefore allows for inclusion of these areas. This then requires deeper sequencing, which means the sample is sequenced more times than normal, perhaps even hundreds or thousands of times, and would mean resequencing the genomes. The feasibility of *de novo* assembly on *M. tb* genomes has been demonstrated before by Bryant *et al*. [102] and therefore is a viable solution In the context of the samples that they used, however, Bryant *et al.* found no SNPs within the PE, PPE, and PE-PGRS genes and thus we may not within our data.

Alternatively, using different pipelines with different filtering thresholds will potentially increase the number of SNPs found in the data [187], as explored in Chapter 3. In addition, mixed bases i.e. potential sites of SNPs that are present only in certain coinfecting strains, could be represented using IUPAC codes instead of ‘N’, emphasising within-host diversity.

5.5.3 STRENGTHS

Whilst the resolution of finding a singular tree that could adequately describe the phylogenetic data was not reached, the analysis possesses several strengths. Firstly, the data was assessed for quality at several stages and SNPs of a low quality were removed using filtering techniques commonly used for bioinformatics analysis of TB genomes.

The settings for the maximum likelihood tree were chosen using a model finding program which involves comparison between numerous different models instead of choosing a model at random. Bootstrapping was also used to determine the confidence in the tree.

Another strength of this analysis is the use of path sampling and Bayes Factor analysis to compare multiple model settings in BEAST. This approach then allows for a rigorous, unbiased decision for which MCMC settings are best for the data, as opposed to merely choosing one at random without comparison.

Additionally, the convergence and mixing of chains was checked and optimised through lengthened MCMC chains and combining multiple chains, resulting in well-mixed and converged MCMC chains with high ESS values.

5.5.4 LIMITATIONS

Although the choice of models and settings in BEAST was done through thorough model comparison, ideally, the choice of models and settings should be made such that they explain the data well. The premise of comparing the models to each other and choosing the best one does not necessarily mean that the chosen one is a good fit, merely that it is not as bad as the others.

Another limitation is that not every setting choice is possible in BEAST, for example, the substitution models are relatively limited and the TMVe model, that was determined as the best for the data by ModelFinder, is not available in BEAST.

5.6 CONCLUSION

In this chapter, I aimed to create a timed phylogenetic tree from the raw WGS data for 415 samples available from a London TB outbreak. The resulting tree would then be available for use with the statistical inference technique as per Chapter 4 to reveal the transmission dynamics amongst the cases. The sequencing data that was available for analysis was overall good quality with only 12 samples that were unable to be sequenced. The bioinformatics pipeline used revealed a total 261 SNPs amongst the samples, with a maximum distance of 20 SNPs between any two samples. After determining the SNPs in the dataset, ML phylogenetic analysis was performed and used to assess the presence of a temporal signal in the data. Finally, after determining there was temporal signal in the data, BEAST was used to produce timed phylogenetic trees, following some analyses to determine the best settings.

Following the analysis, a summary tree was produced in order to run TransPhylo analyses on, however, the posterior support for the tree was very low for many of the nodes in the tree, suggesting little confidence in the tree. After examining the posterior sample of trees, it is likely that there was too much uncertainty in the placement of the samples, to produce a confident summary tree. More investigation would be needed in order to determine the exact reason for the large amount of uncertainty, but one possibility is that the genomes were too closely related, leaving not enough diversity to make confident assumptions about their relationships. The result of this is that there are limitations to using WGS data to track transmission if cases are too closely related.

6 MODELLING AND INTERVENTIONS (OBJECTIVE 4)

6.1 INTRODUCTION

Whilst WGS data can help us to understand patterns of transmission retrospectively [73], mathematical modelling provides a way to investigate hypotheses surrounding long-term future outbreak dynamics and potential interventions. The ability to do so has led mathematical modelling to become a valuable tool within the skillset of public health teams and outbreak investigations worldwide [66, 188].

In this chapter, I describe the construction and analysis of an original compartmental model used to describe the transmission and disease progression of TB. With the model calibrated using data from the London outbreak introduced in Chapter 5, I then introduce three interventions, including them from the beginning of the outbreak, and determine which of them could be considered the most effective in reducing the number of TB cases for this outbreak.

6.2 METHODS

6.2.1 DATA

In addition to using the TB modelling literature, there were parameters for the mathematical modelling that needed to be derived from the public health data.

The Field Epidemiological Services at PHE Victoria have been involved in the management and monitoring of the outbreak throughout and have a bespoke database created by identifying cases from various sources: an earlier, abandoned database specifically for this cluster before the London TB Register (LTBR) was started; a cluster investigation database of cases from all clusters and cases in ETS and LTBR with the correct cluster ID, as determined by their MIRU-VNTR type. The database contains 532 cases diagnosed from 1995 to the end of 2013. In the database, there is information on social risk factors as well as the date symptoms started, the date they were diagnosed, the date they started treatment and the date they finished treatment. From these variables, the length of time between symptoms starting and being diagnosed can be calculated and used within the model, as well as the time spent on treatment.

6.2.1.1 ETHICS

Due to the use of sensitive surveillance data for this work, I applied for ethical approval and this was granted by the UCL Research Ethics Committee (UCL Ethics Project ID number 6255/001).

6.2.2 COMPARTMENTAL MODEL

6.2.2.1 DESCRIPTIVE SUMMARY

A compartmental model was built for the TB outbreak in London. Individuals are "born" into the population susceptible to TB (i.e. into compartment S) at a rate equal to the baseline death rate (μ). We assume everyone is susceptible when entering the population as we are only interested in individuals who are infected with the outbreak TB strain and we assume that having an existing TB infection with another strain does not afford any immunity/increase in obtaining a second infection or progression to active TB. Individuals are then infected and progress to a latent state (L_f) at rate $\frac{\beta I}{N}$, where β represents the transmission parameter, N represents the total population and I represents all the individuals who are infectious and therefore contribute to the transmission potential (i.e. those present in the early active disease, late active disease and lost to follow-up stages). Whilst N is described as the total population size, the underlying population for this outbreak is not straightforward to define, i.e. should it be the entire population of London or just the size of the hard-to-reach population of London. As a result, the "birth rate" describes the rate of people entering the outbreak population as opposed to the real birth rate.

The first latent state compartment represents individuals who have been recently infected, defined as the period two years after infection due to the fact that during the first two years after infection individuals are at the highest risk of developing active disease [189] thus the average duration in the early latent compartment is defined as 2 years. From there they can either progress to active disease or contain the infection and enter a second latent state that can last decades (L_s), eventually resulting in disease via reactivation. Individuals in the second latent state can also return to the first latent state via re-infection at rate $k\beta I$, where k represents immunity to re-infection given that they are already infected [190]. Having two latent phases is common practice in TB modelling [136] and serves the purpose of approximately modelling the phenomenon of a very variable latent period by splitting individuals into two groups: those progressing to disease very quickly or much later in life [139].

In the model there are two stages to active disease: "early" active (A_e) where the individual is symptomatic and infectious but not sick enough to seek care and is therefore not on treatment and can transmit the infection; and "late" active (A_l), where the symptoms are much worse and seeking care is inevitable. This phenomenon, where active disease is considered a spectrum, has been biologically described by Dowdy [191] and used in modelling by Dowdy [192]. Those in the early active stage contribute to transmission at a lower proportion than those in the late active stage, as they are defined as less infectious, but could be important in the spread and containment of disease as they represent an infectious source that is harder to target. Once in the late active stage, individuals progress

to being on treatment (T). This transition accounts for the time to diagnosis and time from diagnosis to commencing an effective treatment regimen (as it is assumed that the time between diagnosis and treatment is negligible).

Given that once an individual is on treatment their ability to transmit the infection is considered to diminish within two weeks [193] those at this stage are not considered to contribute to the transmission potential. Those on treatment can progress to one of three states: susceptible (i.e. we assume no immunity to subsequent infection) having completed treatment successfully; lost to follow-up (L) where they are no longer being treated and are as infectious as before they started treatment; or active late disease representing a relapse i.e. treatment failure. The difference between becoming lost to follow-up and relapsing depends on where the individual is in the health system i.e. if they are lost to follow-up we do not know their location but assume they are no longer on treatment, whereas a return to active TB means that they are still in the health system, albeit switched to another (effective) treatment.

Whilst in the lost to follow-up state, individuals can return to being on treatment. This was included as there were examples of individuals who were lost to follow up at 12 months but returned to complete treatment (Table 6.5) and individuals who were lost to follow up at final outcome but had a second notification several years later, possibly the same episode.

All states experience a baseline death rate of μ, and those in the late active and lost to follow-up states experience an additional death rate due to TB, μ_{tb}.

6.2.2.2 MODEL EQUATIONS AND PARAMETERS

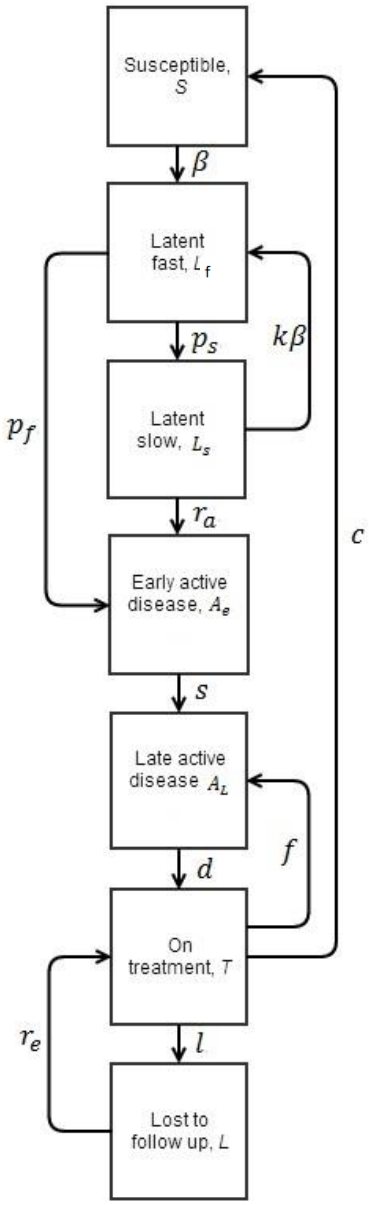

Figure 6.1 Diagram of compartmental model

Rates for moving from one compartment to another (excluding death rates and case finding rate) are denoted by a letter next to an arrow. Case-finding would be represented as an arrow from active early disease (A_e) to on treatment (T) with rate c_f. Each compartment experiences a baseline death of μ. Additionally, A_L and L experience a death rate due to tb, μ_{tb}

In this section the mathematical description of the model is presented alongside the definition of the parameters in the model (Table 6.1). The code used to build the model and can be found at GitHub (https://github.com/holliehatherell/Thesis/London Outbreak Model). The determination of the parameter values was made using a combination of the TB literature, epidemiological data analysis, Bayesian inference and modelling rules (described in Section 6.2.2.3). All rates are per person per year.

$$\frac{dS}{dt} = \mu N - \mu S + cT - \frac{\beta S(A_l + L + \phi A_e)}{N}$$

$$\frac{dL_f}{dt} = \frac{\beta S(A_l + L + \phi A_e)}{N} - \mu L_f - (p_f + p_s)L_f + \frac{k\beta L_s(\phi A_e + A_l + L)}{N}$$

$$\frac{dL_s}{dt} = p_s L_f - \mu L_s - \frac{k\beta L_s(\phi A_e + A_l + L)}{N} - r_a L_s$$

$$\frac{dA_e}{dt} = r_a L_s + p_f L_f - \mu A_e - sA_e$$

$$\frac{dA_l}{dt} = sA_e - dA_l - (\mu + \mu_{tb})A_l + fT$$

$$\frac{dT}{dt} = dA_l + r_e L - \mu T - cT - fT - lT$$

$$\frac{dL}{dt} = lT - r_e L - (\mu + \mu_{tb})L$$

Parameter	Definition	How it is calculated
$\boldsymbol{\beta}$	Rate of transmission	Bayesian inference
$\boldsymbol{\phi}$	Relative rate of transmission by individuals in early active disease compared to late active disease	Literature/Bayesian inference
$\boldsymbol{\mu}$	Baseline death rate (due to causes other than TB)	Modelling rules/outbreak data
$\boldsymbol{c}$	Cure rate	Modelling rules/outbreak data
$\boldsymbol{k}$	Immunity parameter (relative probability of reinfection for a latently infected individual versus susceptible individual)	Modelling TB literature
$\boldsymbol{p_s}$	Rate of slow progression from latent to active disease	Bayesian inference
$\boldsymbol{p_f}$	Rate of fast progression from latent to active disease	Modelling rules/outbreak data
$\boldsymbol{r_a}$	Rate of re-activation	Bayesian inference
$\boldsymbol{\mu_{tb}}$	Death rate due to TB	Bayesian inference
$\boldsymbol{d}$	Rate of diagnosis	Modelling rules/outbreak data
$\boldsymbol{f}$	Rate of treatment failure	Modelling rules/outbreak data
$\boldsymbol{s}$	Rate of becoming symptomatic	Modelling rules/outbreak data
$\boldsymbol{l}$	Rate of LFU	Modelling rules/outbreak data

r_e	Rate of re-engagement with health services	Modelling rules/outbreak data

Table 6.1 Parameters featured in the model and what they represent as well as the method of how their value has been determined

The rate of transmission is traditionally defined as the product of contact rate between susceptible and infectious individuals and the probability of infection given a contact event [68].

6.2.2.3 DETERMINING PARAMETER VALUES

The parameter values used in the model were calculated using three methods: simple rules of deterministic compartmental models that govern the rates in conjunction with data on the outbreak (see Subchapter 6.2.1), the TB literature and Bayesian inference.

Two rules used to calculate rates are:

- The average duration spent in a compartment is equal to the inverse of the sum of the rates out of the compartment
- The proportion that moves from compartment i to compartment j is equal to the ratio of the rate from i to j over the sum of all the rates out of compartment i

The immunity parameter k was fixed to a value of 0.5 [190].

As determined in the data analysis (Table 6.4), the average time spent on treatment is 352.9 days (0.97 years). Using this and known proportions of outcomes (Table 6.5) will determine cure rate (c), LFU rate (l), relapse rate (f) and death rate (μ).

$$\frac{1}{c + l + f + \mu} = 0.97$$

Rearranging gives

$$c + l + f + \mu \approx 1.031$$

453 have a final outcome of either death, treatment completion, LFU or treatment stopped (other outcomes such as unknown, still on treatment or transferred not accounted for). 371 are listed as completing treatment at final outcome (not LFU at any point) and 363 of those completed without relapsing, defined as cure, thus

$$\frac{c}{c + l + f + \mu} = 363/453$$

$$\Longrightarrow c = 0.826$$

The remaining 8 completed treatment at some point but had a second episode with the same TB strain, thus

$$\frac{f}{c+l+f+\mu} = 8/453$$

$$\Longrightarrow f = 0.018$$

45 are listed as lost to follow up or treatment stopped at 12 months and another 17 are listed as LFU at final outcome (not LFU or treatment stopped at 12 months). Thus, 62/453 individuals were listed as LFU at some point regardless of final outcome. Thus

$$\frac{l}{c+l+f+\mu} = 62/453$$

$$\Longrightarrow l = 0.141$$

The remaining 20 individuals died without being lost to follow up or curing.

$$\frac{\mu}{c+l+f+\mu} = 20/453$$

$$\Longrightarrow \mu = 0.045$$

This is roughly equal to $\frac{1}{22}$ thus representing that individuals in the outbreak population stay in the community for an average of 22 years.

With the period of early latent infection defined as 2 years, constraints can be ascertained for the rates of fast and slow progression to active disease using the rules on leaving the early latent infection compartment i.e.

$$\frac{1}{p_f + p_s + \mu} = 2$$

Rearranging and substituting in the value for μ implies

$$p_f \approx 0.454 - p_s$$

The duration spent in the early active disease compartment was assumed to be 9 months (0.75 years), based on previous modelling work [191]. This and the death rate determine the rate of becoming symptomatic (progressing to late disease), as:

$$\frac{1}{s+\mu} = 0.75$$

Rearranging and substituting in the value for μ implies

$$s \approx 1.288$$

As determined in the data analysis, the average time between onset of symptoms and starting treatment is 117 days (0.32 years). This then defines how long individuals spend in the late active disease compartment i.e. determines the diagnosis rate, d:

$$\frac{1}{d + \mu + \mu_{tb}} = 0.32$$

Rearranging and substituting in the value for μ implies

$$d \approx 3.079 - \mu_{tb}$$

Average time spent lost to follow up is 1 year, thus using rule 1

$$\frac{1}{r_e + \mu_{tb} + \mu} \approx 1$$

Rearranging and substituting in the value for μ implies

$$r_e \approx 0.954 - \mu_{tb}$$

6.2.2.4 BAYESIAN PARAMETER INFERENCE

Once relationships for the parameters had been determined using the outbreak data and modelling rules, Bayesian parameter inference with the relationships built in was used to estimate the remaining free parameters. The analysis was performed using R package *deBInfer* [194] using an MCMC method to sample from the target distribution, similar to in Chapter 4. The MCMC visits different values within a parameter space and the result is a posterior distribution of visited parameter values. A Gaussian likelihood model with a Gaussian random walk proposal distribution was used. The ranges for the free parameters i.e. $\beta,\ \phi,\ p_s,\ r_a,\ N,\ \mu_{tb}$, were then set so that the pre-determined values could not be negative e.g. as $p_f = 0.454 - p_s$, the maximum value for p_s is 0.454. If it is any larger than this then p_f becomes negative. The ranges for β and N were set wide enough that it would encompass all most likely values.

For Bayesian inference, it is necessary to propose a prior distribution from which the algorithm samples parameter values. Uniform priors were chosen for all the parameters essentially implying that there is little prior knowledge about these parameter values.

Parameter	Chain 1	Chain 2	Chain 3	Chain 4	Chain 5
$\boldsymbol{\beta}$	40	20	25	30	35

$\boldsymbol{\phi}$	1	0.1	0.3	0.5	0.8
$\boldsymbol{\mu_{tb}}$	0.45	0.1096	0.1596	0.2096	0.3096
$\boldsymbol{p_s}$	0.4	0.2	0.25	0.3	0.35
$\boldsymbol{r_a}$	0.024	0.00049	0.00099	0.0159	0.0199
$\boldsymbol{N}$	3000	100	800	1500	2500

Table 6.2 Initial parameter values for each chain

Multiple chains were run using different starting values (see Table 6.2) to ensure this was not affecting the result. Manual tuning of the proposal variances was done until the chains were mixing sufficiently as measured by the acceptance rate i.e. the proportion of values proposed by the MCMC that are accepted. The chains were run for 100,000 samples and then were combined using the $mcmc.list$ function from the coda package [195]. The convergence of the chains was then examined using the Gelman-Rubin diagnostic [196]. The calculation of the Gelman-Rubin diagnostic requires multiple chains with different starting values and random seed numbers, which are then combined. By using the *gelman.plot* function in the R package *coda* on the combined MCMC chain, the chains are compared in order to determine if they have converged. If the median and 97.5% lines decrease to 0 and remain there i.e. there are no large peaks seen over the iterations, then the chain can be considered to have converged

6.2.2.5 UNCERTAINTY ANALYSIS

Uncertainty can be approached by using the posterior distributions generated by the MCMC; it is possible to look at the 95% highest posterior density (HPD) interval for all the model variables. The model is run 100 times using 100 different parameter samples taken from the posterior parameter distributions. The 95% HPD interval is calculated for all the variables across the 100 model simulations at each time point; it tells us the credible value for the variable and the range of the variable. This can reveal how uncertainty in the model parameters can propagate through the model by looking at the range of values for the model variables over a range of input parameter values. For example, if the model variables do not vary much when using a range of different input parameter values, then it would signify that uncertainty in the parameter values does not have a large impact on the output, and thus we have robust interpretations.

This analysis was performed using the *post_sim* function from the *deBInfer* package.

6.2.2.6 SENSITIVITY ANALYSIS

In order to assess the sensitivity of certain outputs to the parameters and hence uncover which parameters have the largest effect on the outputs, sensitivity analysis was performed. The two outputs examined here were R_0 (the basic reproduction number) and incidence over time. These outputs were chosen because they are two important measures of the severity of outbreaks: the basic reproduction number gives a way of computing if an outbreak will die out or become an epidemic if introduced into a fully susceptible population [197]; and the incidence (the number of cases per unit time) gives a measure of the outbreak that is more easily interpretable from a public health point of view.

6.2.2.6.1 CALCULATION OF R0

An expression for the basic reproduction number was determined using the next generation method detailed by Yang [198]. Firstly, the states-at-infection and infectiousness are defined, in this case (L_s, L_f) and (A_e, A_L, L) respectively. The disease-free equilibrium of the ODE system is $(N, 0,0,0,0,0,0)$. Then, following Yang:

$$f = \begin{pmatrix} \frac{\beta S(\phi A_e + A_l + L}{N} \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$v = \begin{pmatrix} \mu L_f + p_f L_f + p_s L_f + \frac{k\beta L_s(\phi A_e + A_l + L)}{N} \\ \mu L_s + r_a L_s + \frac{k\beta L_s(\phi A_e + A_l + L)}{N} - p_s L_f \\ -r_a L_s + \mu A_e + s A_e - p_f L_f + c_f A_e \\ -s A_e + (\mu + \mu_{tb}) A_l + d A_l - fT \\ (\mu + \mu_{tb}) L - lT + r_e L \end{pmatrix}$$

Differentiating f and v with respect to each variable yields:

$$F = \left.\frac{\partial f}{\partial x}\right|_{(N,0,0,0,0,0,0)} = \begin{pmatrix} 0 & 0\beta\phi & \beta\beta \\ 0 & 0\ 0 & 0\,0 \\ 0 & 0\ 0 & 0\,0 \\ 0 & 0\ 0 & 0\,0 \\ 0 & 0\ 0 & 0\,0 \end{pmatrix}$$

$$V = \left.\frac{\partial v}{\partial x}\right|_{(N,0,0,0,0,0,0)} = \begin{pmatrix} \mu + p_s + p_f & 0 & 0 & 0 & 0 \\ -p_s & \mu + r_a & 0 & 0 & 0 \\ -p_f & -r_a & \mu + s + c_f & 0 & 0 \\ 0 & 0 & -s & \mu + \mu_{tb} + d & 0 \\ 0 & 0 & 0 & 0 & \mu_{tb} + \mu + r_e \end{pmatrix}$$

$$E = F * inv(V)$$
$$= \begin{pmatrix} \frac{\beta(\mu p_f + p_f r_a + p_s r_a)x}{(\mu + r_a)(c_f + \mu + s)(\mu + p_f + p_s)y} & \frac{\beta r_a x}{(\mu + r_a)(c_f + \mu + s)y} & \frac{\beta x}{(c_f + \mu + s)y} & \frac{\beta}{y} & \frac{\beta}{\mu + \mu_{tb} + r_e} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

where $x = s + d\phi + \mu\phi + \mu_{tb}\phi$ and $y = d + \mu + \mu_{tb}$. The non-zero eigenvalue of the matrix E is then the definition of R_0:

$$R_0 = \frac{\beta(\mu p_f + p_f r_a + p_s r_a)(s + d\phi + \mu\phi + \mu_{tb}\phi)}{(\mu + r_a)(\mu + p_f + p_s)(c_f + \mu + s)(\mu + d + \mu_{tb})} \quad (6.1)$$

6.2.2.6.2 DEFINITION OF INCIDENCE

Incidence in terms of the model is defined as dA_l i.e. the rate of diagnosis, d, multiplied by the number of individuals with late active disease, overall this expresses the number of individuals being diagnosed. This definition is used because it is necessary to have a measure from the model that can be compared to a quantity that was recorded for the outbreak. As it is not possible to know the true incidence for the outbreak because not everyone will be diagnosed/microbiologically confirmed/genotyped, the best available measure of disease incidence is the number of cases that were confirmed as part of the outbreak using genotyping.

6.2.2.6.3 RELATIONSHIPS BETWEEN OUTPUTS AND PARAMETER VALUES

To qualitatively get an idea of the relationships between the outputs (R_0 and incidence) and the model parameters, scatterplots were generated from parameter sets determined by Latin hypercube sampling (LHS) [199]. In general, LHS works as follows: if there are N parameters for which there needs to be M sets of values sampled and each parameter has a range within which sampling must be done then each parameter range would be split into M smaller equal length intervals, each with equal probability. One parameter value is then chosen at random from each of the M sampling intervals. LHS was performed using the *pse* package in R, set to sample 200 different sets of parameters (for bootstrapping) from a uniform distribution where the ranges were the HPD intervals from the MCMC chains. As incidence is a time series variable, scatterplots were plotted at multiple time points to see whether different parameters are important at different time points in the outbreak.

Partial rank correlation coefficient (PRCC) analysis was then used for sensitivity analysis using the *plotprcc* function on the same LHS samples used for the scatterplots. The PRCC is more informative than the scatterplots as it considers the effect of varying the other parameters simultaneously, rather than keeping them fixed [200]. The result, a value

between -1 and 1, signifies the strength of the correlation between the input and output, meaning the model is more sensitive to parameters with a PRCC value closer to -1 or 1.

6.2.2.7 INTERVENTIONS

The model was then used to evaluate the effectiveness of three interventions: an active case finding service, a reduction in the rate of loss to follow-up and an increase in the rate of re-engagement. Active case finding was modelled by a transition from early active disease to on treatment $(A_e \rightarrow T)$ at rate c_f. This is assuming that a service that actively goes into the community to screen for TB, such as Find and Treat [37], would identify disease at an earlier stage than passive case finding, i.e. before an individual's symptoms were serious enough to seek care and potentially even before they become symptomatic, therefore hopefully treating the individual before they become fully infectious or infectious at all. This premise was demonstrated by Storey *et al* [201] where those found via active case finding were more likely to be sputum smear negative than those presenting via passive case finding. By introducing an active case finding rate to model the effect of an organisation like Find and Treat we are ignoring the inevitable increase in rate of diagnosis and re-engagement that would come from screening the community. Reducing loss to follow-up was modelled by a decrease in the parameter l, and increasing re-engagement was modelled by an increase in the parameter r_e.

To examine the impact of interventions on R_0 and incidence, a second sensitivity analysis was performed on the model, solely varying the intervention parameters, $r_e, l,$ and c_f [189]. Ranges for the parameters are detailed in Table 6.3.

Parameter	Baseline (no intervention)	Intervention value	Justification of intervention value
c_f	0	1.513	Curtis *et al* [202]
r_e	0.563	1.127	
l	0.141	0	Maximum effect of stopping LFU (i.e. no one becoming LFU) vs LFU rate determined from data

Table 6.3 Intervention parameters and their range of values and the source of the values.

6.3 RESULTS

In the following chapter, the results of sensitivity and uncertainty analyses for model are presented, alongside the results of the MCMC analysis. Finally, the results of the interventions in the model are presented.

6.3.1 OUTBREAK EPIDEMIOLOGY

In this section, the results on the time from symptom onset to treatment initiation and treatment outcome as determined from the outbreak data are presented.

The average time spent on treatment for the cases in this outbreak, 352.9 days (Table 6.4), is longer than the standard 6 months for drug-susceptible TB [203] because the outbreak strain is isoniazid-resistant meaning isoniazid cannot be used in the drug regimen and a longer regimen with an alternative drug must be used instead. The recommendation for this outbreak was 1 year of treatment [147].

Mean duration between onset of symptoms and starting treatment	117 days (0.32 years)
Mean treatment duration	352.9 days (0.97 years)

Table 6.4 Mean time symptoms were present before starting treatment and average time on treatment before final outcome within the outbreak.

PHE determined the median length of time from diagnosis to starting treatment for pulmonary TB cases diagnosed in 2017 to be 79 days (interquartile range 39-143) [1], showing that the cases in this outbreak took slightly longer to start treatment.

68 records (12.8%) were missing both a start and end of treatment date and 148 records were missing a symptom onset date (27.8%).

The treatment outcomes after 1 year of treatment and the final known outcome for individuals in the outbreak are shown in Table 6.5. Overall, the rate of completing treatment within the outbreak is high with 70% of individuals in the outbreak having finished treatment at the final known outcome. Roughly 7% of individuals were LFU at 1 year, although 5.6% of those then went on to finish treatment having re-engaged. 21 individuals (3.9%) had a final outcome of death. 49 individuals had no data on their 1 year or final outcome.

Outcome at 1 year	N (%)	Final outcome	N (%)	Final outcome	N (%)	Outcome at 1 year	N (%)
Completed	213 (40)	Completed	213 (100)	Completed	374 (70.2)	Completed	213 (57)
Unknown	177 (33.2)	Completed	99 (55.9)			Unknown	99 (26.5)
		Unknown	49 (27.7)			On treatment	59 (15.8)
		Lost to follow up	13 (7.3)			Lost to follow up	2 (0.5)
		Died	6 (3.4)			Treatment stopped	1 (0.3)
		Transferred out	5 (2.8)	Unknown	49 (9.2)	Unknown	49 (100)
		Treatment stopped	3 (1.7)	Lost to follow up	44 (8.26)	Lost to follow up	30 (68.2)
		On treatment	2 (1.1)			Unknown	13 (29.5)
On treatment	72 (13.5)	Completed	59 (82)			On treatment	1 (2.3)
		On treatment	11 (15.3)	Died	21 (3.9)	Died	14 (66.7)
		Treatment stopped	1 (1.4)			Unknown	6 (28.6)
		Lost to follow up	1 (1.4)			Lost to follow up	1 (4.8)
Lost to follow up	36 (6.8)	Lost to follow up	30 (83.3)	Transferred out	18 (3.4)	Transferred out	11 (61.1)
		Transferred out	2 (5.6)			Unknown	5 (27.8)
		Completed	2 (5.6)			Lost to follow up	2 (11.1)
		Died	1 (2.8)	On treatment	13 (2.4)	On treatment	11 (84.6)
		Treatment stopped	1 (2.8)			Unknown	2 (15.4)
Died	14 (2.6)	Died	14 (100)	Treatment stopped	13 (2.4)	Treatment stopped	8 (61.5)
Transferred out	11 (2.1)	Transferred out	11 (100)			Unknown	3 (23.1)
Treatment stopped	9 (1.7)	Treatment stopped	8 (88.9)			On treatment	1 (7.7)

Completed	1 (100)	Lost to follow up	1 (7.7)

Table 6.5 Treatment outcomes for outbreak cases.

In the first column, the outcomes known at one year for the cases are listed. The second column lists the final known outcomes broken down within each outcome at one year i.e. it states the final outcome for each case that was 'on treatment' at one year and so on. The third and fourth columns are similar but the opposite: the final outcome is listed in the third column and then broken down by outcome at one year.

6.3.2 BAYESIAN INFERENCE

During the first round of MCMC runs it became apparent that the chain was sticking and not sampling the full parameter space. After some investigation, it was identified that the proposal function was often proposing values outside of the prior distributions for the parameters and thus many proposals were not being accepted. By changing the standard deviation of the Gaussian random walk proposal distribution, there was a vast improvement in the mixing of the chain as evidenced in Figure 6.2.

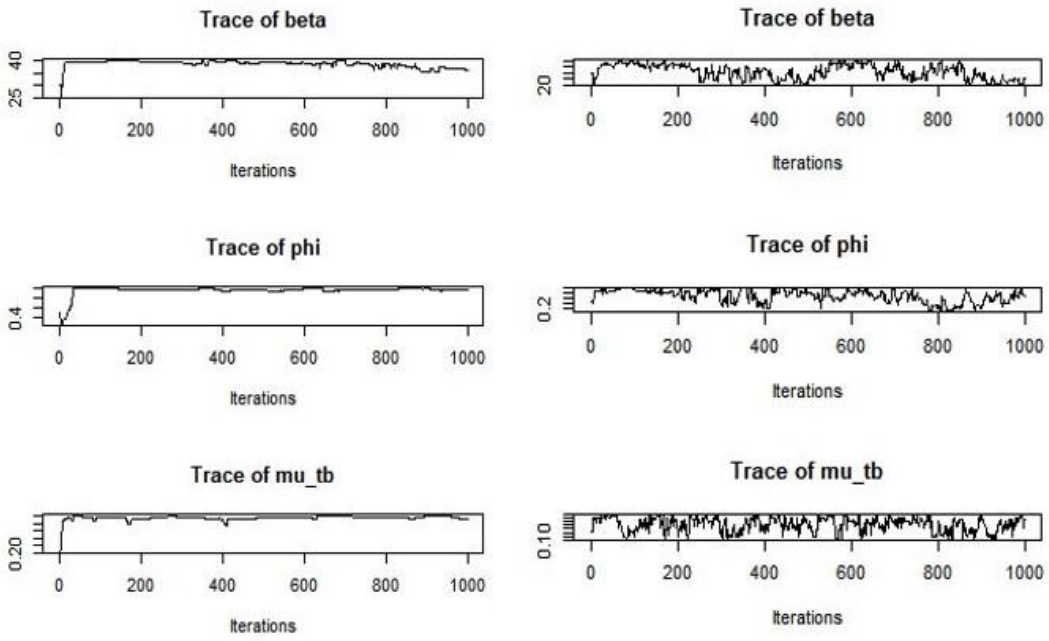

Figure 6.2 Resolution of chain sticking in the initial outbreak model
Comparison of chain mixing with standard deviation of the proposal distribution set to 1 (left) and 8 (right).

In order to fine tune the standard deviation, the value was also sampled in the first set of MCMC chains. The inferred value for the standard deviation that provided the best chain mixing was determined to be ≈ 8. The chains were then re-run with the standard deviation fixed at 8.

Parameter	Value	Acceptance rate	Parameter	Value	Acceptance rate
$\boldsymbol{\beta}$	27.458	0.451	p_s	0.259	0.343
$\boldsymbol{\phi}$	0.430	0.351	r_a	0.012	0.363
$\boldsymbol{\mu_{tb}}$	0.391	0.358	N	408	0.466

Table 6.6 Results from the combined MCMC chain.

Value is the most likely value of each parameter as determined by the MCMC chain (the chain 'visits' this value the most when sampling the parameter space) and the final column shows the acceptance rate of each parameter, which correlates to how well the chain is mixing for that parameter (i.e. the proportion of times the chain accepts and moves to the proposed value)

Once the standard deviation was fixed, the MCMC was re-run and the results are tabulated in Table 6.6. The 'value' column shows the most likely values of the parameters, i.e. the value the chain 'visits' most when sampling the parameter space. This then provides the parameter values for the model.

The next step was to check how well the chain was mixing, i.e. how well it sampled the parameter space, using the acceptance rate. A 'good' acceptance rate is generally considered to be between 0.2 and 0.3, corresponding to 20-30% of proposed values being accepted. This is to facilitate the right amount of mixing: a high acceptance rate would mean the chain is jumping around too much and not revealing the underlying distribution and a low acceptance rate would mean the chain would often get stuck and mix slowly. Whilst not all the parameters have acceptance rates between 0.2 and 0.3, the rates are sufficiently far away from both extremes.

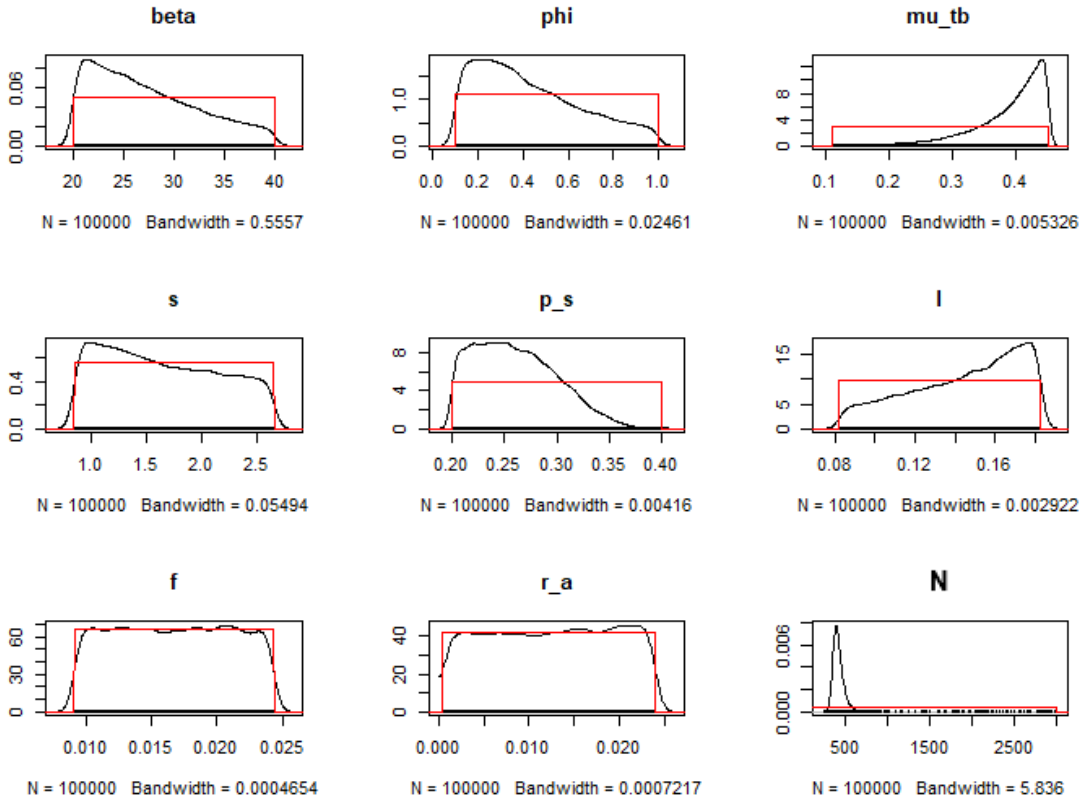

Figure 6.3 Graphs showing the prior density (red line) and posterior density (black line) for each parameter sampled by the MCMC (chain 1).

The prior densities are all uniform, which is why the density remains the same for all values over the possible range. The posterior densities vary across the parameters: f and r_a both have fairly uniform posterior densities which suggests there is some uncertainty around these parameters, whereas μ_{tb} and N have clearly defined peaks (Figure 6.3).

Figure 6.4 shows that all the chains converged.

6.3.3 UNCERTAINTY ANALYSIS

As can be seen in Figure 6.5, the variables S, L_f and L_s all have wide 95% highest density intervals, meaning that there is more uncertainty around these variables as opposed to the others.

6.3.4 SENSITIVITY ANALYSIS

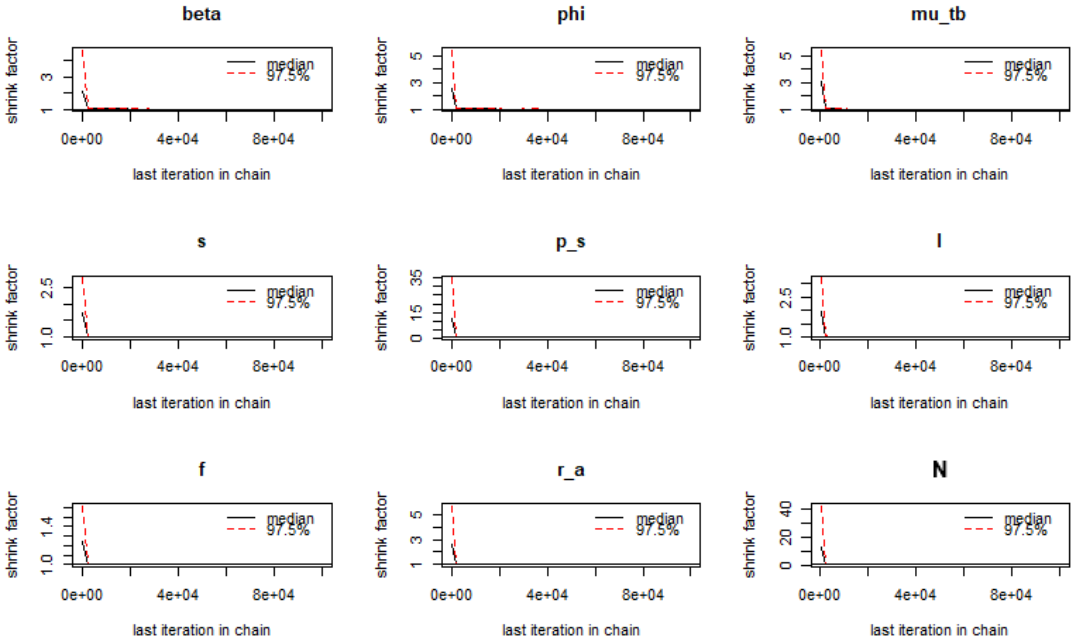

Figure 6.4 Gelman-Rubin plot showing convergence between the five MCMC chains

Both scatterplots and PRCC values expressing the relationship between the parameters and R_0 both demonstrate that β, r_a and ϕ all have strong positive correlations with R_0, meaning that an increase in the effective contact rate, reactivation rate and the infectiousness of individuals with early active TB disease would increase the chances of an epidemic occurring (Figure 6.6 and Figure 6.7). On the other hand, p_s and s are both strongly negatively correlated with R_0 meaning that increasing the rate of slow progression and becoming symptomatic would decrease the chances of an epidemic occurring. If the rate of slow progression is increased and r_a remains small this would mean that more individuals have extremely long latent periods, potentially dying of non-TB related causes before reactivating and transmitting the infection, hence the strong negative correlation.

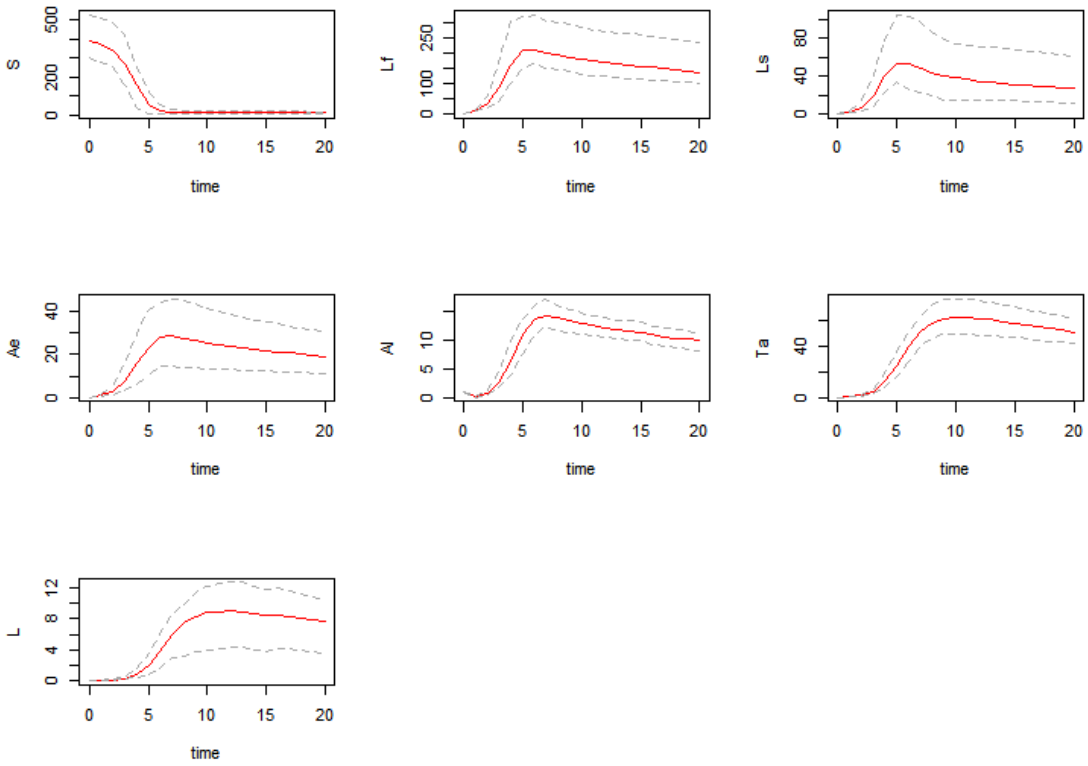

Figure 6.5 Sensitivity analysis of model using MCMC posterior samples for parameters

The red lines indicate the median posterior trajectory (the trajectory using the median values) and the dashed grey lines indicate the 95% highest posterior density interval for all model variables from 100 simulations using 100 posterior parameter samples (chain 4)

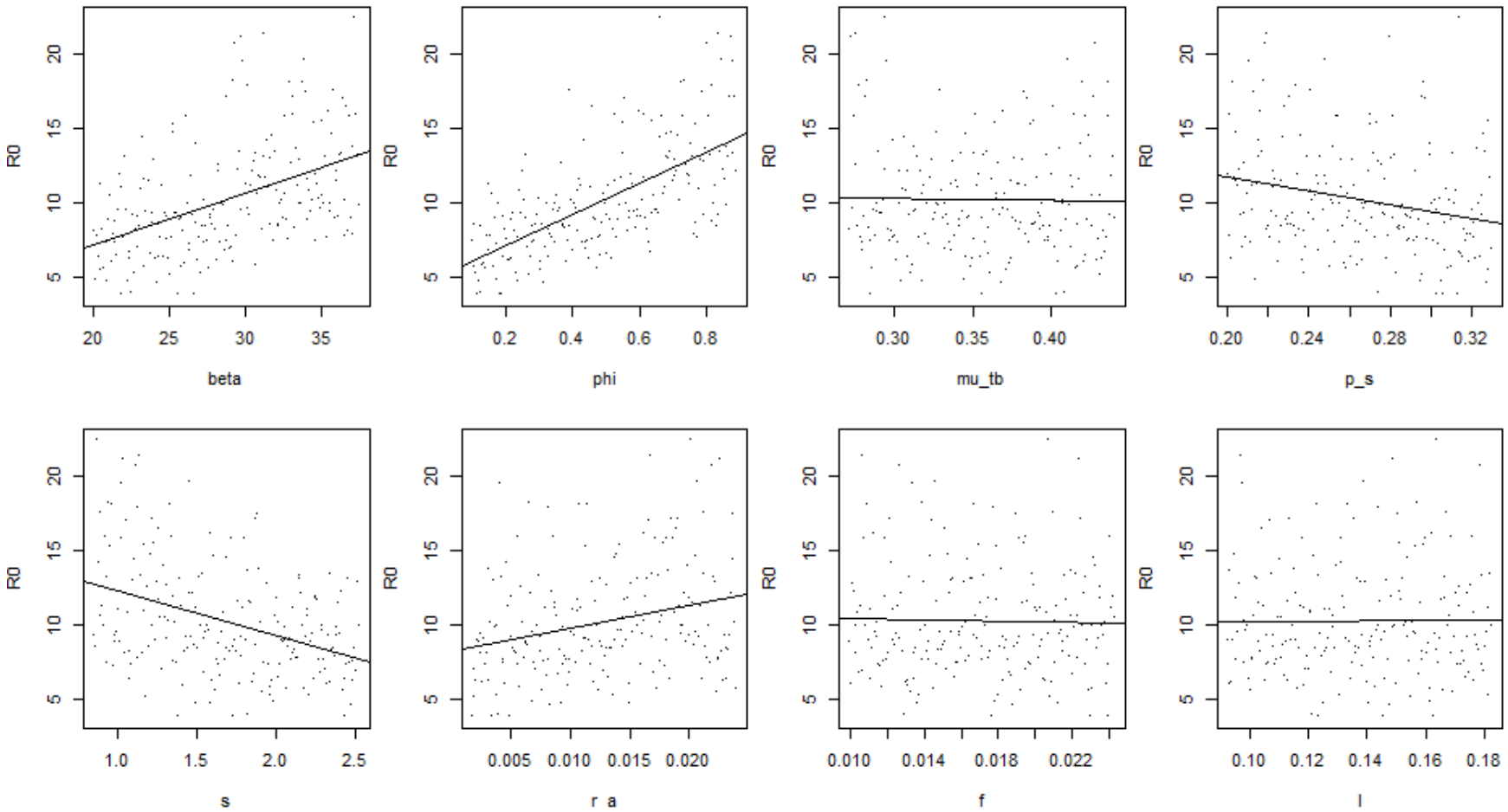

Figure 6.6 Scatterplots for all parameters versus R_0

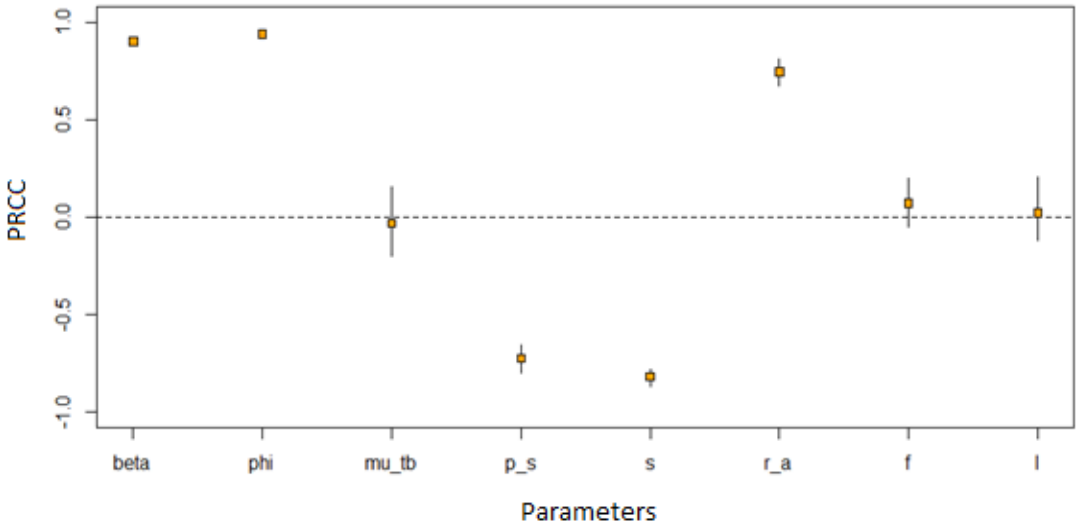

Figure 6.7 PRCCs for each parameter with respect to the basic reproduction number.

The confidence intervals shown in this plot by lines are generated by bootstrapping

Figure 6.8 shows how the parameters affect incidence over time. A number of the parameters, namely the reactivation rate, r_a, and the treatment failure rate, f, have very little correlation with incidence at any time. Other parameters, such as the transmission rate, β, relative transmission rate, ϕ, and the progression to late active disease rate, s, are initially positively correlated with incidence but over time become uncorrelated. Interestingly, the positive correlation between s and incidence is converse to the negative correlation found with R_0. Similarly to R_0, incidence is negatively correlated with the rate of slow progression to

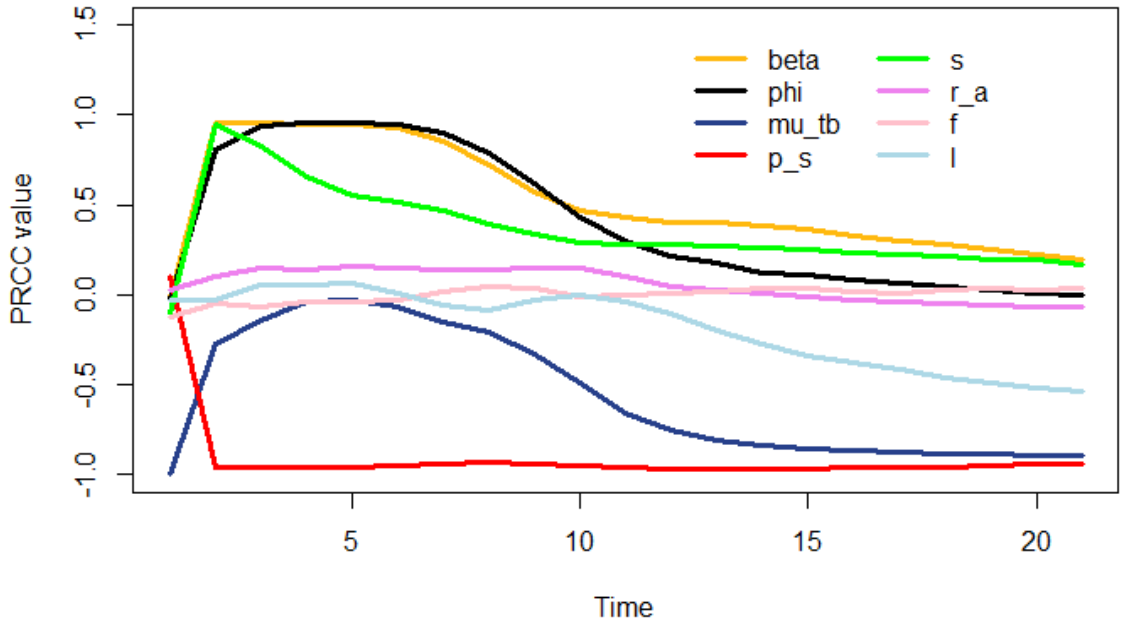

Figure 6.8 PRCCs for each parameter over time with respect to incidence

active disease, p_s, throughout the course of the simulation.

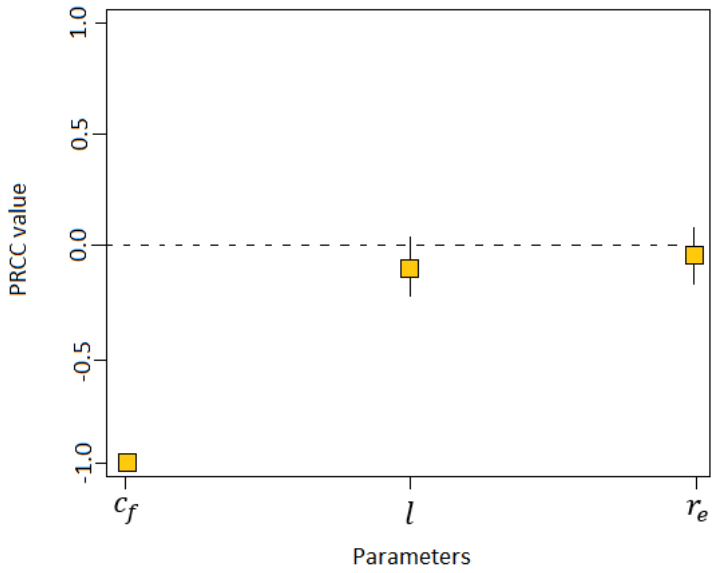

Figure 6.9 PRCCs for the three intervention parameters with respect to R_0, all other parameters held fixed.
Confidence intervals determined via bootstrapping are demonstrated with vertical lines

The results of the sensitivity analysis performed on the intervention parameters with respect to R_0 can be seen in Figure 6.9. The active case finding rate, c_f, is extremely negatively correlated with R_0, with a PRCC value of almost 1, implying that an increase in the case finding rate will decrease R_0. The LFU rate, l, and re-engagement rate, r_e, have little to no correlation with R_0, which is expected as they do not appear in the expression for R_0.

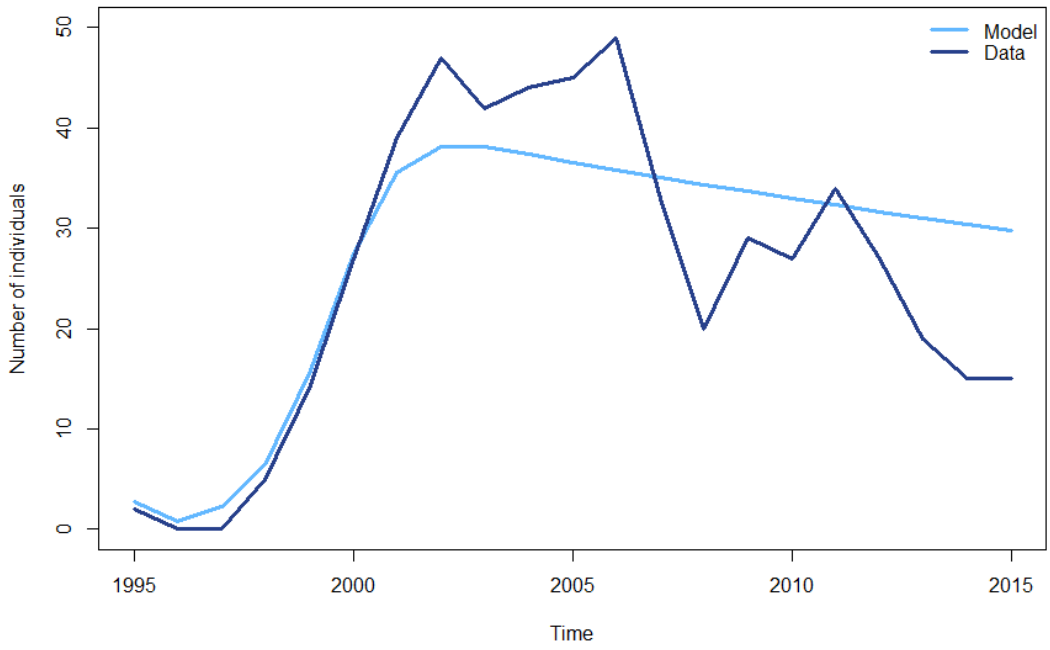

Figure 6.10 Incidence from model compared to number of cases confirmed as part of the outbreak

6.3.5 MODEL

Firstly, it is important to examine the fit of the model to data, which can be examined using Figure 6.10.

The model seems to fit the data well in the first five years during the growth of the outbreak (1995-2000), suggesting the transmission rate has been well fitted. However, the model then appears to diverge slightly from the data, not quite recapitulating the peak seen. The incidence levels (number of new cases) in 2016 are 15 (data) and 29 (model).

When examining the different model compartments over time without interventions (Figure 6.11) it is apparent that the population are quickly infected but, due to the low rate of progression to disease, the majority of individuals remain latent.

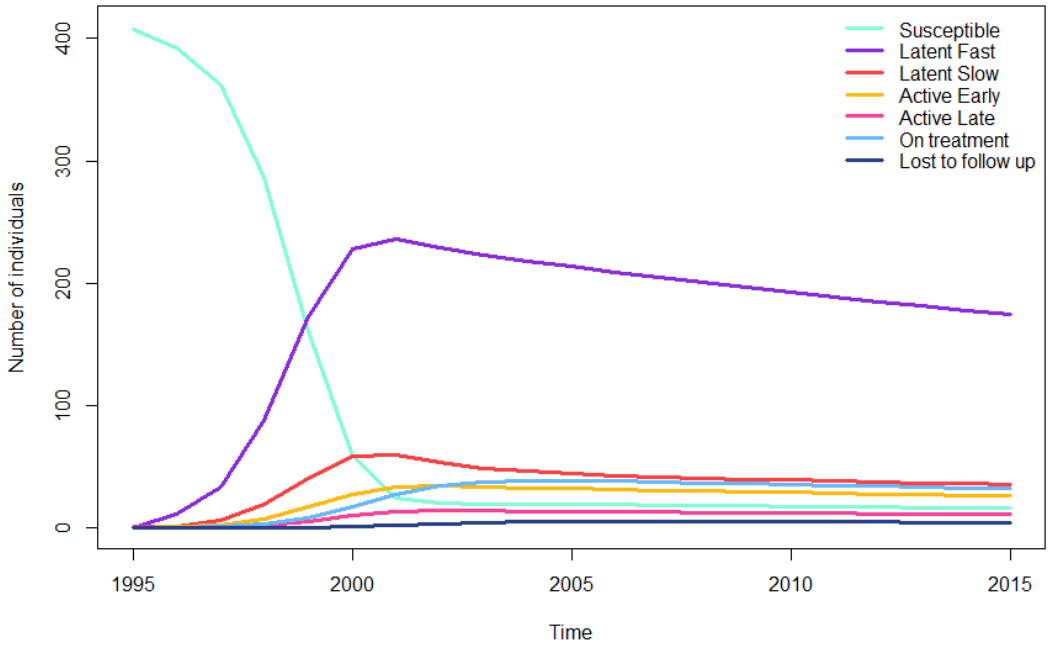

Figure 6.11 Different compartmental groups using baseline parameter values

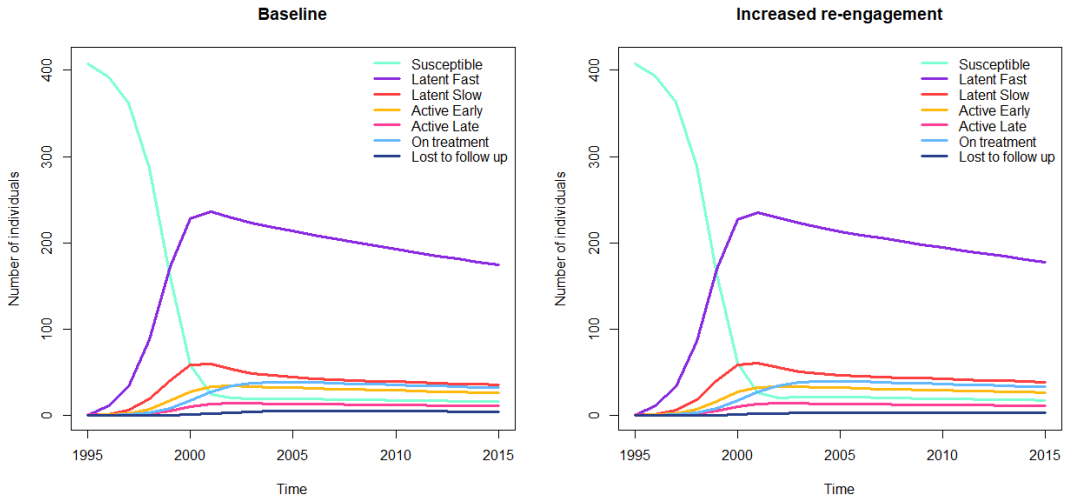

Figure 6.12 Model results with the re-engagement intervention parameter r_e set to the baseline value (0.563) and increased re-engagement value (1.127)

After the above work, a model had been established into which the interventions could be added. These are now explored in the following sections.

Firstly, the re-engagement parameter was altered to explore the effect of increasing re-engagement versus the baseline. Figure 6.12 depicts the model variables when the re-engagement parameter is set to the minimum and maximum values as mentioned in Table 6.3.

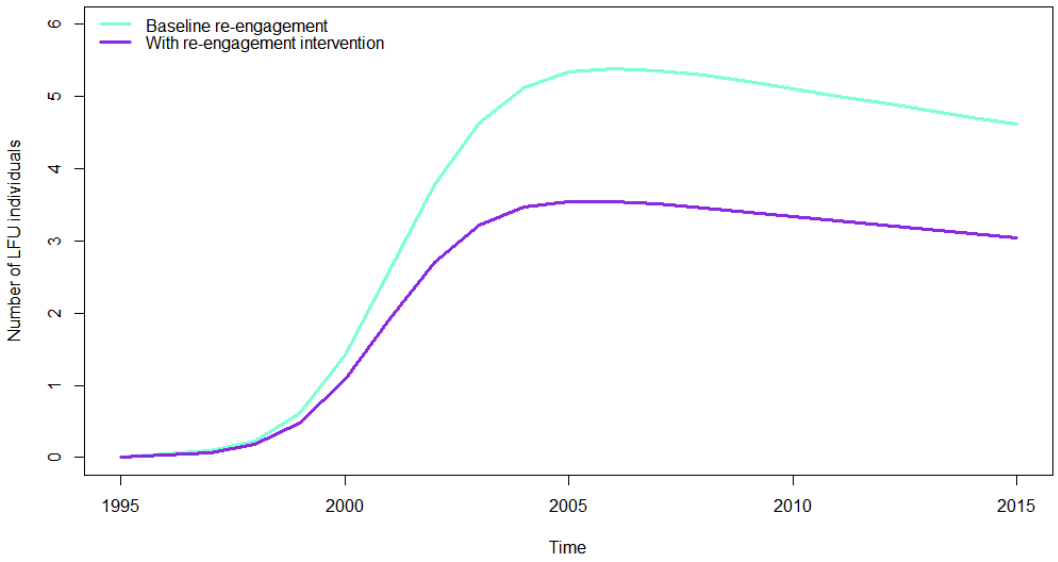

Figure 6.13 Comparison of the number of lost to follow up individuals over time with baseline re-engagement and increased re-engagement as an intervention

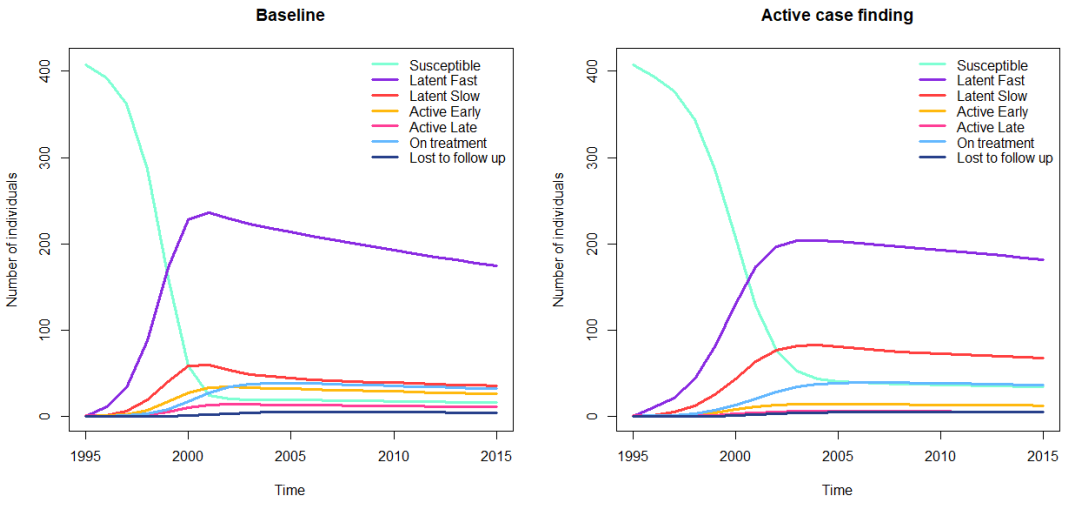

Figure 6.14 Model results without any interventions (left) and with the case finding intervention parameter c_f set to 1.513 (right)

There is a small decrease in the number of LFU individuals when the re-engagement is increased versus baseline re-engagement (5.39 versus 3.54 in 2005, see Figure 6.13) as they are being reintroduced back onto treatment and potentially completing their treatment.

When case finding is introduced as an intervention, the number of individuals in the active early and late disease compartment inevitably is reduced year on year (e.g. active early:

34.15 versus 13.33 in 2002; active late: 14.17 versus 5.40 in 2002, see Figure 6.14), as they are moved at a faster rate from active early onto treatment, circumnavigating the active late stage. The number of individuals in the latent slow and susceptible compartments is increased compared to at baseline (e.g. in 2015 latent slow: 35.42 versus 68.03; susceptible: 16.06 versus 34.51).

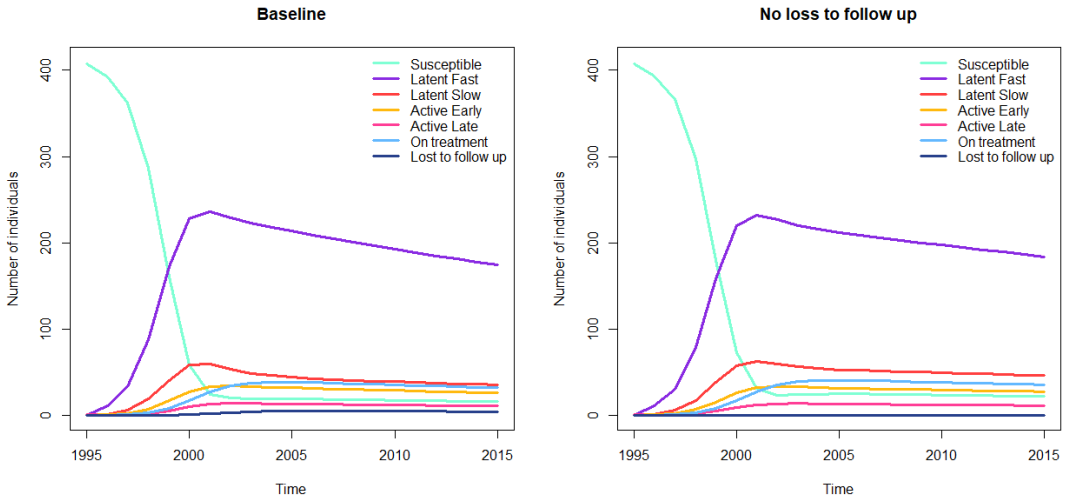

Figure 6.15 Model results without interventions (left) and with the loss to follow up intervention parameter l set to 0 (right)

By reducing the rate of loss to follow up to zero, the number of latent slow, susceptible and on treatment individuals is increased versus at baseline (e.g. in 2015, susceptible: 60.03 versus 16.06, latent slow: 101.35 versus 35.42, on treatment: 35.40 versus 32.05, see Figure 6.15).

As lost to follow up and re-engagement do not appear in the expression for R_0, it is not possible to reduce R_0 to below the threshold for an epidemic (i.e. < 1) through these interventions. However, the active case finding parameter c_f does appear in the expression for R_0, meaning that we can find a value for c_f that would prevent the outbreak from occurring. If we solve the inequality $R_0 < 1$ using Equation 6.1 with all parameter values substituted save c_f, we find $c_f > 12.1871$. Employing a case finding rate of 12.871 is equivalent to having to actively case find 90% of all cases with early active disease within 9 months before they progress to late active disease.

Model	Number of new cases in 2015	Percentage change from baseline (no intervention)
No intervention	26	-
Case finding intervention	15	-42%

Lost to follow up intervention	31	+19%
Re-engagement intervention	30	+15%

Table 6.7 Comparison of the number of new cases in 2015 predicted by the model under different intervention scenarios

Given the results displayed in Table 6.7, we can conclude that active case finding is more effective than reducing the LFU rate or increasing re-engagement rate of LFU individuals.

6.4 DISCUSSION

In this chapter, a compartmental model was developed and fit to incidence data from an ongoing outbreak centred in London. Using sensitivity and uncertainty analyses, I determined which parameters most affected the outcomes of R_0 and incidence. Finally, I looked at the effect of including three different interventions: active case-finding, reducing loss to follow up, and increasing re-engagement into the model and examined the effect on incidence.

6.4.1 KEY FINDINGS

When fitting the model to incidence data, the model does not recapitulate the second peak seen in the real-life incidence curve (Figure 6.10). This is not possible when using such a simple model and would require perhaps a time-dependent transmission parameter or multiple populations being modelled however there is not enough evidence in the data to suggest what the mechanism behind the second peak might be, without further investigation. Smith *et al.* [153] suggested that the initial peak may represent infections which rapidly progressed to disease and the second represents infections with a longer period of latency or alternatively, the second wave of cases may have resulted from a second period of intensive transmission. These hypotheses could be better tested if a transmission tree could be constructed. As a result, this was not explored, and the model essentially averages out the peaked data.

The main finding of the intervention analysis is that increasing re-engagement and reducing loss to follow up seem to have very little impact on the control of the outbreak. This is likely because the number of individuals who are lost to follow up are quite a small proportion of the population. Increasing the active case finding rate makes the largest difference to the incidence of cases, likely because it impacts upon the early active disease compartment and helps to avoid a period of infectiousness.

Increasing re-engagement and case-finding rates results in an increase in the number of individuals who are in the latent slow (L_s) compartment (and susceptible, but only very slightly for re-engagement). Presumably this is because increasing the re-engagement and case-finding leads to an increased number of individuals completing treatment and returning to the susceptible compartment, where they are then able to be re-infected and this then leads to a build-up in the latent slow compartment as individuals are more likely to progress slowly than quickly and stay there for a long amount of time.

In the early stages of the outbreak, active case finding is associated with an increase in incidence, as intuitively, it is increasing the number of individuals who are diagnosed. However, this quickly drops off after a few years and becomes associated with a decrease in incidence as more individuals are being treated and becoming susceptible and the number of infectious individuals decreases.

When examining the results of the intervention analysis the sensitivity and uncertainty analysis results must be considered. Rate of failure of treatment (f) and rate of re-activation (r_a) are uncertain as the posteriors from the Bayesian inference are relatively uniform over a large period of values meaning those values are equally likely. However, both R_0 and incidence seem relatively insensitive to both parameters (Figure 6.7 and Figure 6.8) most likely because they are such small rates, thus we can be confident that they should have little effect on the outcome of the interventions. This is similarly concluded by Fojo *et al*. [204] who also employ a Bayesian parameter inference method for a TB transmission model but for New York City and find that for most of their rates of progression to active disease after long latency periods (2+ years) their posterior distributions are similarly as uniform. Fojo *et al.*'s PRCC analysis shows rate of progression to active plays a small role in driving TB incidence (for foreign-born at least).

6.4.2 STRENGTHS

One strength of the study is that the design of the compartmental model has been built specifically for the outbreak and therefore includes the key patterns seen in the data e.g. re-infection, loss to follow up and re-engagement. The structure was also formed based on key features of good TB models, i.e. the implementation of latency was determined after taking into consideration the findings of Menzies *et al.* [67], who determined that models that included two latency compartments in series outperformed models that included them in parallel or no latency at all, in terms of model predictions versus empirical data. Even though it was built specifically for the London outbreak, TB outbreaks and populations usually share similar dynamics and structures that the model should be suitable for other TB outbreaks. Another strength of the study was the determination of parameter values directly from the data instead of exclusively from the literature.

Where there were no data to directly estimate parameter values, Bayesian inference was performed to sample within a range of parameter values. The inference was performed rigorously, with multiple MCMC chains and convergence and mixing assessed. The use of MCMC provided the opportunity to investigate the uncertainty of the parameter values and test the robustness of the model predictions [205].

As well as assessing uncertainty, it is important to assess the sensitivity of the important model outputs to the parameter values. This not only helps to test the robustness of the model but also highlights which parameters may be key targets for interventions [205]. PRCC sensitivity analysis was chosen as it accounts for the variability of other parameters to help temper interaction effects, provides a quantitative measure that can be easily compared between parameters and it allows for temporal assessment of parameters too.

Finally, another strength of the study is the investigation of realistic interventions i.e. interventions that are currently implemented as part of the strategy the UK uses to tackle TB. This then allows us to effectively translate the findings of this analysis directly into public health outbreak investigations tools.

6.4.3 LIMITATIONS

Some limitations to the modelling study stem from the parameters and their values. Firstly, it is difficult to evaluate the "accuracy" of the value used for β. However, Fojo *et al.*'s values for β ranged from 5 to 68.74 per person with active TB per year in their various scenarios, with a value of 31·5 per person with active TB per year for their scenario which modelled the 1980s increase closest to the value used in this study. Given that New York City is comparable to London in that it is a high TB incidence metropolis with high immigration rates within a low TB incidence country and that the value found from the Bayesian inference sits within their range, it seems to be a valid estimate.

Additionally, birth rate is not equal to overall mortality rate as death due to TB is not included in the birth rate. This helps when fitting to incidence data as the model is not forced mathematically to reach equilibrium but does mean that the total population number will fluctuate over the course of time.

Another potential issue with data fitting is the assumption that the number of individuals starting on treatment is equal to the number diagnosed. This is not always true as some diagnosed individuals do not start treatment, thus the number of cases within the outbreak is underestimated and these cases are likely to be infectious as they remain untreated. In addition, it is possible that individuals in the outbreak were initially treated with an ineffective treatment (with isoniazid) if the resistance to isoniazid had not been tested for. In the data, 12.8% were without a recorded start or end of treatment date, defining the upper limit on the proportion of individuals who did not have any treatment at all (we cannot rule out that the

dates may just not have been recorded). There is no information in the data on whether a case was initially given the incorrect treatment regimen. The effect of an untreated contingent could be investigated in the model by including a rate from active late disease to lost to follow up. This would of course increase the number of lost to follow up individuals. A similar structure has been explored in the TB literature: Mandal *et al.* [206] included a compartment for individuals who never seek treatment.

The model structure was determined by a need to include elements of the stages of TB infection and disease as well as the characteristics of the outbreak, i.e. lost to follow up and re-engagement. Other additions to TB models include some form of immunity to TB either innate or post-treatment and development of further drug resistance but the lack of available data on these factors did not justify the additional model complexity. The inclusion of immunity is known to reduce the incidence as it provides protection from infection however there is debate around the extent of protection provided by previous infection and measures such as the Bacille Calmette-Guerin vaccination, which makes implementing it in a model difficult. Including further drug resistance requires twice the number of compartments to represent an individual at every stage but with a different strain, this obviously increases complexity. Given that there is only knowledge of a very small number of cases with an MDR version of the strain (11 patients), it was not deemed a necessary inclusion.

There are also issues with the data that could impact on the outcome of the model. Symptom start date was self-reported and therefore relies on the memory of the individual, thus the calculation of the time from becoming symptomatic to starting treatment, which determines the rate of diagnosis, is likely to be imprecise.

The definition of “lost to follow-up” here refers to those who are known to have had an interruption in their treatment. This therefore relies on accurate recording of loss to follow up by diagnosticians. Because of the low granularity of the treatment outcomes (only outcomes at 12 month and final outcome available) it is possible that individuals had treatment interruptions in between the recorded datapoints. There is also a matter of individuals who were non-adherent but may not have been lost. Whilst there were data on this recorded in the bespoke database for the outbreak from handwritten notes made by nurses, the data are difficult to categorise and mostly incomplete, especially in later years. Non-adherence could perhaps have been crudely included by splitting the population into those who are adherent and those who are not and giving those who are not adherent some infection potential, but this would have required including additional complexity based on limited data.

6.4.4 FURTHER WORK

In addition to the above, extensions to this modelling work could be undertaken. Another intervention used to tackle the TB epidemic in the UK is treatment for LTBI, which consists of either rifampicin and isoniazid for three months or isoniazid only for six months [19]. It is

usually only administered to close contacts of patients with TB, healthcare workers, immunosuppressed patients or migrants for high-incidence countries [207]. One consideration for this outbreak might have been implementing an LTBI screening programme in areas such as homeless shelters and prisons in order to seek out some of the hard-to-reach LTBI cases with this outbreak strain. This can be included in the model via a rate from the latent slow and latent fast compartments into the susceptible compartment or even into a separate compartment which denotes some immunity to those who have been on LTBI treatment, that would eventually wane, returning individuals to the susceptible state. The rate of LTBI treatment (rate from latently infected to the susceptible/immune compartment) would need to account for the proportion of LTBI individuals who would be started on treatment, complete it, and the efficacy. It must be noted that for this outbreak, due to the isoniazid-resistant nature of the strain, the LTBI treatment would ideally be a rifampicin-based regimen.

The inclusion of LTBI treatment would impact the number of individuals with active disease by reducing the number of individuals who progress from latently infected to active disease. However, by returning to the susceptible compartment there is the possibility they can be re-infected. Thus, unless LTBI can be highly effective and employed with a high coverage, it may be unlikely to end the outbreak.

7 DISCUSSION

In this thesis, I set out to achieve four objectives to contribute to our understanding of how we can use whole genome sequencing data, statistical inference and mathematical modelling to understand certain aspects of tuberculosis transmission and use these insights in a public health setting to help control outbreaks. Firstly, I explored the current literature to uncover what could be determined about the approaches available for interpreting whole genome sequencing in the setting of tuberculosis transmission (Chapter 3). In light of the need for considering genomic diversity when interpreting WGS data for transmission highlighted by the literature review, a statistical inference method, TransPhylo, was used to analyse WGS data from a real-world outbreak (Chapter 4). Using bioinformatic and phylogenetic methods I then analysed WGS data from a large isoniazid-resistant TB outbreak to determine if a timed-tree could be produced for analysis with TransPhylo and the implications of this for public health interventions in this setting (Chapter 5). Finally, I developed a novel mathematical model to describe TB transmission for the same outbreak, with a population that experiences loss-to-follow-up (Chapter 6) to examine the effects of three different public health interventions for combatting the outbreak. The main findings of the thesis can be summarised under three themes:

1. WGS data analysis: results around how to analyse and interpret WGS in order to extract meaningful information for use in TB transmission.
2. TB transmission: findings directly related to TB transmission, such as how we can use WGS data to determine the possibility of transmission between cases.
3. Public health: important results for TB public health.

7.1.1 WGS DATA ANALYSIS FINDINGS

From my systematic review, a key finding that emerged from the examination of studies involving bioinformatics analysis of TB WGS data is that bioinformatics pipelines, i.e. the programs and methods used to firstly assemble genomes and then analyse them, vary hugely. The evidence can be seen in Chapter 3 where the settings used by the studies in the review, such as filtering thresholds, are compared. The impact of different bioinformatics pipelines on the number of SNPs determined was explored by Altmann *et al* [187] who found a difference of up to 20,000 SNPs between different pipelines. With some studies relying on the finding of singular SNPs in order to distinguish transmission from not transmission, it is important that the full effect of bioinformatic choices are considered, and if possible, a unified code of practise is drawn up for future studies.

The second result related to WGS analysis was that producing a timed phylogenetic tree from London TB outbreak genomes for use in phylogenetic inference of transmission trees proved untenable. Both trees produced with the ML and Bayesian methods had little

confidence, shown through low bootstrap values and posterior values. One possible cause could be that there is a lack of variation amongst the data and therefore a lack of information. The lack of information in the data could be down to much of the variation being in genes that were removed for analysis (a standard procedure) due to the potential introduction of errors from mapping to highly repetitive areas of the genome. Alternatively, it may be that the molecular clock for this strain is much too slow and therefore not enough divergence had occurred over time, or at least in comparison to the number of cases that were infected and sampled. To compare, Roetzer *et al.* [91], when analysing WGS data from TB cases in an outbreak from 1997 to 2010, found 85 SNPs amongst 86 samples, versus 261 amongst 403 samples in the London outbreak, suggesting very limited diversity in this setting.

Given that our interpretations of WGS data are dependent upon the ability to produce reliable consensus sequences, we should be aware of the effect of analytical choices we make upon our findings and that some data may not be suited to certain types of analyses.

7.1.2 TUBERCULOSIS TRANSMISSION FINDINGS

As identified through the different studies in this thesis, there are clearly numerous ways in which WGS and epidemiological data can be used to enhance our understanding of TB transmission. Many such methods were identified in the systematic review (Chapter 3) and two methods were explored in more detail in Chapters 4 and 6. One important finding that was highlighted from the systematic review of these methods was that the use of fixed SNP thresholds had been used multiple times to either exclude transmission or identify possible transmission. Despite this being a very attractive prospect as it is easy to apply, especially over large, complex datasets, the data and findings identified by studies looking at diversity of TB with WGS data drew attention to the fact that a TB strain can mutate significantly in a fairly short time, potentially complicating the use of a fixed threshold across all situations. This limitation therefore suggests caution should be exercised when interpreting data in this way.

7.1.3 PUBLIC HEALTH FINDINGS

An important public health finding from this thesis was that we can use WGS data and statistical inference to determine the infection timings of the outbreak cases and use that information to establish whether recent cases have been a result of recent infection or reactivation. In British Columbia, the most recent cases within an ‘outbreak’ were demonstrated to be the result of reactivation, suggesting that there is no more transmission occurring and allowing for the outbreak to be declared as complete.

The caveat to this finding is that there could be unsampled cases, which if not found before becoming infectious would lead to a re-emergence of the outbreak. As Kelowna, the area in which the outbreak took place, is quite small the public health team were able to get detailed information of each of the outbreak cases and track the outbreak very carefully, meaning they were confident all cases had been sampled. However, this is unlikely to be the case in many other circumstances, such as in a much larger city. For example, Kühnert *et al* [208] investigated an outbreak of a strain imported into California from Thailand, with an estimate of only 9% sampling coverage. Even though the method can be adapted to impute where there are missing cases, the method is unable to impute onward transmission from a case, i.e. it can only impute where there are missing ancestors not descendants, and thus it would have to be used in conjunction with contact tracing to ensure all contacts of the most recent case had been tested for active/latent TB to bring the outbreak to a full close.

A second important public health finding from the thesis was that the use of active case finding within a large TB outbreak in London was considered a more effective intervention in terms of decreasing the number of cases in 2015 than reducing loss to follow up or increasing re-engagement with treatment after loss to follow up as determined by the modelling study (Chapter 6). Such use of mathematical modelling for comparing interventions not only helps to direct public health efforts and reduce the amount of resources that are wasted on ineffective interventions, but also reveals more about the transmission dynamics of the outbreak, e.g. we can assume that LFU individuals were not contributing greatly to the infectious potential of the outbreak compared to the individuals with late active disease, mostly like due to a smaller magnitude of individuals.

7.2 STRENGTHS

Throughout the thesis a diverse set of research tools and approaches have been employed, each with their own strengths. In Chapter 3 a systematic literature review was undertaken. Systematic literature reviews are performed with the intention of capturing information from all possible studies that are within the scope of the review subject, as opposed to a general literature review which does not require undertaking a comprehensive search of online databases and therefore may miss key studies. This rigour (i.e. multiple search engines used, general keywords and synonyms used, extraction and inclusion done independently by multiple individuals, pre-determined data points for extraction) means that the results of the literature review undertaken in Chapter 3 are thorough.

The work undertaken using TransPhylo used a novel version of the method which improved the inferences of timings of infection. The ability to assess the infection dates and determine the recent transmission dynamics for the outbreak allowed the public health investigation team to declare the outbreak over, a first for TB.

Within the modelling chapter, uncertainty in model assumptions and parameters was addressed through a range of sensitivity and uncertainty analyses. These analyses provided a rigorous way to assess the effects of parameter choices and hence allow the dimensionality of the problem to be reduced should certain parameters be relatively insensitive, i.e. to an extent, parameters may make little to no difference to the outcome. In addition, it can be reassuring in the case of unknown parameter values, should the effect of altering the parameter value not have an impact on the outcome of the model, e.g. as was seen for the reactivation rate.

An additional strength of these studies is the availability of data from multiple sources to inform the parameterisation of models and to help understand transmission. Combination of literature reviews (modelling and epidemiological studies) and data.

Another significant strength is the application of the phylogenetic, bioinformatic and inference tools to real-world data, firstly the Canadian outbreak and then the London outbreak. Being able to analyse real-world data means that it is possible to assess whether the tools are practical for the use in which they are intended. We are attempting to assess tools for use in public health settings meaning they must be usable in real-world settings (specifically must be able to deal with uncertainty, missing data etc.) and ideally in real-time, i.e. are relatively fast at providing results, not just for idealised simulated data.

7.3 LIMITATIONS

Limitations of each individual study have been summarised in the relevant chapter. This section reflects on some of these limitations and highlights potential approaches that may improve future analyses.

Furthermore, there were limitations associated with data availability, which is not uncommon in the analysis of observational data largely collected to support service provision, including paucity of appropriate parameters for the London TB outbreak, e.g. 9.2% had no available information on their treatment outcome and 12.8% lack information about their symptom onset date. This is linked to the method of data collection, which largely was via a paper questionnaire sent to TB nurses and doctors who had seen the patient; these were often not returned. Public health officials and outbreak and surveillance staff should endeavour to collect better quality data to support investigations in future studies.

7.4 FUTURE WORK

As discussed in the relevant chapters there are numerous avenues for expanding upon the work undertaken in this thesis. Firstly, the TransPhylo method could be re-performed on the Canadian outbreak data (Chapter 4) after some adjustments such as including epidemiological data that may improve the ability to determine transmission, as seen in the

original TransPhylo study [99]. It would be a useful exercise to compare the results with and without including epidemiological data to see if this has an impact on the outcome of timing of infection. In addition, it may be worthwhile to use the version of TransPhylo that allows imputation of potentially unsampled cases to assess if there were any missing cases and if this has any effect on the results of infection timing.

Secondly, an attempt at re-analysing the London outbreak genomes (Chapter 5) using *de novo* assembly and/or including SNPs in the PE/PPE genes to see if that helps resolve any issues with the trees would be recommended. If this should produce a phylogenetic tree with greater confidence, then it would be an interesting exercise to perform TransPhylo analysis on the tree and potentially reveal transmission dynamics that could inform the model in Chapter 6. Alternatively, if a timed phylogenetic tree still cannot be reliably produced with these methods, attempting to use a method that simultaneously infers the phylogenetic and transmission tree would be the next method to try.

Finally, the mathematical model presented in Chapter 6 can be modified by considering immunity (innate or acquired), multiple strains with different drug resistance patterns to account for the cases that were found to be multi-drug resistant [209], or demography [210]. Other interventions could also be assessed, for instance, LTBI treatment.

7.5 CONCLUSION

Overall, in this thesis I have demonstrated ways in which WGS and mathematical modelling can both be used to inform public health practices with respect to TB transmission. WGS can be used alone or alongside epidemiological data to determine whether transmission has occurred or distinguish between re-infection and relapse and thereby enlighten public health departments as to where transmission is coming from and therefore what interventions might be most appropriate. When used in combination with statistical inference, more detailed information can be inferred around the timing of transmission, which in turn can inform public health departments as to whether transmission is still occurring or if recent cases are a result of reactivation, again informing practices. Finally, mathematical modelling can help inform public health practices by providing a tool for intervention evaluation, thus helping to decide which would be the most effective in an outbreak.

8 BIBLIOGRAPHY

[1] Public Health England, “Tuberculosis in England 2018 report: presenting data to the end of 2017,” 2018. [Online]. Available: https://www.gov.uk/government/publications/tuberculosis-in-england-annual-report.

[2] Public Health England, “Tuberculosis in London: Annual review (2017 data),” 2018. [Online]. Available: https://www.gov.uk/government/publications/tuberculosis-tb-regional-reports.

[3] R. Riley, “Airborne infection,” *Am J Med,* vol. 57, pp. 466-75, 1974.

[4] M. Golden and H. Vikram, “Extrapulmonary Tuberculosis: An Overview,” *Am Fam Physician,* vol. 72, no. 9, pp. 1761-1768, 2005.

[5] J. Flynn and J. Chan, “Immunology of tuberculosis,” *Annu Rev Immunol,* vol. 19, pp. 93-129, 2001.

[6] M. Behr, P. Edelstein and L. Ramakrishnan, “Revisiting the timetable of tuberculosis,” *BMJ,* vol. 362, 2018.

[7] R. Sloot, M. van der Loeff, P. Kouw and M. Borgdorff, “Risk of Tuberculosis after Recent Exposure. A 10-Year Follow-up Study of Contacts in Amsterdam,” *American Journal of Respiratory and Critical Care Medicine,* vol. 190, no. 9, 2014.

[8] J. Flynn and J. Chan, “Tuberculosis: Latency and Reactivation,” *Infection and Immunity,* vol. 69, no. 7 , pp. 4195-4201, 2001.

[9] A. Zumla, P. Malon, J. Henderson and J. Grange, “Impact of HIV infection on tuberculosis,” *Postgraduate Medical Journal,* vol. 76, pp. 259-268, 2000.

[10] M. Lerm and M. Netea, “Trained immunity: a new avenue for tuberculosis vaccine,” *Journal of Internal Medicine,* vol. 279, no. 4, pp. 337-46, 2016.

[11] A. Bandera, A. Gori, L. D. E. A. Catozzi, G. Marchetti, C. Molteni, G. Ferrario, L. Codecasa, V. Penati, A. Matteeli and F. Franzetti, “Molecular Epidemiology Study of Exogenous Reinfection in an Area with a Low Incidence of Tuberculosis,” *Journal of Clinical Microbiology,* vol. 39, no. 6, pp. 2213-2218, 2001.

[12] C.-Y. Chiang and L. Riley, “Exogenous reinfection in tuberculosis,” *Lancet Infect Dis,*

vol. 5, p. 629–36, 2005.

[13] L. Campos, M. Rocha, D. Willers and D. Silva, “Characteristics of Patients with Smear-Negative Pulmonary Tuberculosis (TB) in a Region with High TB and HIV Prevalence,” *PLoS One,* vol. 11, no. 1, 2016.

[14] W.-C. Chao, Y.-W. Huang, M.-C. Yu, W.-T. Yang, C.-J. Lin, J.-J. Lee, R.-M. Huang, C.-C. Shieh, S.-T. Chien and J.-Y. Chien, “Outcome correlation of smear-positivity but culture-negativity during standard anti-tuberculosis treatment in Taiwan,” *BMC Infectious Diseases,* vol. 15, no. 67, 2015.

[15] M. Asghar, S. Mehta, H. Cheema, R. Patti and W. Pascal, “Sputum smear and culture-negative tuberculosis with associated pleural effusion: a diagnostic challenge,” *Cureus,* vol. 10, no. 10, 2018.

[16] J. Shaw and N. Wynn-Williams, “Infectivity of pulmonary tuberculosis in relation to sputum status,” *Am Rev Tuberc,* vol. 69, pp. 724-32, 1954.

[17] S. Grzybowski, G. Barnett and K. Styblo, “Contacts of cases of active pulmonary tuberculosis,” *Bull Int Union Tuberc,* vol. 50, pp. 90-106, 1975.

[18] A. Nachiappan, K. Rahbar, X. Shi, E. Guy, E. Mortani Barbosa Jr, G. Shroff, D. Ocazionez, A. Schlesinger, S. Katz and M. Hammer, “Pulmonary Tuberculosis; Role of radiology in Diagnosis and Management,” *RadioGraphics,* vol. 37, no. 1, 2017.

[19] National Institute for Health and Care Excellence, “Tuberculosis [NG33],” 2016. [Online]. Available: https://www.nice.org.uk/guidance/ng33.

[20] World Health Organization, *Treatment of Tuberculosis: Guidelines for National Programmes, fourth edition,* 2010.

[21] R. Shi, N. Itagaki and I. Sugawara, “Overview of Anti-Tuberculosis (TB) Drugs and Their Resistance Mechanisms,” *Mini-reviews in Medicinal Chemistry,* vol. 7, no. 11, pp. 1177-1185, 2007.

[22] J. Zeyland and E. Piasecka-Zeyland, “Antituberculous immunity produced by BCG vaccine,” *Acta Paediatrica,* vol. 27, no. 3, pp. 393-401, 1940.

[23] Y. J. Ryu, “Diagnosis of Pulmonary Tuberculosis: Recent Advances and Diagnostic Algorithms,” *Tuberculosis and Respiratory Diseases,* vol. 78, no. 2, pp. 64-71, 2015.

[24] World Health Organisation, “Global tuberculosis report 2017,” [Online]. Available:

http://www.who.int/tb/publications/global_report/en/. [Accessed 1 6 2018].

[25] M. Berry and O. M. Kon, “Multidrug- and extensively drug-resistant tuberculosis: an emerging threat,” *European Respiratory Review,* vol. 18, no. 114, pp. 195-197, 2009.

[26] H. Esmail, C. Barry, D. B. Young and R. J. Wilkinson, “The ongoing challenge of latent tuberculosis,” *Philosophical Transactions of the Royal Society B,* vol. 369, no. 1645, pp. 20130437-20130437, 2014.

[27] S. Valway, M. Sanchez, T. Shinnick, I. Orme, T. Agerton, D. Hoy, J. Jones, H. Westmoreland and I. Onorato, “An outbreak involving extensive transmission of a virulent strain of Mycobacterium tuberculosis,” *N Engl J Med,* vol. 338, no. 24, p. 1783, 1998.

[28] M. Uplekar, D. Weil, K. Lönnroth, E. Jaramillo, C. Lienhardt, H. M. Y. Dias, D. Falzon, K. Floyd, G. Gargioni, H. Getahun, C. Gilpin, P. Glaziou, M. Grzemska, F. Mirzayev, H. Nakatani and M. Raviglione, “WHO's new end TB strategy.,” *The Lancet,* vol. 385, no. 9979, pp. 1799-1801, 2015.

[29] “The Zero TB Initiative,” [Online]. Available: https://www.zerotbinitiative.org. [Accessed 5 10 2018].

[30] “Stop TB Partnership - Home Page,” , . [Online]. Available: http://www.stoptb.org/. [Accessed 2 7 2018].

[31] J. A. Caylà and A. Orcau, “Control of tuberculosis in large cities in developed countries: an organizational problem,” *BMC Medicine,* vol. 9, no. 1, pp. 127-127, 2011.

[32] G. de Vries, R. W. Aldridge, J. A. Caylà, W. Haas, A. Sandgren, N. van Hest and I. Abubakar, “Epidemiology of tuberculosis in big cities of the European Union and European Economic Area countries,” *Eurosurveillance,* vol. 19, no. 9, p. 20726, 2014.

[33] M. Kruijshaar, I. Abubakar, M. Dedicoat, G. H. Bothamley, H. Maguire, J. Moore, J. Crofts and M. Lipman, “Evidence for a national problem: continued rise in tuberculosis case numbers in urban areas outside London,” *Thorax,* vol. 67, pp. 275-277, 2012.

[34] K. Lönnroth, G. Migliori, I. Abubakar, L. D'Ambrosio, G. de Vries, R. Diel, P. Douglas, D. Falzon, M. Gaudreau, D. Goletti, E. González Ochoa, P. LoBue, A. Matteelli, H. Njoo, I. Solovic, A. Story, T. Tayeb, M. van der Werf, D. Weil and J. Zellweger, “Towards tuberculosis elimination: an action framework for low-incidence countries,” *Eur Respir J,* vol. 45, no. 4, pp. 928-52, 2015.

[35] Public Health England, “Collaborative tuberculosis strategy for England: 2015-2020,” 2015. [Online]. Available: https://www.gov.uk/government/publications/collaborative-tuberculosis-strategy-for-england.

[36] A. Shaghaghi, R. Bhopal and S. A, “Approaches to recruiting 'hard-to-reach' populations into research: A review of the literature,” *Health Promot Perspect,* vol. 1, no. 2, pp. 86-94, 2011.

[37] M. Jit, H. R. Stagg, R. W. Aldridge, P. White and I. Abubakar, “Dedicated outreach service for hard to reach patients with tuberculosis in London: observational study and economic evaluation,” *BMJ,* vol. 343, 2011.

[38] C. Mulder, E. Klinkenberg and D. Manissero, “Effectiveness of tuberculosis contact tracing among migrants and the foreign-born population,” *Eurosurveillance,* vol. 14, no. 11, 2009.

[39] N. Bock, R. Sales, T. Rogers and B. DeVoe, “A spoonful of sugar...: improving adherence to tuberculosis treatment using financial incentives [Notes from the Field],” *IJTLD,* vol. 5, no. 1, pp. 96-98, 2011.

[40] N. Martin, P. Morris and P. Kelly, “Food incentives to improve completion of tuberculosis treatment: randomised controlled trial in Dili, Timor-Leste,” *BMJ,* 2009.

[41] Q. Liu, K. Abba, M. Alejandria, V. Balanag, R. Berba and M. Lansang, “Reminder systems and late patient tracers in diagnosis and management of tuberculosis,” *Evidence-based Child Health: A Cochrane Review Journal,* vol. 5, no. 3, pp. 1206-1245, 2010.

[42] K. Rainwater-Lovett, I. Rodriguez-Barraquer and W. Moss, “Viral Epidemiology: Tracking Viruses with Smartphones and Social Media,” in *Viral Pathogenesis: From Basics to Systems Biology*, Academic Press, 2016, pp. 241-252.

[43] B. Mathema, J. R. Andrews, T. Cohen, M. W. Borgdorff, M. A. Behr, J. R. Glynn, R. Rustomjee, B. J. Silk and R. Wood, “Drivers of Tuberculosis Transmission,” *The Journal of Infectious Diseases,* vol. 216, 2017.

[44] A. C. Schürch, K. Kremer, O. Daviena, A. Kiers, M. J. Boeree, R. J. Siezen and D. van Soolingen, “High-Resolution Typing by Integration of Genome Sequencing Data in a Large Tuberculosis Cluster,” *Journal of Clinical Microbiology,* vol. 48, no. 9, pp. 3403-3406, 2010.

[45] J. M. Bryant, A. C. Schürch, H. van Deutekom, S. R. Harris, J. de Beer, V. de Jager, K.

Kremer, S. A. F. T. van Hijum, R. J. Siezen, M. W. Borgdorff, S. D. Bentley, J. Parkhill and D. van Soolingen, “Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data.,” *BMC Infectious Diseases,* vol. 13, no. 1, pp. 110-110, 2013.

[46] S. Duchêne, K. Holt, F.-X. Weill, S. Le Hello, J. Hawkey, D. Edwards, M. Fourment and E. Holmes, “Genome-scale rates of evolutionary change in bacteria,” *Microb Genom,* vol. 2, no. 11, 2016.

[47] C. B. Ford, P. L. Lin, M. R. Chase, R. R. Shah, O. Iartchouk, J. E. Galagan, N. Mohaideen, T. R. Ioerger, J. C. Sacchettini, M. Lipsitch, J. L. Flynn and S. M. Fortune, “Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection,” *Nature Genetics,* vol. 43, no. 5, pp. 482-486, 2011.

[48] T. Walker, C. Ip, R. Harrell, J. Evans, G. Kapatai, M. Dedicoat, D. Eyre, D. Wilson, P. Hawkey, C. DW and e. al, “Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study,” *The Lancet Infectious diseases,* vol. 13, no. 2, pp. 137-146, 2013.

[49] I. Hershkovitz, H. Donoghue, D. Minnikin, G. Besra, O. Y.-C. Lee, A. Gernaey, E. Galili, V. Eshed, C. Greenblatt, E. Lemma, G. Bar-Gal and M. Spigelman, “Detection and Molecular Characterization of 9000-Year-Old Mycobacterium tuberculosis from a Neolithic Settlement in the Eastern Mediterranean,” *PLoS One,* vol. 3, no. 10, 2008.

[50] S. Gagneux, “Host–pathogen coevolution in human tuberculosis,” *Philos Trans R Soc Lond B Biol Sci,* vol. 367, no. 1590, p. 850–859, 2012.

[51] D. Brites and S. Gagneux, “Co-evolution of Mycobacterium tuberculosis and Homo sapiens,” *Immunological Reviews,* vol. 264, no. 1, pp. 6-24, 2015.

[52] L. Fenner, M. Egger, T. Bodmer, H. Furrer, M. Ballif, M. Battegay, P. Helbling, J. Fehr, T. Gsponer, H. Rieder, M. Zwahlen, M. Hoffmann, E. Bernasconi, M. Cavassini, A. Calmy, M. Dolina, R. Frei, J. Janssens, S. Borrell, D. Stucki, J. Schrenzel and Bottger, “HIV Infection Disrupts the Sympatric Host–Pathogen Relationship in Human Tuberculosis,” *PLoS Genetics,* vol. 9, no. 3, 2013.

[53] D. Stucki, M. Ballif, M. Egger, H. Furrer, E. Altpeter, M. Battegay, S. C. Droz, T. Bruderer, M. Coscolla, S. Borrell, K. Zürcher, J.-P. Janssens, A. Calmy, J. M. Stalder, K. Jaton, H. L. Rieder, G. E. Pfyffer, H. H. Siegrist, M. Hoffmann, J. Fehr, M. Dolina, R. Frei, J. Schrenzel, E. C. Böttger, S. Gagneux and L. Fenner, “Standard genotyping overestimates transmission of Mycobacterium tuberculosis among immigrants in a low

incidence country," *Journal of Clinical Microbiology,* vol. 54, no. 7, pp. 1862-1870, 2016.

[54] N. A. Rosenberg, A. G. Tsolaki and M. M. Tanaka, "Estimating change rates of genetic markers using serial samples: applications to the transposon IS6110 in Mycobacterium tuberculosis," *Theoretical Population Biology,* vol. 63, p. 347–363, 2003.

[55] R. R. Kao, D. T. Haydon, S. Lycett and P. R. Murcia, "Supersize me: how whole-genome sequencing and big data are transforming epidemiology," *Trends in Microbiology,* vol. 22, no. 5, pp. 282-291, 2014.

[56] A. C. Schürch and D. van Soolingen, "DNA fingerprinting of Mycobacterium tuberculosis: from phage typing to whole-genome sequencing.," *Infection, Genetics and Evolution,* vol. 12, no. 4, pp. 602-609, 2012.

[57] T. M. Walker, P. Monk, E. G. Smith and T. Peto, "Contact investigations for outbreaks of Mycobacterium tuberculosis: advances through whole genome sequencing," *Clinical Microbiology and Infection,* vol. 19, no. 9, pp. 796-802, 2013.

[58] R. Doyle, C. Burgess, R. Williams, R. Gorton, H. Booth, J. Brown, J. Bryant, J. Chan, D. Creer, J. Holdstock, H. Kunst, S. Lozewicz, G. Platt, E. Romer, G. Speight, S. Tiberi, I. Abubakar, M. Lipman, T. McHugh and J. Breuer, "Direct Whole-Genome Sequencing of Sputum Accurately Identifies Drug-Resistant Mycobacterium tuberculosis Faster than MGIT Culture Sequencing," *J Clin Microbiol,* vol. 56, no. 8, 2018.

[59] C. Pareek, R. Smoczynski and A. Tretyn, "Sequencing technologies and genome sequencing," *J Appl Genet,* vol. 52, no. 4, p. 413–435, 2011.

[60] E. van Dijk, H. Auger, Y. Jaszczyszyn and C. Thermes, "Ten years of next-generation sequencing technology," *Trends in Genetics,* vol. 30, no. 9, pp. 418-426, 2014.

[61] N. Casali, A. Broda, S. R. Harris, J. Parkhill, T. Brown and F. Drobniewski, "Whole Genome Sequence Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective Observational Study," *PLOS Medicine,* vol. 13, no. 10, 2016.

[62] D. Baum, "Reading a Phylogenetic Tree: The Meaning of Monophyletic Groups," *Nature Education,* vol. 1, no. 1, p. 190, 2008.

[63] Z. Yang and B. Rannala, "Molecular phylogenetics: principles and practice," *Nature Reviews Genetics,* vol. 13, pp. 303-314, 2012.

[64] Centers for Disease Control and Prevention, “Tuberculosis Outbreak Associated With a Homeless Shelter - Kane County, Illinois, 2007-2011,” *MMWR,* vol. 61, pp. 186-189, 2012.

[65] G. Garnett, S. Cousens, T. Hallett, R. Steketee and N. Walker, “Mathematical models in the evaluation of health programmes,” *The Lancet,* vol. 378, no. 9790, pp. 515-525, 2011.

[66] J. Lessler and D. Cummings, “Mechanistic Models of Infectious Disease and Their Impact on Public Health,” *American Journal of Epidemiology,* vol. 183, no. 5, 2016.

[67] N. A. Menzies, E. Wolf, D. Connors, M. Bellerose, A. N. Sbarra, T. Cohen, A. N. Hill, R. Yaesoubi, K. Galer, P. White, I. Abubakar and J. A. Salomon, “Progression from latent infection to active disease in dynamic tuberculosis transmission models: a systematic review of the validity of modelling assumptions,” *Lancet Infectious Diseases,* 2018.

[68] J. Blackwood and L. Childs, “An introduction to compartmental modeling for the budding infectious disease modeler,” *Letters in Biomathematics,* vol. 5, no. 1, pp. 195-221, 2018.

[69] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, V. Colizza, L. Isella, C. Régis, J.-F. Pinton, N. Khanafer, W. Van den Broeck and P. Vanhems, “Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees,” *BMC Medicine,* vol. 9, 2011.

[70] R. Almeida, A. Brito da Cruz, N. Martins and M. Monteiro, “An epidemiological MSEIR model described by the Caputo fractional derivative,” *International Journal of Dynamics and Control,* vol. 7, no. 2, pp. 776-784, 2019.

[71] European Centre for Disease Prevention and Control, “Monitoring the use of whole-genome sequencing in infectious disease surveillance in Europe 2015-2017,” ECDC, Stockholm, 2018.

[72] Public Health England, “England world leaders in the use of whole genome sequencing to diagnose TB,” 2017. [Online]. Available: https://www.gov.uk/government/news/england-world-leaders-in-the-use-of-whole-genome-sequencing-to-diagnose-tb.

[73] H.-A. Hatherell, C. Colijn, H. R. Stagg, C. Jackson, J. R. Winter and I. Abubakar, “Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review,” *BMC Medicine,* vol. 14, no. 1, p. 21, 2016.

[74] Y. H. Grad and M. Lipsitch, “Epidemiologic data and pathogen genome sequences: a powerful synergy for public health,” *Genome Biology,* vol. 15, no. 11, pp. 538-538, 2014.

[75] S. Sreevatsan, X. Pan, K. E. Stockbauer, N. D. Connell, B. N. Kreiswirth, T. S. Whittam and J. M. Musser, “Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination,” *Proceedings of the National Academy of Sciences of the United States of America,* vol. 94, no. 18, 1997.

[76] C. B. Ford, K. Yusim, T. Ioerger, S. Feng, M. R. Chase, M. Greene, B. T. Korber and S. M. Fortune, “Mycobacterium tuberculosis – Heterogeneity revealed through whole genome sequencing,” *Tuberculosis,* vol. 92, no. 3, pp. 194-201, 2012.

[77] D. J. Wilson, “Insights from Genomics into Bacterial Pathogen Populations,” *PLOS Pathogens,* vol. 8, no. 9, 2012.

[78] E. R. Robinson, T. M. Walker and M. J. Pallen, “Genomics and outbreak investigation: from sequence to consequence,” *Genome Medicine,* vol. 5, no. 4, pp. 36-36, 2013.

[79] J. A. Guerra-Assunção, R. M. Houben, A. C. Crampin, T. Mzembe, K. Mallard, F. Coll, P. Khan, L. Banda, A. Chiwaya, R. P. Pereira, R. McNerney, D. Harris, J. Parkhill, T. G. Clark and J. R. Glynn, “Recurrence due to Relapse or Reinfection With Mycobacterium tuberculosis: A Whole-Genome Sequencing Approach in a Large, Population-Based Cohort With a High HIV Infection Prevalence and Active Follow-up,” *The Journal of Infectious Diseases,* vol. 211, no. 7, pp. 1154-1163, 2015.

[80] M.-L. Lambert, E. Hasker, A. V. Deun, D. Roberfroid, M. Boelaert and P. van der Stuyft, “Recurrence in tuberculosis: relapse or reinfection?,” *Lancet Infectious Diseases,* vol. 3, no. 5, pp. 282-287, 2003.

[81] M. Wlodarska, J. C. Johnston, J. L. Gardy and P. Tang, “A Microbiological Revolution Meets an Ancient Disease: Improving the Management of Tuberculosis with Genomics,” *Clinical Microbiology Reviews,* vol. 28, no. 2, pp. 523-539, 2015.

[82] T. Jombart, A. Cori, X. Didelot, S. Cauchemez, C. Fraser and N. M. Ferguson, “Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data,” *PLOS Computational Biology,* vol. 10, no. 1, 2014.

[83] C. U. Köser, M. T. G. Holden, M. J. Ellington, E. J. Cartwright, N. Brown, A. Ogilvy-Stuart, L. Y. Hsu, C. Chewapreecha, N. J. Croucher, S. R. Harris, M. Sanders, M. C. Enright, G. Dougan, S. D. Bentley, J. Parkhill, L. Fraser, J. R. Betley, O. Schulz-

Trieglaff, G. P. Smith and S. J. Peacock, “Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak.,” *The New England Journal of Medicine,* vol. 366, no. 24, pp. 2267-2275, 2012.

[84] C. W. Knetsch, T. R. Connor, A. Mutreja, S. van Dorp, I. Sanders, H. P. Browne, D. Harris, L. Lipman, E. Keessen, J. Corver, E. J. Kuijper and T. D. Lawley, “Whole genome sequencing reveals potential spread of Clostridium difficile between humans and farm animals in the Netherlands, 2002 to 2011,” *Eurosurveillance,* vol. 19, no. 45, pp. 20954-20954, 2014.

[85] “The PRISMA statement,” [Online]. Available: www.prisma-statement.org. [Accessed 2014 11 10].

[86] N. Noah, “Strengthening the reporting of molecular epidemiology for infectious diseases (STROME-ID): an extension of the STROBE statement,” *Epidemiology and Infection,* vol. 142, no. 07, pp. 1343-1343, 2014.

[87] C. Luchini, B. Stubbs, M. Solmi and N. Veronese, “Assessing the quality of studies in meta-analyses: Advantages and limitations of the Newcastle Ottawa Scale,” *World J Meta-Anal,* vol. 5, no. 4, pp. 80-84, 2017.

[88] J. L. Gardy, J. C. Johnston, S. J. H. Sui, V. J. Cook, L. Shah, E. Brodkin, S. Rempel, R. G. Moore, Y. Zhao, R. A. Holt, R. Varhol, I. Birol, M. Lem, M. K. Sharma, K. Elwood, R. C. Brunham and P. Tang, “Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak,” *The New England Journal of Medicine,* vol. 364, no. 8, pp. 730-739, 2011.

[89] T. Luo, C. Yang, Y. Peng, L. Lu, G. Sun, J. Wu, X. Jin, J. Hong, F. Li, J. Mei, K. DeRiemer and Q. Gao, “Whole-genome sequencing to detect recent transmission of Mycobacterium tuberculosis in settings with a high burden of tuberculosis.,” *Tuberculosis,* vol. 94, no. 4, pp. 434-440, 2014.

[90] C. Mehaffy, J. L. Guthrie, D. C. Alexander, R. Stuart, E. Rea and F. B. Jamieson, “Marked Microevolution of a Unique Mycobacterium tuberculosis Strain in 17 Years of Ongoing Transmission in a High Risk Population,” *PLOS ONE,* vol. 9, no. 11, 2014.

[91] A. Roetzer, R. Diel, T. A. Kohl, C. Rückert, U. Nübel, J. Blom, T. Wirth, S. Jaenicke, S. Schuback, S. Rüsch-Gerdes, P. Supply, J. Kalinowski and S. Niemann, “Whole Genome Sequencing versus Traditional Genotyping for Investigation of a Mycobacterium tuberculosis Outbreak: A Longitudinal Molecular Epidemiological Study,” *PLOS Medicine,* vol. 10, no. 2, 2013.

[92] J. A. Guerra-Assunção, A. C. Crampin, R. Houben, T. Mzembe, K. Mallard, F. Coll, P. Khan, L. Banda, A. Chiwaya, R. Pereira, R. McNerney, P. Fine, J. Parkhill, T. G. Clark and J. R. Glynn, "Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area," *eLife,* vol. 4, 2015.

[93] R. Lee, "Re-emergence and Amplification of Tuberculosis in the Canadian Arctic.," *The Journal of infectious diseases,* 2015.

[94] A. A. Witney, K. A. Gould, A. Arnold, D. Coleman, R. Delgado, J. Dhillon, M. Pond, C. F. Pope, T. Planche, N. G. Stoker, C. A. Cosgrove, P. D. Butcher, T. S. Harrison and J. Hinds, "Clinical Application of Whole-Genome Sequencing To Inform Treatment for Multidrug-Resistant Tuberculosis Cases," *Journal of Clinical Microbiology,* vol. 53, no. 5, pp. 1473-1483, 2015.

[95] O. M. Williams, T. Abeel, N. Casali, K. A. Cohen, A. S. Pym, S. B. Mungall, C. A. Desjardins, A. K. Banerjee, F. Drobniewski, A. M. Earl and G. S. Cooke, "Fatal Nosocomial MDR TB Identified through Routine Genetic Analysis and Whole-Genome Sequencing," *Emerging Infectious Diseases,* vol. 21, no. 6, pp. 1082-1084, 2015.

[96] T. M. Walker, M. K. Lalor, A. Broda, L. S. Ortega, M. Morgan, L. Parker, S. Churchill, K. Bennett, T. Golubchik, A. Giess, C. d. O. Elias, K. Jeffery, I. Bowler, I. Laurenson, A. Barrett, F. Drobniewski, N. D. McCarthy, L. F. Anderson, I. Abubakar, H. L. Thomas, P. Monk, E. G. Smith, A. S. Walker, D. W. Crook, D. W. Crook, T. Peto and C. Conlon, "Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study.," *The Lancet Respiratory Medicine,* vol. 2, no. 4, pp. 285-292, 2014.

[97] T. G. Clark, K. Mallard, F. Coll, M. D. Preston, S. A. Assefa, D. Harris, S. Ogwang, F. Mumbowa, B. Kirenga, D. M. O'Sullivan, A. Okwera, K. D. Eisenach, M. Joloba, S. D. Bentley, J. J. Ellner, J. Parkhill, E. C. Jones-López and R. McNerney, "Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing," *PLOS ONE,* vol. 8, no. 12, 2013.

[98] M. Kato-Maeda, C. Ho, B. Passarelli, N. Banaei, J. Grinsdale, L. L. Flores, J. Anderson, M. Murray, G. Rose, L. M. Kawamura, N. Pourmand, M. A. Tariq, S. Gagneux and P. C. Hopewell, "Use of Whole Genome Sequencing to Determine the Microevolution of Mycobacterium tuberculosis during an Outbreak," *PLOS ONE,* vol. 8, no. 3, 2013.

[99] X. Didelot, J. L. Gardy and C. Colijn, "Bayesian Inference of Infectious Disease Transmission from Whole-Genome Sequence Data," *Molecular Biology and Evolution,*

vol. 31, no. 7, pp. 1869-1879, 2014.

[100] D. Stucki, M. Ballif, T. Bodmer, M. Coscolla, A.-M. Maurer, S. C. Droz, C. Butz, S. Borrell, C. Längle, J. Feldmann, H. Furrer, C. Mordasini, P. Helbling, H. L. Rieder, M. Egger, S. Gagneux and L. Fenner, “Tracking a Tuberculosis Outbreak Over 21 Years: Strain-Specific Single-Nucleotide Polymorphism Typing Combined With Targeted Whole-Genome Sequencing,” *The Journal of Infectious Diseases,* vol. 211, no. 8, pp. 1306-1316, 2015.

[101] P. Smit, “Enhanced tuberculosis outbreak investigation using whole genome sequencing and IGRA,” *European Respiratory Journal,* vol. 45, no. 1, pp. 276-279, 2015.

[102] J. M. Bryant, S. R. Harris, J. Parkhill, R. Dawson, A. H. Diacon, P. D. van Helden, A. Pym, A. A. Mahayiddin, C. Chuchottaworn, I. M. Sanne, C. Louw, M. J. Boeree, M. Hoelscher, T. D. McHugh, A. Bateson, R. D. Hunt, S. Mwaigwisya, L. Wright, S. H. Gillespie and S. D. Bentley, “Whole-genome sequencing to establish relapse or re-infection with Mycobacterium tuberculosis: a retrospective observational study,” *The Lancet Respiratory Medicine,* vol. 1, no. 10, pp. 786-792, 2013.

[103] L. Pérez-Lago, I. Comas, Y. Navarro, F. González-Candelas, M. Herranz, E. Bouza and D. García-de-Viedma, “Whole Genome Sequencing Analysis of Intrapatient Microevolution in Mycobacterium tuberculosis: Potential Impact on the Inference of Tuberculosis Transmission,” *The Journal of Infectious Diseases,* vol. 209, no. 1, pp. 98-108, 2014.

[104] N. Casali, V. Nikolayevskyy, Y. Balabanova, S. R. Harris, O. Ignatyeva, I. Kontsevaya, J. Corander, J. M. Bryant, J. Parkhill, S. Nejentsev, R. D. Horstmann, T. Brown and F. Drobniewski, “Evolution and transmission of drug-resistant tuberculosis in a Russian population,” *Nature Genetics,* vol. 46, no. 3, pp. 279-286, 2014.

[105] T. R. Ioerger, Y. Feng, X. Chen, K. M. Dobos, T. C. Victor, E. M. Streicher, R. M. Warren, N. C. G. van Pittius, P. D. van Helden and J. C. Sacchettini, “The non-clonality of drug resistance in Beijing-genotype isolates of Mycobacterium tuberculosis from the Western Cape of South Africa.,” *BMC Genomics,* vol. 11, no. 1, pp. 670-670, 2010.

[106] F. Lanzas, P. C. Karakousis, J. C. Sacchettini and T. R. Ioerger, “Multidrug-Resistant Tuberculosis in Panama Is Driven by Clonal Expansion of a Multidrug-Resistant Mycobacterium tuberculosis Strain Related to the KZN Extensively Drug-Resistant M. tuberculosis Strain from South Africa,” *Journal of Clinical Microbiology,* vol. 51, no. 10, pp. 3277-3285, 2013.

[107] S. M. Regmi, A. Chaiprasert, S. Kulawonganunchai, S. Tongsima, O. O. Coker, T. Prammananan, W. Viratyosin and I. Thaipisuttikul, “Whole genome sequence analysis of multidrug-resistant Mycobacterium tuberculosis Beijing isolates from an outbreak in Thailand,” *Molecular Genetics and Genomics,* vol. 290, no. 5, pp. 1933-1941, 2015.

[108] O. Ocheretina, L. Shen, V. E. Escuyer, M.-M. Mabou, G. Royal-Mardi, S. Collins, J. W. Pape and D. W. Fitzgerald, “Whole Genome Sequencing Investigation of a Tuberculosis Outbreak in Port-au-Prince, Haiti Caused by a Strain with a “Low-Level” rpoB Mutation L511P – Insights into a Mechanism of Resistance Escalation,” *PLOS ONE,* vol. 10, no. 6, 2015.

[109] M. Kato-Maeda, J. Z. Metcalfe and L. L. Flores, “Genotyping of Mycobacterium tuberculosis: application in epidemiologic studies.,” *Future Microbiology,* vol. 6, no. 2, pp. 203-216, 2011.

[110] M. Egger, P. Juni, C. Bartlett, F. Holenstein and J. Sterne, “How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical Study,” *Health Technol Assess,* vol. 7, pp. 1-76, 2003.

[111] S. Niemann, C. U. Köser, S. Gagneux, C. Plinke, S. Homolka, H. R. Bignell, R. J. Carter, R. K. Cheetham, A. J. Cox, N. A. Gormley, P. Kokko-Gonzales, L. Murray, R. Rigatti, V. P. Smith, F. P. Arends, H. S. Cox, G. Smith and J. A. Archer, “Genomic Diversity among Drug Sensitive and Multidrug Resistant Isolates of Mycobacterium tuberculosis with Identical DNA Fingerprints,” *PLOS ONE,* vol. 4, no. 10, 2009.

[112] R. J. Ypma, W. M. van Ballegooijen and J. Wallinga, “Relating Phylogenetic Trees to Transmission Trees of Infectious Disease Outbreaks,” *Genetics,* vol. 195, no. 3, pp. 1055-1062, 2013.

[113] S. H. Gillespie, A. M. Crook, T. D. McHugh, C. M. Mendel, S. Meredith, S. Murray, F. Pappas, P. P. J. Phillips and A. Nunn, “Four-Month Moxifloxacin-Based Regimens for Drug-Sensitive Tuberculosis,” *The New England Journal of Medicine,* vol. 371, no. 17, pp. 1577-1587, 2014.

[114] R. Colangeli, V. L. Arcus, R. T. Cursons, A. Ruthe, N. Karalus, K. Coley, S. D. Manning, S. Kim, E. Marchiano and D. Alland, “Whole genome sequencing of Mycobacterium tuberculosis reveals slow growth and low mutation rates during latent infections in humans.,” *PLOS ONE,* vol. 9, no. 3, 2014.

[115] C. J. Worby, H.-H. Chang, W. P. Hanage and M. Lipsitch, “The Distribution of Pairwise Genetic Distances: A Tool for Investigating Disease Transmission,” *Genetics,* vol. 198,

no. 4, pp. 1395-1404, 2014.

[116] Q. Liu, Y. Guo, J. Li, J. Long, B. Zhang and Y. Shyr, “Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data.,” *BMC Genomics,* vol. 13, 2012.

[117] H. David, “Probability Distribution of Drug-Resistant Mutants in Unselected Populations of Mycobacterium-Tuberculosis.,” *Appl Microbiol,* vol. 20, no. 5, p. 810, 1970.

[118] G. Kaplan, A. Moreira, H. Wainwright, B. Kreiswirth, M. Tanverdi, B. Mathema, S. Ramaswamy, G. Walther, L. Steyn and C. Barry 3rd, “Mycobacterium tuberculosis growth at the cavity surface: a microenvironment with failed immunity.,” *Infect Immun,* vol. 71, no. 12, pp. 7099-108, 2003.

[119] Q. Liu, L. Via, T. Luo, L. Liang, X. Liu, S. Wu, Q. Shen, W. Wei, X. Ruan, X. Yuan and e. al, “Within patient microevolution of Mycobacterium tuberculosis correlates with heterogeneous responses to treatment.,” *Sci Rep,* vol. 5, 2015.

[120] R. van den Berg, COMMUNICABLE MEDICAL DISEASES: A holistic and social medicine perspective for healthcare providers, Balboa Press, 2014.

[121] T. Cohen, P. van Helden, D. Wilson, C. Colijn, M. McLaughlin, I. Abubakar and R. Warren, “Mixed-strain mycobacterium tuberculosis infections and the implications for tuberculosis treatment and control.,” *Clin Microbiol Rev,* vol. 25, no. 4, pp. 708-719, 2012.

[122] P. Black, M. de Vos, G. Louw, R. van der Merwe, A. Dippenaar, E. Streicher, A. Abdallah, S. Sampson, T. Victor and T. Dolby, “Whole genome sequencing reveals genomic heterogeneity and antibiotic purification in Mycobacterium tuberculosis isolates,” *BMC Genomics,* vol. 16, 2015.

[123] D. Warner, A. Koch and V. Mizrahi, “Diversity and disease pathogenesis in Mycobacterium tuberculosis.,” *Trends in Microbiology,* vol. 23, no. 1, pp. 14-21, 2015.

[124] C. Köser, L. Fraser, A. Ioannou, J. Becq, M. Ellington, M. Holden, S. Reuter, M. Estée Török, S. Bentley, J. Parkhill, N. Gormley, G. Smith and S. Peacock, “Rapid single-colony whole-genome sequencing of bacterial pathogens,” *J Antimicrob Chemother,* vol. 69, no. 5, p. 1275–1281, 2014.

[125] C. Worby, M. Lipsitch and W. Hanage, “Shared genomic variants: identification of transmission routes using pathogen deep sequence data.,” *bioRxiv,* 2015.

[126] S. Waffenschmidt, E. Hausner, W. Sieben, T. Jaschinski, M. Knelangen and I. Overesch, “Effective study selection using text-mining or a single-screening approach: a study protocol,” *Systematic Reviews,* vol. 7, no. 166, 2018.

[127] H.-A. Hatherell, X. Didelot, S. Pollock, P. Tang, A. Crisan, J. C. Johnston, C. Colijn and J. L. Gardy, “Declaring a tuberculosis outbreak over with genomic epidemiology,” *Microbial Genomics,* vol. 2, no. 5, 2016.

[128] K. Toman, “Tuberculosis case-finding and chemotherapy: questions and answers,” 1979.

[129] Center for Disease Control, “Controlling Tuberculosis in the United States: Recommendations from the American Thoracic Society, CDC, and the Infectious Diseases Society of America,” *Morbidity And Mortality Weekly Report: Recommendations and Reports,* vol. 54, no. RR12, pp. 1-81, 2005.

[130] E. M. Cottam, G. Thébaud, J. Wadsworth, J. Gloster, L. Mansley, D. J. Paton, D. P. King and D. T. Haydon, “Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus,” 2008. [Online]. Available: http://rspb.royalsocietypublishing.org/content/275/1637/887. [Accessed 24 9 2018].

[131] M. J. Morelli, G. Thébaud, J. Chadœuf, D. P. King, D. T. Haydon and S. Soubeyrand, “A Bayesian Inference Framework to Reconstruct Transmission Trees Using Epidemiological and Genetic Data,” *PLOS Computational Biology,* vol. 8, no. 11, 2012.

[132] T. D. Lieberman, D. Wilson, R. Misra, L. L. Xiong, P. Moodley, T. Cohen and R. Kishony, “Genomic diversity in autopsy samples reveals within-host dissemination of HIV-associated Mycobacterium tuberculosis,” *Nature Medicine,* vol. 22, no. 12, pp. 1470-1474, 2016.

[133] E. Kenah, T. Britton, M. E. Halloran and I. M. Longini, “Molecular Infectious Disease Epidemiology: Survival Analysis and Algorithms Linking Phylogenies to Transmission Trees,” *PLOS Computational Biology,* vol. 12, no. 4, 2016.

[134] A. Drummond and A. Rambaut, “BEAST: Bayesian evolutionary analysis by sampling trees,” *BMC Evolutionary Biology,* vol. 7, p. 214, 2007.

[135] O. Krylova and D. J. Earn, “Effects of the infectious period distribution on predicted transitions in childhood disease dynamics.,” *Journal of the Royal Society Interface,* vol. 10, no. 84, pp. 20130098-20130098, 2013.

[136] R. Ragonnet, J. M. Trauer, N. Scott, M. T. Meehan, J. T. Denholm and E. S. McBryde,

"Optimally capturing latency dynamics in models of tuberculosis transmission," *Epidemics,* vol. 21, pp. 39-47, 2017.

[137] J. Cheng, L. Hiscoe, S. Pollock, P. Hasselback, J. Gardy and R. Parker, "A clonal outbreak of tuberculosis in a homeless population in the interior of British Columbia, Canada, 2008-2015.," *Epidemiology and Infection,* vol. 143, no. 15, pp. 3220-3226, 2015.

[138] N. C. Grassly and C. Fraser, "Mathematical models of infectious disease transmission," *Nature Reviews Microbiology,* vol. 6, no. 6, pp. 477-487, 2008.

[139] M. W. Borgdorff, M. M. G. G. Sebek, R. B. Geskus, K. Kremer, N. A. Kalisvaart and D. van Soolingen, "The incubation period distribution of tuberculosis estimated with a molecular epidemiological approach," *International Journal of Epidemiology,* vol. 40, no. 4, pp. 964-970, 2011.

[140] J. Kingman, "The coalescent," *Stoch Proc Appl,* vol. 13, no. 3, pp. 235-248, 1982.

[141] Y. Xu, H. Topliffe, J. Stimson, H. Stagg, I. Abubakar and C. Colijn, "Transmission analysis of a large TB outbreak in London: mathematical modelling study using genomic data," *bioRxiv,* 2019.

[142] X. Didelot, C. Fraser, J. Gardy and C. Colijn, "Genomic Infectious Disease Epidemiology in Partially Sampled and Ongoing Outbreaks," *Mol Biol Evol,* vol. 34, no. 4, pp. 997-1007, 2017.

[143] D. Ayabina, J. Ronning, K. Alfsnes, N. Debech, O. Brynildsrud, T. Arnesen, G. Norheim, A.-T. Mengshoel, R. Rykkvin, U. Dahle, C. Colijn and V. Eldholm, "Genome-based transmission modelling separates imported tuberculosis from recent transmission within an immigrant population," *Microb Genom,* vol. 4, no. 10, 2018.

[144] D. Klinkenberg, J. A. Backer, X. Didelot, C. Colijn and J. Wallinga, "Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks.," *PLOS Computational Biology,* vol. 13, no. 5, 2017.

[145] M. Hall, M. Woolhouse and A. Rambaut, "Epidemic reconstruction in a phylogenetics framework: transmission trees as partitions of the node set," *PLoS Comput Biol,* vol. 11, no. 12, 2015.

[146] N. De Maio, C. J. Worby, D. J. Wilson and N. Stoesser, "Bayesian reconstruction of transmission within outbreaks using genomic variants," *PLOS Computational Biology,* vol. 14, no. 4, p. 213819, 2018.

[147] H. Maguire, S. Brailsford, J. Carless, M. Yates, L. Altass, S. Yates, S. Anaraki, A. Charlett, S. Lozewicz, M. Lipman and G. Bothamley, “Large outbreak of isoniazid-monoresistant tuberculosis in London, 1995 to 2006: case-control study and recommendations.,” *Eurosurveillance,* vol. 16, no. 13, p. 19830, 2011.

[148] F. Neely, H. Maguire, F. Le Brun, A. Davies, D. Gelb and S. Yates, “High rate of transmission among contacts in large London outbreak of isoniazid mono-resistant tuberculosis,” *Journal of Public Health,* vol. 32, no. 1, pp. 44-51, 2010.

[149] M. Ruddy, A. Davies, M. Yates, S. Balasegaram, Y. Drabu, B. Patel, S. Lozewicz, S. Sen, M. Bahl, E. James, M. Lipman, G. Duckworth, J. Watson, M. Piper, F. Drobniewski and H. Maguire, “Outbreak of isoniazid resistant tuberculosis in north London,” *Thorax,* vol. 59, no. 4, pp. 279-285, 2004.

[150] M. Ruddy, A. Davies, M. Yates, F. Drobniewski, B. Patel, S. Yates, S. Balasegaram, S. Lozewicz, S. Sen, Y. Drabu, G. Duckworth, J. Watson, M. Piper and H. Maguire, “A continuing outbreak of isoniazid resistant tuberculosis in North London,” *Journal of Infection,* vol. 44, no. 2, p. 108, 2002.

[151] T. Weniger, J. Krawczyk, P. Supply, S. Niemann and D. Harmsen, “MIRU-VNTRplus: a web tool for polyphasic genotyping of Mycobacterium tuberculosis complex bacteria,” *Nucleic Acids Research,* vol. 38, pp. W326-W331, 2010.

[152] H. Maguire, M. Ruddy, G. Bothamley, B. Patel, M. Lipman, F. Drobniewski, M. Yates and T. Brown, “Multidrug resistance emerging in North London outbreak,” *Thorax,* vol. 61, no. 6, pp. 547-548, 2006.

[153] C. Smith, S. Trienekens, C. Anderson, M. Lalor, T. Brown, A. Story, H. Fry, A. Hayward and H. Maguire, “Twenty years and counting: epidemiology of an outbreak of isoniazid-resistant tuberculosis in England and Wales, 1995 to 2014,” *Eurosurveillance,* vol. 22, no. 8, 2017.

[154] B. F. F. Ouellette, “The GenBank Sequence Database,” 2006. [Online]. Available: http://bioon.com/book/biology/bioinformatics/chapter-3.pdf. [Accessed 6 1 2019].

[155] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. B. Gabriel, M. J. Daly and M. A. DePristo, “The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data,” *Genome Research,* vol. 20, no. 9, pp. 1297-1303, 2010.

[156] G. Van der Auwera, M. Carneiro, C. Hartl and e. al, “From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.,” *Current*

Protocols in Bioinformatics, vol. 11, 2013.

[157] S. Andrews, “FastQC: A Quality Control tool for High Throughput Sequence Data,” [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. [Accessed 6 1 2019].

[158] R. Nielsen, J. Paul, A. Albrechtsen and Y. Song, “Genotype and SNP calling from next-generation sequencing data,” *Nat Rev Genet,* vol. 12, no. 6, pp. 443-451, 2011.

[159] H. Li, “Improving SNP discovery by base alignment quality,” *Bioinformatics,* vol. 27, no. 8, pp. 1157-1158, 2011.

[160] J. Phelan, F. Coll, I. Bergval, R. Anthony, R. Warren, S. Sampson, N. Gey van Pittius, J. Gylnn, A. Crampin, A. Alves, T. Bessa, S. Campino, K. Dheda, L. Grandjean, R. Hasan, Z. Hasan, A. Miranda, D. Moore, S. Panaiotov, J. Perdigao and I. Portugal, “Recombination in pe/ppe genes contributes to genetic variation in Mycobacterium tuberculosis lineages,” *BMC Genomics,* vol. 17, no. 151, 2016.

[161] D. Posada, “jModelTest: phylogenetic model averaging,” *Mol Biol Evol,* vol. 25, no. 7, pp. 1253-6, 2008.

[162] D. Darriba, G. Taboada, R. Doallo and D. Posada, “jModelTest 2: more models, new heuristics and parallel computing,” *Nature Methods,* vol. 9, p. 772, 2012.

[163] S. Kalyaanamoorthy, B. Minh, T. Wong, A. von Haeseler and L. Jermiin, “ModelFinder: fast model selection for accurate phylogenetic estimates,” *Nature Methods,* vol. 14, pp. 587-589, 2017.

[164] S. Ho, “The molecular clock and estimating species divergence,” *Nature Education,* 2008.

[165] A. Drummond and R. Bouckaert, Bayesian evolutionary analysis with BEAST, Cambridge: Cambridge University Press, 2015.

[166] M. Suchard, R. Weiss and J. Sinsheimer, “Bayesian selection of continuous-time Markoc chain evolutionary models,” *Mol Biol Evol,* vol. 18, pp. 1001-1013, 2001.

[167] N. Friel and A. Pettitt, “Marginal Likelihood Estimation via Power Posteriors,” *Journal of the Royal Statistical Society. Series B (Statistical Methodology),* vol. 70, no. 3, pp. 589-607, 2008.

[168] G. Baele, P. Lemey, T. Bedford, A. Rambaut, M. Suchard and A. Alekseyenko, “Improving the accuracy of demographic and molecular clock model comparison while

accommodating phylogenetic uncertainty," *Mol. Biol. Evol.,* vol. 29, no. 9, pp. 2157-2167, 2012.

[169] G. Baele, W. Li, A. Drummond, M. Suchard and P. Lemey, "Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics," *Mol. Biol. Evol.,* vol. 30, no. 2, pp. 239-243, 2013.

[170] R. Kass and A. Raftery, "Bayes Factors," *Journal of the American Statistical Association,* vol. 90, no. 430, pp. 773-795, 1995.

[171] J. O'Reilly and P. Donoghue, "The Efficacy of Consensus Tree Methods for Summarizing Phylogenetic Relationships from a Posterior Sample of Trees Estimated from Morphological Data," *Syst Biol,* vol. 67, no. 2, pp. 354-362, 2018.

[172] A. Rambaut, A. Drummond, D. Xie, G. Baele and M. Suchard, "Posterior summarisation in Bayesian phylogenetics using Tracer 1.7," *Systematic Biology,* vol. 67, no. 5, pp. 901-904, 2018.

[173] E. Nummelin, "MC's for MCMC'ists," *International Statistical Review,* vol. 70, no. 2, pp. 215-240, 2002.

[174] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows–Wheeler Transform," *Bioinformatics,* vol. 25, no. 14, p. 1754–1760, 2009.

[175] J. M. Lew, A. Kapopoulou, L. Jones and S. T. Cole, "TubercuList - 10 years after," *Tuberculosis,* vol. 91, no. 1, pp. 1-7, 2011.

[176] A. Rambaut, T. T.-Y. Lam, L. M. Carvalho and O. G. Pybus, "Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen)," *Virus Evolution,* vol. 2, no. 1, 2016.

[177] V. Rangannan and M. Bansal, "PromBase: a web resource for various genomic features and predicted promoters in prokaryotic genomes," *BMC Research Notes,* vol. 4, no. 257, 2011.

[178] A. Cornish-Bowden, "Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations," *Nucleic Acids Res,* vol. 13, pp. 3021-3030, 1985.

[179] S. Tavaré, "Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences," *Lectures on Mathematics in the Life Sciences,* vol. 17, pp. 57-86, 1986.

[180] M. Hasegawa, H. Kishino and T. Yano, "Dating of human-ape splitting by a molecular clock of mitochondrial DNA," *Journal of Molecular Evolution,* vol. 22, no. 2, pp. 160-

174, 1985.

[181] K. Tamura and M. Nei, “Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees,” *Mol Biol Evol,* vol. 10, no. 3, pp. 512-526, 1993.

[182] T. Stadler, “Birth-Death with Serial Samples,” *J Theor Biol,* vol. 267, pp. 396-404, 2010.

[183] A. Drummond, G. Nicholls, A. Rodrigo and W. Solomon, “Serially Sampled Data,” *Genetics,* vol. 161, pp. 1307-1320, 2002.

[184] M. Gill, P. Lemey, N. Faria, A. Rambaut, B. Shapiro and M. Suchard, “SkyGrid Coalescent,” *Mol Biol Evol,* vol. 30, pp. 713-724, 2013.

[185] R. Griffiths and S. Tavare, “Parametric Coalescent,” *Phil Trans R Soc Lond B Biol Sci,* vol. 344, pp. 403-410, 1994.

[186] J. Heled and R. Bouckaert, “Looking for trees in the forest: summary tree from posterior samples,” *BMC Evolutionary Biology,* vol. 13, no. 221, 2013.

[187] A. Altmann, P. Weber, D. Bader, M. Preuss, E. Binder and B. Mueller-Myhsok, “A beginners guide to SNP calling from high-throughput DNA-sequencing data,” *Hum Genet,* vol. 131, pp. 1541-1554, 2012.

[188] Z. Dembek, T. Chekol and A. and Wu, “Best practice assessment of disease modelling for infectious disease outbreaks,” *Epidemiology and Infection,* vol. 146, p. 1207–1215, 2018.

[189] H. Hartman-Adams, K. Clark and G. Juckett, “Update on Latent Tuberculosis Infection,” *American Family Physician,* vol. 89, no. 11, pp. 889-896, 2014.

[190] R. Ragonnet, J. M. Trauer, E. S. McBryde, R. Houben, J. T. Denholm, A. Handel and T. Sumner, “Is IPT more effective in high-burden settings? Modelling the effect of tuberculosis incidence on IPT impact.,” *International Journal of Tuberculosis and Lung Disease,* vol. 21, no. 1, pp. 60-66, 2017.

[191] D. W. Dowdy, S. Basu and J. R. Andrews, “Is passive diagnosis enough? The impact of subclinical disease on diagnostic strategies for tuberculosis.,” *American Journal of Respiratory and Critical Care Medicine,* vol. 187, no. 5, pp. 543-551, 2013.

[192] D. W. Dowdy, J. R. Andrews, P. J. Dodd and R. H. Gilman, “A user-friendly, open-source tool to project impact and cost of diagnostic tests for tuberculosis,” *eLife,* vol. 3,

no. 3, 2014.

[193] D. Ahmad and W. Morgan, “How long are TB patients infectious,” *Canadian Medical Association Journal,* vol. 163, no. 2, pp. 157-158, 2000.

[194] P. H. Boersch-Supan, S. J. Ryan and L. R. Johnson, “deBInfer: Bayesian inference for dynamical models of biological systems in R,” *Methods in Ecology and Evolution,* vol. 8, no. 4, pp. 511-518, 2017.

[195] M. Plummer, N. Best, K. Cowles and K. Vines, “CODA: Convergence Diagnosis and Output Analysis for MCMC,” *R News,* vol. 6, pp. 7-11, 2006.

[196] A. Gelma and D. Rubin, “Inference from Iterative Simulation using Multiple Sequences,” *Statistical Science,* vol. 7, pp. 457-511, 1992.

[197] K. Dietz, “The estimation of the basic reproduction number for infectious diseases,” *Statistical Methods in Medical Research,* vol. 2, no. 1, pp. 23-41, 1993.

[198] H. Yang, “The basic reproduction number obtained from Jacobian and next generation matrices - A case study of dengue transmission modelling,” *Biosystems,* 2014.

[199] M. D. McKay, R. J. Beckman and W. Conover, “A comparison of three methods for selecting values of input variables in the analysis of output from a computer code,” *Technometrics,* vol. 42, no. 1, pp. 55-61, 2000.

[200] S. Marino, I. Hogue, C. Ray and D. Kirschner, “A methodology for performing global uncertainty and sensitivity analysis in systems biology,” *J Theor Biol,* vol. 254, no. 1, pp. 178-96, 2008.

[201] A. Story, R. Aldridge, I. Abubakar, H. Stagg, M. Lipman, J. Watson and A. Hayward, “Active case finding for pulmonary tuberculosis using mobile digital chest radiography: an observational study,” *International Journal of Tuberculosis and Lung Disease,* vol. 16, no. 11, pp. 1461-1467, 2012.

[202] J. Curtis, “Impact of x-ray screening programmes for active tuberculosis in homeless populations: a systematic review of original studies,” *Journal of Public Health,* vol. 38, no. 1, pp. 106-114, 2016.

[203] W. W. Yew, C. Lange and C. C. Leung, “Treatment of tuberculosis: update 2010,” *European Respiratory Journal,* vol. 37, no. 2, pp. 441-462, 2011.

[204] A. T. Fojo, N. Stennis, A. S. Azman, E. A. Kendall, S. Shrestha, S. D. Ahuja and D. W. Dowdy, “Current and future trends in tuberculosis incidence in New York City: a

dynamic modelling analysis," *The Lancet. Public health,* vol. 2, no. 7, 2017.

[205] J. Wu, R. Dhingra, M. Gambhir and J. Remais, "Sensitivity analysis of infectious disease models: methods, advances and their application," *Journal of the Royal Society Interface,* vol. 10, no. 86, 2013.

[206] S. Mandal and N. Arinaminpathy, "Transmission modeling and health systems: the case of TB in India," *International Health,* vol. 7, no. 2, pp. 114-20, 2015.

[207] Public Health England, "Latent TB Testing and Treatment for Migrants: A practical guide for commissioners and practitioners," 2015. [Online]. Available: https://www.gov.uk/government/publications/latent-tb-infection-ltbi-testing-and-treatment.

[208] D. Kuhnert, M. Coscolla, D. Stucki, J. Metcalfe, L. Fenner, S. Gagneux and T. Stadler, "Tuberculosis outbreak investigation using phylodynamic analysis," *Epidemics,* vol. 25, pp. 47-53, 2018.

[209] J. Trauer, J. Denholm and E. McBryde, "Construction of a mathematical model for tuberculosis transmission in highly endemic regions of the Asia-pacific," *Journal of Theoretical Biology,* vol. 358, pp. 74-84, 2014.

[210] S. Arregui, M. Iglesias, S. Samper, D. Marinova, C. Martin, J. Sanz and Y. Moreno, "Data-driven model for the assessment of the Mycobacterium tuberculosis transmission in evolving demographic structures," *Proc Natl Acad Sci,* vol. 115, no. 14, 2018.

9 APPENDICES

9.1 APPENDIX 1 – SYSTEMATIC REVIEW TABLES

9.1.1 SEARCH STRATEGIES FOR EACH DATABASE

Order of search	Search terms	Number of results
#1	Tuberculosis[MeSH exploded] or Mycobacterium Tuberculosis[MeSH exploded] or tuberculosis or TB	230591
#2	Disease Transmission, Infectious[MeSH exploded] or Disease Outbreaks[MeSH exploded] or Epidemics[MeSH exploded] or epidemiology or Epidemiology[MeSH exploded] or transmi* or outbreak* or pandemic* or spread* or epidemic* or endemic	926226
#3	((whole genome OR full genome OR entire genome OR complete genome OR next generation) ADJ3 sequenc*) OR NGS OR WGS	22826
#4	#1 AND #2 AND #3	116

Table 9.1 Search strategy for MEDLINE (14.07.15)

Order of search	Search terms	Number of results
#1	Tuberculosis[MeSH exploded] or TB or tuberculosis or Mycobacterium tuberculosis[MeSH exploded]	317110
#2	Bacterial transmission[MeSH exploded] or Disease transmission[MeSH exploded] or transmi* or spread* or pandemic* or outbreak* or endemic or epidemic* or Epidemic[MeSH exploded] or Epidemiology[MeSH exploded] or epidemiolog*	3372564
#3	((whole genome OR full genome OR entire genome OR complete genome OR next generation) ADJ3 sequenc*) OR NGS OR WGS	29589
#4	#1 AND #2 AND #3	160

Table 9.2 Search strategy for EMBASE+classic EMBASE (14.07.15)

Order of search	Search terms	Number of results

#1	epidemiolog* or transmi* or spread* or epidemic* or endemic or pandemic* or outbreak* or epidemiology[MeSH terms] or disease transmission, infectious[MeSH terms] or disease outbreaks[MeSH terms] or epidemics[MeSH terms] or pandemics[MeSH terms]	2511303
#2	tuberculosis[MeSH Terms] or tuberculosis or TB or mycobacterium tuberculosis[MeSH terms] or mycobacterium tuberculosis	231531
#3	((full genome or complete genome or entire genome or next generation or whole genome) and (sequencing or sequence or sequences)) or NGS or WGS	101329
#4	#1 AND #2 AND #3	197

Table 9.3 Search strategy for PubMed (14.07.15)

Order of search	Search terms	Number of results
#1	Epidemiolog* or outbreak* or transmi* or pandemic* or epidemic* or endemic	1489972
#2	Tuberculosis or TB or “mycobacterium tuberculosis”	151403
#3	((“full genome” OR “whole genome” OR “complete genome” OR “entire genome” OR “next generation”) NEAR/3 sequenc*) OR NGS OR WGS	30603
#4	#1 AND #2 AND #3	184

Table 9.4 Search strategy for Web of Science Core collection (14.07.15)

Order of search	Search terms	Number of results
#1	Disease Transmission[MeSH exploded] or Disease Outbreaks[MeSH] or Epidemiology[MeSH exploded] or transmi* or epidemiolog* or spread* or outbreak* or epidemic* or endemic	632323
#2	Tuberculosis[MeSH exploded] or Mycobacterium tuberculosis[MeSH] or TB or tuberculosis	17827

#3	((full genome or complete genome or whole genome or entire genome or next generation) N3 sequenc*) OR NGS OR WGS	720
#4	#1 AND #2 AND #3	12

Table 9.5 Search strategy for CINAHL (14.07.15)

Order of search	Search terms	Number of results
#1	((“full genome” or “whole genome” or “complete genome” or “entire genome” or “next generation”) W/3 sequenc*) or WGS or NGS	893820
#2	Tuberculosis or TB or “mycobacterium tuberculosis”	189310
#3	Epidemiolog* or outbreak* or spread* or pandemic* or epidemic* or transmi* or endemic	3244319
#4	TITLE-ABSTR-KEY(#1 AND #2 AND #3)	16

Table 9.6 Search strategy for ScienceDirect (14.07.15)

Order of search	Search terms	Number of results
#1	((whole genome OR full genome OR entire genome OR complete genome OR next generation) AND sequenc*) OR WGS OR NGS	311853
#2	Tuberculosis or tb or “mycobacterium tuberculosis”	221222
#3	Epidemiolog* or transmi* or spread* or epidemic* or pandemic* or outbreak* or endemic	1770174
#4	Abstract(#1 AND #2 AND #3)	9

Table 9.7 Search strategy for WILEY (14.07.15)

9.1.2 DATA ITEMS FOR EXTRACTION

General	Bioinformatics	Phylogenetic tree	Mixed infections	Relapses	Direction	Limitations
Aim	% of the genome covered by reads	Method (maximum likelihood, Bayesian etc.)	How were mixed infections defined	How were relapses defined?	How was direction determined?	Small sample
Theme	Reference genome	Software	What was the effect on transmission?			Culturing method
Number of individuals in study	Software	How were SNPs used				Missing samples
How were the samples identified	Sequencing machine					Other
Country	Other information					
Incidence rate classification (High ≥40 cases per 100,000, Low <40 cases per 100,000)						
Type of study						
Sample type						
Was epidemiological/contact tracing data used						
Population type (convenience, representative, epidemiologically/genotypically clustered, other)						
Exclusion criteria						
When and how were samples collected						
Length of period of collection						
How long was follow-up (for recurrent disease)						

Table 9.8 Predetermined data for extraction

9.1.3 QUALITY ASSESSMENT

	Was the infectious-disease case definition appropriate? Did they used appropriate diagnosis methods?	Were measures taken to minimise and measure cross-contamination?	Was the timeframe of the study appropriate? (3 years minimum set as a threshold where transmission examined)	Were the participants representative?	If the study investigates molecular clusters, did they state the sampling fraction?	Were methods used to detect multiple-strain infections appropriate? Was their effect on the study findings included?	Were efforts made to address discovery or ascertainment bias?	Did the study consider alternative explanations for findings when transmission chains are being investigated, and report the consistency between molecular and epidemiological evidence?	Was follow-up time long enough for outcomes to occur? (≥1year)	Was sample size justified, where a number was decided before the study was undertaken? (for hypothesis driven studies)
Bryant *et al*. (BMC Infectious Diseases, 2013)	Adequate	Unknown	Adequate	Adequate	Inadequate	Unknown	Adequate	Adequate/Adequate	N/A	N/A
Bryant *et al.* (The Lancet Resp Med, 2013)	Adequate	Adequate	N/A	Unknown	N/A	Adequate/Adequate	Adequate	N/A	Adequate	N/A
Casali *et al.*	Adequate	Unknown	Inadequate	Adequate	N/A	Unknown	Adequate	N/A/Adequate	N/A	N/A
Clark *et al.*	Adequate	Adequate	Adequate	Unknown	Adequate	Adequate/Inadequate	Adequate	N/A/Adequate	N/A	N/A
Didelot *et al.*	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Adequate/Adequate	N/A	N/A
Gardy *et al.*	Adequate	Unknown	Adequate	Adequate	Adequate	Inadequate/Inadequate	Adequate	Adequate/Inadequate	N/A	N/A
Guerra-Assuncao *et al*. (2015)	Adequate	Adequate	Adequate	Adequate	Adequate	Adequate/N/A	Adequate	Adequate/Adequate	Adequate	N/A
Guerra-Assuncao *et al.* (2014)	Adequate	Adequate	N/A	Adequate	N/A	Adequate/Adequate	Adequate	N/A	Adequate	N/A

Ioerger *et al.*	Adequate	Unknown	Unknown	Unknown	Inadequate	Unknown	Adequate	N/A	N/A	N/A
Kato-Maeda *et al.*	Adequate	Unknown	Inadequate	Adequate	Adequate	Adequate/Adequate	Adequate	Adequate/Adequate	N/A	N/A
Lanzas *et al.*	Adequate	Adequate	Adequate	Unknown	N/A	Unknown	Adequate	N/A	N/A	N/A
Lee *et al.*	Adequate	Adequate	Adequate	Adequate	Adequate	Unknown	Adequate	Inadequate/Adequate	N/A	N/A
Luo *et al*.	Adequate	Unknown	Inadequate	Adequate	Adequate	Unknown	Adequate	Inadequate/Adequate	N/A	N/A
Martin Williams *et al.*	Adequate	Unknown	Adequate	N/A	N/A	Unknown	Adequate	Inadequate/Adequate	N/A	N/A
Mehaffy *et al.*	Adequate	Unknown	Adequate	Adequate	Adequate	Adequate/N/A	Adequate	Inadequate/Inadequate	N/A	N/A
Ocheretina *et al.*	Inadequate	Unknown	Adequate	Adequate	Adequate	Unknown	Adequate	Inadequate/Adequate	N/A	N/A
Perez-Lago *et al.*	Adequate	Unknown	Adequate	Adequate	Adequate	Unknown	Adequate	Inadequate/Adequate	N/A	N/A
Regmi *et al.*	Adequate	Unknown	Adequate	Inadequate	Adequate	Unknown	Adequate	N/A	N/A	N/A
Roetzer *et al.*	Adequate	Unknown	Adequate	Adequate	Adequate	Unknown	Adequate	Inadequate/Inadequate	N/A	N/A
Schürch *et al.*	Adequate	Unknown	Adequate	Adequate	Adequate	Unknown	Inadequate	Adequate/Inadequate	Adequate	N/A
Smit *et al.*	Adequate	Unknown	Inadequate	Adequate	Adequate	Unknown	Adequate	Inadequate/Adequate	N/A	N/A
Stucki *et al.*	Adequate	Unknown	Adequate	Adequate	Adequate	Adequate/Inadequate	Adequate	Inadequate/Inadequate	N/A	N/A
Walker *et al.* (2014)	Adequate	Adequate	Adequate	Adequate	Adequate	Unknown	Adequate	Inadequate/Adequate	N/A	N/A
Walker *et al.* (2013)	Adequate	Unknown	Adequate	Adequate for clusters/Unknown for cross-sectional/longitudinal patients	Inadequate	Unknown	Adequate	Inadequate/Adequate	N/A	N/A
Witney *et al.*	Adequate	Unknown	Adequate	Adequate	N/A	Adequate/N/A	Adequate	N/A/Adequate	N/A	N/A

Table 9.9 Quality assessment of included studies

Unknown = not mentioned, Adequate = considered to not be at risk of bias, Inadequate = considered to be at risk of bias, N/A = due to the nature of the study this is not able to be considered

9.1.4 INCLUDED STUDIES AND EXTRACTED DATA

Journal article	Participants	Country (TB burden*)	Sample size	Type of study	Length of study	Focus	Sequencing machine	Reference genome	Patient characteristics	Lineages	Quality of SNP	Read length	Max no. of SNPs
Bryant *et al*.	RFLP clusters with epidemiological links	Netherlands (Low)	199	Retrospective	-	Confirmation	Illumina Genome Analyzer IIx	H37Rv	Drug resistant	Four Global lineages (Euro-American, East-African Indian, East Asian, Indo-Oceanic)	Alleles need support of ≥75% of reads on each strand, base quality score ≥ 50 and mapping quality score ≥30. Repetitive regions are avoided.	76/108 bp	Pairwise SNP distances: range 0-149, mean 3.42. 11,879 variable positions found over all samples
Bryant *et al*.	RCT participants: previously untreated, drug-sensitive, smear-positive pulmonary TB without severe co-morbidities	Malaysia, South Africa and Thailand (High)	47 pairs	Retrospective observational	Patients were observed for 18 months, including treatment and follow-up	Recurrences and Diversity	Illumina HiSeq	H37Rv	No severe co-morbidities. Drug sensitive	Four Global lineages (Euro-American, East-African Indian, East Asian, Indo-Oceanic)	SNPs in the PE and PPE genes that differed between the relapse pairs were discounted. SNP quality as above.	100 bp	Pairwise SNP distances: range 0-1419, mean 113.278 (relapse/re-infection pairs). 10,354 variable positions.
Casali *et al*.	Representative sample of patients with pulmonary disease. Culture-proven.	Russia (High)	1000	Prospective	2 years	Resistance	Illumina Genome Analyzer IIx or HiSeq 2000	H37Rv	Drug resistance (MDR and XDR)	Beijing, Central Asian Strain, Euro-American and East-African Indian	Alleles need support of >70% of reads, including ≥ 5 in each direction and mapping quality ≥ 45. Repetitive regions were avoided.	54/75/ 100 bp	SNP distances between linked cases: range: 0-183 SNPs. 32,445 variable positions

Clark *et al.*	Convenience sample of treatment experienced TB patients (69% of MDR-TB cases in Uganda)	Uganda (High)	51 samples (41 patients)		4 years	Resistance and Confirmation	Illumina HiSeq 2000	H37Rv	HIV present. Age: 19-50, Males and females. MDR	Central Asian Strain, Beijing, East-African Indian	Only variants of high quality (≥Q30) and supported by bi-directional reads were retained. Variants in PPE/PE loci were excluded.	76 bp	Range 0-1060 (compared to reference). 6857 variable positions in total. SNP distances between linked cases: 0-32
Didelot *et al.*	Outbreak cases, defined by the same MIRU-VNTR and contact tracing	Canada (Low)	33			Direction	Illumina HiSeq	CDC1551	-	-	Retained positions called with quality score of 222, genotype quality of 99, and no indication of strand basis or low depth of coverage. SNV excluded if located within 50bp of another SNV.		
Gardy *et al.*	Outbreak cases, defined by the same MIRU-VNTR and contact tracing	Canada (Low)	32 sequenced	Retrospective	2 years	Confirmation and Diversity	Illumina Genome Analyzer II	CDC1551	Age: 1-71, Males and Females	-	Excluded: i) SNPs with quality scores <30; ii) SNPs occurring in clusters (i.e. within 10 bp of each other); iii) SNPs identical across all 36 samples; and iv) 15 SNP positions at which one or more samples displayed an ambiguous	50 bp	204 SNPs amongst all samples

												residue call		
Guerra-Assunção *et al.*	Culture confirmed cases in Karonga district	Malawi (High)	1687 sequenced with high quality data		15 years	Diversity, Direction, Recurrences and confirmation	Illumina HiSeq 2000	H37Rv	Age: <20-50+. HIV present. Males and females	East Asian, Euro American, Indo-Oceanic, East-African Indian	Removed low-quality sequences and low-quality 3′ ends of reads, retaining only reads ≥ 50 bp long, with nucleotides above quality score Q27. Excluded samples with coverage less than 10-fold or with >15% missing genotypes. Excluded genome positions with >15% missing genotypes and those in highly repetitive regions.	100 bp	Paired SNP distances: 0-almost 2000
Guerra-Assunção *et al.*	Laboratory confirmed TB cases who had completed treatment	Malawi (High)	60 pairs with WGS	Population-based	14 years	Recurrences and Diversity	Illumina HiSeq 2000	H37Rv	HIV present. Age: <30 – 50+, Males and Females	East Asian, Euro American, Indo-Oceanic, East-African Indian	Removed low-quality sequences and low-quality 3′ ends of reads, retaining only reads ≥ 50 bp long, with nucleotides above quality score Q27. Excluded SNPs with >15% missing genotypes and those in highly repetitive regions.	100 bp	Paired SNP distances: 0-1000+
Ioerger *et al.*	Two RFLP drug resistant clusters	South Africa (High)	14			Resistance	Illumina Genome	H37Rv/HN878	Drug resistance.	Beijing		36 bp	1546 SNPs in sample

							Analyzer II						
Kato-Maeda *et al.*	Individuals found through contact tracing to be involved in a transmission chain	USA (Low)	9	Population-based	22 months	Direction	Illumina Genome Analyzer	H37Rv	HIV absent. Hispanic males. Age: 18 – 34. Drug susceptible.	-	SNPs in PE, PE-PGRS, PPE genes and mobile elements were excluded. 25 putative SNPs, (≥85% of reads supported one base call and ≥ 12 reads depth), were analysed with PCR Sanger method. 7 confirmed as true SNPs.		7 SNPs between all samples
Lanzas *et al.*	66 MDR and 31 drug sensitive patients	Panama (High)	97		10 years	Resistance	Illumina Genome Analyzer IIx	H37Rv	HIV present. Age: 14 – 81. Males and females. MDR and drug susceptible.	Mainly Latin American-Mediterranean	Needed depth of coverage ≥25% of the mean, and the majority nucleotide represented in >70% of reads; gaps and regions with clusters of SNPs were excluded.	36-54 bp	6,890 variable positions
Lee *et al.*	Outbreak cases	Canada (Low)	78 sequenced (out of 82)		22 years	Confirmation	Illumina MiSeq 250	H37Rv		Euro-American	Excluded SNPs with Phred score <50	50+ bp	
Luo *et al.*	Two clusters based on MIRU-VNTR and SNP typing	China (High)	32 sequenced	Population-based	1 year	Confirmation and Direction	Illumina HiSeq	H37Rv	Age: 17 – 79. Males and females. MDR and non-MDR.	Beijing	SNPs with coverage <3 and SNPs in the PE/PPE, PE-PGRS and drug-resistance associated genes	300 bp	SNP distances for linked cases: 0-100+

										were filtered		
Martin Williams *et al.*	Patients with identical MIRU-VNTR to first identified case	UK (Low)	4 (plus outbreak strain and 36 South Africa strains for comparison)		Confirmation	Illumina MiSeq	H37Rv					
Mehaffy *et al.*	Cluster based on spoligotyping and MIRU-VNTR	Canada (Low)	56 samples (53 patients)	17 years	Direction, Diversity and Confirmation	Illumina	H37Rv	Age: 20 – 74. Males and females. HIV present. All drug susceptible.		SNPs required a minimum read depth of 20X and a variant frequency of at least 75. SNPs in the PE, PPE and PE_PGRS gene were excluded.		722 SNPs compared to H37Rv
Ochereti na *et al.*	Samples sharing the same drug-resistance mutation	Haiti (High)	7 sequenced	5 years	Resistance	Illumina HiSeq 2000	H37Rv			Excluded SNPs in PPE, PE-PGRS and wag22 genes and where one or more samples displayed an ambiguous residue with over 20% match with reference alleles	50 bp	755 variant positions compared to H37Rv, 22 SNPs and 1 deletion between 6 samples
Pérez-Lago *et al.*	Epidemiologically supported MIRU-VNTR and RFLP clusters with at least one clonal variant	Spain (Low)	36	7 years	Diversity and Direction	Illumina HiSeq	MRCA of the MTBC	-	Euro-American	SNP calls of low quality: minimum coverage 10, minimum mapping quality of the SNP 20	51-101 bp	Within cluster SNP distances: 0-18
Regmi *et al.*	Cluster define by MIRU-VNTR and	Thailand (High)	4 samples sequence	6 years	Resistance	Illumina HiSeq 2000	H37Rv	-	Beijing	Phred quality score of ≤20 and SNVs with	100 bp	1242 common SNPs

	spoligotyping		d (54 total)								coverage of fewer than 10 reads were discarded. Additionally, heterozygous SNVs with allele frequencies of <75 % that were commonly present in all four samples were discarded		between outbreak samples and reference
Roetzer *et al.*	Large strain cluster (Haarlem lineage), identified by RFLP and MIRU-VNTR	Germany (Low)	86	Prospective population-based	14 years	Confirmation and Direction	Illumina	H37Rv	HIV present. Age: 2 – 83. Males and females. Drug susceptible.	Haarlem	SNPs needed a minimum coverage of 10 reads and a minimum allele frequency of 80% as thresholds for detection.		85 SNPs in sample. SNP distances between linked cases: 0-3
Schürch *et al.*	Harlingen cluster (RFLP with contact tracing)	Netherlands (Low)	3 sequenced (104 checked for 8 SNPs)		16 years	Direction and Recurrences	GS FLX Titanium		-	-	8 polymorphic SNPs were verified by subsequent resequencing on an ABI 3730xl sequencer	400 bp	8 SNPs between 3 samples
Smit *et al.*	Clustered with spoligotype and MIRU-VNTR	Finland (Low)	12 outbreak + 7 historical sequenced (14 in total)		1 year	Direction	-	-	Age: 16-23 years	-	Single-nucleotide polymorphisms (SNPs) were considered valid if supported by at least two and 70% of mapped reads on each strand with a minimum mapping quality of 45		
Stucki *et*	Cluster samples identified with SNP	Switzerland	69 samples		20 years	Direction	Illumina	Inferred common	Age: 34-53	-	SNPs with a coverage of ≥10	-	133 variable

al.	typing	(Low)	sequenced					ancestor of all MTBC lineages	years. HIV present. Males and females.		reads and Phred-score≥20. SNPs in "PE/PPE/PGRS," "maturase," "phage," "insertion sequence," or "13E12 repeat family protein" genes or with missing nucleotide calls in at least 3 samples were excluded. The short-read alignment tool SMALT was also used to call SNPs. Only positions called by both after filtering were included.		positions amongst the 69 samples
Walker *et al.*	Random cross-sectional and longitudinal samples from single patients. Samples from community MIRU-VNTR and household clusters	UK (Low)	390 samples (254 patients)	Retrospective observational	Archived between 1994 and 2011	Confirmation and Diversity	Illumina HiSeq	H37Rv	-	Beijing, European American, Central Asian, East-African Indian	>75% of reads needed to support variant calls, which had to be homozygous in a diploid model. Only variants supported by ≥5 reads, including one in each direction that did not occur at sites with unusual depth and were not within 12 bp of another nucleotide variant, were accepted.	75 bp	Pairwise SNP distances: 0-5 for linked cases, 0-150 for unlinked cases 1,096 SNPs was the largest pairwise distance between longitudinal samples
Walker *et*	Unselected,	UK (Low)	247	Observational	6 years	Confirmati	Illumina	H37Rv	Age: 1-89	-	Variant calls in		SNP

al.	geographically restricted population					on	HiSeq				non-repetitive regions were made providing they were supported by ≥5 reads, including one in each direction. Sites where minority variants represented >10% of read depth were defined as mixed and no base called.	distances between linked cases: 0-7 (median 1). Median pairwise SNP distances 1106 (857-1715) without secondary cases from each genomic cluster
Witney *et al.*	Six hospital patients with suspected XDR-TB	UK (Low)	16 samples (6 patients)		7 years	Confirmation	Ion Torrent personal genome machine	H37Rv	-	Beijing	Mapping quality of >30, site quality score of >30, ≥4 reads covering each site with ≥2 reads mapping to each strand but with a maximum depth of coverage of 200x, ≥75% of reads supporting the site, and an allelic frequency of 1.	33-297 pairwise SNP distances

Table 9.10 Data extraction for the included studies

bp = base pairs, MRCA = Most recent common ancestor, MTBC = *M. tb* complex. *TB burdens of countries were taken from World Health Organization [24] with high burden defined as >40 cases/100,000

9.1.5 FACTORS AFFECTING THE NUMBER OF POLYMORPHISMS DETECTED IN SEQUENCES

Study factors	Effect on number of SNPs detected
Study duration	Assuming that mutations occur and become fixed as time evolves, the longer the duration of study the more polymorphisms that will have occurred and been fixed in the population (so this affects the number of SNPs found in the study overall not between related cases)
Strain diversity	If there are highly diverse strains in the population then large SNP distances will be found between pairs of sequences
Sequencing machine	Sequencing machines (e.g. Illumina) require the sample to be cultured before it is sequenced. This can reduce the number of polymorphisms detected by causing a bottleneck
Length of reads	The longer the read, the more SNPs found [211]
Coverage	The deeper the coverage, the more polymorphisms likely to be found
Definition of quality read/SNP	The definition of a quality read will affect the number of SNPs 'confirmed' as the definition relies on support from a certain number of reads. Thus, more stringent rules on quality reads will mean fewer reads to support variants. Stringent definitions of SNPs requiring high confidence in variants will result in fewer SNPs found.
Bioinformatics software	Factors such as the internal filtering criteria may affect the number of polymorphisms found [212]
Number of amplification steps	The more amplification steps, the more errors are likely to be introduced [55] resulting in polymorphisms

Table 9.11 The effect of study specific factors on the number of polymorphisms detected in sequences

9.2 APPENDIX 2 - TRANSPHYLO WITH EPIDEMIOLOGICAL DATA

When including epidemiological data on the individuals, such as locations and infectiousness, Equation 4.1 becomes:

$$P(T, \varepsilon, Neg|G, A) \propto P(G, A|T, \varepsilon, Neg)P(\varepsilon, Neg, T)$$

Assuming that G and A are independent given T, this can be re-written as

$$P(T, \varepsilon, Neg|G, A) \propto P(G|T, Neg)P(A|T, \varepsilon)P(T|\varepsilon)\pi(\varepsilon, Neg)$$

The second term represents the probability of the epidemiological data given that the transmission events are fixed in the tree, for example, the probability that i and j are in locations A and B respectively given they infected each other. Our description of $P(A|T, \varepsilon)$assumes an open population and merely that we are dealing with the observed samples.

This could be implemented into the MCMC inference in the same way Didelot *et al*. [99] did using a penalty system for when transmission events in the proposed transmission tree do not correspond well with the known epidemiological information.

9.3 APPENDIX 3 – BIOINFORMATIC AND PHYLOGENETIC ANALYSIS

9.3.1 IQ-TREE COMMANDS

In order to determine the best substitution model and produce a ML tree, IQ-TREE was run via the web server provided by the Center for Integrative Bioinformatics Vienna, Austria at http://iqtree.cibiv.univie.ac.at/. The command that was run using the FASTA file containing the SNPs across all the genomes (londonOutbreakSNPs.fa) was the following:

path/to/iqtree -s londonOutbreakSNPs.fa -st DNA -m TEST+ASC -bb 1000 -alrt 1000

9.3.2 IQ-TREE RESULTS

Rate parameter R:

Substitution	Rate
A-C	1.3066
A-G	2.9982
A-T	0.1512
C-G	1.3563
C-T	2.9982
G-T	1.0000

Table 9.12 Substitution rates determined by IQ-TREE for the genomic data

State frequencies: (equal frequencies)

Rate matrix Q:

A	-0.9084	0.2664	0.6112	0.03082
C	0.2664	-1.154	0.2765	0.6112
G	0.6112	0.2765	-1.092	0.2039
T	0.03082	0.6112	0.2039	-0.8459

Model of rate heterogeneity: Uniform

9.3.3 BEAST RESULTS

Operator (parameter)	Chain 1	Chain 2	Chain 3
Scale (kappa)	0.2376	0.2364	0.2367
Frequencies	0.2322	0.2319	0.2317
pInv	0.2332	0.2335	0.2335
Scale (uced.mean)	0.235	0.2354	0.2348
Up:nodeHeights(treeModel) down:uced.mean	0.2316	0.2318	0.2314
swapOperator(branchRates.categories)	0.7439	0.7442	0.744
uniformInteger(branchRates.categories)	0.824	0.8236	0.8235
subtreeSlide(treeModel)	0.2334	0.2332	0.2329
Narrow Exchange(treeModel)	0.4486	0.4487	0.4489
Wide Exchange (treeModel)	0.0441	0.0435	0.0451
WilsonBalding(treeModel)	0.1237	0.1231	0.125
Scale(treeModel.rootHeight)	0.2403	0.2402	0.2402
Uniform(nodeHeights(treeModel))	0.7419	0.7418	0.7421
Scale(constant.popSize)	0.2343	0.2344	0.2344

Table 9.13 Acceptance rates for the parameters of the MCMC model for each chain using the optimal model settings

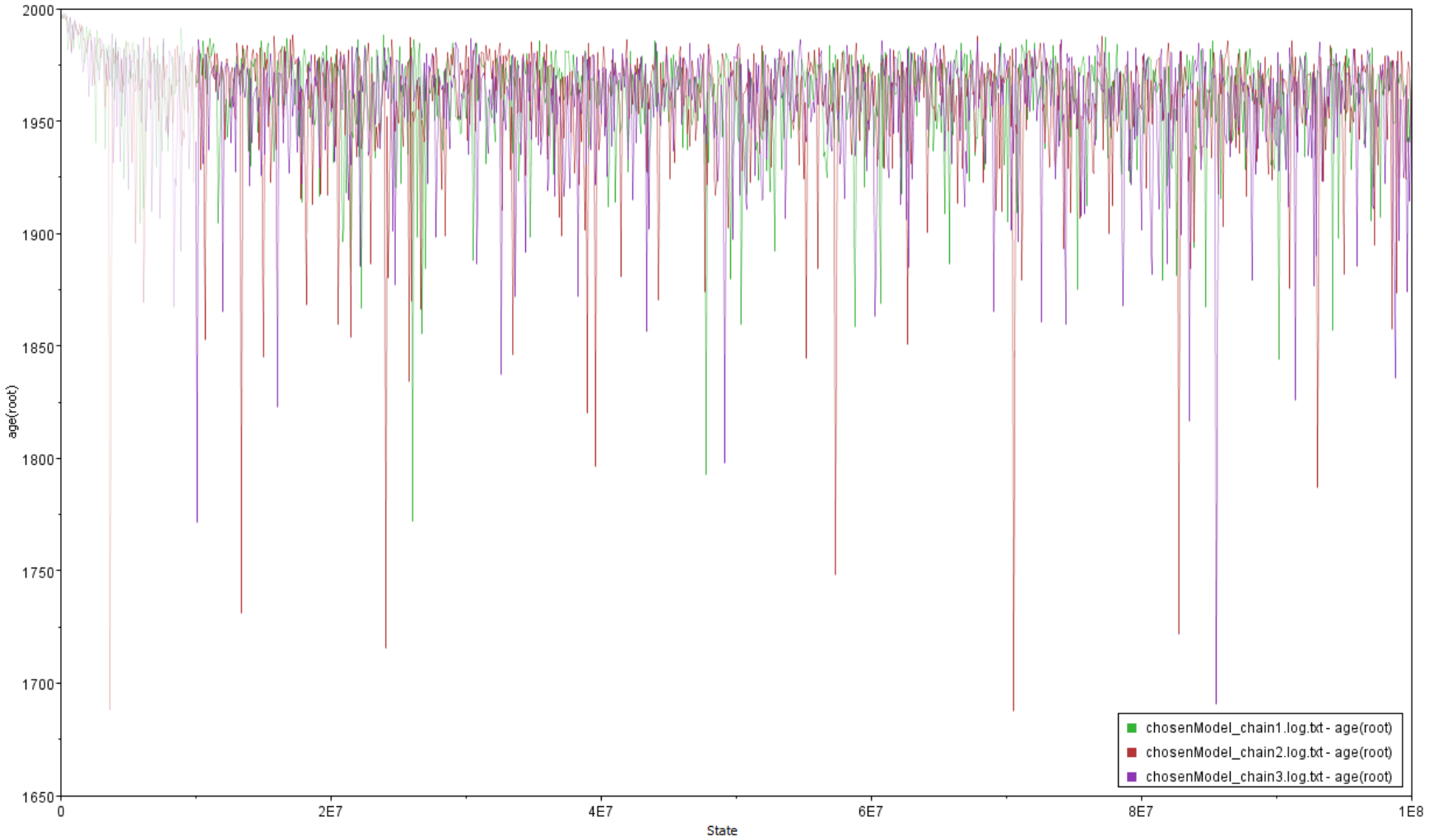

Figure 9.1 Trace plot for the parameter age(root) for all three chains

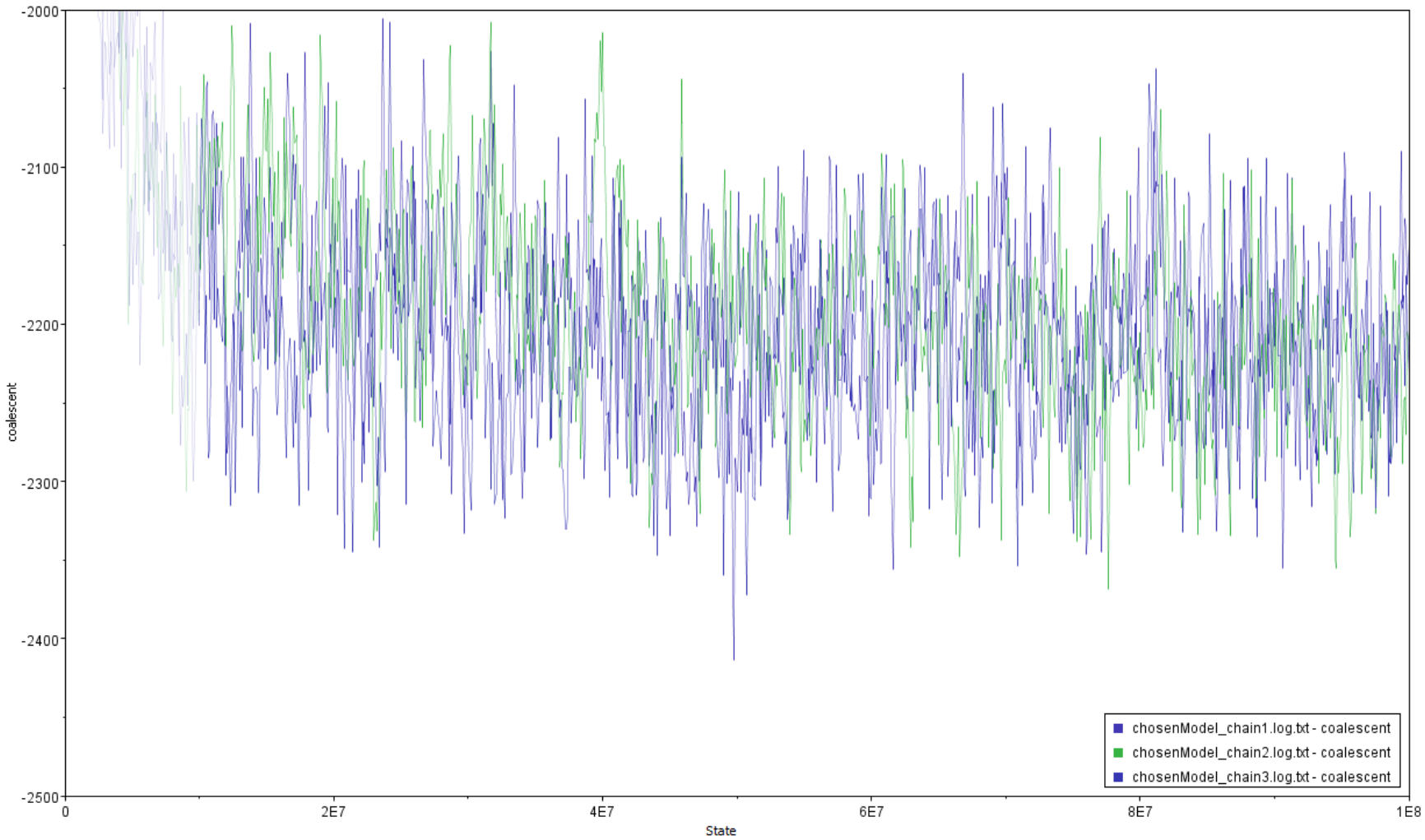

Figure 9.2 Trace plot for the parameter coalescent for all three chains

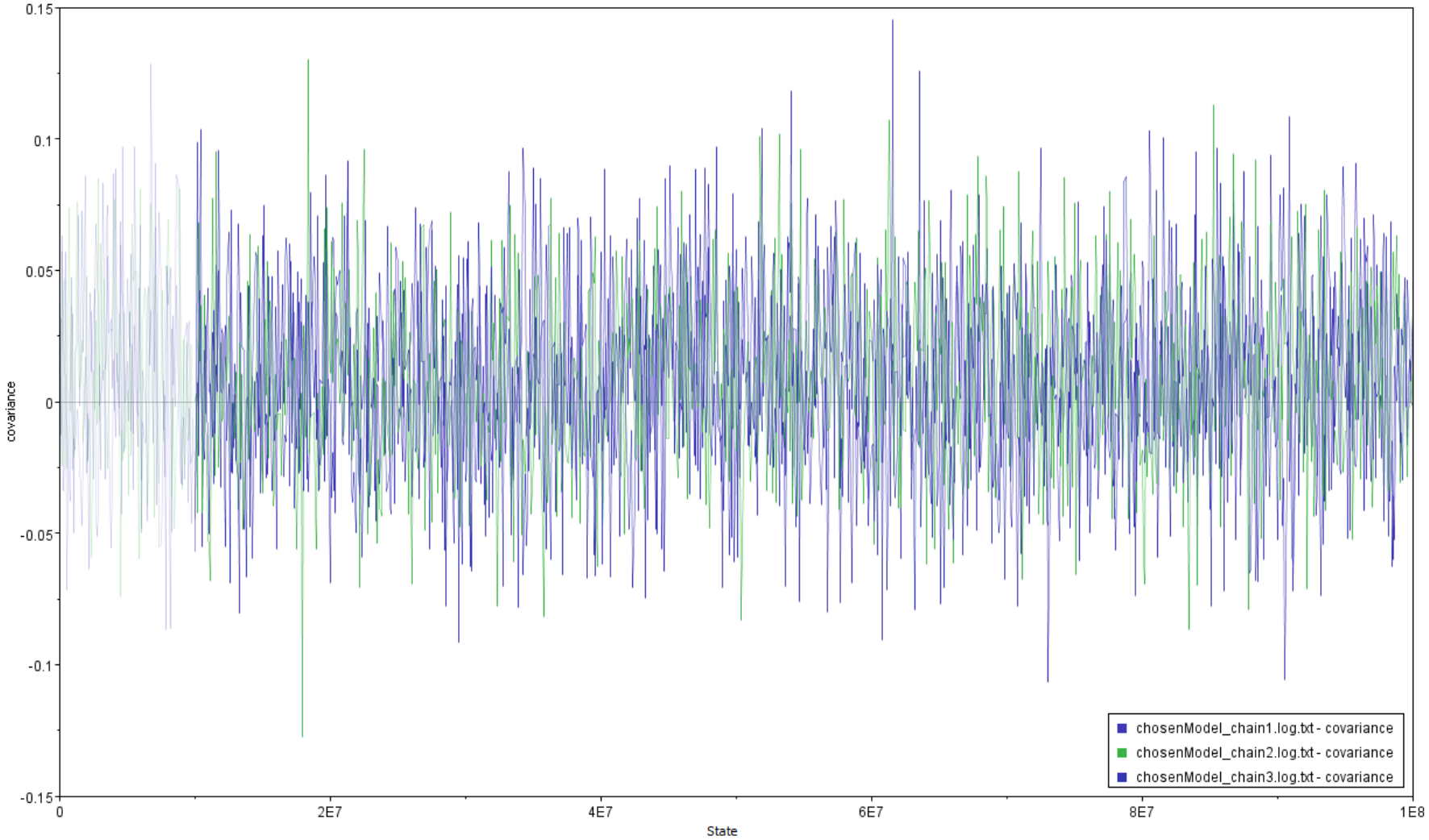

Figure 9.3 Trace plot for the parameter covariance for all three chains

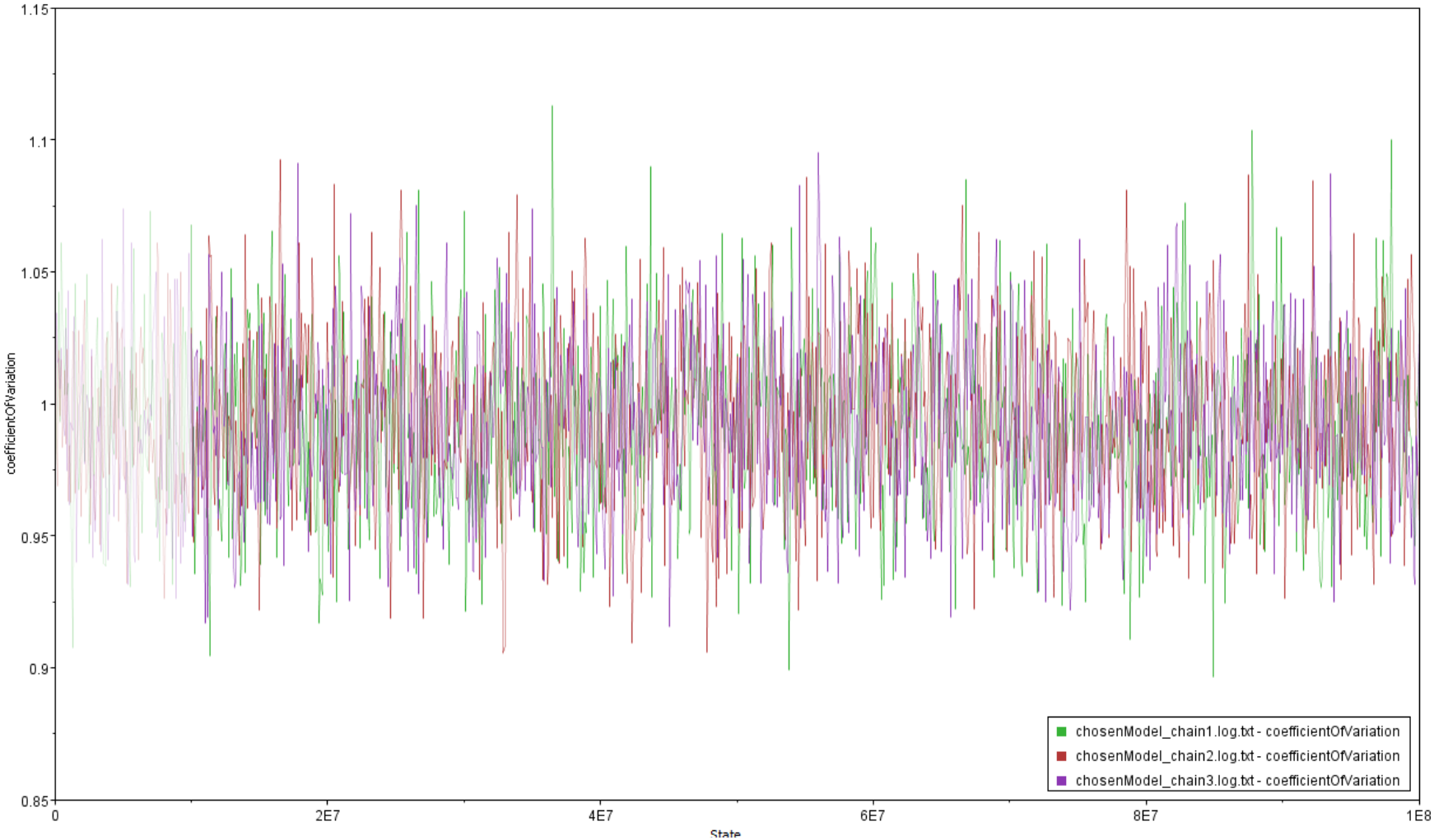

Figure 9.4 Trace plot for the parameter coefficientOfVariation for all three chains

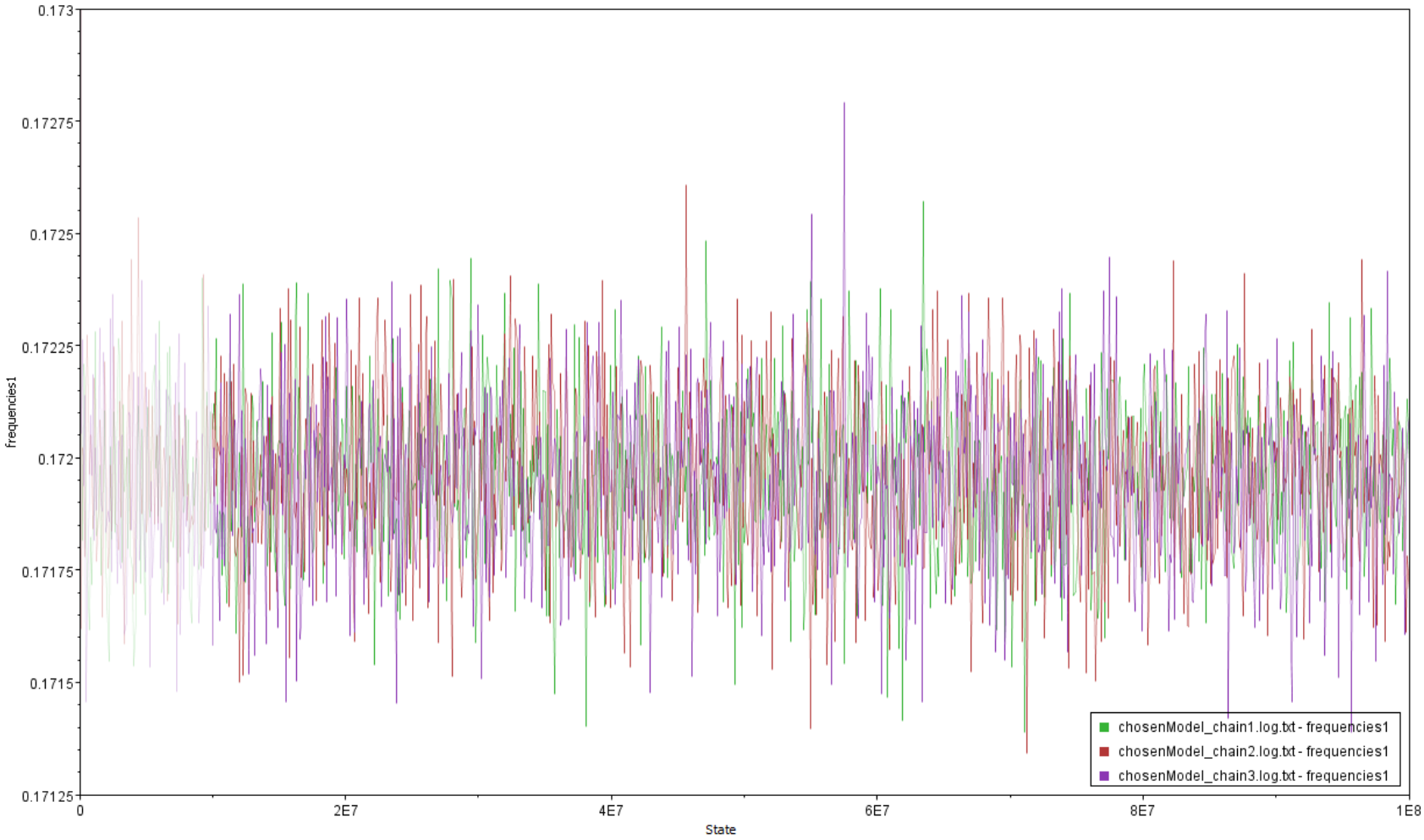

Figure 9.5 Trace plot for the parameter frequencies1 for all three chains

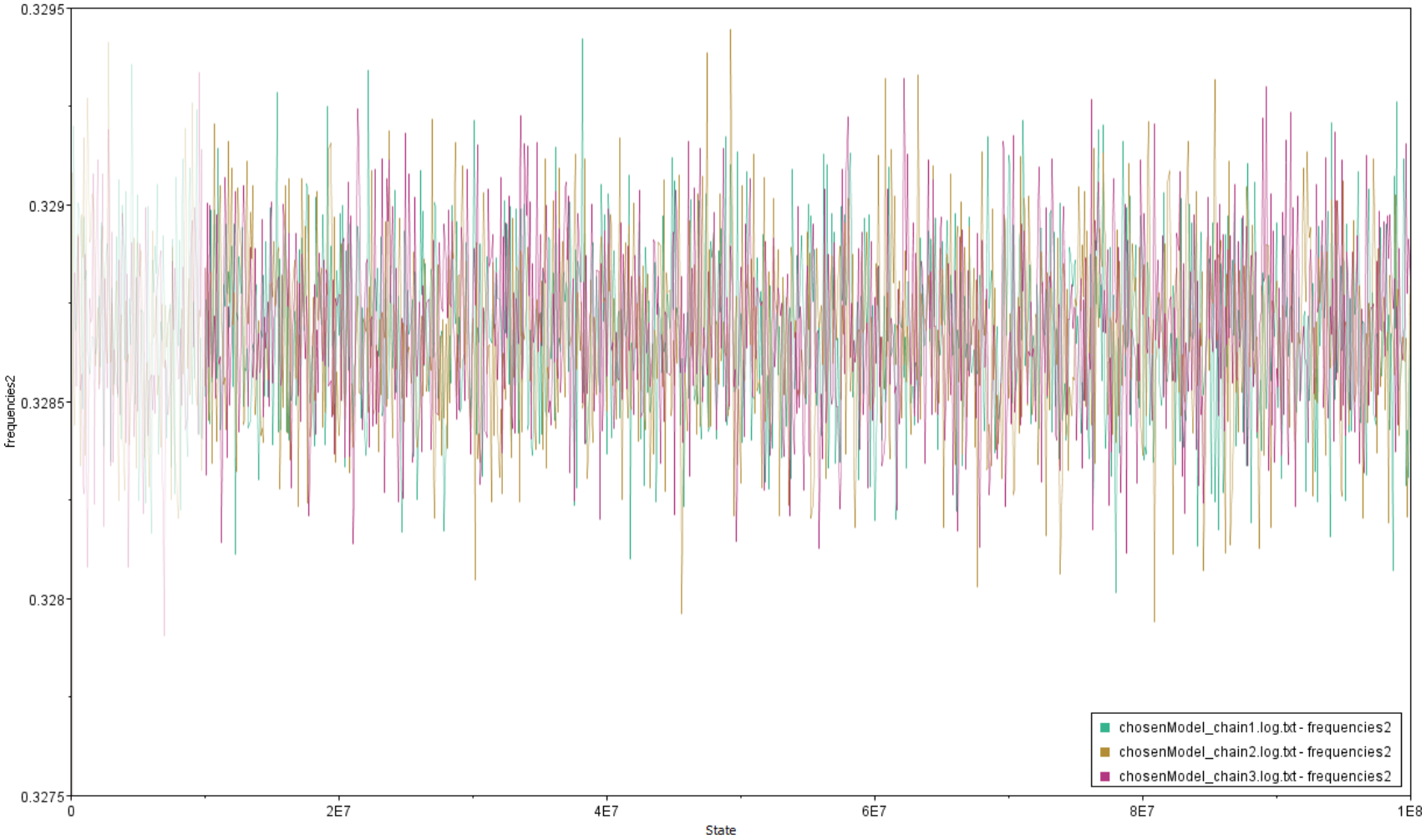

Figure 9.6 Trace plot for the parameter frequencies2 for all three chains

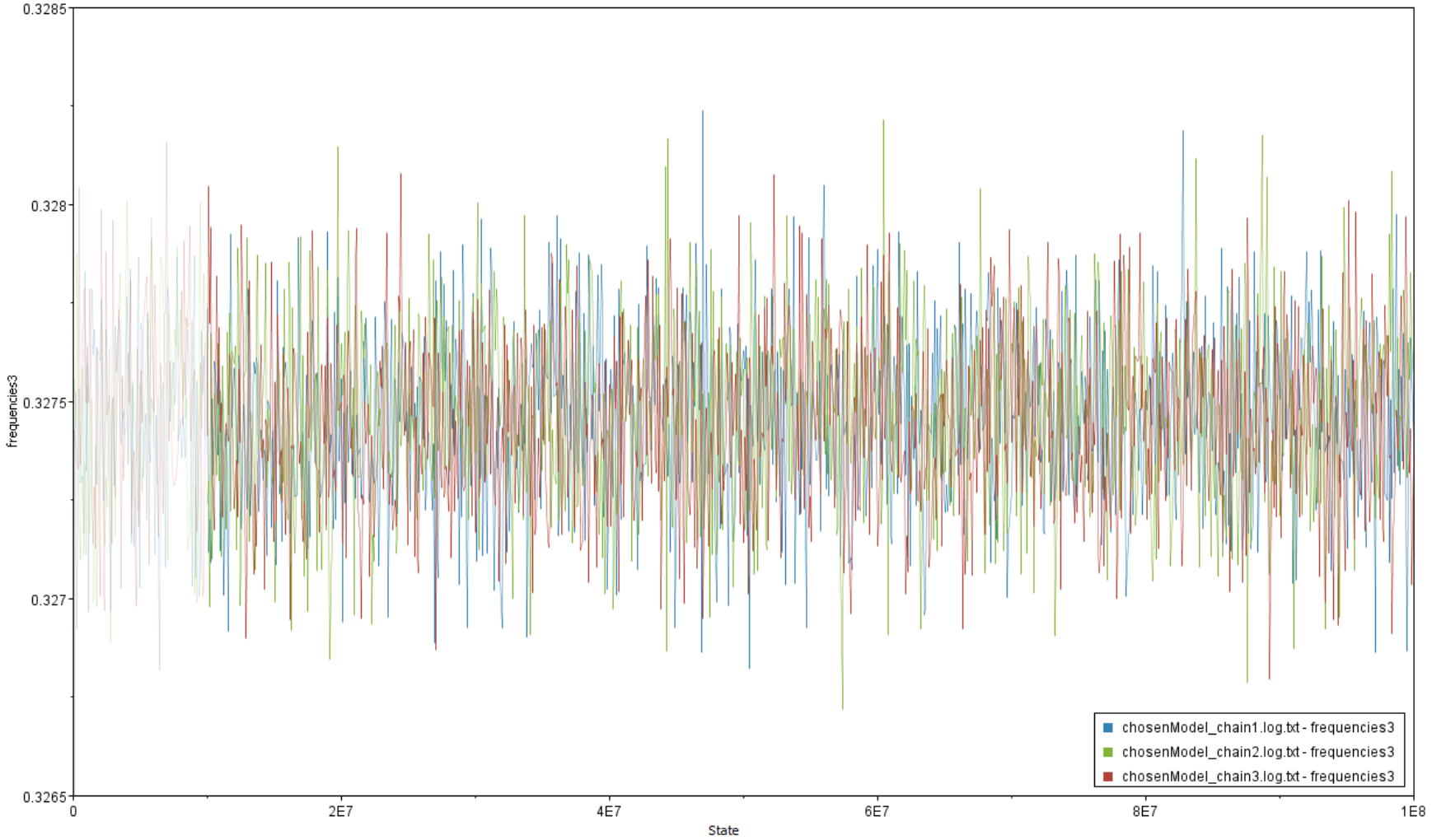

Figure 9.7 Trace plot for the parameter frequencies3 for all three chains

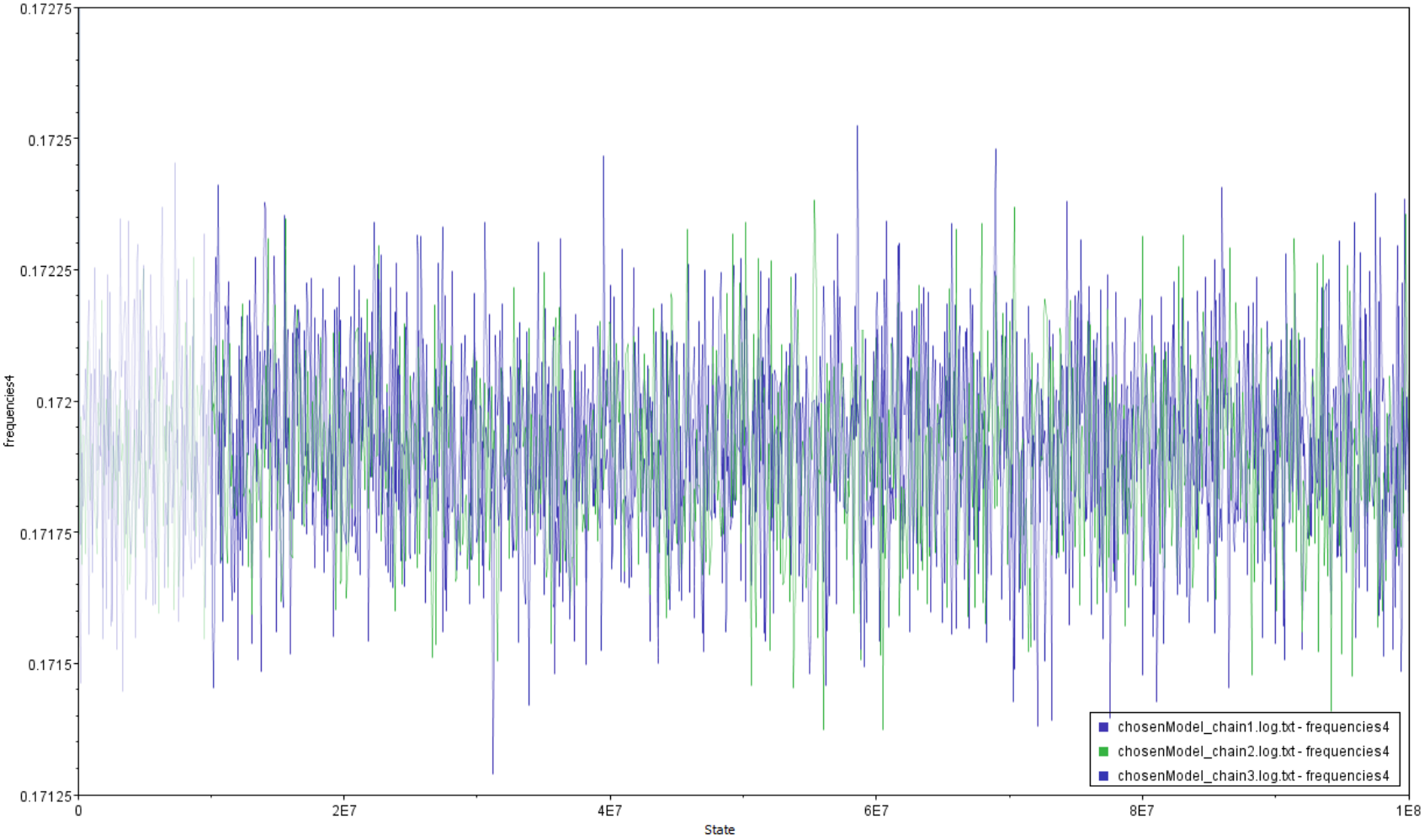

Figure 9.8 Trace plot for the parameter frequencies4 for all three chains

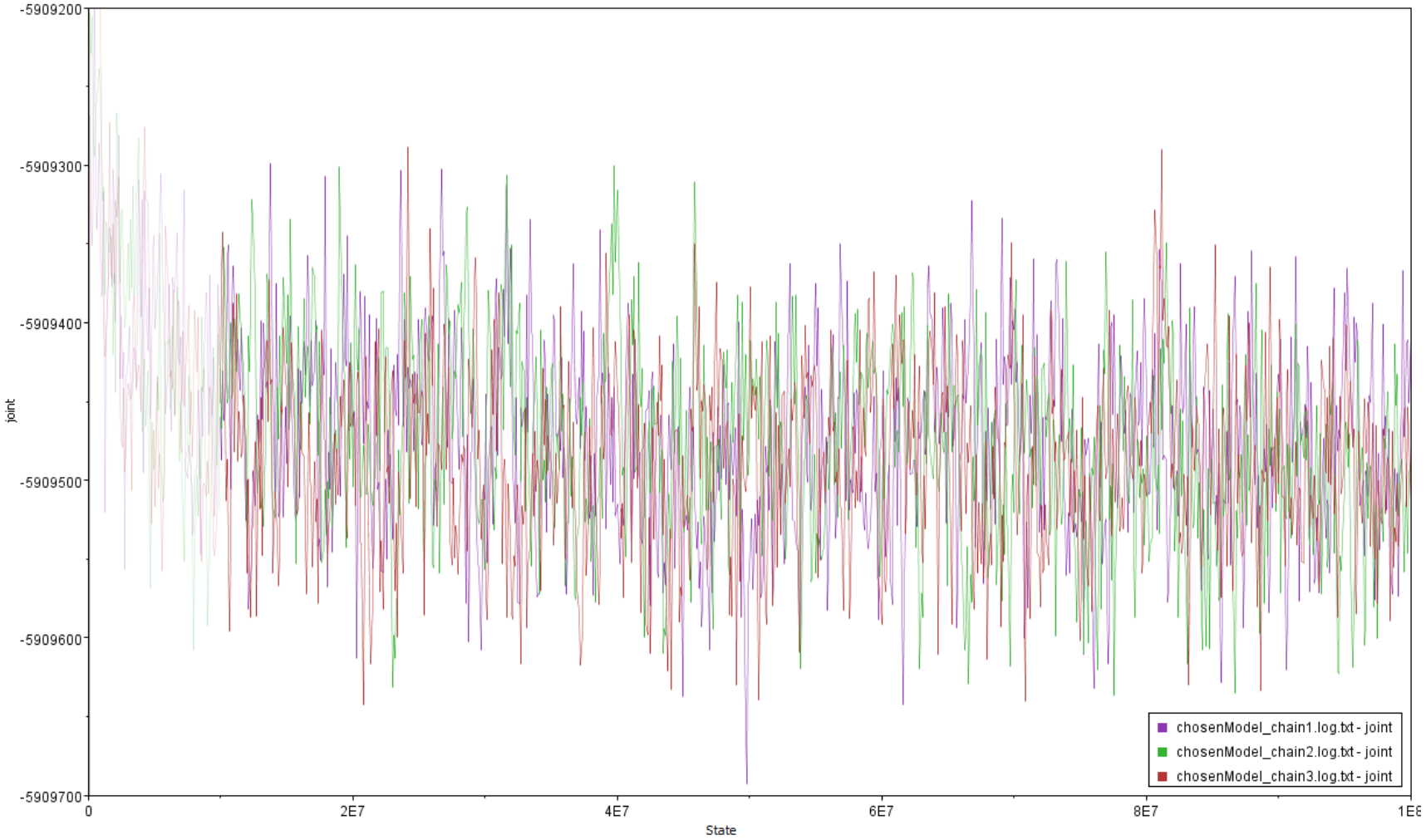

Figure 9.9 Trace plot for the parameter joint for all three chains

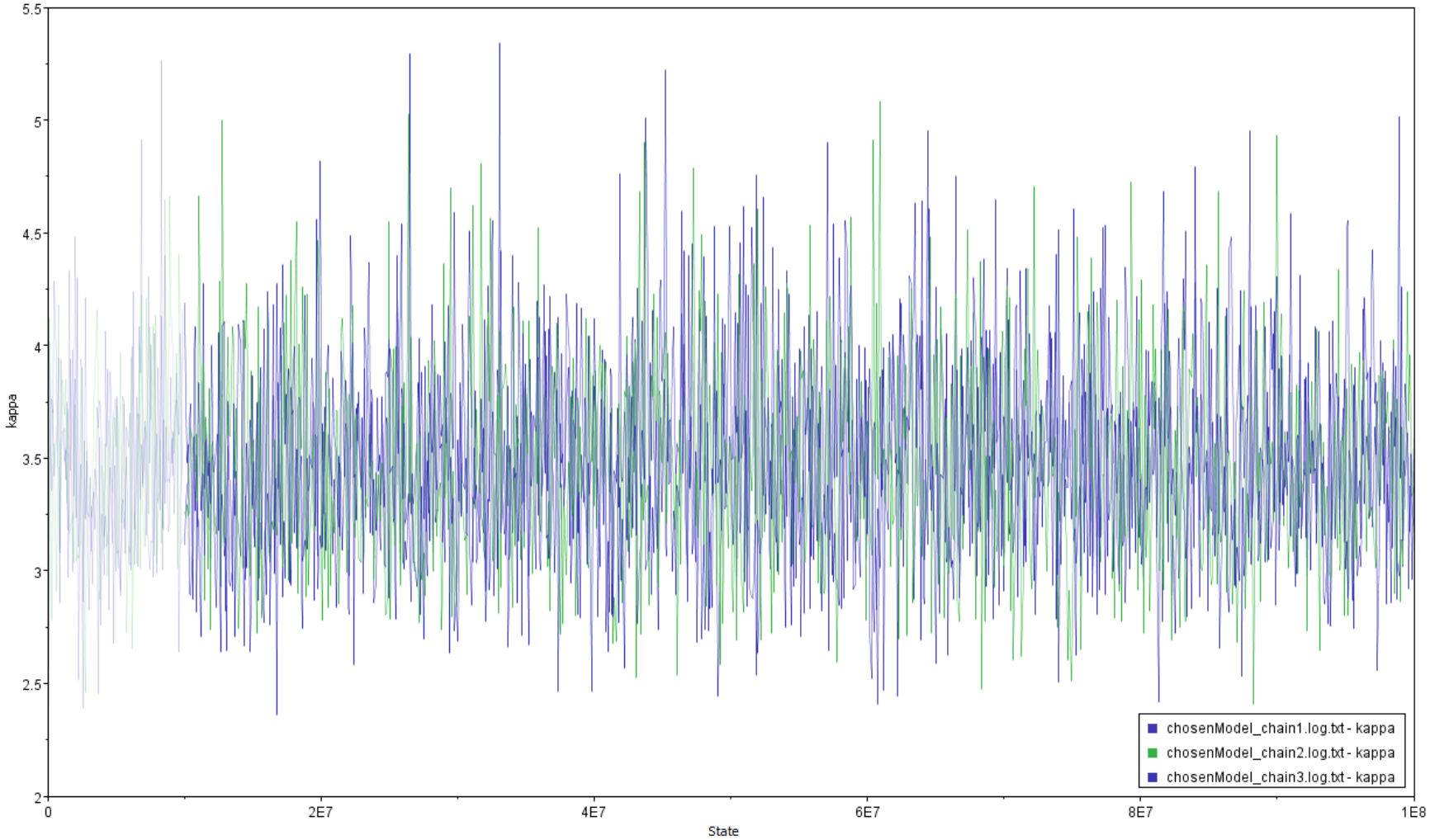

Figure 9.10 Trace plot for the parameter kappa for all three chains

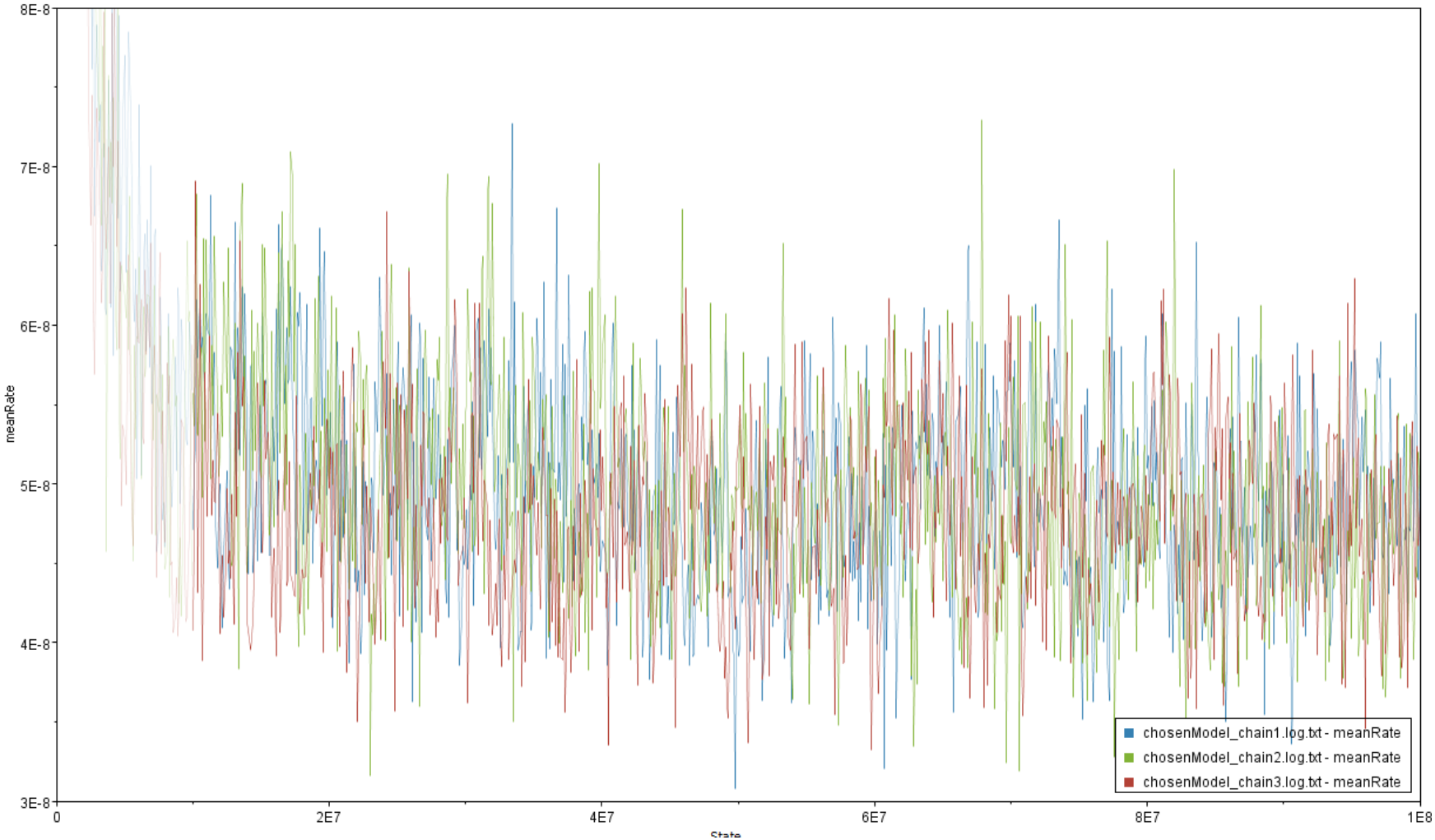

Figure 9.11 Trace plot for the parameter meanRate for all three chains

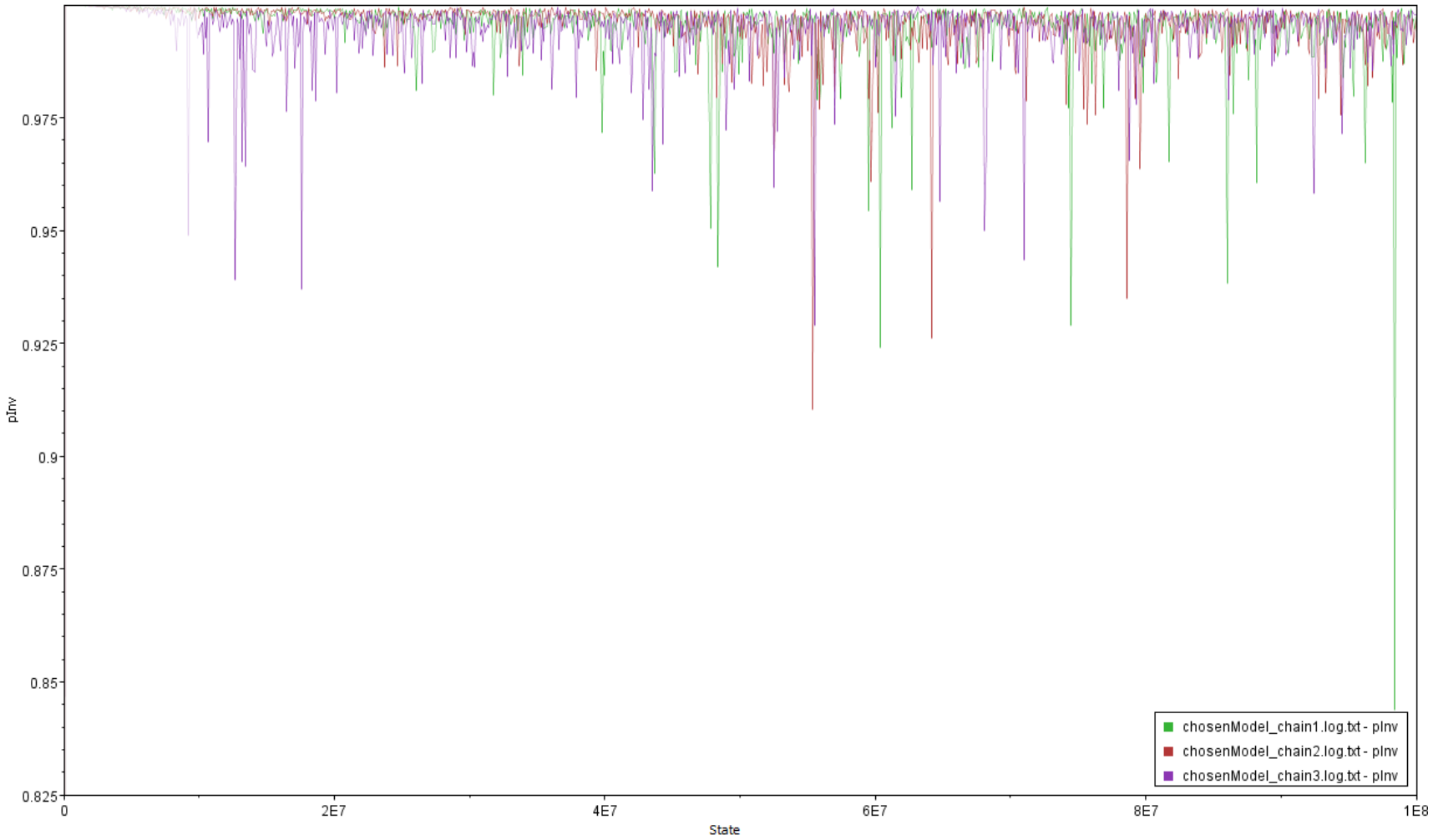

Figure 9.12 Trace plot for the parameter pInv for all three chains

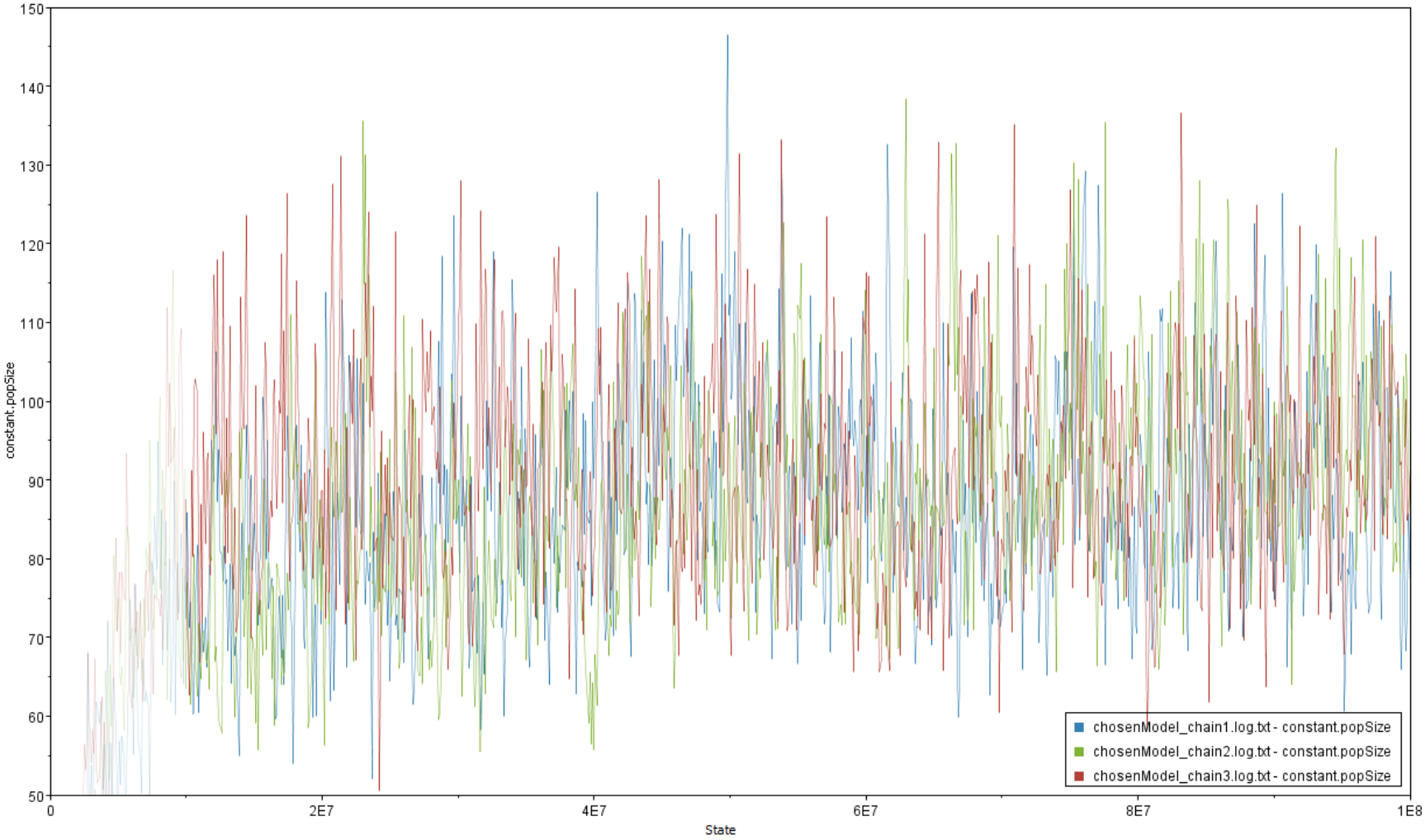

Figure 9.13 Trace plot for the parameter constant.popSize for all three chains

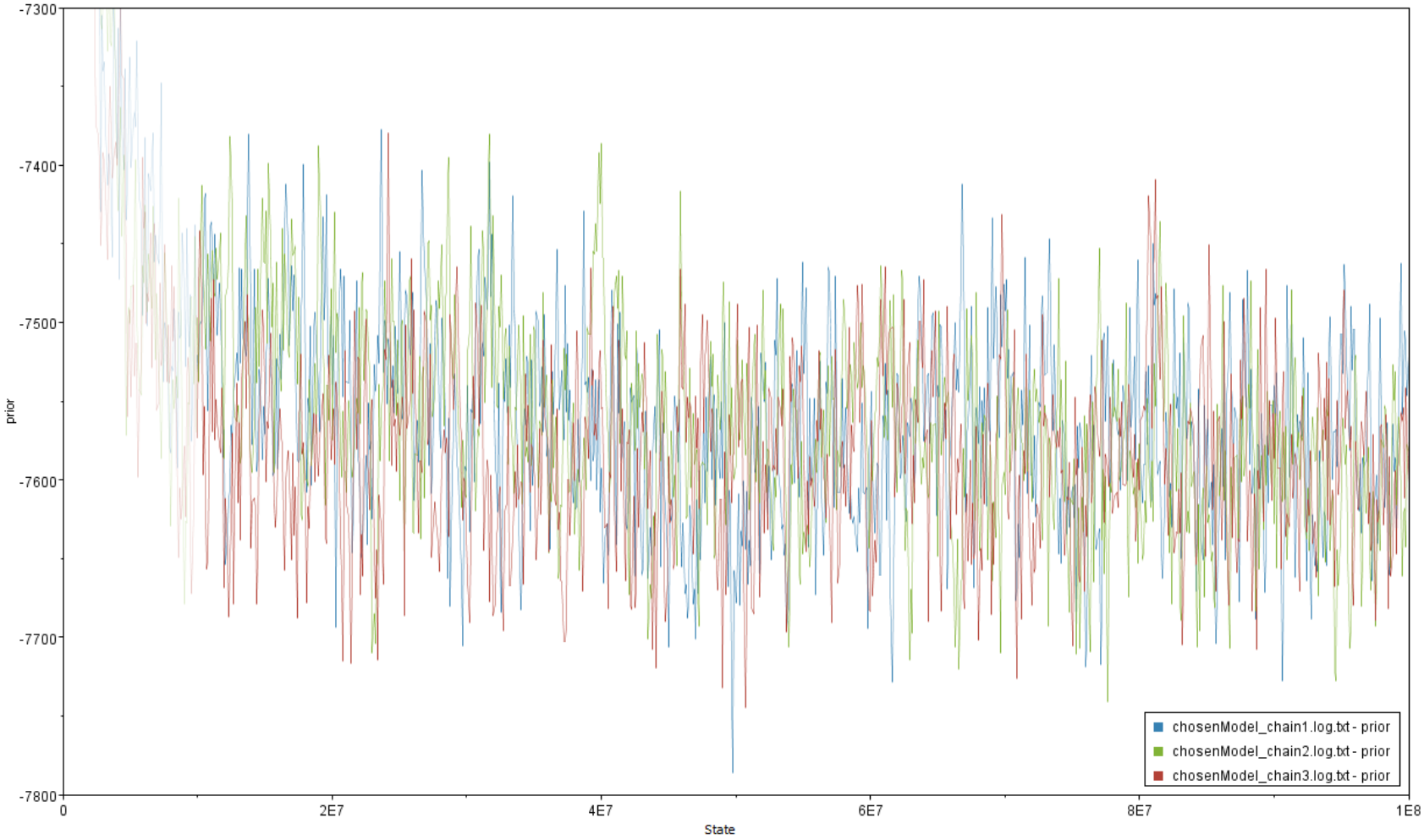

Figure 9.14 Trace plot for the parameter prior for all three chains

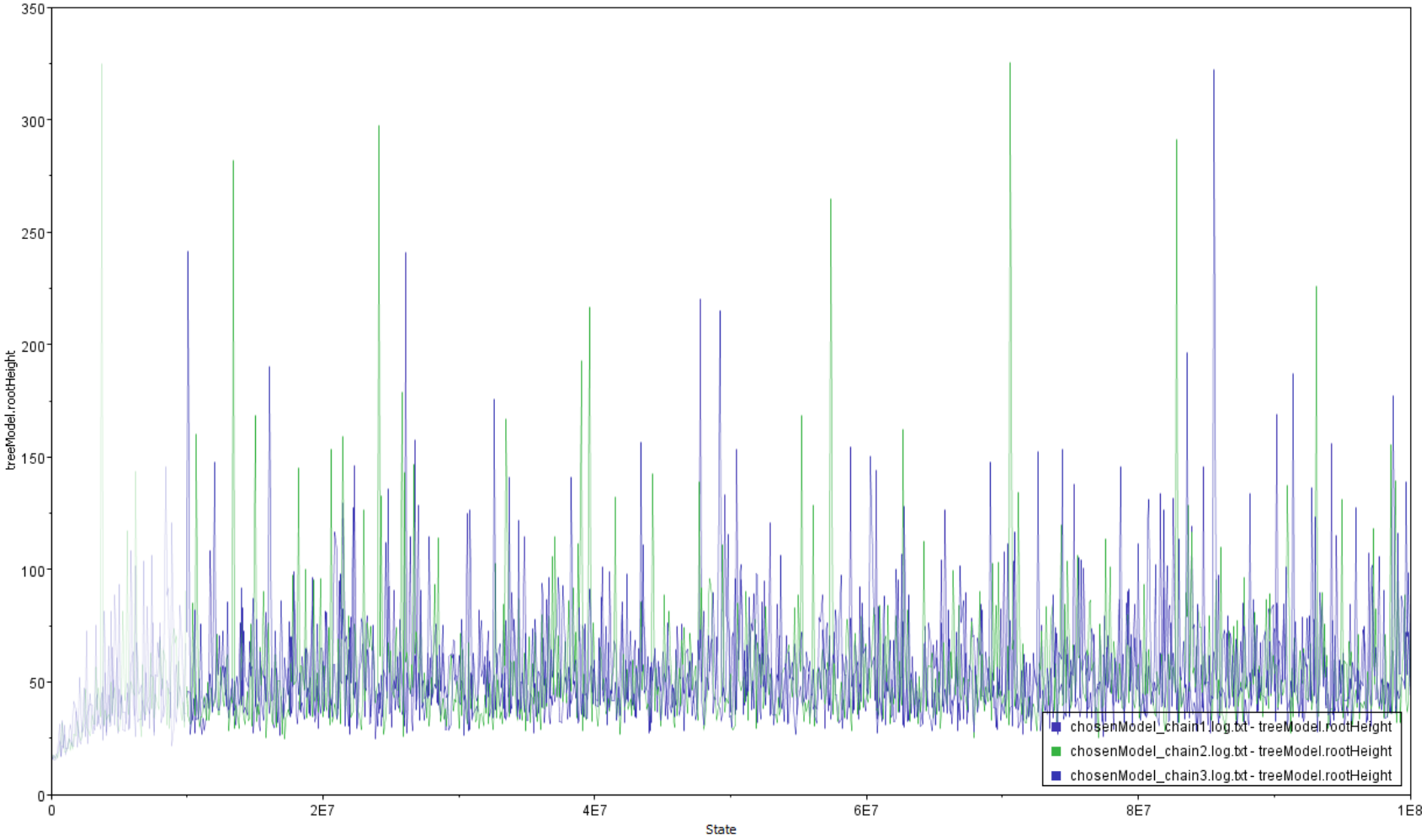

Figure 9.15 Trace plot for the parameter treeModel.rootHeight for all three chains

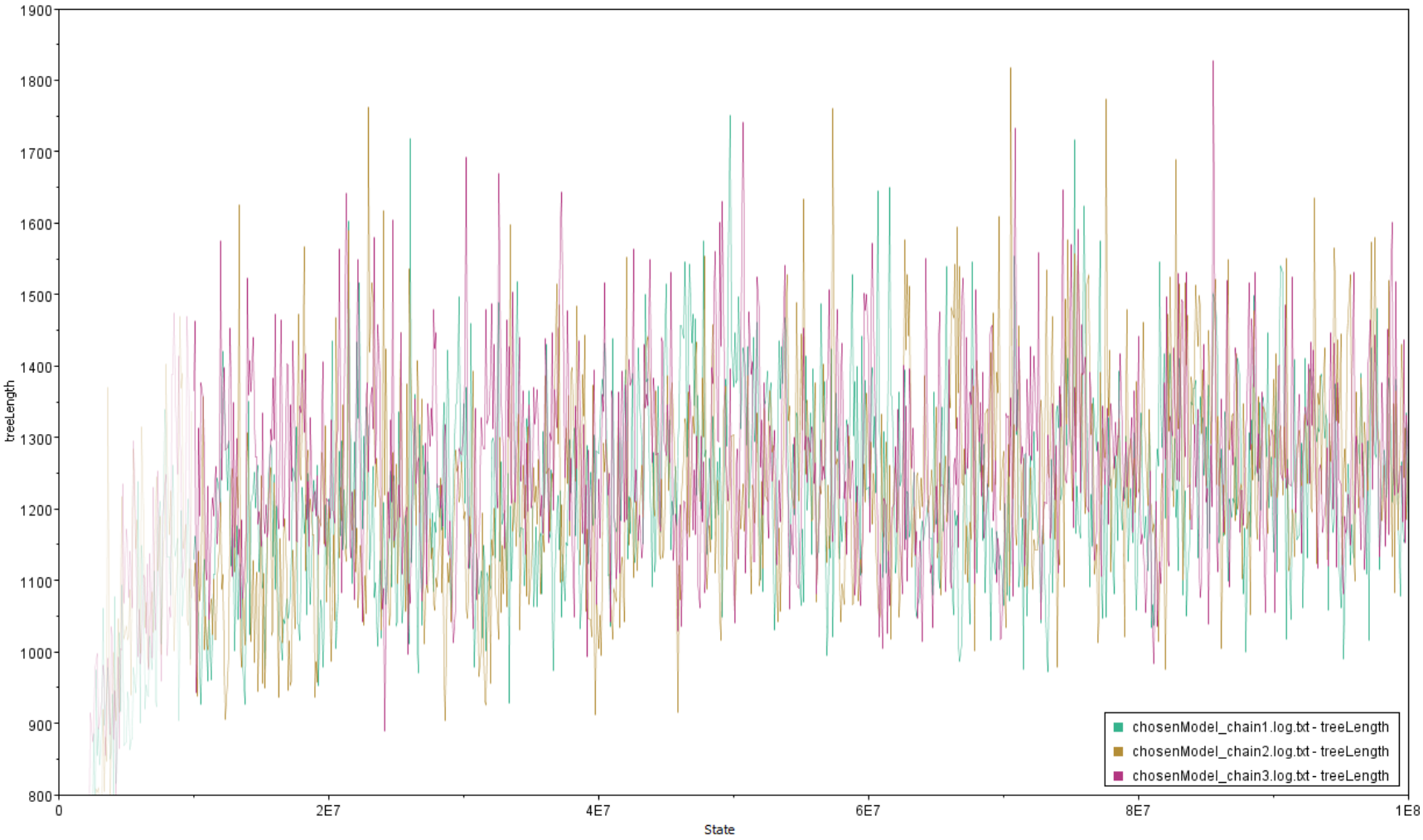

Figure 9.16 Trace plot for the parameter treeLength for all three chains

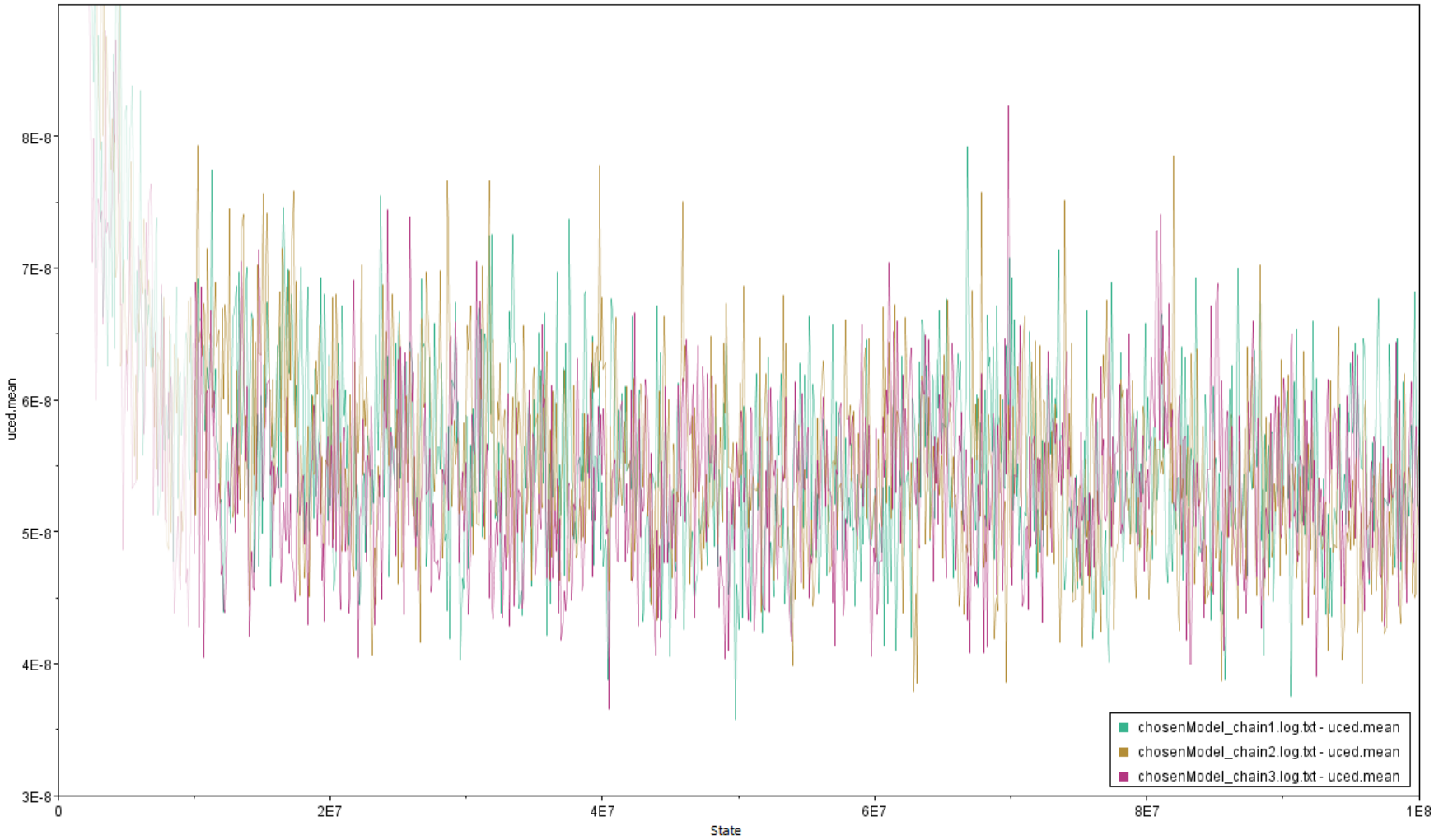

Figure 9.17 Trace plot for the parameter uced.mean for all three chains

Operator (parameter)	Chain 1	Chain 2	Chain 3	Operator (parameter)	Chain 1	Chain 2	Chain 3
joint	198	97	288	frequencies3	901	901	789
prior	36	13	268	frequencies4	901	828	774
llikelihood	11	7	374	pInv	301	46	801
treeModel.rootHeight	537	901	901	uced.mean	98	23	333
age(root)	537	901	901	meanRate	52	22	343
treelength	45	21	348	coefficientOfVariation	624	725	827
constant.popSize	44	14	320	covariance	845	901	790
kappa	901	901	640	treeLikelihood	11	7	374
frequencies1	901	901	901	branchRates	-	-	-
frequencies2	901	901	901	coalescent	36	13	267

Table 9.14 Effective Sample Size (ESS) values for the parameters of the MCMC model for each chain using the optimal model setting