

MODELLING AND GENOMIC ANALYSIS OF
COMPETITION AND DIVERSITY IN
Mycobacterium tuberculosis

A THESIS PRESENTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY OF IMPERIAL COLLEGE LONDON

AND THE
DIPLOMA OF IMPERIAL COLLEGE

BY
DIEPREYE AYABINA

DEPARTMENT OF MATHEMATICS
IMPERIAL COLLEGE
180 QUEEN'S GATE, LONDON SW7 2BZ

FEBRUARY 2018

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Signed: Diepreye Ayabina

COPYRIGHT

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

ABSTRACT

Numerous studies have identified tuberculosis patients in whom more than one distinct strain of *Mycobacterium tuberculosis* (*M. tuberculosis*) is present. The diversity of *M. tuberculosis* can have dramatic effects on disease dynamics. This thesis focuses on the study of diversity of *M. tuberculosis* and competition between its strains by analysing mathematical models and applying statistical techniques to clinical, genetic and epidemiological data.

Mathematical models of *M. tuberculosis*, both in-vitro and within host are developed and analysed. Single strain models are analysed and then extended to incorporate the interaction of two or more *M. tuberculosis* strains. We find that during active disease, competition between strains is not as severe as during latency. Analysis of the within host models using approaches from data science identify key model parameters that affect the outcome of infection. These models are further explored using virtual experiments to answer questions such as how does re-infection affect disease progression?

Evolutionary tools, especially phylogenetic trees, are increasingly being used to study short-term variation in *M. tuberculosis*. Some regions of a genome sequence may be disruptive in a phylogenetic framework. We propose a phylogeny-based method to detect phylogenetically disruptive sites along a multiple sequence alignment and illustrate the effect excluding these sites has on onward inference of the phylogeny.

In many Western countries tuberculosis (TB) incidence is low and largely shaped by immigrant populations originating from high-burden countries. We combine whole genome sequence data, times of arrival in Norway and case presentation times to estimate the time of transmission for individual patients. We focus on genomic clusters of patients originally from the horn of Africa. We find that there is strong evidence of ongoing TB transmission in Norway within these populations. These results show how genomic and epidemiological data can be combined to provide useful information for public health.

ACKNOWLEDGMENTS

I would first like to express my gratitude to my supervisor, Dr. Caroline Colijn, who inspired this project. She is an amazing mentor and it has been a real pleasure working with her and gaining from her vast knowledge.

Thanks to our collaborators: Joanna Bacon, Louis Grandjean, Vegard Eldholm who provided data and also worked with us on different projects.

I would also like to thank the members of Caroline's research group at Imperial College for many stimulating discussions and for sharing their work at group meetings.

I am grateful to Imperial College for its welcoming spirit, and for providing the right space for research. Thanks to the Nigerian universities commission for providing funding for this work.

Finally I would like to thank my family and friends for their support, patience and encouragement. Also a warm thank you goes to my dear friend Emeka for proofreading this work.

To Jehovah Rohi.

LIST OF FIGURES

2.1	<i>Plot showing how bacterial cell numbers change under fast (green) and slow (blue) dilution rates</i>	27
2.2	<i>Growth chamber of the chemostat</i>	28
2.3	<i>Model simulations for system 2.3, for both dilution rates. . . .</i>	33
2.4	<i>Grey lines (A) show the fits from 500 parameter sets randomly drawn from the posterior distribution of viable bacterial numbers for slow and fast dilution rates, the red lines (B) are the B_1 population (INH sensitive), the blue lines are the B_2 population (INH resistant), and the cyan (C) lines are plots of the resource concentration.</i>	39
2.5	<i>Plots of the resource-dependent growth rate $\phi_i(R)$ for fast and slow dilution rates over time. Note that growth is frequently greater than the dilution rate due to the departure from steady-state conditions.</i>	40
2.6	<i>Histograms of 1000 random samples of posterior parameter distributions for slow and fast dilution rates.</i>	41
2.7	<i>Heatmaps (A) showing how culture recovery depends on fitness costs and histograms (B) of relative fitness $f \sim \frac{\lambda_2/k_2}{\lambda_1/k_1}$ of 1000 posterior parameter sets . Recovery strength ranges from dark blue (weakest recovery) to bright yellow (strongest recovery). . .</i>	41
2.8	<i>Grey lines show the fits from 500 parameter sets randomly drawn from the posterior distribution resulting from running the MCMC separately on the individual replicated cultures . .</i>	42
2.9	<i>Boxplots of 5000 random samples of posterior parameter distributions for slow and fast dilution rates (for individual cultures).</i>	43
2.10	<i>Correlation plots of 1000 random samples of posterior parameter distributions for slow and fast dilution rates. Dark blue and dark red colors display parameters with high positive and negative linear dependencies, respectively</i>	44

2.11	<i>Heat maps of the joint posterior distributions of parameters that are significantly correlated (with a correlation coefficient $r \geq 0.5$). Lighter, brighter areas correspond to areas of higher posterior density.</i>	45
2.12	<i>Heat maps of the joint posterior distributions of parameters that are significantly correlated (with a correlation coefficient $r \geq 0.5$). Lighter, brighter areas correspond to areas of higher posterior density.</i>	46
2.13	<i>Proportion plots and plots of the Shannon diversity index H' for 1000 parameter sets of the multi-population model: with twenty resistant sub-populations competing after the susceptible sub-population(black) dies off. The fittest sub-population (green) maintains the highest proportion under both dilution rates.</i>	47
3.1	<i>Outcomes of exposure to Mycobacterium tuberculosis.</i>	55
3.2	<i>Schematic of model 3.1 describing immune response to tuberculosis infection. The three model compartments are depicted as M (macrophages), B (bacteria) and T (T-cells).</i>	59
3.3	<i>Histogram showing final bacterial load: 10000 parameter sets were drawn from the ranges in Table 3.1</i>	61
3.4	<i>Model simulation time course for latency (green) and active (yellow) classes.</i>	61
3.5	<i>Boxplots of $\alpha_T, \mu_T, i, a_1, a_2, a_3, a_4, N_1, c_1, c_2, c_3, \mu_M$ and N_2 for the three disease outcomes.</i>	62
3.6	<i>Correlation matrices for parameters in the different disease outcomes: (a) clearance, (b) latency (c) active disease, colors range from dark red ($r = -1$) to dark blue ($r = 1$), where r is the correlation coefficient. Parameter pairs with $r > 0.1$ are colored blue ($r > 0$) and pink ($r < 0$).</i>	63
3.7	<i>DAPC scatter plot for model 3.1. The x-axis is a linear combination of parameters corresponding to the first discriminant function, and the y-axis corresponds to the second.</i>	64
3.8	<i>Contributions of parameters to the first and second discriminant functions of the DAPC of model 3.1. The height of each bar is proportional to the contribution of the corresponding parameter to the discriminant function. Only parameters whose contribution is above an arbitrary threshold (grey horizontal line) are indicated for the sake of clarity.</i>	65

3.9	<i>Heatmaps illustrating how disease outcome depends on combinations of key parameters (as determined by DAPC)</i>	65
3.10	<i>Importance of each parameter to the random forest classifier with each parameter on the y-axis, and their importance on the x-axis, ordered top-to-bottom as most- to least-important. A: Mean decrease in accuracy shows much the model accuracy decreases if that parameter is dropped from the model and B: the mean decrease in Gini score shows the measure of parameter importance based on the Gini index. The higher the value of both of these, the more the importance of the variable in the model.</i>	66
3.11	<i>Illustration of Euler's method to quantify competition. y_i is reduced to obtain $\hat{y}_i$. Then Euler's method is used to obtain the predicted values y_{i+1}, y_{i+2}, $\hat{y}_{i+1}$ and $\hat{y}_{i+2}$. Then the values are compared using the formula. This is repeated for all time steps. $R > 0$ means that the perturbation does not favor strain 2 i.e. there is less of strain 2, when there is less of strain 1. On other hand when $R < 0$, it implies that there is competition between the strains. The more negative R is, the greater the competition.</i>	69
3.12	<i>Histogram of final bacterial load for model 3.2 simulations</i>	70
3.13	<i>Boxplots of the parameters of Model 3.2 grouped according to the disease outcomes.</i>	71
3.14	<i>Plot of final the log of population size of both strains coloured by the disease outcome class. The lines show the separation of the region into the different disease classes.</i>	72
3.15	<i>Boxplots of effective competition values R for latent and active disease outcomes. The more negative the values the stiffer the competition between the strains.</i>	73
3.16	<i>Densities of simulations of model 3.2 plotted on the discriminant function. 'coex' means coexistence between the strains, while 'comp_ex' means one strain out-competes the other.</i>	74
3.17	<i>Contributions of parameters to the discriminant function of the DAPC of model 3.2. The height of each bar is proportional to the contribution of the corresponding parameter to the discriminant function. Only parameters whose contribution is above an arbitrary threshold (grey horizontal line) are indicated for the sake of clarity.</i>	75

3.18	<i>Proportions of simulations in scenario 1 that fall into the different outcome classes for the sets: A: the host infected with both strains $\{h, B_1, B_2\}$, B: the bacterial load of strain 1 after re-infection $\{B_1\}$, C: the bacterial load of strain 2 after re-infection $\{B_2\}$, D: the host infected with only strain 2 $\{h, B_2\}$</i>	77
3.19	<i>Proportions of simulations in scenario 2 that fall into the different outcome classes for the sets: A: the host infected with both strains $\{h, B_1, B_2\}$, B: the bacterial load of strain 1 after re-infection $\{B_1\}$, C: the bacterial load of strain 2 after re-infection $\{B_2\}$, D: the host infected with only strain 2 $\{h, B_2\}$</i>	78
4.1	<i>An example of multiple alignment of four sequences.</i>	85
4.2	<i>Flow chart of the steps in the approach</i>	88
4.3	<i>Figure illustrating how the Kendall-Colijn metric works. The distance between tip 1 and 3 in T_a is 1, but in T_b, its 2.</i>	89
4.4	<i>The random tree used in the simulation</i>	90
4.5	<i>A: The true tree and B: reference tree constructed with simulated data and noisy sites. This shows the topological differences between the two trees. For each tip, the function colours it grey if its ancestors in tree1 has the same partition of tip descendants in tree2, otherwise the tip gets coloured from pale orange to red to reflect the dissimilarity. Trees are plotted using plotTreeDiff, a function in the R package treescape.</i>	92
4.6	<i>Distance from reference tree and true tree of the simulated data. Nine trees have distance greater than zero from the reference tree. These trees also have distance zero from the true tree.</i>	93
4.7	<i>Distances of trees with window of 10 sites removed from the reference tree</i>	94
4.8	<i>Distance of each site along full alignment, with the genes of the top 15 sites (sites with the top 15 distances) labelled. Each dot represents a site and the size of the dots are mapped to the distance associated with each site.</i>	95
4.9	<i>Influence values of sites along the entire alignment with the genes of the top 15 outlier sites highlighted. Each dot represents a site and the size of the dots are mapped to the distance associated with each site.</i>	96

4.10	<i>A. MDS projection of distances between these trees coloured by their ranking, B. MDS projection of distances between these trees coloured by their groups (median tree of each group is annotated by MT‘group number’)</i>	96
4.11	<i>A. Distance from reference tree of trees with top m disruptive sites removed ($m = 1 : 1043$), B. Mean bootstrap values of the nodes of median trees of the 5 groups</i>	97
5.1	<i>simple transmission tree. Here, $v_{5,7} = 2$ because the MRCI of cases 5 and 7 is case 4, which is 2 steps from the source case 1. But $v_{2,6} = v_{2,5} = v_{2,7} = 0$, and so on for pairs of cases whose MRCI is the source case, 1.</i>	107
5.2	<i>Illustration of the trapezoid method. The area of a trapezoid is equal to the sum of the area of the triangle and the rectangle, the integral is the sum of the areas of the trapezoids.</i>	112
5.3	<i>Left: A coloured phylogenetic tree, each colour corresponds to a host and each lineage is coloured by the host it was in at the time. Right: the subtree in host A, with the $d_{A,j}$ shown. . . .</i>	114
5.4	<i>A: An example of a coloured phylogeny on 7 tips: each host has a unique colour to depict the evolution within each host, the red stars depict transmission events. B : Corresponding transmission tree: vertical arrows represent transmission between cases, and the sampled individuals have a red circle on them</i>	115
5.5	<i>Schematic diagram of migration and tuberculosis burden . . .</i>	118
5.6	<i>Clinical <i>M. tuberculosis</i> isolates and phylogenetic reconstruction, with clades of interest highlighted.</i>	120
5.7	<i>Posterior distribution of timed phylogenetic trees with the maximum credibility clade tree (MCC) and median tree (MT) highlighted in red. The median tree is obtained using the Kendall-Colijn metric metric described in section 4.6</i>	121
5.8	<i>The smaller clades with the cases, sampling dates and country of birth highlighted.</i>	122
5.9	<i>Posterior generation time distribution for Clade a. The prior distribution is highlighted in blue and the posterior fits are in red.</i>	123

5.10	<i>Posterior sampling time distribution for Clade a. The prior distribution is highlighted in blue and the posterior fits are in red.</i>	124
5.11	<i>Arrival times (in blue) plotted on top of estimated infection times for all cases of interest with available data. The case numbers are colored by clade assignment (clade a in green, clade e in orange and clade b in grey). The blue shaded area covers the time from earliest and latest possible arrival times, whereas a dotted single line indicates the latest possible arrival time. P-values indicate probability of infection after arrival in Norway, averaged over 10 different TransPhylo inference procedures. The country of origin of patients not originating from the Horn of Africa is annotated in black boxes.</i>	126
5.12	<i>Infection times of cases in clade a using the same priors but different input trees: MCC - maximum credibility clade tree, MED - medium tree and three random trees (RAN1, RAN2, RAN3)</i>	127
5.13	<i>Arrival times of cases in clades of interest plotted on their infection times obtained using six different prior parameter sets</i>	128
5.14	<i>Multidimensional scaling plot of 1,000 transmission trees obtained from using 5 different prior parameters 1 – 5 in Table 4.2, coloured according to prior parameter set.</i>	129
5.15	<i>Probability of infection after arrival in Norway for 22 cases included in TransPhylo analyses. The lines annotate the number of cases with probabilities equal to 0.5 and 0.9 in the cumulative distribution plot. With $y_1 = 16$, $y_2 = 16$, $y_3 = y_4 =$</i>	131
5.16	<i>Median transmission trees of the posterior distribution of transmission trees of the selected clades. Each horizontal line represents a case with the darkness of the line representing changing infectivity over time, the red circles indicate the sampled cases and time of sampling, the vertical arrows represent transmission from case to case. The cases are illustrated as C_n, where n is the case number.</i>	132

LIST OF TABLES

2.1	<i>Definitions of all states and parameters</i>	51
2.2	<i>Baseline values of parameter fits (slow dilution)</i>	52
2.3	<i>Baseline values of parameter fits (fast dilution)</i>	52
3.1	<i>Table of parameter definitions and ranges. Parameter ranges are specified with a — separating upper and lower ranges, and are obtained from [126]. Where a single value is given, the parameter was fixed.</i>	81
4.1	<i>Table the genes, distances of top distance sites along alignment. The resistance column is compiled from the TB drug resistance mutation database [159]</i>	100
5.1	<i>Tables showing (a) prior and (b) posterior parameter values for the generation time distribution on clade a (shape is G_{sh} scale is G_{sc}) and sampling time distribution (shape is S_{sh} and scale is S_{sc})</i>	123
5.2	<i>Table summarizing main results of study. Probabilities of infection after arrival in Norway (Not Penalized : $P_{NP}(t_{inf} > t_{arr})$ and Penalized: $P_P(t_{inf} > t_{arr})$) are shown where available. We conclude that the cases are likely infected in Norway (LNorway) based on these probabilities</i>	130
5.3	<i>Table summarizing additional inference: PIN means period in Norway, P/E means Pulmonary or Extra-pulmonary TB, MPD means mean pairwise distance</i>	135
5.4	<i>Table summarizing additional inference: PIN means period in Norway, P/E means Pulmonary or Extra-pulmonary TB, MPD means mean pairwise distance</i>	136
5.5	<i>Table summarizing additional inference: PIN means period in Norway, P/E means Pulmonary or Extra-pulmonary TB, MPD means mean pairwise distance</i>	137

5.6	<i>Some available Bayesian methods for inferring transmission trees using genetic and epidemiological data. If ‘Yes’, then method uses or has the property and otherwise if ‘No’. If ‘Est’, then property is estimated by the method. If ‘Lim’, then the method includes a limited version of the property.</i>	138
-----	---	-----

CONTENTS

1	INTRODUCTION	20
2	CHEMOSTAT MODELS OF <i>Mycobacterium tuberculosis</i> UNDER THE ACTION OF ISONIAZID	25
2.1	The Chemostat	27
2.2	Chemostat model of a single bacterial population	28
2.3	Chemostat model of a single bacterial population under action of isoniazid	32
2.4	Two populations under the action of isoniazid	33
2.5	Parameter estimation	35
2.5.1	Bayesian parameter estimation	35
2.5.2	Initial concentration of resistant strains	36
2.6	Results	37
2.6.1	Fits to pooled data	37
2.6.2	Growth rates	38
2.6.3	Posterior parameter values	38
2.6.4	Fitness costs of INH resistance and ‘bounceback’ of cultures	39
2.6.5	Fits to individual cultures	41
2.7	Parameter identifiability	42
2.8	Multipopulation model	44
2.9	Competition in the chemostat	47
2.10	Diversity in the system 2.8	48
2.11	Conclusion	48
3	WITHIN HOST MODEL OF MYCOBACTERIUM TUBERCULOSIS	53
3.1	Introduction	53

3.2	Immune response to tuberculosis	54
3.3	Tuberculosis progression	54
3.4	Discriminant analysis principal component analysis	55
3.5	Random forests	57
3.6	One strain model	58
3.7	Analysis of Model 3.1	59
3.8	Two strain model	66
3.9	Measuring effective competition and co-existence	67
3.10	Analysis of model 3.2	69
3.11	Virtual experiments	72
3.11.1	Scenario One	76
3.11.2	Scenario Two	77
3.12	Conclusion	78
4	DETECTING DISRUPTIVE SITES IN TUBERCULOSIS GENOME	82
4.1	Introduction	82
4.2	Inference of phylogenetic trees	83
4.3	<i>Mycobacterium tuberculosis</i> Genome	85
4.4	Current methods	86
4.5	The method	87
4.6	Comparing phylogenetic trees	88
4.7	Simulation	89
4.8	Tuberculosis data	91
4.9	Results	91
4.9.1	Simulated data	91
4.9.2	Tuberculosis data	91
4.10	Obtaining a more robust tree	94
4.11	Conclusion	97

5	INFERRING TRANSMISSION EVENTS USING GENOMIC AND EPIDEMIOLOGICAL DATA: IS THERE ONGOING TRANSMISSION IN NORWAY?	101
5.1	Genetic data and transmission inference	102
5.1.1	Threshold of genetic similarity	102
5.1.2	Genetic distances	103
5.1.3	Genetic distances and epidemiological data	103
5.1.4	Using phylogenetic trees	104
5.1.5	Bayesian inference of transmission trees	105
5.2	Comparing Transmission trees	106
5.3	TransPhylo	107
5.3.1	Model for epidemiological process between hosts	108
5.3.2	Probability of a Transmission tree $P(T \theta)$	109
5.3.3	Probability of the phylogenetic tree given the transmission tree and within host parameter $P(\mathcal{P} N_{eg}, T)$. .	113
5.4	Penalized tuberculosis transmission	115
5.5	Migration and tuberculosis in low incidence countries	116
5.6	Determining transmission events among immigrants in Norway	118
5.6.1	Transmission reconstruction	119
5.6.2	Probability of infection prior to arrival	122
5.6.3	Results	124
5.7	Conclusion	131
6	CONCLUSION	139
	REFERENCES	168

Do not despise the days of little beginnings.

The book of Zechariah

1

INTRODUCTION

Tuberculosis (TB) is a global health problem and the battle to eradicate this deadly disease has been ongoing for more than half a century. This is despite the availability of effective chemotherapy and allocation of massive resources to develop better intervention strategies ([186],[203]). *Mycobacterium tuberculosis* (*M.tuberculosis*), the causative agent of this disease has persisted in the human population for tens of thousands of years and continues to evolve in order to successfully infect and transmit between humans. Tuberculosis most commonly affects the lungs (pulmonary TB), but it can also cause disease in other organs [64]. The symptoms of pulmonary TB include coughing, often with sputum or blood, chest pain, fatigue, unintentional weight loss, fever and night sweats and chills.

The World Health Organization (WHO) has defined a new strategy for tuberculosis prevention, care and control: the End TB Strategy, with the goal of reducing incidence of the disease by 90% by 2035 [74, 63]. Although progress has been made in global TB control, it is not fast enough to reach this target [58]. There is thus increased urgency in the need for improved TB control especially in developing countries where control is becoming difficult due to increasing drug resistance and in some populations HIV co-infection [150]. In the 2017 global tuberculosis report, it was estimated that there were a total of 10.4 million new cases and 1.3 million deaths in 2016 [1] with

much of the burden concentrated in high-burden settings in Asia and Africa [147].

M. tuberculosis is carried in airborne particles called, droplet nuclei, which are generated when an infectious person coughs or sneezes. Flynn and Chan [64] state that once an individual has been exposed to *M. tuberculosis*, there are four possible outcomes:

- The immune response can be very effective and clear the bacilli such that the individual has no chance of developing tuberculosis.
- The host may be unable to control the growth of bacilli immediately after infection leading to clinical disease known as primary tuberculosis.
- The immune response may be able to contain, but not clear the bacilli such that the host has latent tuberculosis.
- Changes in the immune response following the establishment of latent infection may result in clinical disease, known as reactivation tuberculosis.

The outcome of infection depends on very complex host-pathogen interactions. During latency, genes hypothesized to be involved in host-pathogen interactions are expected to be under more selection pressure and therefore are more likely to undergo genetic changes [193]. As such one can expect several nucleotide changes to accrue within a single host, with different individual lineages being transmitted onward to secondary cases. Also the long- and variable latent period makes it challenging to control tuberculosis because it is hard to identify when people were originally infected, and individuals with latent TB form a large reservoir of potential cases.

The genetic diversity of a pathogen can have dramatic effects on the dynamics of human diseases. Firstly, poorly understood diversity of a pathogen can lead to diagnostics that cannot detect the entire set of pathogens causing a particular disease [183]. Another issue is the emergence of antibiotic resistance and the difficulties associated with the development of vaccines against antigenically diverse pathogens [81]. Pathogen diversity poses as a huge barrier to disease control and surveillance and in principle can either increase or decrease pathogen transmission and disease risk [144]. This genotypic

variation can be exploited as a rich source for inference on how pathogens are spreading, evolving and adapting to the introduction of vaccination and antibiotics [50] .

M.tuberculosis is a member of a group of closely related bacteria known as *Mycobacterium tuberculosis* complex (MTBC) which comprises seven distinct strongly geographically associated lineages ([150],[10],[195]). In the 1900's, there was a general belief that genetic diversity within MTBC was too limited to account for the differences in outcomes following exposure [37]. However recent availability of next generation sequencing techniques have led to the refuting of this belief. Subsequent studies [125, 178] have provided evidence of heterogeneity in virulence and transmissibility.

In recent years, numerous studies have identified tuberculosis patients in whom there are more than one distinct strain of *M. tuberculosis*. This may be as a result of re-infection with a new strain especially in settings where the incidence of TB is high [29]. Such mixed infections can have a dramatic effect on the relative frequency of these strains at the population level as shown by [163]. The fact that an individual can simultaneously harbour more than one strain of *M. tuberculosis* can have adverse effects on disease dynamics. For instance, it could lead to poor treatment outcomes i.e. development of multidrug resistant tuberculosis especially in cases where one strain is sensitive and the other is not [141].

The presence of more than one type of organism in an environment generally leads to competition for resources in a bid to survive. When multiple pathogen strains are maintained in a population, they interact at various levels for the exploitation of susceptible hosts or nutrients all in the struggle for survival. Competition can also occur within a host, when the host is infected simultaneously by different pathogen sub-populations. The emergence and spread of novel pathogen strains depend on whether they are able to out-compete pre-existing strains. When a new, more competitive pathogen strain or population emerges, it can rapidly displace the pre-existing strain or population and thus affect disease dynamics. Also when strains are in strict competition with one another on a population level, i.e. a host can be infected with only one strain at a time, then the strain with the highest reproduction number persists. When competition is not fierce, then strains are allowed to co-exist for longer periods of time.

Understanding the diversity of *M. tuberculosis* will help elucidate how it will evolve in response to interventions. Also short-term genetic diversity accruing on the time scale of an outbreak can inform us of who infected whom, hopefully allowing better outbreak control. No single approach or class of models is likely to explain completely a phenomenon as complex as pathogen diversity and competition. In this study, two major approaches have been utilized: analysis of mathematical models of *M. tuberculosis* populations and genomic analysis of *M. tuberculosis* genome sequences.

Mathematical modelling can be used to investigate the behaviour of biological systems, test hypothesis where carrying out a study may not be feasible, and make predictions that would inform health policies. Modelling may provide us with a better understanding of the underlying principles that govern biological systems, offering explanations and insights. Mathematical models have previously been used to study diversity and competition in *M. tuberculosis*. These models are mostly the susceptible-infected (SI) type models that study the effects of diversity and competition on a population level. Cohen et al [29] developed a model which incorporates a latent phase in which co-infection is possible, to study the effect of strain diversity and competition on vaccine effectiveness. Colijn et al [34] proposed a similar model to study the effect of latent infection on strain diversity. Sergeev et al [163] proposed a model similar to [34] to study the population-level impact of mixed infections, where one strain is resistant and the other is not.

Two chapters in this work focus on mathematical models of mycobacterium tuberculosis. In chapter 2, we present a mathematical model of diverse subpopulations of mycobacterium tuberculosis growing in a chemostat under the action of isoniazid, an antibiotic used during tuberculosis chemotherapy. This was published in [8]. Chapter 3 presents within host models with the incorporation of competition between two strains to study the immune response to tuberculosis infection.

Analysing pathogen sequences offer a way to measure diversity and using such information can help to understand resistance and improve TB control. *Nature* defines genomic analysis as the identification, measurement or comparison of genomic features such as DNA sequence, structural variation and gene expression. As sequencing technology has become more accessible and affordable, genetic analysis has played an increasingly important role in infec-

tious disease research [116]. Pathogen genome sequences can help to confirm suspected cases of an infectious disease and also discriminate between different strains [190]. Multiple pathogen sequences can be analysed together using phylogenetic methods to elucidate genetic diversity and evolutionary history.

Phylogenetic trees can be seen as graphical representation of the evolutionary history of a set of taxa. The phylogenetic tree can thus reveal aspects of the epidemiology of a disease outbreak. These trees are not directly observed but are inferred from data and have been used repeatedly in tuberculosis research. Sheen et al [165] applied phylogenetic trees to delineate the strain diversity of mycobacterium tuberculosis in Peru and its association with drug resistance, Muller et al [139] used phylogenetic analysis to decipher the genetic diversity of ancient tuberculosis strains, Benavente et al [13] developed a web based tool, *PhyTB*, which contextualises mycobacterium tuberculosis genomic variation within epidemiological and phylogenetic settings. Given the rising computational challenges in tree reconstruction, our research does not focus on inference methods. Rather in chapter 4, we discuss and present a method to detect sites along the genome of *M. tuberculosis* that are phylogenetically disruptive. Parts of this work was published in [75].

Pathogen sequences have also been used to make inference about who infected whom, alongside traditional contact tracing methods [96, 48, 98]. Since sequence and surveillance data contain information regarding the transmission process, analysing both datasets at the same time would give more accurate estimates of epidemiological parameters than when they are analysed separately [153]. In Chapter 5, a Bayesian method of inferring who infected whom, TransPhylo, proposed by [48] is applied to a tuberculosis dataset from Norway to infer infection times of migrants. Combining the results of this analysis with epidemiological data elucidates the time of infection. Parts of this work was published in [100].

The race is not to the swift, nor the battle to the strong...

King Solomon

2

Chemostat models of *Mycobacterium tuberculosis* under the action of isoniazid

INTRODUCTION

Drug resistance to tuberculosis (TB) has become more widespread in the last decade. As such understanding the emergence and fitness of antibiotic-resistant strains is crucial for the development of new interventions. The management of human TB relies heavily on vaccination, case finding, and antibiotic treatment. Current treatments for tuberculosis are complex and lengthy leading to incomplete treatment, non-compliance, and the development of multi-drug resistance. Isoniazid (INH) and rifampicin (RIF) form the core of standard treatment regimens, and resistance to them is a keystone of multi-drug resistant (MDR) tuberculosis. Given the impact of tuberculosis on global health and the seemingly rapid decline of effectiveness of anti-tuberculosis agents, understanding the growth response of *Mycobacterium tuberculosis* cells to anti-tuberculosis drugs is critical.

It has been demonstrated [93] that the reduction in the number of viable *M. tuberculosis* cells in the sputum in response to antibiotic treatment is biphasic. The first phase, which is mediated by INH, is characterised by a rapid reduction in the bacterial population, while the second phase involves

a slower reduction in the bacterial population. It was generally believed that this rapid reduction is as a result of a depletion of bacilli in the exponential phase of growth. However Ahmad et al [3] suggest that the reduction in the bactericidal activity of INH is as a result of the selection of phenotypically tolerant ‘persisters’, while Gumbo et al [80] suggest that it is as a result of the development of antibiotic resistance. Antibiotic resistance in *M. tuberculosis* may be as a result of development of tolerance to hostile environments other than exposure to antibiotics [92] or as a result of the presence of slow growing non replicating subpopulations that are protected from antibiotic action [202].

Bacilli in pulmonary cavities are thought to be growing in an aerobic environment and therefore behave in a way that can be imitated by in-vitro systems like the chemostat [72]. The chemostat has been used to study and capture the response of *M. tuberculosis* cells to specific challenges such as carbon limitation ([85] [16]), oxygen limitation [9], and nutrient limitation [17]. Chemostat models of *M. tuberculosis* growing under fast and slow dilution are thought to be good models for the different phases of tuberculosis: an acute phase and an asymptomatic/persistent phase respectively [16]. In the acute phase, bacterial populations are growing at their maximum rate while during the persistent phase, bacterial populations are either slow growing or non-replicating.

Jeeves et al [91] in a recent study interrogated populations of *M. tuberculosis* cells cultured under fast ($0.03h^{-1}$) and slow ($0.01h^{-1}$) dilution rates in a chemostat in order to understand how the slow growth of *M. tuberculosis* contributes to INH tolerance. *M. tuberculosis* cells were grown to steady state and INH was introduced to the system and maintained at minimum inhibitory concentration (MIC). Three replicate cultures were performed at the different dilution rates to steady state. They found that bacterial numbers recover at both dilution rates after an initial rapid decline in population size. However, the cells under slow dilution rate do not only recover but are able to bounce back and regain their original population size (Figure 2.1). The mechanisms permitting this complete recovery under the slow dilution rate but not under the fast dilution rate are not known but raise interesting questions which may be answered by mathematical modeling.

In this chapter, a series of mathematical models of chemostat systems are

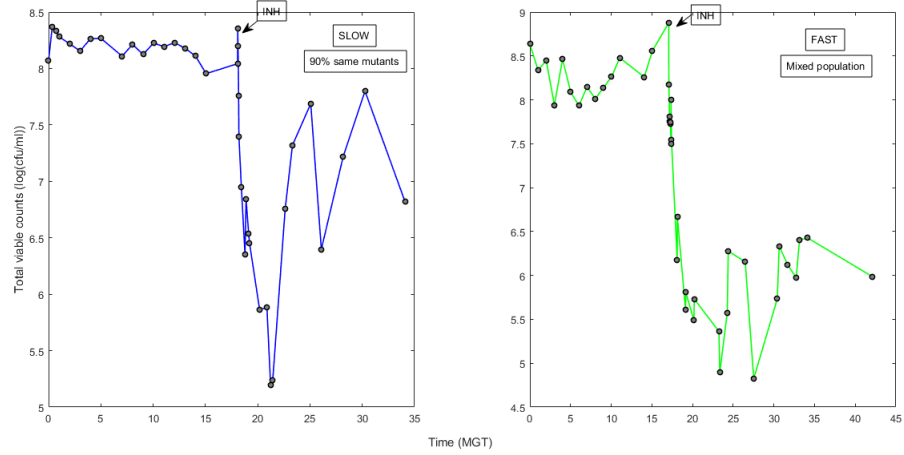


Figure 2.1: Plot showing how bacterial cell numbers change under fast (green) and slow (blue) dilution rates

developed to explain the differences in the response to INH of *M. tuberculosis* cells cultured under two growth rates in a chemostat. The first model consists of two state variables: a population of bacteria and the resource concentration in the chemostat. Two different parameters are used to model the action of INH: a bactericidal and bacteriostatic term. This model is then modified to incorporate a resistant bacterial sub-population. We obtain posterior distributions of the modified model parameters consistent with data (all three cultures pooled together) using a Markov chain Monte Carlo (MCMC) method. We also analyse replicate cultures separately. Using the posterior distribution of the model parameters, we explore the dynamics of diverse INH-resistant sub-populations in a multi-population model.

2.1 THE CHEMOSTAT

The chemostat is a well stirred apparatus used in the study of microbial physiology. The chemostat consists of three compartments: a nutrient reservoir, a vessel containing bacterial culture (growth chamber), and a tank to collect waste. Via an inflow from the nutrient reservoir fresh nutrition is added and from an outflow to the tank, bacteria are harvested. The nutrient reservoir contains all the necessary nutrients required for growth of the bacteria, all in excess except one which is called the limiting nutrient.

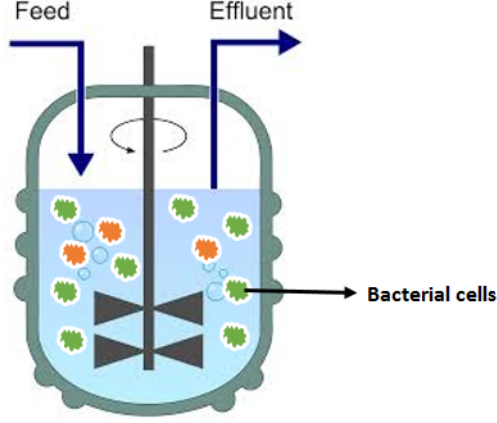


Figure 2.2: *Growth chamber of the chemostat*

2.2 CHEMOSTAT MODEL OF A SINGLE BACTERIAL POPULATION

Modeling of the chemostat is based on compartmental analysis, laws of mass action, and mass balance [26]. Bacteria B grow in a chemostat feeding on a limiting nutrient R . The growth rate of the bacteria is controlled by a constant dilution rate D . This is the rate of inflow and outflow from the chemostat. It represents the fraction of the volume in the chemostat being replaced per unit time. The rate of change of bacterial density in chemostat is given by the difference between growth and washout. Also the rate of change of nutrient concentration in the growth chamber is given by the difference between the input and the sum of consumption and washout.

The growth rate of the bacteria in the medium depends on the limiting resource. Jacques Monod carried out a series of experiments in the 1940s to determine the resource-dependent growth rate of bacteria feeding on a limiting nutrient R . He obtained the following relation

$$\phi(R) = \lambda_{max} \frac{R}{K + R},$$

where λ_{max} is the maximum resource independent growth rate and K is the

half saturation constant i.e. the concentration of the resource at which the bacterial population is growing at half of its maximum rate. Models of continuous culture are well-established and typically follow this pattern ([53], [182], [87]). The consumption of the limiting nutrient in the medium is proportional to the resource dependent growth rate and a conversion efficiency parameter ϵ , which is the resource required produce a single new bacterial cell [6]. Thus the consumption rate is given as $\epsilon\phi(R)$. We represent the concentration of the limiting resource entering the growth chamber as C , therefore the system of equations governing bacterial growth in a chemostat is

$$\begin{aligned} dB/dt &= \phi(R)B - DB. \\ dR/dt &= DC - \epsilon\phi(R)B - DR. \end{aligned} \tag{2.1}$$

The models discussed in this chapter are all are nonlinear and cannot be solved analytically for $B(t)$ and $R(t)$, however, we can gain insights by investigating the steady state behaviour. Let $f(B, R) = \frac{dB}{dt}$ and $g(B, R) = \frac{dR}{dt}$ and let B^* and R^* be the fixed points of the system. The equilibrium values of system 2.1 are found by solving the equations

$$\begin{aligned} B^*(\phi(R^*) - D) &= 0 \\ D(C - R^*) - \epsilon\phi(R^*)B^* &= 0 \end{aligned}$$

The trivial steady state $(0, C)$ also called the wash out equilibrium, always exists, however this is of no biological importance. The other fixed point of the system $(\frac{C-R^*}{\epsilon}, \frac{KD}{\lambda_{max}-D})$ is called the survival equilibrium. If $C > R^*$, and $\lambda_{max} > D$, then the ‘survival equilibrium’ exists i.e. $B^* > 0$.

In chemostat cultures, bacteria cannot survive unless the dilution rate is less than the maximum resource independent growth rate and for this reason, the steady value of the resource concentration, R^* , is known as the population’s

‘break-even’ concentration. This is because it is the resource concentration the population requires to support growth that balances death and dilution.

Linear stability analysis approximates a nonlinear system as linear in the neighbourhood of a fixed point. Linearization about the fixed points allows us to determine the stability of the steady states. If we define new variables $x_1 = B - B^*$ and $x_2 = R - R^*$ and let $f = \frac{dB}{dt}$, $g = \frac{dR}{dt}$, then the linear system that approximates the chemostat model near a fixed point is

$$\dot{\mathbf{x}} = \mathbf{J}\mathbf{x} \quad (2.2)$$

where $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ and $\mathbf{J} = \begin{pmatrix} f_B & f_R \\ g_B & g_R \end{pmatrix} \Big|_{B^*, R^*}$ is the Jacobian matrix evaluated at the fixed point (B^*, R^*) .

For system 2.1 the Jacobian matrix is

$$\mathbf{J} = \begin{pmatrix} \frac{\lambda_{max}R}{K+R} - D & \frac{K\lambda_{max}B}{(K+R)^2} \\ -\frac{\epsilon\lambda_{max}R}{K+R} & -\left(\frac{\epsilon\lambda_{max}B}{(K+R)^2} + D\right) \end{pmatrix} \Big|_{B^*, R^*}$$

For the fixed point $(\frac{C-R^*}{\epsilon}, \frac{KD}{\lambda_{max}-D})$,

$$\mathbf{J} = \begin{pmatrix} 0 & (\lambda_{max} - D) \frac{C(\lambda_{max}-D)-KD}{\epsilon KD} \\ -\epsilon D & -((\lambda_{max} - D) \frac{C(\lambda_{max}-D)-KD}{K\epsilon\lambda_{max}} + D) \end{pmatrix}$$

To analyse the stability of the fixed points we need to calculate the eigenvalues of the matrix $\mathbf{J}$. To do this we use the characteristic equation $|\mathbf{J} - wI| = 0$.

$$\begin{vmatrix} j_{11} - w & j_{12} \\ j_{21} & j_{22} - w \end{vmatrix} = 0.$$

$$w^2 - w(j_{11} + j_{22}) + (j_{11}j_{22} - j_{21}j_{12}) = 0.$$

Therefore,

$$w_i = \frac{(j_{11} + j_{22}) \pm \sqrt{(j_{11} + j_{22})^2 - 4(j_{11}j_{22} - j_{21}j_{12})}}{2},$$

where $i = 1, 2$.

w_1 and w_2 could either be real numbers or complex conjugates. For a fixed point to be stable, the eigenvalues have to be negative if they are real numbers or the real parts of the eigenvalues have to be negative if they are complex conjugates. When $j_{11} + j_{22} < 0$, then if w_1 and w_2 are complex numbers then they are negative and the fixed points are stable. If w_1 and w_2 are real numbers, $j_{11} + j_{22} < 0$ is not sufficient to guarantee that $w_1, w_2 < 0$. The eigenvalues will only be negative when $j_{11} + j_{22} < 0$ and $|j_{11} + j_{22}| > \sqrt{(j_{11} + j_{22})^2 - 4(j_{11}j_{22} - j_{21}j_{12})}$, which means $(j_{11}j_{22} - j_{21}j_{12}) > 0$. Thus the necessary and sufficient conditions for the eigenvalues to be negative is for

1. $j_{11} + j_{22} = \text{Trace}(J) < 0$
2. $j_{11}j_{22} - j_{21}j_{12} = |J| > 0$

For system 2.1,

$|J| = \left(\frac{D(\lambda_{max}-D)(C(\lambda_{max}-D)-KD)}{K\lambda_{max}} \right)$ and the trace, $\text{Trace}(J) = -((\lambda_{max} - D) \frac{C(\lambda_{max}-D)-KD}{K\epsilon\lambda_{max}} + D)$. For conditions 1 and 2 to be satisfied, λ_{max} must be greater than D and C must be greater than R^* , which are the same conditions for the existence of the fixed point. Therefore whenever the fixed point exists, it is locally asymptotically stable.

2.3 CHEMOSTAT MODEL OF A SINGLE BACTERIAL POPULATION UNDER ACTION OF ISONIAZID

We now consider a situation where bacterial cells growing in the presence of an antibiotic A , which is added to the chemostat at minimum inhibitory concentration (MIC). We quantify the effect of the antibiotic with the use of two parameters: $\hat{A}$ and P to represent reduction in growth by a factor $(1 - \hat{A})$ and bactericidal effect respectively. The concentration of the antibiotic is assumed to be kept constant.

We have the following system of equations.

$$\begin{aligned}\frac{dB}{dt} &= B(\phi(R) - \phi(R)\hat{A} - P - D) \\ \frac{dR}{dt} &= D(C - R) - \epsilon\phi(R)B\end{aligned}\tag{2.3}$$

When $\hat{A} = P = 0$, this model reduces to system 2.1. Let $x = 1 - \hat{A}$ and $y = P + D$. The equilibrium values of system 2.3 are found by solving the equations

$$\begin{aligned}B^*(\phi(R^*)x - y) &= 0 \\ D(C - R^*) - \epsilon\phi(R^*)B^* &= 0.\end{aligned}$$

The washout equilibrium is the same as obtained in system 2.1. The survival equilibrium of the system is $(\frac{D(C-R^*)}{\epsilon\phi(R^*)}, \frac{Ky}{\lambda_{max}x-y})$. As in system 2.1, we see that $B^* > 0$, only if $C > R^*$ and λ_{max} must be bigger than $\frac{y}{x}$, $\frac{y}{x}$ represents the depletion of cells in this system. For this fixed point, we have

$$J = \begin{pmatrix} 0 & \frac{D(\lambda_{max}x-y)(C(\lambda_{max}x-y)-yK)}{\epsilon y K \lambda_{max}} \\ -\frac{\epsilon y}{x} & -(D + \frac{D(\lambda_{max}x-y)(C(\lambda_{max}x-y)-yK)}{Kx\lambda_{max}y}) \end{pmatrix}.$$

$|J| = \left(\frac{D(\lambda_{max}x-y)(C(\lambda_{max}x-y)-yK)}{xk\lambda_{max}} \right)$ and $Trace(J) = -(D + \frac{D(\lambda_{max}x-y)(C(\lambda_{max}x-y)-yk)}{Kx\lambda_{max}y})$. Like in system 2.1, whenever the fixed point exists, it is stable because if the conditions for existence are satisfied, then conditions 1 and 2 are met.

Figure 2.3 shows the time course for the system 2.3. The model is able to show the qualitative behaviour of the fast growing bacteria i.e. rapid death and recovery to a lower population size. The model however does not capture the behaviour observed in the slow-growing bacteria.

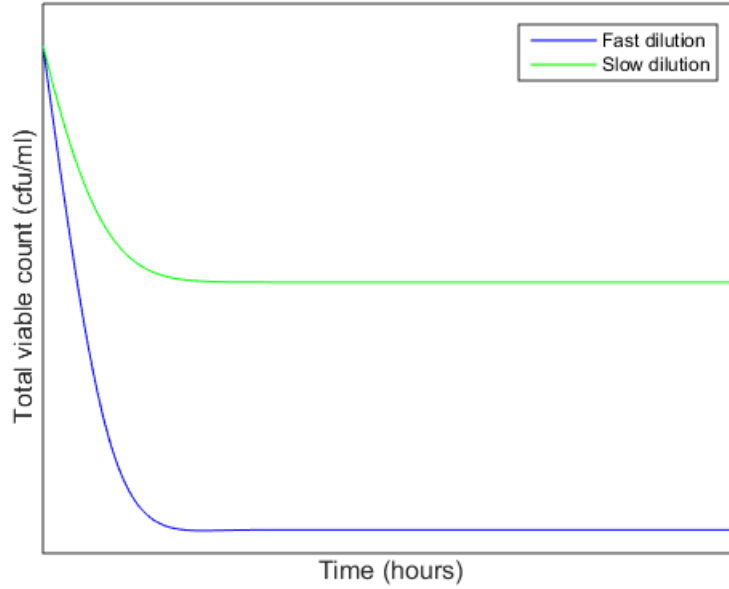


Figure 2.3: *Model simulations for system 2.3, for both dilution rates.*

2.4 TWO POPULATIONS UNDER THE ACTION OF ISONIAZID

The reason for recovery under both fast and slow dilution rates may be as a result of the presence of a resistant sub-population on which isoniazid has little or no bactericidal effect. In vitro models of *M. tuberculosis* have shown that drug resistance is often associated with a fitness cost [66]. The presence of antibiotics in the chemostat constitutes a strong and persistent selective pressure favouring the evolution of antibiotic-resistant sub-populations. Re-

sistance mutations may be expected to impart a fitness cost because they target important biological functions in the cell. For example. Fitness cost can also be seen as a reduction of the rate of replication under prevailing environmental conditions. For this chapter, we consider fitness costs in terms of the parameters that affect the growth of the bacterial cell i.e. λ_{max} , ϵ , and K

We now include two bacterial populations: B_1 , sensitive to INH, and B_2 , resistant to INH. We assume that the resistant subpopulation arise from the sensitive population by mutation at rate μ (we do not consider reverse mutation). Both subpopulations grow by consumption of the resources and are washed out at the (constant) dilution rate. Like in systems 2.1 and 2.3, the resources are consumed by the bacteria at a rate proportional to their resource-dependent growth rate and conversion efficiency parameter ϵ .

The sensitive and resistant strains have different maximum growth rates (λ_i), different yield constants (ϵ_i) and different half saturation constants (k_i ; the concentration of resource at which the strain reaches half its maximum growth rate), so resistance can confer a fitness cost. The resource-dependent growth rate, $\phi_i(R)$ is thus different for the two strains.

The model equations are:

$$\begin{aligned}\frac{dB_1}{dt} &= B_1((1 - \hat{A})\phi_1(R) - \mu - P - D), \\ \frac{dB_2}{dt} &= B_2(\phi_2(R) - D) + \mu B_1, \\ \frac{dR}{dt} &= D(C - R) - \epsilon_1\phi_1(R)B_1 - \epsilon_2\phi_2(R)B_2.\end{aligned}\tag{2.4}$$

The system is set up in such a way that the sensitive population dies off and is being replaced by a population that is resistant to isoniazid. When the sensitive population dies off, the system is reduced to system 2.1 and thus has the same steady state analysis.

2.5 PARAMETER ESTIMATION

We fit the parameters related to antibiotic activity ($\hat{A}$, P), half saturation constant (k), conversion efficiency parameter (ϵ), maximum growth rate λ_{max} and mutation rate (μ) of system 2.4 to experimental data. *M. tuberculosis* cells are generally characterized as being slow growers with a maximum growth rate not exceeding $1/16 \text{ days}^{-1}$ in optimum conditions [15]. We choose values of λ_m within the range $(1/23 - 1/20)$ [65]. For the fitting procedure, the smallest mutation rate of *M. tuberculosis* cells during exposure to INH is chosen to 2.56×10^{-8} per cell per generation [43]. We do not include the dilution rate D and nutrient concentration C in the MCMC procedure as these are known and fixed.

2.5.1 BAYESIAN PARAMETER ESTIMATION

Bayesian inference grows out of the simple formula known as Bayes rule. It involves inferring the posterior probability density function over model parameters (θ) given the data (d), for some model (M). The problem is formulated as

$$P(\theta | d) \propto P(d | \theta)P(\theta)$$

where $P(\theta | d)$ is the posterior probability distribution of the parameters, $P(d | \theta)$ is referred to as the likelihood and $P(\theta)$ is the prior probability of the parameters. The prior probability represents our subjective uncertainty or subjective belief about the parameters.

In most cases, it is usually difficult to estimate the posterior distribution in a closed form. The Markov Chain Monte (MCMC) algorithm provides an efficient way to explore the parameter space. It generates samples from the posterior distribution by constructing a reversible Markov chain that has as its equilibrium distribution the target posterior distribution [167].

We applied a Bayesian MCMC approach to find posterior sets of parameters of system 2.4 using the MATLAB MCMC package by Haario et al [83]. We use a uniform prior for all parameters within reasonable ranges as found in literature; Table 1 shows the prior parameter choices. We based our

likelihood on the weighted squared error function $\chi^2(\theta)$ in the usual way, with

$$\chi^2(\theta) = \sum_{i=1}^m \sum_{j=1}^n \left(\frac{B_{ij}^d - B_i(\theta, t_j)}{\sigma_{ij}^d} \right)^2,$$

where B_{ij}^d are the n data points for each observable m , $B(\theta, t_i)$ is the solution of the m -dimensional dynamical system at time point i , and σ_{ij}^d are the corresponding measurement errors [179]. This is equivalent to using the likelihood $L(\theta)$ [154], where

$$\chi^2(\theta) = -2 \log(L(\theta)) \quad (2.5)$$

under the assumption of Gaussian noise. The MCMC procedure produces a posterior collection of parameter sets corresponding to model trajectories that fit the data.

2.5.2 INITIAL CONCENTRATION OF RESISTANT STRAINS

Prior to the addition of INH to the chemostat cultures, the *M. tuberculosis* cells are grown to steady state. During this period of growth, mutation can occur, and it can be expected that small numbers of different sub-populations will arise from these large populations undergoing continuous turnover [33]. Accordingly, it is very unlikely that there would be absolutely no cells of minority sub-populations in a system with 10^8 CFU/mL. Despite this, we avoid directly introducing resistant sub-populations, and instead initialize the model with sensitive cells only, at their steady-state values prior to addition of INH. We allow the resistant sub-population to arise through mutation and persist (or not, depending on its fitness in the system without INH) to reach its steady state before we add INH. We also use the initial the value of R to be the pre-drug steady-state value. This can be calculated by setting $\hat{A} = P = 0$ in the system 2.4 and calculating the steady state value of B_2 as shown below:

$$\begin{aligned}
B_1(\phi_1(R) - \mu - D) &= 0, \\
B_2(\phi_2(R) - D) + \mu B_1 &= 0, \\
D(C - R) - \epsilon_1 \phi_1(R) B_1 - \epsilon_2 \phi_2(R) B_2 &= 0.
\end{aligned}
\tag{2.6}$$

We solve to obtain $R^* = \frac{k_1(\mu+D)}{\lambda_1-(\mu+D)}$ and $R^* = \frac{k_2(D-\mu B_1/B_2)}{\lambda_2-(D-\mu B_1/B_2)}$ which are the break even concentration of B_1 and B_2 respectively. The parameters are set up in such a way that $\lambda_1 > \lambda_2$ and $k_2 > k_1$, as B_1 's are fitter than the B_2 's and as such have a lower break-even concentration. We therefore note that in the absence of mutation, B_2 die off and B_1 win eventually. However the fact that the B_1 's are mutating continuously ensures a steady supply of B_2 's in the chemostat. Using the break even concentration of the B_1 's, we obtain an expression for B_2^* to be :

$$B_2^* = \frac{\mu}{D - \phi_2(R^*)} \left[\frac{D(C - R^*)}{\epsilon_1 \phi_1(R^*) + \epsilon_2 \phi_2(R^*) \frac{\mu}{D - \phi_2(R^*)}} \right]$$

where $\phi_1(R^*) = \mu + D$ and $\phi_2(R^*) = \frac{\lambda_2 k_1 (\mu + D)}{k_2 (\lambda_1 - (\mu + D)) + k_1 (\mu + D)}$.

2.6 RESULTS

2.6.1 FITS TO POOLED DATA

Figure 2.4A shows the system 2.4 growth curves alongside observed data (all three cultures pooled together). The sensitive population B_1 dies off rapidly upon addition of INH, leading to a temporary excess of glycerol. Figure 2.4B shows that the resistant sub-population begins to grow immediately. The initial decline in the population size of the total population size causes an apparent lag phase in the rise of the resistant population count because the majority of bacteria are sensitive to INH when the antibiotic is first added. The population size begins to increase again in both the fast and slow dilution rates, however under slow dilution, the bacteria are washed out

at a slower rate and thus have time to maximize the excess nutrient with an apparent increase in growth rate. This in turn enables the population to re-establish the original population size. In contrast, a lower but fairly constant population size of resistant bacteria is eventually maintained under the fast dilution rate (as observed by Jeeves et al [91]). Beste et al [16] previously determined that under fast dilution, mycobacterial cells grow at a rate that is very close to the maximum physiologically achievable growth rate for *M. tuberculosis* i.e. increasing the dilution rate to higher than $0.03h^{-1}$ resulted in wash-out of the cultures. Also from the steady state analysis of system 2.3, we recall that for the non-trivial fixed point to exist $\lambda_{max} > D$. We believe that when INH was added to the cultures, the cells were already dividing at their maximum rate. If this were the case then an increase in the growth rate (and therefore an increase in biomass) would not be possible, even in the presence of excess glycerol. However, we note that if left to run for a longer time period, the cultures under fast dilution may eventually increase, possibly up to similar levels to those attained by the cultures under slow dilution.

2.6.2 GROWTH RATES

Under steady-state conditions, the imposed slow and fast dilution rates ($0.01h^{-1}$ and $0.03h^{-1}$ respectively) in the chemostat correspond to slow and fast bacterial growth, and this is a motivation for the use of continuous cultures as models for chronic disease (slow dilution rates) and acute disease (fast dilution rates) of *M. tuberculosis* infection ([14], [131]). The comparison of two distinctly different growth rates in a controlled and defined system enabled us to measure the direct effect of growth rate on the response of *M. tuberculosis* to INH exposure. However, once the system departs from steady-state (in this case when INH was added), growth rates and dilution rates can become uncoupled. At times the growth rate under slow dilution reached the growth rates imposed by fast dilution rates (Figure 2.5).

2.6.3 POSTERIOR PARAMETER VALUES

There are wide ranges of posterior parameters that allow the model match the data, and most posterior parameters are similar under fast and slow

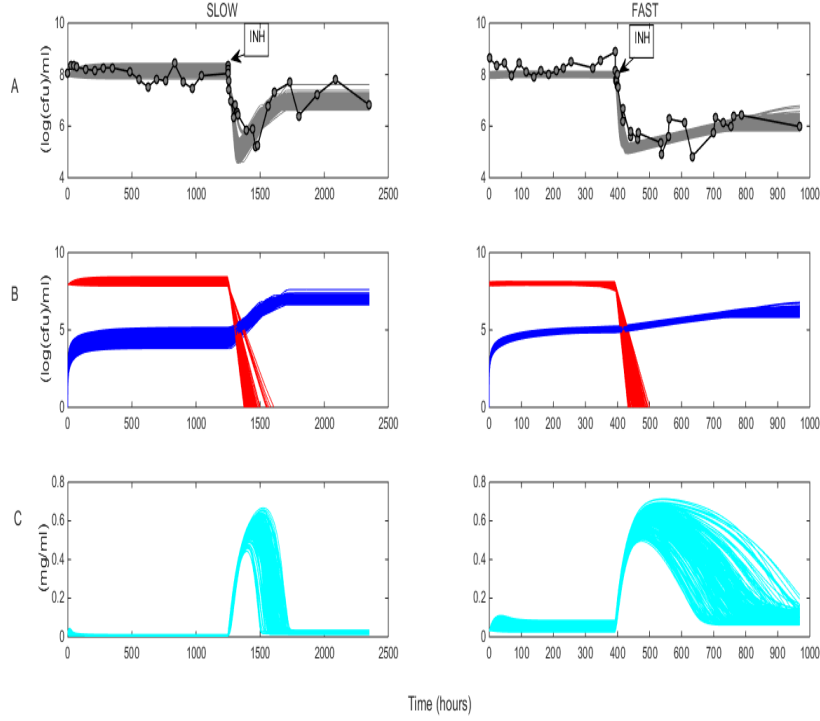


Figure 2.4: Grey lines (A) show the fits from 500 parameter sets randomly drawn from the posterior distribution of viable bacterial numbers for slow and fast dilution rates, the red lines (B) are the B_1 population (INH sensitive), the blue lines are the B_2 population (INH resistant), and the cyan (C) lines are plots of the resource concentration.

dilution (Figure 2.6). A few show marked differences: λ_2 , k_1 , P , μ and ϵ_2 are higher under fast dilution and there are some correlations. The kill rate P and mutation rate μ have a difference of about an order of magnitude between the two dilution rates, consistent with Jeeves et al.'s measurement of mutation rates of 10^{-7} and 10^{-6} per cell per generation in slow dilution and fast dilution respectively. Tables 2.2 and 2.3 show the baseline values of parameter fits for slow and fast dilution rates.

2.6.4 FITNESS COSTS OF INH RESISTANCE AND ‘BOUNCEBACK’ OF CULTURES

The posterior parameters defining the resistant sub-population allow us to explore the estimated fitness cost of isoniazid resistance. We define the rela-

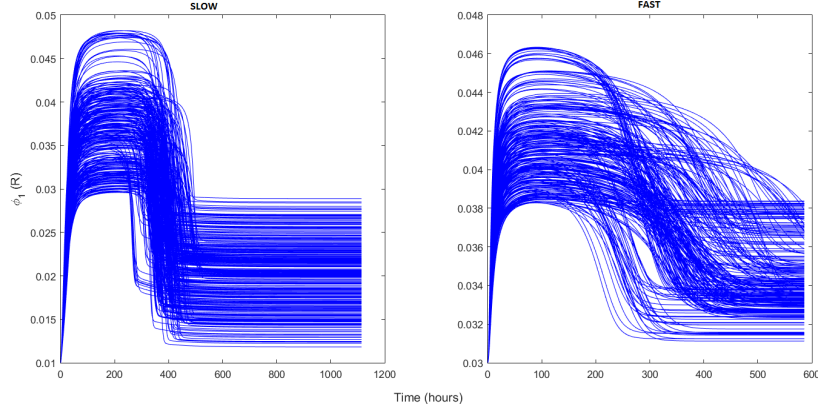


Figure 2.5: *Plots of the resource-dependent growth rate $\phi_i(R)$ for fast and slow dilution rates over time. Note that growth is frequently greater than the dilution rate due to the departure from steady-state conditions.*

tive fitness (f) in this context as the ratio of growth rates between resistant and sensitive sub-populations : $f = \frac{\phi_2(R)}{\phi_1(R)} = \frac{\lambda_2/(k_2+R)}{\lambda_1/(k_1+R)}$. This depends on the level of resource R , this is intuitive as relative fitness is specific to an environment. When $R \gg k_i$, $f \approx \frac{\lambda_2}{\lambda_1}$, and when $R \ll k_i$, $f \approx \frac{\lambda_2/k_2}{\lambda_1/k_1}$. Figure 2.7 shows how the recovery is affected by λ_2 and k_2 , under slow and fast dilution rates. Under fast dilution, the recovery depends very sensitively on the fitness costs whereas under slow dilution, recovery is much more robust to small changes in the growth parameters. The lower panels of Figure 2.7 show the posterior relative fitness values $f \approx \frac{\lambda_2/k_2}{\lambda_1/k_1}$ for the two dilution rates. While slow dilution allowed more rapid recovery of the bacterial population, the relative fitness values of the minority sub-population that are performing this recovery are notably lower than the corresponding fast-dilution sub-population. This is because under slow dilution, sub-populations can stay in the system long enough to consume the resource; competition for this now-depleted resource means that the most fit quickly dominates. Under fast dilution, all remaining sub-populations can grow quickly enough to out-pace the rapid dilution, but they cannot consume all of the resource before being washed out (they do eventually, but not until quite a bit later than their slow-dilution counterparts).

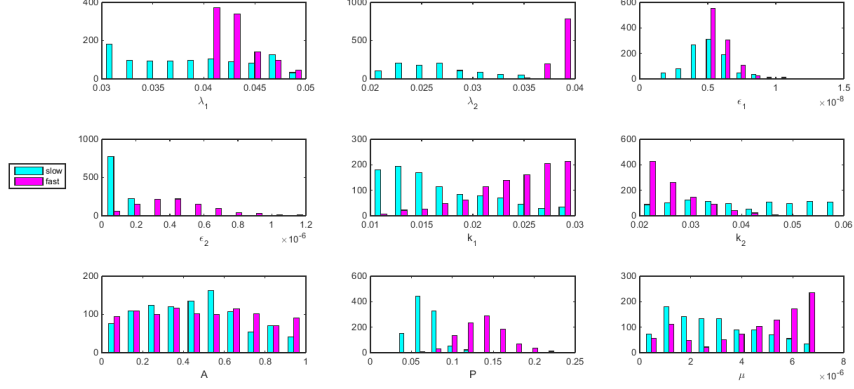


Figure 2.6: *Histograms of 1000 random samples of posterior parameter distributions for slow and fast dilution rates.*

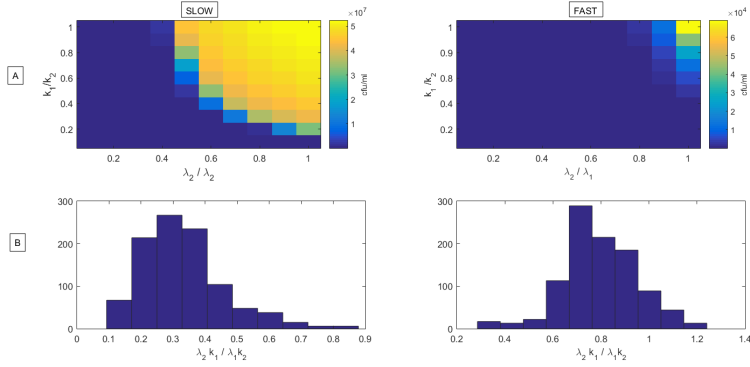


Figure 2.7: *Heatmaps (A) showing how culture recovery depends on fitness costs and histograms (B) of relative fitness $f \sim \frac{\lambda_2/k_2}{\lambda_1/k_1}$ of 1000 posterior parameter sets. Recovery strength ranges from dark blue (weakest recovery) to bright yellow (strongest recovery).*

2.6.5 FITS TO INDIVIDUAL CULTURES

Figure 2.9 shows that most of the posterior parameter values are similar across the individual cultures for each dilution rate. As was observed in the fits to combined replicates, the values of P are higher under slow dilution rates. However, some of the patterns observed in the combined fits (i.e. ϵ and μ greater under fast than slow dilution) are not observed in the individual

fits. This may be as a result of the fact that the numbers of data points for individual cultures are not sufficient to capture these trends.

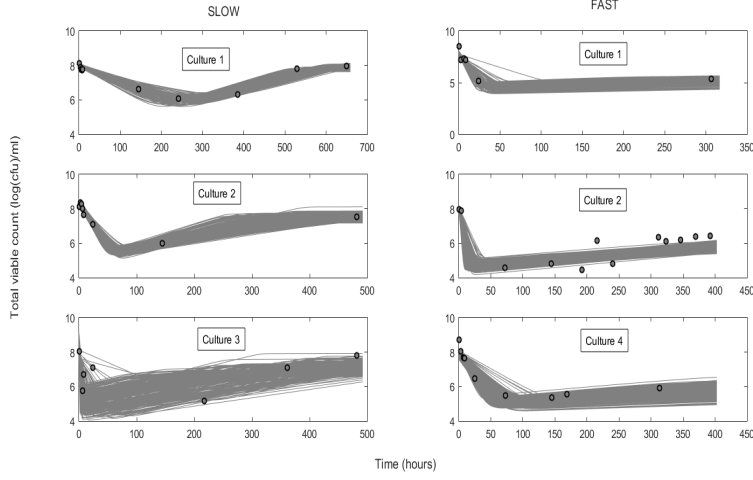


Figure 2.8: Grey lines show the fits from 500 parameter sets randomly drawn from the posterior distribution resulting from running the MCMC separately on the individual replicated cultures

2.7 PARAMETER IDENTIFIABILITY

When a model is fitted to experimental data, the question of how reliable parameter estimates are arises. Strong correlations and functional relationships between parameters may prevent them from being uniquely identified. There may exist parameters which can vary several orders of magnitude without affecting the quality of fit. Our aim in this work is not to use the model to identify values of the parameters, but to find the origins of the recovery dynamics and diversity patterns observed in slow and fast dilution. For these purposes, a model whose fits to observation are not highly sensitive to individual values of unknown parameters is an advantage, as the model's structure robustly captures the observed dynamics. In contrast, for parameter estimation, such sensitivity would be desirable. Despite this difference in aims, in this section we explore parameter identifiability.

Figure 2.10 shows correlation plots of the parameters under slow and fast dilution rates; several parameter pairs are significantly correlated. Under both dilution rates, there is a negative correlation between $\hat{A}$ and P . This

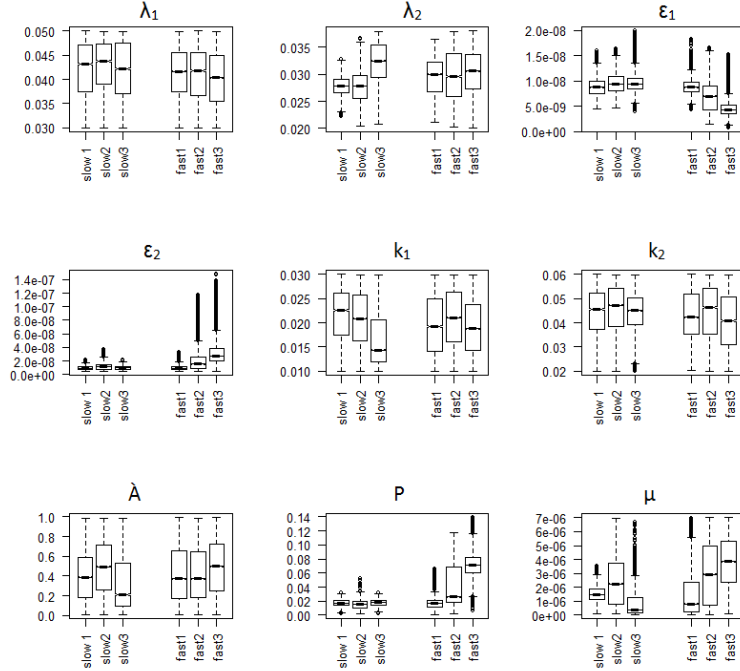


Figure 2.9: *Boxplots of 5000 random samples of posterior parameter distributions for slow and fast dilution rates (for individual cultures).*

means that they are not independent, and indeed their effects are related (they model INH reducing bacterial numbers through growth inhibition and bactericidal action). The dependence of the parameters λ and k has been noted previously [106, 120]; both affect the growth rate.

Bayesian inference provides distributions of parameter values that are consistent with data. The Bayesian approach to fitting is well-suited to the situation where parameters are not identifiable; many combinations of parameters may suffice for the model to have a specified behaviour. Hines et al [86] provide a framework which uses the MCMC traces of Bayesian fits to check for identifiability. The structure of the posterior distribution reveals whether regions of very high posterior distribution are localized to specific regions in the parameter space. They propose that if the posterior parameter values in best agreement with the data are confined to a small region of the parameter space, then the model parameters are identifiable. Figures 2.11 and 2.12 show this for the parameters that are significantly correlated, under

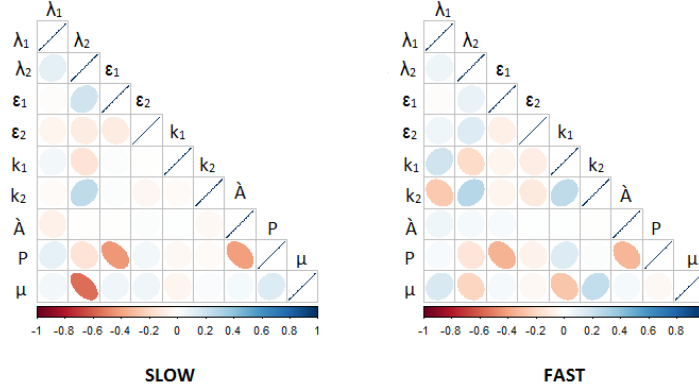


Figure 2.10: *Correlation plots of 1000 random samples of posterior parameter distributions for slow and fast dilution rates. Dark blue and dark red colors display parameters with high positive and negative linear dependencies, respectively*

both dilution rates, the high likelihood are localized to specific regions (in yellow).

2.8 MULTIPOPULATION MODEL

We extend the two-strain model to a multi-strain model using the posterior parameter values of the MCMC chain to define a posterior collection of resistant strains, each of which is consistent with the observed growth curves. We model m distinct mutations that could cause an *M. tuberculosis* bacillus to evade the action of INH. The susceptible strain B_1 mutates to m strains at the different rates. Each mutant subpopulation exhibits different maximum growth rates λ_i , different yield constants ϵ_i , and different half saturation constants, k_i , as we assume that some mutations are more expensive than others.

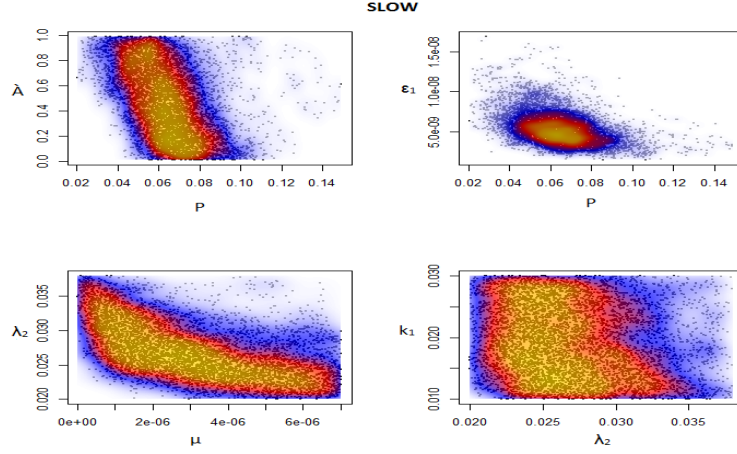


Figure 2.11: Heat maps of the joint posterior distributions of parameters that are significantly correlated (with a correlation coefficient $|r| \geq 0.5$). Lighter, brighter areas correspond to areas of higher posterior density.

The differential equations are:

$$\begin{aligned}
 \frac{dB_1}{dt} &= B_1((1 - \hat{A})\phi_1(R) - \mu - P - D) \\
 \frac{dB_i}{dt} &= B_i(\phi_i(R) - D) + \mu_i B_1, \quad i = 2, \dots, m \\
 \frac{dR}{dt} &= D(C - R) - \sum_{i=1}^n \epsilon_i \phi_i(R) B_i.
 \end{aligned} \tag{2.7}$$

where

$$\mu = \sum_{i=2}^n \mu_i.$$

Drug resistance in *M. tuberculosis* occurs through chromosomal mutations as is common with other bacterial populations. Several genes such as *KatG*, *inhA*, *oxyR-aphC*, *kasA* and *ndh* have been associated with resistance to INH and mutations found in association with drug-resistant *M. tuberculosis* cells may result in resistance to different levels of antibiotic [135]. Also the mutation rate for individual genes varies significantly between and within genes [72]. In order to capture this, we use different values of μ_i here. If the strains with the least fitness costs are assumed to appear faster than other strains, then this gives them an advantage. Therefore in our model simulation, we

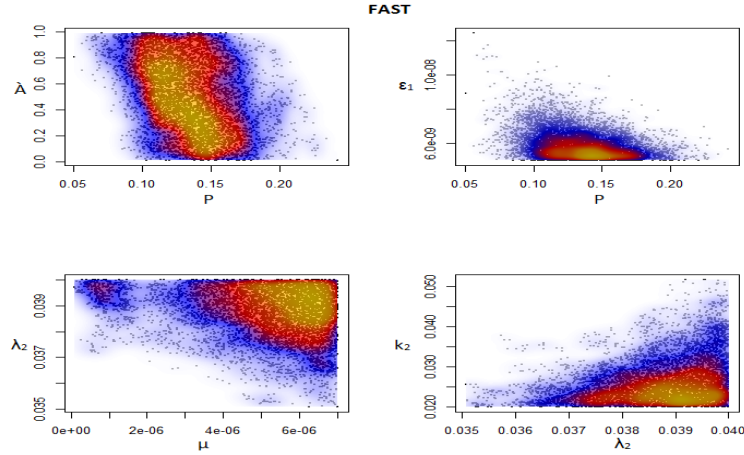


Figure 2.12: Heat maps of the joint posterior distributions of parameters that are significantly correlated (with a correlation coefficient $|r| \geq 0.5$). Lighter, brighter areas correspond to areas of higher posterior density.

assume that mutations that confer higher net growth occur at a lower rate than mutations with lower net growth. Removing this assumption does not change the population dynamics, likely because the ultimate fate of each strain depends more on its fitness cost than on the rate at which it appears in the system.

We explored system 2.7 using the posterior parameter fits of system 2.4 to obtain diverse resistant sub-populations. The diversity patterns differ markedly under fast and slow dilution rates (Figure 2.13), with fast dilution allowing a wider diversity of sub-populations for a much longer period than slow dilution. This correlates with experimental findings where it was observed that greater diversity in the *katG* gene was observed under a fast dilution than slow [91]. This may be surprising given that fast dilution provides very strong selective pressure for rapid division. However, the model predicts that there is an excess of the growth resource under fast dilution, with the effect that there is reduced competition for resources between sub-populations. The effect of dilution rate on diverse population interactions in chemostat systems has been noted previously [175, 107], without the focus on resistance and diversity.

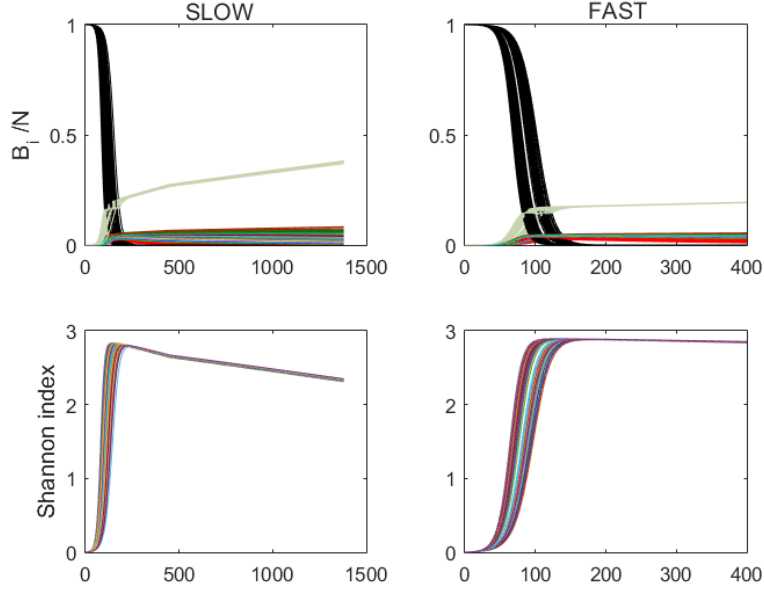


Figure 2.13: *Proportion plots and plots of the Shannon diversity index H' for 1000 parameter sets of the multi-population model: with twenty resistant sub-populations competing after the susceptible sub-population (black) dies off. The fittest sub-population (green) maintains the highest proportion under both dilution rates.*

2.9 COMPETITION IN THE CHEMOSTAT

After the sensitive subpopulation of system 2.7 dies off (due to mutation and action of isoniazid), the system reduces model of several subpopulations of mycobacterium tuberculosis resistant to isoniazid shown below.

$$\begin{aligned} \frac{dB_i}{dt} &= B_i(\phi_i(R) - D), \quad i = 1, \dots, n \\ \frac{dR}{dt} &= D(C - R) - \sum_{i=1}^n \epsilon_i \phi_i(R) B_i. \end{aligned} \quad (2.8)$$

The break even concentration for the populations are $\alpha_i = \frac{K_i D}{\lambda_{max_i} - D}$ ($i = 1 \dots n$). It has been previously shown [87, 113, 166, 88], that in chemostat models of n competing populations ($n \geq 2$), the population with the small-

est break even concentration always wins. The break even concentration depends on the dilution rate and two parameters intrinsic to the population: λ_i and k_i , which we draw from posterior parameter values to reflect different fitness costs. The fitter a population is, the lower its break-even concentration, and thus the greater its chances of surviving in the chemostat.

2.10 DIVERSITY IN THE SYSTEM 2.8

A range of diversity indices have been used with bacterial communities [52, 129, 27]. Here we analyse diversity using Shannon’s index H' [164], which is defined as $H' = -\sum_{i=1}^n P_i \ln P_i$, where P_i is the proportion of the total population occupied by sub-population i . The proportion of the total population occupied by a sub-population varies inversely with its fitness cost. The resistant sub-population with the least fitness cost will have the highest proportion of the total population (Figure 2.13). As we do not consider onward mutation, cultures under fast and slow dilution rates will eventually reach a steady state with one dominant sub-population. However, simulations indicate that it would take very long times to reach this equilibrium under fast dilution as the diversity index is maintained at a constant level for longer (Figure 2.13).

2.11 CONCLUSION

In this chapter, we modelled *M. tuberculosis* growth and diversity in continuous culture under the action of INH. We obtained posterior parameter distributions for which system 2.4 matches experimental data on the recovery of the population during addition of INH, and explored the dynamics of diverse INH-resistant sub-populations consistent with these data in system 2.7. System 2.4 robustly matches observed growth curves and provides a platform for exploring the emergence of resistance to INH in continuous culture. Even though fast dilution provides very strong selective pressure in favour of rapid cell division, fast dilution resulted in excess resource compared to slow dilution, and the diversity of model sub-populations that could be maintained over long periods was greater under fast dilution. This is consistent with the observations of Jeeves et al [91] regarding diversity in *katG*

mutations. We interpret this result in the context of fitness and competition in a chemostat.

An organism's fitness is often framed as a one-dimensional variable capturing an organism's overall capability for growth. This depends on environmental factors affecting the organism as well as the organism's ability to thrive in the environment. In our chemostat model, there are three distinct elements of fitness of the sub-populations: the ability to divide before being washed out, the strength of resistance to INH, and the ability to consume and efficiently use the carbon resource before competitors. The relative importance of these aspects of fitness is different under the two dilution rates. Fast dilution confers strong selection in favour of rapid division, and might, on that basis, be expected to permit less diversity than slow dilution. However, fast dilution also results in excess glycerol in the medium, weakening competition for resources. It appears that the effects of reduced competition are a stronger driver of diversity in this system than the effects of strong selection for rapid growth. The fact that the conversion parameter ϵ_2 was inferred to be much larger under fast dilution than under slow dilution supports the notion that selection for rapid carbon usage is lower under fast dilution.

An increased capacity of cultures to give rise to mutants (reflected in the mutation rate) increases the adaptability of these cultures to harsh conditions [124]; this could be reflective of the adaptation of clinical populations to the host environment. As the posterior values of the mutation rate are notably higher under fast dilution than under slow dilution, one might expect that cultures in fast dilution would be able to recover more easily. However, the potential advantages conferred by rapid mutation may be outweighed by the pressure of fast dilution and the fact that the fast dilution rate may reach the physiological limit of *M. tuberculosis*' growth capacity. All of these factors combine to affect the outcome of mutation-selection balance in this system, resulting in higher long-term diversity under fast dilution than under slow dilution.

This chapter raises questions that can be answered in further chemostat studies: does high resource availability delay the end course of exploitative competition? What determines the mutation rate and how does it affect the evolution and fitness of resistance? Studies combining whole-genome sequencing with monitoring of the resource concentration will provide tests

of the predictions we have made, as well as insights as to the continuous generation of diversity after the drug-sensitive cells have died out.

It is reasonable to assume that upon emergence, resistant sub-populations compete with their drug-susceptible progenitors for resources. If continuous culture systems are a model for in-host infection, our results suggest that drug selection occurs when host resources are likely to be plentiful – during slow growth periods of the infection or when bacterial numbers have declined after early bactericidal action – giving rise to diverse resistant sub-populations. Our model does not require slow-growing persisters in response to antibiotic action or other advantages beyond simple mutation in order to match observed complete recovery of bacterial populations. Even in the very simple, constrained and highly selective environment of nutrient-limited continuous culture systems, *M. tuberculosis* can rapidly generate fit drug-resistant mutants that can establish long-term survival.

Table 2.1: *Definitions of all states and parameters*

Symbol	Definition		
States			
B_1	Susceptible bacterial sub-population (cfu/mL)		
B_2	Resistant sub-population (cfu/mL)		
R	Resource concentration (mg/mL)		
Variables		Priors $\sim U(a, b)$	
		a	b
μ	Mutation rate (h^{-1})	2.56×10^{-8}	2×10^{-6}
$\hat{A}$	Bacteriostatic action of INH	0.01	0.99
P	Antibiotic bactericidal rate (h^{-1})	0	4
ϵ_1	Conversion efficiency parameter of B_1 (mg/cfu)	5×10^{-10}	2×10^{-8}
ϵ_2	Conversion efficiency parameter of B_2 (mg/cfu)	5×10^{-9}	2×10^{-7}
D	Dilution rate (h^{-1})	-	-
C	Resource input concentration (mg/L)	-	-
λ_1	Maximum growth rate of susceptible sub-population (h^{-1})	0.03	0.05
λ_2	Maximum growth rate of resistant sub-population (h^{-1})	0.02	0.04
k_1	Half saturation constant of susceptible sub-population (mg/L)	0.01	0.03
k_2	Half saturation constant of resistant sub-population (mg/L)	0.02	0.06

Table 2.2: *Baseline values of parameter fits (slow dilution)*

Parameter	Mean	2.5 th percentile	97.5 th percentile
μ	3.01×10^{-6}	2.32×10^{-7}	6.68×10^{-6}
$\hat{A}$	0.04	0.027	0.95
P	0.06	0.03	0.10
ϵ_1	5.46×10^{-8}	2.61×10^{-9}	1.07×10^{-8}
ϵ_2	8.46×10^{-8}	2.69×10^{-8}	1.77×10^{-7}
λ_1	0.041	0.031	0.049
λ_2	0.027	0.021	0.037
k_1	0.019	0.01	0.03
k_2	0.04	0.02	0.05

Table 2.3: *Baseline values of parameter fits (fast dilution)*

Parameter	Mean	2.5 th percentile	97.5 th percentile
μ	5.33×10^{-6}	2.14×10^{-6}	6.94×10^{-6}
$\hat{A}$	0.50	0.03	0.96
P	0.06	0.03	0.09
ϵ_1	7.29×10^{-9}	5.06×10^{-9}	1.26×10^{-8}
ϵ_2	4.83×10^{-7}	1.07×10^{-7}	1.08×10^{-6}
λ_1	0.03	0.04	0.05
λ_2	0.03	0.036	0.04
k_1	0.02	0.01	0.03
k_2	0.03	0.02	0.04

If there is no struggle, there is no progress.

Frederick Douglass

3

Within host model of mycobacterium tuberculosis

3.1 INTRODUCTION

The immune response to the presence of *M. tuberculosis* cells in a host is quite complex and involves many components of the immune system [64]. In this chapter, mathematical models of human tuberculosis infection in the lungs developed by [56] are studied. Models of this kind are usually complicated with many parameters and a lot of uncertainty with respect to the parameter values. Due to the high non-linearity of these models, traditional analysis is not feasible and as such we typically rely on a numerical scheme to study them.

The aim of this chapter is to explore model dynamics and identify components of the immune system that affect the outcome of infection using tools from data science. Tools such as discriminant analysis of principal components (DAPC) and random forests are used to identify which model parameters contribute to disease outcome. First of all, a single strain model is studied and this is extended to a two strain model to study the effects of competition on the outcome of disease. Using these models, virtual experiments that mimic re-infection of a host with latent tuberculosis are performed

and the results are analysed.

3.2 IMMUNE RESPONSE TO TUBERCULOSIS

When a host is infected by any disease, the first line of action of the immune system is carried out by the innate (non-specific) immune response. The majority of patients who develop TB have symptoms that are restricted to the lung (pulmonary TB) [126]. The initial site of *Mtb* infection is the alveoli of the lung, where it comes into contact with lung macrophages, dendritic cells and monocytes recruited from the bloodstream [18]. Macrophages are found in every tissue of the body and respond to bacterial infection by engulfing them (phagocytosis) [155]. Macrophages could be activated (by adaptive immune system) or unactivated, with activated macrophages having the ability to kill bacterial cells more effectively. Schluger et al [160] speculate that if the macrophages kill the bacteria quickly enough, the disease can be cleared. *M. tuberculosis* cells however can reproduce in infected macrophages, causing it to burst open and release more bacterial cells. As such to clear the infection, the adaptive immune system develops. The adaptive immune system consists mainly of the lymphocytes which are classified into two broad groups: the *B* lymphocytes (*B* cells responsible for antibody responses) and *T* lymphocytes (*T* cells responsible for cell-mediated immunity). For tuberculosis, a cell-mediated immune response is required to fight infection [64] and as such the *T* cells form the adaptive immune response for tuberculosis.

3.3 TUBERCULOSIS PROGRESSION

Tuberculosis disease progression is relatively slow compared to other bacterial and viral infections. It is believed that 90% of all infections remain latent, while 5% progress rapidly to active disease with the remaining 5% being able to clear the disease [35]. Gammack et al [69] show that the progression of tuberculosis can be split into three types of outcomes:

- controlled via innate immunity (clearance)
- controlled via cell-mediated immunity (latency): latency refers to the situation where an individual has been exposed to *M. tuberculosis* and

subsequently develops an asymptomatic infection. It can be considered as state where bacterial growth is slow or absent.

- granuloma growth with cell-mediated immunity (active disease): Granulomas are multicellular lesions that form in the lungs of infected individuals in response to persistent antigenic stimulation. They usually contain bacteria, macrophages and other immune cells. It is speculated that disease progression may be indicative of the formation and function of granulomas [162, 69].

Although changes in the immune response following the establishment of latent infection may result in clinical disease, known as reactivation tuberculosis (Figure 3.1), we do not consider this here.

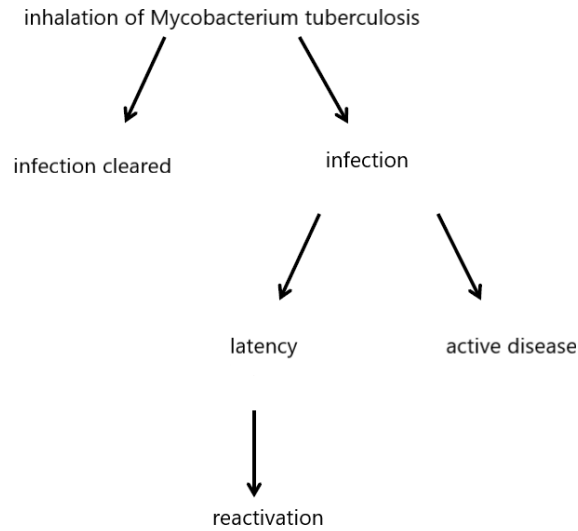


Figure 3.1: *Outcomes of exposure to Mycobacterium tuberculosis.*

3.4 DISCRIMINANT ANALYSIS PRINCIPAL COMPONENT ANALYSIS

Given an $n \times p$ data matrix X , where n is the number of observations (which belong to g different groups) and p is the number of features or parameters, discriminant analysis aims to find a combination of parameters that best separates the groups. If we assume that each column is centered to mean zero,

classical linear discriminant analysis seeks linear combinations of parameters of the form:

$$\sum_{j=1}^p X^j \mathbf{w}_j = \mathbf{X} \mathbf{w},$$

that maximize the objective

$$\mathbf{w} = \operatorname{argmax}_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{s}_b \mathbf{w}}{\mathbf{w}^T \mathbf{s}_w \mathbf{w}}$$

where $\mathbf{w} = [\mathbf{w}_1, \dots, \mathbf{w}_p]^T$, the vector of p loadings known as the discriminant coefficients [95], s_w and s_b are the with-in class and between-class scatter matrices respectively [89]. These linear combination of variables are called principal components or discriminant functions. Discriminant analysis thus utilizes the class information to find informative projections that best separate the data.

Principal component analysis (PCA) on the other hand does not consider the classes of the observations and as such may not be suitable for classification [194]. PCA is used to reduce dimensionality and visualize structure in the data [173]. PCA is also used as a feature extraction method that selects features with the largest variance. These features are obtained by the eigenanalysis of the covariance matrix of X . The resulting eigenvectors are the principal components.

Discriminant analysis of principal components introduced by [95], is a combination of both methods described above. Rather than analyzing X directly, a data transformation using principal component analysis (PCA) is carried out, and the best features (parameters) are extracted, then linear discriminant analysis is carried out on these principal components. Imani et al [89] report that when both methods are combined, the accuracy is improved than when each method is used separately.

In this chapter, DAPC is used to systematically identify combinations of parameters that distinguish between disease outcomes. DAPC is implemented in the *ade4* package [95], in R.

3.5 RANDOM FORESTS

A decision tree is a graph that illustrates every possible outcome of a decision. The parameters in our data matrix can be viewed as predictors of the disease outcome. A decision tree stratifies the predictor space into simple regions to classify an observation. To predict the class of a given observation, it typically uses the mean (regression trees) or the mode (classification trees) of the training observations in the region to which the observation belongs [90]. Classification trees are grown using recursive binary splitting based on the *Gini* index defined as

$$Gini = \sum_{k=1}^K \hat{p}_{mk}(1 - \hat{p}_{mk}),$$

where there are K classes and $\hat{p}_{mk}$ is the proportion of observations in the m^{th} region that belong to the k^{th} class. If p_{mk} is close to 1 or 0, then the *Gini* index takes on a small value. As such, the *Gini* index gives a measure of node purity with a small value indicating that a node contains predominantly observations from a single class. Decision trees are quite simple to understand and visualize but are ridden with the problem of high variance. That is if our data matrix is split into two different parts and a decision tree is built for each part, the results may be completely different.

Random forests provide improvement over single decision trees. It reduces the variance by averaging over several decision trees which are built on random independent samples of the data matrix. This is intuitive as taking the average of the variance of n independently drawn samples, reduces the total variance by a factor of n . A number of decision trees are built on bootstrapped training samples. To build any one of these trees, each time a split is considered, only a random subset (of size h) of the predictors (usually $h = \sqrt{p}$) is chosen as split candidates. The rationale behind this is to reduce the chances of trees being highly correlated. When there is a strong predictor in the data, there is a possibility that the decision trees would look very similar because this predictor would mostly be at the root of the trees. As such, averaging over a group of trees would not necessarily reduce the variance except the trees are uncorrelated. One can obtain from a random forest, the overall summary of the importance each predictor in classifying

the observations. This can be done using the *Gini* index i.e. by adding up the total amount that the *Gini* index is decreased by splits over a given predictor, averaged over all trees in the random forest. The mean decrease in *Gini* coefficient is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. Each time a particular predictor is used to split a node, the *Gini* index for the child nodes are calculated and compared to that of the original node. Predictors that result in nodes with higher purity have a higher decrease in *Gini* index.

A random forest is built to systematically identify the importance of parameters in classifying disease outcomes. Random forest is implemented in the randomForest package [117], in R.

3.6 ONE STRAIN MODEL

In this model, we consider a single population of bacterial cells in the lungs of an infected individual. The parameters of the models discussed in this chapter were adopted from [126], the primary citations and plausible ranges are shown in Table 3.3. This bacteria could be extracellular (B) or intracellular (contained in infected macrophages). We assume that the extracellular bacteria follow a logistic growth with a maximum carrying capacity, N_{max} and a maximum growth rate, α_B . This bacterial population is subject to a constant killing rate i by the innate immune system.

Macrophages could be resting, activated (recruited by T-cells) or infected with bacteria. We do not model these populations explicitly but consider a combined population of macrophages in the lung. There is a constant supply of resting macrophages at the rate a_1 , and these macrophages are recruited by the T cells (T) at a rate a_2 . Extracellular bacteria become engulfed by macrophages which then carry N_1 bacteria. Infection of macrophages depends on extracellular bacterial load with c_1 denoting the level of bacteria at which we have half maximal rate of infection. Thus the action of macrophages on extracellular bacteria is represented by the term $N_1(a_1 + a_2)\frac{B}{B+c_1}$. Intracellular bacteria become extracellular when infected macrophages die at a rate μ_M and burst open releasing N_2 bacteria.

The rate at which T cells kill bacteria depends on the amount of T cells available. This is given by the terms $a_3\frac{T}{T+c_2}$, where a_3 is the maximum rate

at which T cells kill bacteria and c_2 is the value of T when the killing rate is half its maximal value. T cells also kill infected macrophages at a rate which depends on the effector cell to target ratio ($T : M$). The rate is given as $a_4 \frac{T}{T+c_3M}$, where a_4 is the maximum rate at which T cells kill infected macrophages and c_3 is value of the ratio $T : M$ when the killing rate is half maximal. The growth of the T cell population is induced by the presence of infected macrophages at a constant rate α_T and they decay at a constant rate μ_T

The model is thus represented by the equations below

$$\begin{aligned}\frac{dB}{dt} &= \alpha_B B \left(1 - \frac{B}{N_{max}}\right) - iB - N_1(a_1 + a_2T) \frac{B}{B+c_1} - a_3B \frac{T}{T+c_2} + N_2\mu_M M \\ \frac{dM}{dt} &= (a_1 + a_2T) \frac{B}{B+c_1} - a_4M \frac{T}{T+c_3M} - \mu_M M \\ \frac{dT}{dt} &= \alpha_T M - \mu_T T\end{aligned}\tag{3.1}$$

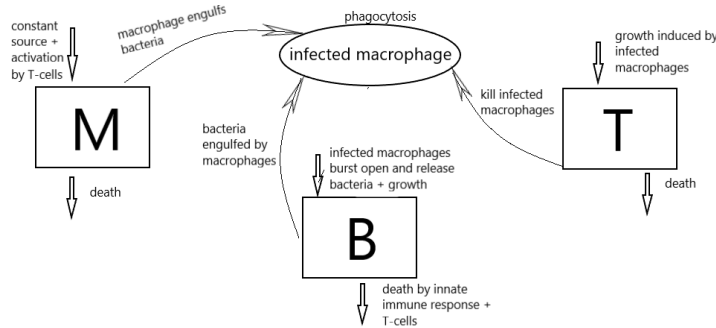


Figure 3.2: *Schematic of model 3.1 describing immune response to tuberculosis infection. The three model compartments are depicted as M (macrophages), B (bacteria) and T (T-cells).*

3.7 ANALYSIS OF MODEL 3.1

To explore the interactions in the model, 20,000 sets of parameters are obtained from ranges previously published [126]. Each parameter is drawn uniformly at random from the parameter ranges as shown in Table 3.3.

Model 3.1 is solved to obtain temporal dynamics for each state variable. We consider time courses for the model in a naive host that has no previous exposure *M. tuberculosis* cells (so $T_0 = 0$ with $B_0 = 25$). Kirschner et al [126] point out that no good markers of tuberculosis progression in humans currently exist. However, a good marker of disease progression commonly used in literature is bacterial load [146, 123]. During latency, the intracellular bacterial load is believed to be 1000-fold less than during active disease. In the simulations, we consider total bacterial load as a marker of disease progression and following [126] define the three groups as follows:

- ≤ 50 : Clearance
- > 50 and $\leq 10^6$: Latency
- $> 10^6$: Active disease

Figure 3.3 shows an histogram of the final bacterial load for the simulations. The model is able to reproduce expected infection outcomes with over 80% of the simulations falling into the latent class, 10% in active disease and the rest in the clearance class.

Figure 3.4 shows the time course for 3.1 using parameters that lead to latency (green) and active disease. For latency, there is an initial round of exponential growth, with some minor oscillations, after which the immune system is able to suppress the bacterial load to a latent equilibrium. Also minor oscillations of the immune response occur in response to oscillations in bacterial load. The model predicts that it takes about one year, on average, to reach latent equilibrium.

For the active class, there is also an initial exponential growth of the bacterial cells, after which the system establishes equilibrium at its bacterial load carrying capacity. The equilibrium levels of macrophages are notably higher than what is obtained during latency.

Figures 3.5 show box-plots of the 20000 parameter sets grouped according to their disease outcome classes. Some parameters such as $a_4, \mu_T, i, N_2, c_2, c_3$ and μ_M have the same distributions across the disease outcome classes, while others are quite different. N_1 and a_1 have much lower values in the clearance class than in the other disease outcomes. This is intuitive as the value of N_1

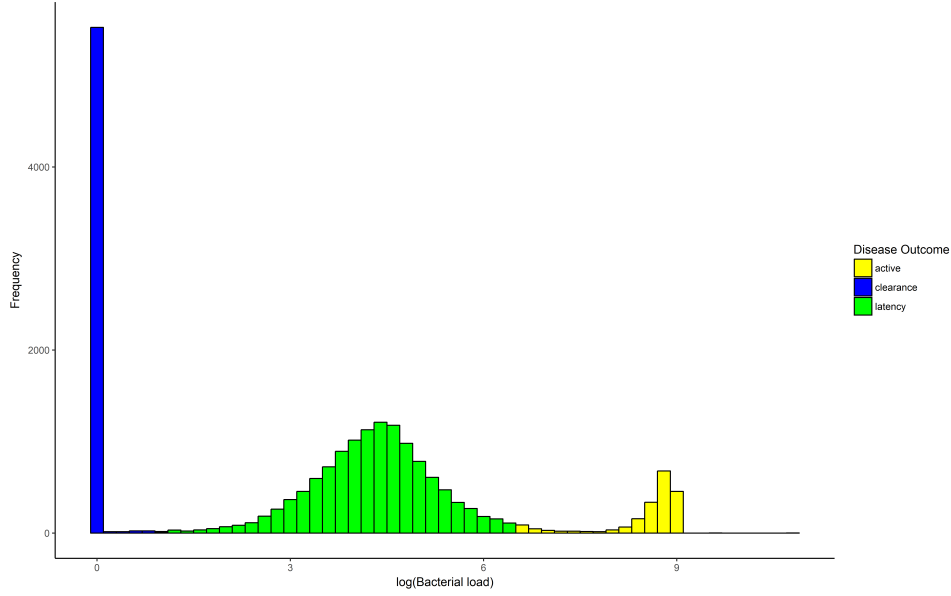


Figure 3.3: *Histogram showing final bacterial load: 10000 parameter sets were drawn from the ranges in Table 3.1*

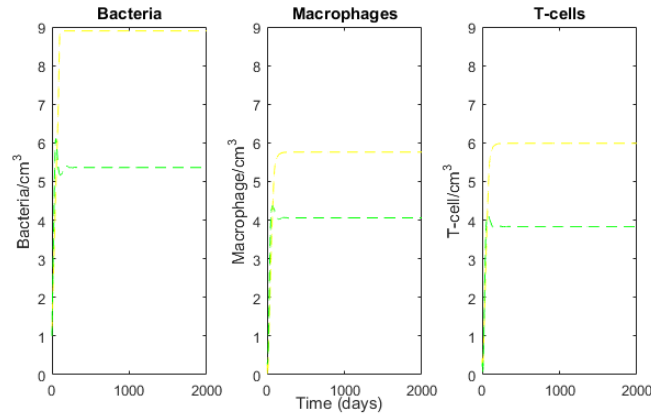


Figure 3.4: *Model simulation time course for latency (green) and active (yellow) classes.*

(which is the amount of bacteria engulfed by macrophages) and a_1 (which is the constant source of macrophages), would need to be high in order to clear the disease. Similarly, the values for bacterial growth rate (α_B) and the level of bacteria at which the rate of infection of macrophages is half maximal (c_1) are much lower in the clearance class than those of latency

and active disease as the smaller these parameters are, the fewer the number of *M. tuberculosis* cells in the host. α_T and a_3 are quite different between active disease and latency classes with lower values in the active class. These parameters represent the growth of T-cells and action of T-cells on bacteria respectively. High efficiency of T-cells at killing bacteria maintains latency, whereas lower levels lead to active disease.

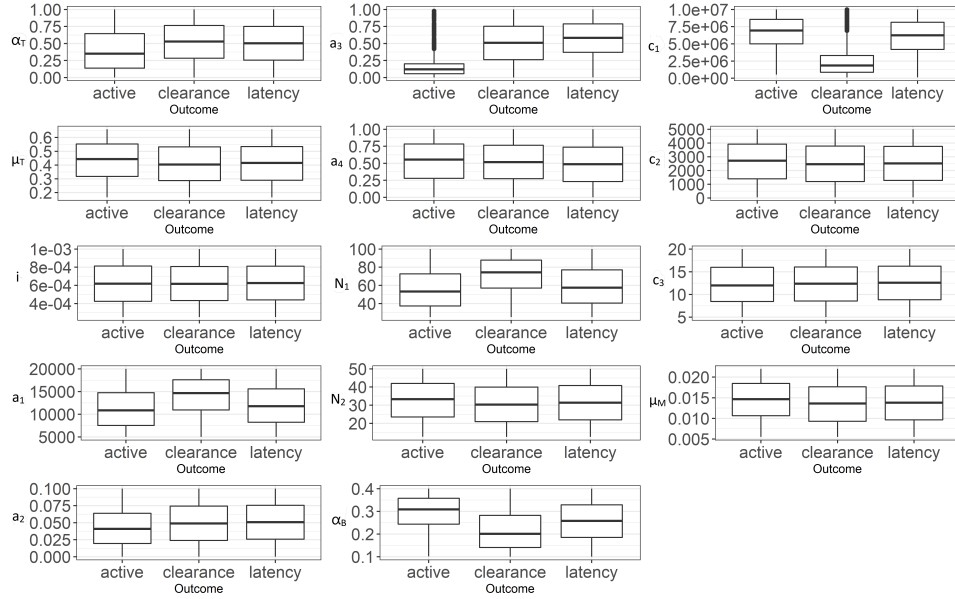


Figure 3.5: *Boxplots of $\alpha_T, \mu_T, i, a_1, a_2, a_3, a_4, N_1, c_1, c_2, c_3, \mu_M$ and N_2 for the three disease outcomes.*

Parameter values are drawn randomly, and as such there is no correlation between them. However when these parameters are separated into the different disease classes, there are correlations among a few of them. Figure 3.5 shows correlation matrices for the parameters in the three disease classes. For clearance, there are only three major correlations: The higher c_1 is, the lesser the ability of the system to clear the bacteria. This is compensated for by higher levels of a_1 and N_1 (which reduce bacterial load and lead to clearance) and lower bacterial growth rate α_B . These correlations also occur in the latency and active disease outcome classes but with lower $|r|$, where r is the correlation coefficient. There are similar correlation pairs in active disease and latency parameter classes but with quite different correlation coefficients. The correlation between c_2 (the value of T cells when the rate

at which T cells kill bacteria is half maximal) and a_3 (the rate at which T cells kill bacteria) is peculiar to latency. The higher c_2 is, the more the bacterial load, but there is a compensation for this in higher levels of a_3 thus causing the system to maintain latency. A correlation that is quite strong compared to the others in the active disease class, is the one between the bacterial growth rate, α_B , and the rate at which T cells kill bacteria, a_3 . As a_3 increases, the bacteria has to grow faster in order to achieve active disease.

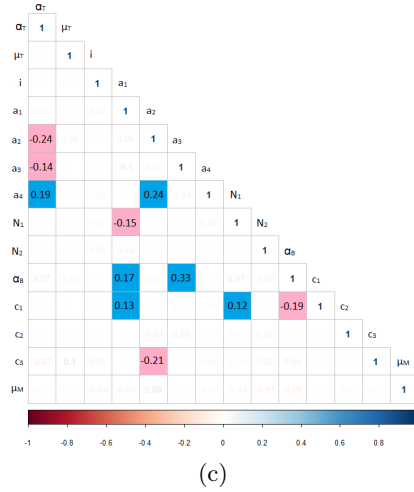
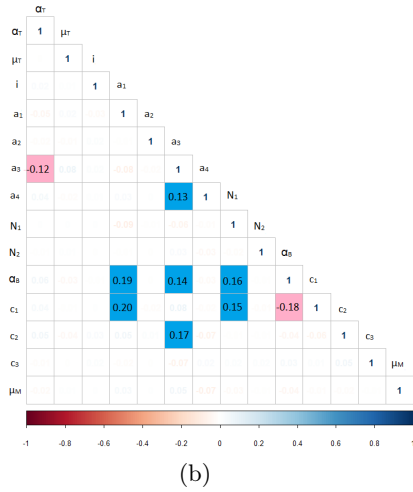
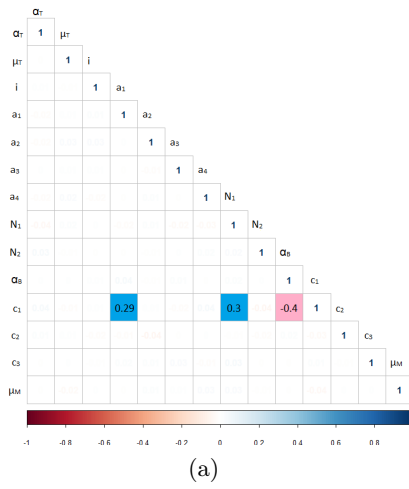


Figure 3.6: Correlation matrices for parameters in the different disease outcomes: (a) clearance, (b) latency (c) active disease, colors range from dark red ($r = -1$) to dark blue ($r = 1$), where r is the correlation coefficient. Parameter pairs with $|r| > 0.1$ are colored blue ($r > 0$) and pink ($r < 0$).

No one single parameter determines the outcome of disease, but a relatively small change in several parameters can move the disease outcome from one class to another. We separated parameter combinations that lead to the three disease outcomes using the DAPC approach described in section 3.4. Figure 3.7 shows the scatter plot of the DAPC analysis. The steady state values of the bacterial population for the model simulation are represented as dots which are coloured by their disease outcome classes. The groups are plotted according to their positions on the DAPC functions. Each discriminant function is a linear combination of the model parameters. Almost all the difference between the groups is captured by the first discriminant function, as seen by the height of the first eigenvalue bar. The ‘loadings’ or coefficients of the discriminant functions show the relative importance of each parameter to group separation. Figure 3.8 shows the contributions of parameters to the first principal component of the DAPC of model 3.1. The level of bacteria at which the infection of macrophages is half maximal, c_1 , makes the highest contribution of over 40%. The other two parameters that make significant contributions are α_B and N_1 , which contribute about 28% and 15% respectively.

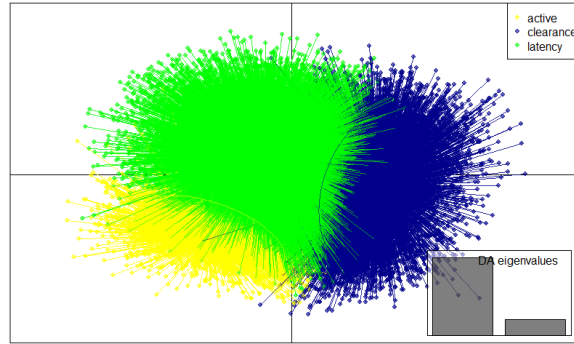


Figure 3.7: *DAPC scatter plot for model 3.1. The x -axis is a linear combination of parameters corresponding to the first discriminant function, and the y -axis corresponds to the second.*

Figure 3.9 is a set of heatmaps that illustrate how combinations of the parameters that contribute the most to the two discriminant functions affect disease outcome. This figure illustrates some intuitive trade-offs. When the rate at which T-cells kill bacteria is low i.e. of $a_3 < 0.2$, the disease outcome is always active disease regardless of the value of N_1 . On the other hand,

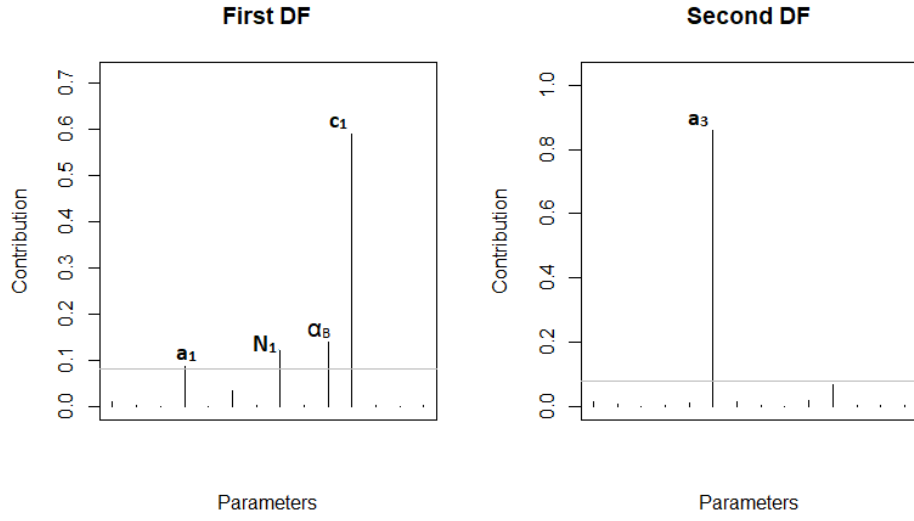


Figure 3.8: Contributions of parameters to the first and second discriminant functions of the DAPC of model 3.1. The height of each bar is proportional to the contribution of the corresponding parameter to the discriminant function. Only parameters whose contribution is above an arbitrary threshold (grey horizontal line) are indicated for the sake of clarity.

when $c_1 < 10^6$, the disease is always cleared regardless of the value of N_1 .

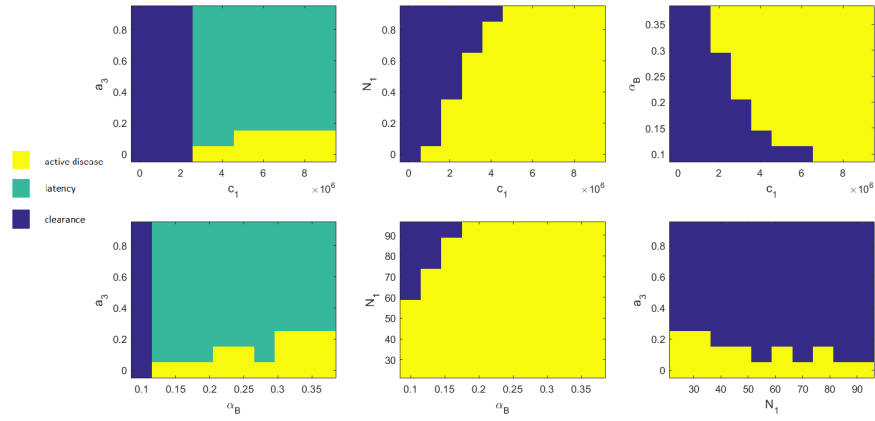


Figure 3.9: Heatmaps illustrating how disease outcome depends on combinations of key parameters (as determined by DAPC)

We also applied the random forests approach explained in section 3.5, to determine which parameters contribute the most to separating the disease outcomes. A random forest made up of 2000 decision trees was built with the simulation results. Figure 3.10 *a* and *b* shows the importance of each parameter in the random forest classifier. c_1 contributes the most in separating the classes. The large break after the first 5 parameters indicate that the other parameters do not contribute much to separating the classes.

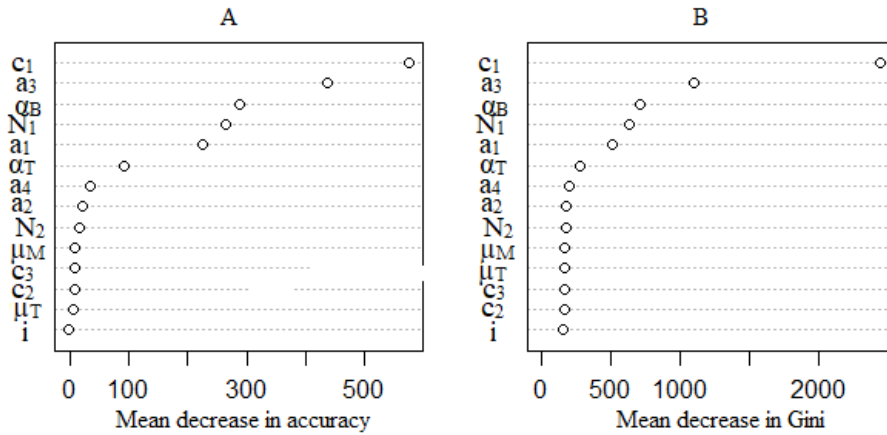


Figure 3.10: *Importance of each parameter to the random forest classifier with each parameter on the y – axis, and their importance on the x – axis, ordered top-to-bottom as most- to least-important. A: Mean decrease in accuracy shows much the model accuracy decreases if that parameter is dropped from the model and B: the mean decrease in Gini score shows the measure of parameter importance based on the Gini index. The higher the value of both of these, the more the importance of the variable in the model.*

3.8 TWO STRAIN MODEL

Recent molecular methods for strain differentiation have revealed that individuals can harbour more than one distinguishable strain of *M. tuberculosis* [30, 47]. We now consider two strains of *M. tuberculosis* populations B_1 and B_2 in the lungs of an infected individual to see how the interaction between them affects disease outcome. Strains may differ by drug resistance determinants or by their ability to evade the immune system [119]. In this

model, we differentiate between the strains by assuming that the immune response is different for the two strains. Also we introduce seven new parameters $\hat{N}_1, \hat{N}_2, \hat{i}, \hat{\alpha}_T, \alpha_{B_2}, \hat{a}_5$, and $\hat{a}_4$ to differentiate between the two strains. These new parameters are defined as: $p_4 N_1 = \hat{N}_1$, $p_3 i_1 = \hat{i}$, $p_1 \alpha_T = \hat{\alpha}_T$, $p_2 \alpha_{B_1} = \alpha_{B_2}$, $p_5 a_3 = \hat{a}_3$, $p_7 N_2 = \hat{N}_2$ and $p_6 a_4 = \hat{a}_4$. The two strain model thus is:

$$\begin{aligned}
\frac{dB_1}{dt} &= \alpha_{B_1} B_1 \left(1 - \frac{B}{N_{max}}\right) - i B_1 - N_1 (a_1 + a_2 T) \frac{B_1}{B + c_1} - a_3 B_1 \frac{T}{T + c_1} + N_2 \mu_{M_1} M_1 \\
\frac{dM_1}{dt} &= (a_1 + a_2 T) \frac{B}{B_1 + c_1} - a_4 M_1 \frac{T}{T + c_3 M} - \mu_{M_1} M_1 \\
\frac{dT}{dt} &= \alpha_T M_1 + \hat{\alpha}_T M_2 - \mu_T T \\
\frac{dB_2}{dt} &= \alpha_{B_2} B_2 \left(1 - \frac{B}{N_{max}}\right) - \hat{i} B_2 - \hat{N}_1 (a_1 + a_2 T) \frac{B_2}{B + c_1} - \hat{a}_3 B_2 \frac{T}{T + c_2} + N_2 \mu_{M_2} M_2 \\
\frac{dM_2}{dt} &= (a_1 + a_2 T) \frac{B_2}{B_2 + c_1} - \hat{a}_4 M_2 \frac{T}{T + c_3 M} - \mu_{M_2} M_2
\end{aligned} \tag{3.2}$$

where $B = B_1 + B_2$, and $M = M_1 + M_2$. We assume that the innate and adaptive immune system have a slightly reduced effect on strain 2, but strain 2 pays a penalty for this in terms of its growth rate. Thus we select $p_i < 1$ such that $p_i \in [0.8, 0.95]$. We note that N_1 and N_2 do not depict differences between the strains but are parameters as defined for model 3.1.

3.9 MEASURING EFFECTIVE COMPETITION AND CO-EXISTENCE

Hosts represent an arena in which pathogens compete for resources and transmission opportunities, with major implications for the evolution of virulence and the structure of populations. There are several studies that focus on virulence evolution and the consequence of whether the different pathogen genotypes remain together within the host, i.e. co-infection, or whether one competitively excludes the other, i.e. super-infection or competitive exclusion [73, 19].

Two strains are said to coexist in a host when they are simultaneously maintained in a population. This means each strain occupies a tangible proportion

of the total population i.e. greater than some percentage value. For model 3.2, since total elimination (clearance) is possible, we assume that the two strains coexist if the total bacterial population is greater than the clearance threshold and both strains occupy at least 5% of the total population. When this condition is not met, one strain is assumed to have out competed the other.

When two strains are competing, an increase or decrease in the population size of one would affect the population size of the other positively or negatively. In order to quantify the level of competition between two strains, direct and indirect effects of the impact of increase or decrease in one strain on the other should be measured. This can be challenging because there may be several ways to decrease the population of a strain, and each of these may have different effects. For example a decrease in B_1 would mean more available space for B_2 to grow since they are both limited by the carrying capacity, N_{max} . One way to quantify competition is by the approach proposed by [32], which applies Euler's method of solving a system of ODEs. For model 3.2, Euler's method can generate a series of values :

$$\begin{aligned}
\frac{dB_1}{dt}|_{t=t_n} &= \frac{dB_1}{dt}|_{t=t_{n-1}} + \frac{dB_1}{dt}\Delta t \\
\frac{dM_1}{dt}|_{t=t_n} &= \frac{dM_1}{dt}|_{t=t_{n-1}} + \frac{dM_1}{dt}\Delta t \\
\frac{dB_2}{dt}|_{t=t_n} &= \frac{dB_2}{dt}|_{t=t_{n-1}} + \frac{dB_2}{dt}\Delta t \\
\frac{dM_2}{dt}|_{t=t_n} &= \frac{dM_2}{dt}|_{t=t_{n-1}} + \frac{dM_2}{dt}\Delta t \\
\frac{dT}{dt}|_{t=t_n} &= \frac{dT}{dt}|_{t=t_{n-1}} + \frac{dT_1}{dt}\Delta t
\end{aligned} \tag{3.3}$$

where Δt is a suitably small step size in the time domain, and the derivatives are obtained from equation 3.2.

To obtain a measure of competition in the model 3.2, it is numerically solved, obtaining values for each of the five model compartments through time. We perturbed (i.e decreased) B_1 and computed the predicted value of B_2 using Euler's method as described in equation 3.3. This was done separately at each time step. We averaged the fractional change in B_2 due to a decrease in B_1 over the trajectory. This method captures the immediate indirect effects

(i.e. two time steps) of a change in B_1 on B_2 .

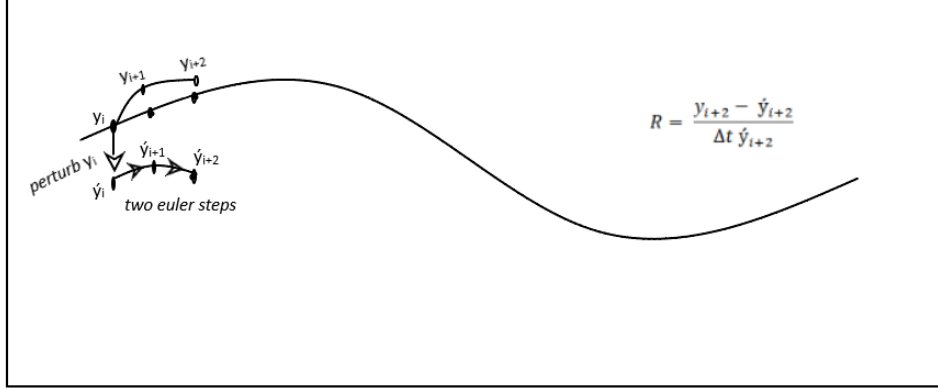


Figure 3.11: *Illustration of Euler's method to quantify competition. y_i is reduced to obtain $\hat{y}_i$. Then Euler's method is used to obtain the predicted values y_{i+1} , y_{i+2} , $\hat{y}_{i+1}$ and $\hat{y}_{i+2}$. Then the values are compared using the formula. This is repeated for all time steps. $R > 0$ means that the perturbation does not favor strain 2 i.e. there is less of strain 2, when there is less of strain 1. On other hand when $R < 0$, it implies that there is competition between the strains. The more negative R is, the greater the competition.*

3.10 ANALYSIS OF MODEL 3.2

As in the analysis for the single strain model, we drew 20,000 parameter sets from the parameter ranges in Table 3.1. The values of p_i are drawn uniformly from the ranges specified in section 3.8. The two-strain model is also able to reproduce expected disease outcomes Figure 3.12. The parameters are separated into the three disease outcome classes using the final bacterial load as in model 3.1. The histograms of these groups of parameters show very similar distributions as observed for the one strain model.

Since the bacterial numbers are very low or cleared out for the clearance class, we do not consider the interaction of the strains in this disease class. Figure 3.14 shows a plot of the final bacterial load of both strains coloured by the disease outcome. The disease outcome is rarely ever active disease whenever one strain is cleared. This may be because there is reduced competition during active disease as the immune system is unable to contain the disease. Using the parameter groups, we measure the effective competition using the

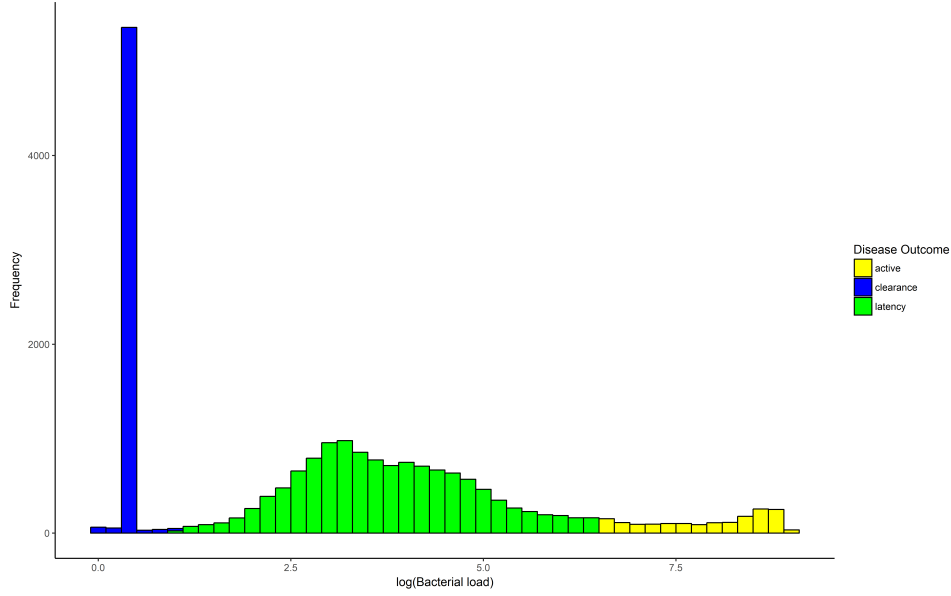


Figure 3.12: *Histogram of final bacterial load for model 3.2 simulations*

method described in section 3.9 for each of the classes. Figure 3.15 shows the box plots of the effective competition measure for active and latency disease outcomes. The values in the latency class are all strictly less than or equal to zero. This suggests that there is more competition in this class compared to the active class which has positive values of R . With less competition between the strains, coexistence is quite generic in the active disease outcome. This may be because of the pressure on the immune system to keep the bacterial numbers low is higher under latency.

For the two strain model, we focus on determining parameters of the model that determine the outcome of the interaction between the strains, which could be competitive exclusion or coexistence. We assume that both strains coexist in the host if the final population of each strain is greater than 5% of the total bacterial load, otherwise we conclude that one strain is out-competed by the other. Using the DAPC approach like in model 3.1, we identify parameters that separate these groups. Figure 3.16 shows the densities of model simulations plotted on the discriminant function with different colours for the two groups. The two groups overlap probably due to the way we define co-existence between the strains. However, the simulations that belong to the coexistence group are more to the left of the discriminant

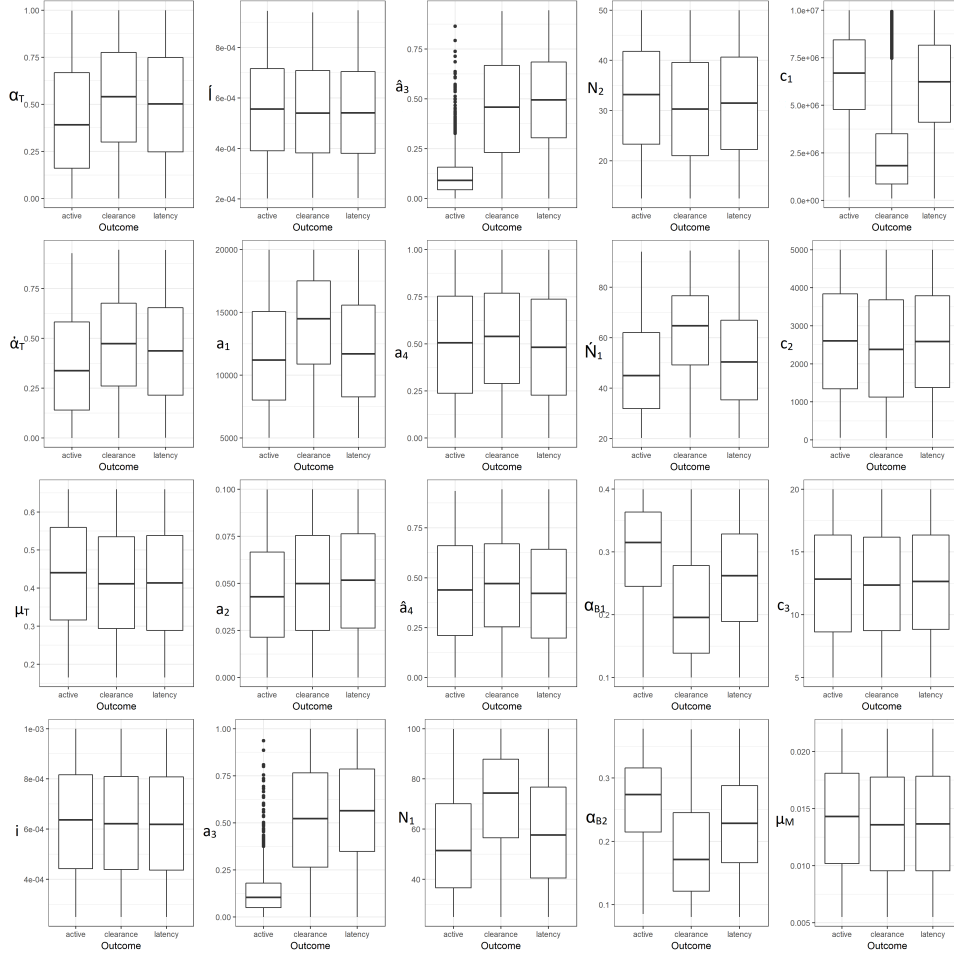


Figure 3.13: *Boxplots of the parameters of Model 3.2 grouped according to the disease outcomes.*

function while others are more too the right of the discriminant function. The parameters that contribute most to the discriminant function are the growth rate and action of T-cells on both strains. When $\hat{a}_3$ and a_3 are high, the host will be able to clear both strains, on the other hand when α_{B_1} and α_{B_2} are high, then both strains grow. The outcome of the interaction between the strains would thus depend on the trade-offs between these parameters.

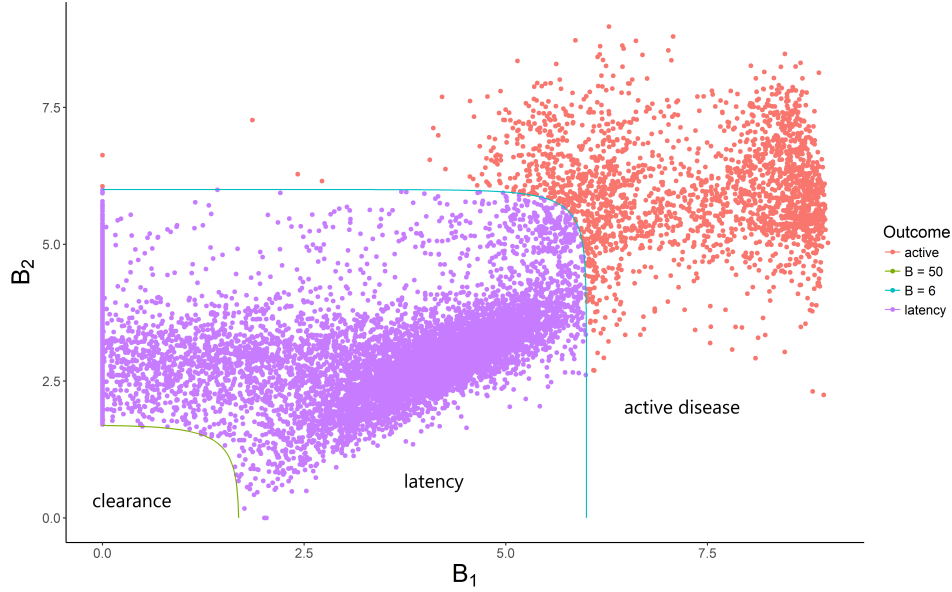


Figure 3.14: Plot of final the log of population size of both strains coloured by the disease outcome class. The lines show the separation of the region into the different disease classes.

3.11 VIRTUAL EXPERIMENTS

The power of the models developed here is that it can be manipulated in a variety of ways to ask questions about different scenarios and disease outcomes. This would give room to explore experimental outcomes that may be otherwise difficult to analyse clinically. At the most fundamental level, latent tuberculosis can be viewed as an equilibrium between host and bacteria [64]. The host response prevents active disease from occurring, and *M. tuberculosis* cells avoid elimination. However, occasionally the immune response fails in some way and reactivation occurs which leads to active disease. There have been numerous studies that investigate the reasons for reactivation [23, 4]. Here we consider the introduction of a second strain into a host with latent tuberculosis. This can be seen as mimicking exogenous re-infection which is speculated to play an important role in recurrent TB in a population [71]. With this approach, we would be able to have insight to a number of important questions. Does strain replacement or co-infection occur? Does the presence of a second strain trigger the disease outcome to a

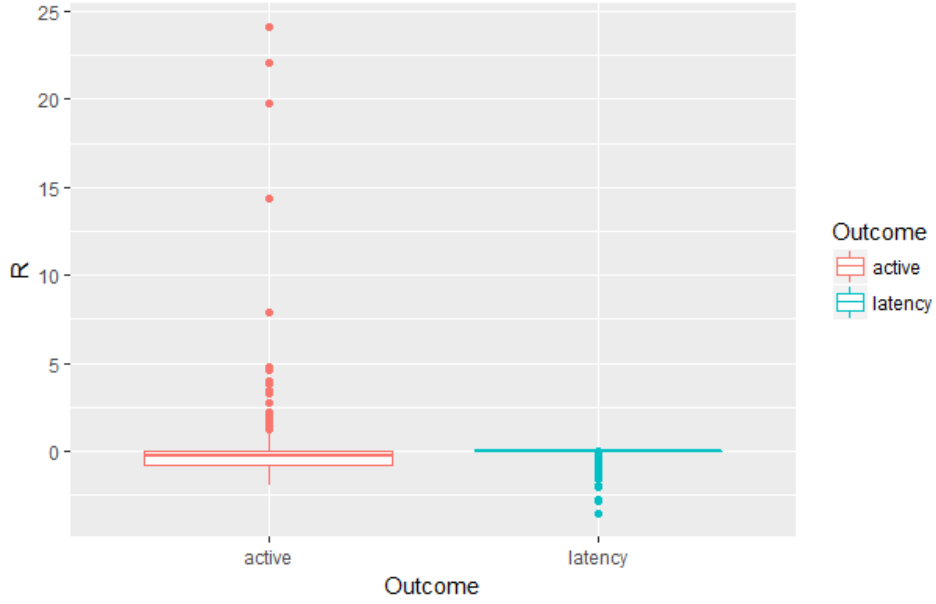


Figure 3.15: *Boxplots of effective competition values R for latent and active disease outcomes. The more negative the values the stiffer the competition between the strains.*

different class? How does competition between the two strains affect disease outcome?

In order to distinguish the strains (i.e. the strain that caused the initial infection and the one introduced during re-infection), we consider the splitting the parameters of the one-strain model into those that are ‘intrinsic’ to a host and those that are ‘intrinsic’ to a bacterial strain. The parameters N_1, N_2, α_B and c_1 are considered as ‘strain’ parameters while the others are ‘host’ parameters.

In our virtual experiments, we start the simulation with only one strain and at some time point when latency is achieved, a second strain is introduced. We consider two scenarios (in terms of the immune response) for the virtual experiments:

- A scenario where both strains elicit different immunological host responses i.e. the parameters that govern the immune response to infection (innate: i and adaptive: a_3, a_4) are different for the two strains.

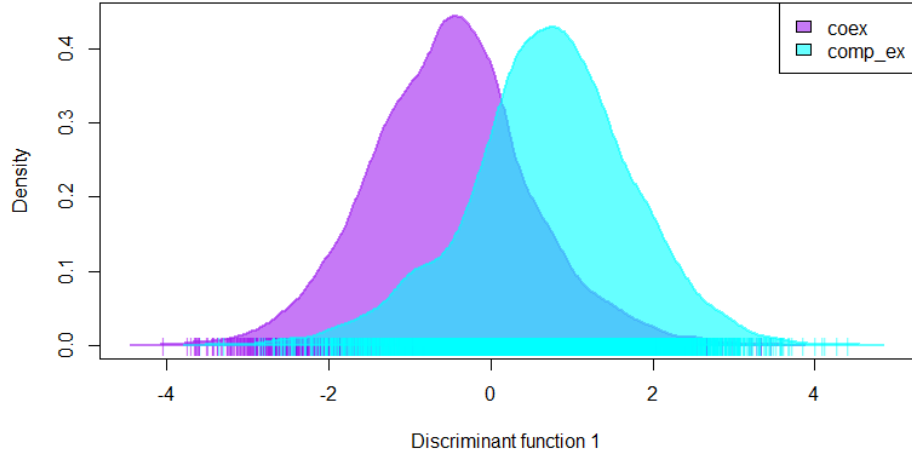


Figure 3.16: *Densities of simulations of model 3.2 plotted on the discriminant function. 'coex' means coexistence between the strains, while 'comp_ex' means one strain out-competes the other.*

As such when a new strain is introduced, seven parameters are added to the one-strain model. The resulting system is the two-strain model (Equation 3.2).

- In the second scenario, both strains elicit similar immunological host responses i.e. there is no coupling between the immune response and the strain type. When the second strain is introduced, only four parameters are added to the one-strain model. The resulting system of equations is given in equation 3.4

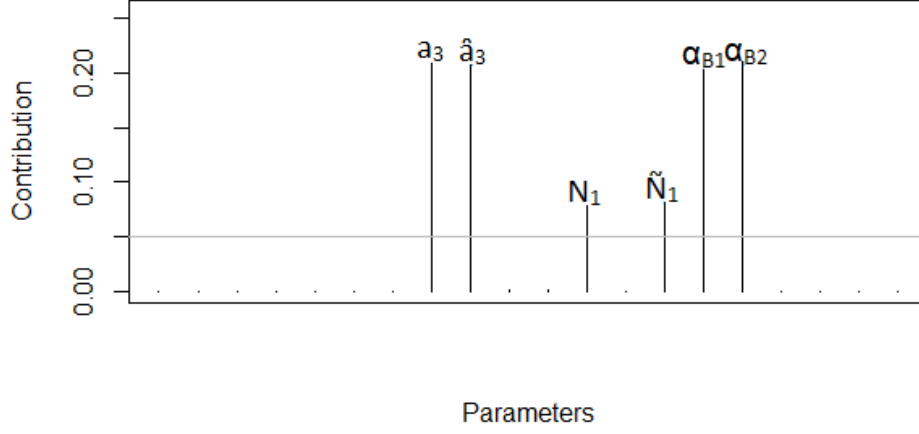


Figure 3.17: Contributions of parameters to the discriminant function of the DAPC of model 3.2. The height of each bar is proportional to the contribution of the corresponding parameter to the discriminant function. Only parameters whose contribution is above an arbitrary threshold (grey horizontal line) are indicated for the sake of clarity.

$$\begin{aligned}
\frac{dB_1}{dt} &= \alpha_{B_1} B_1 \left(1 - \frac{B}{N_{max}}\right) - iB_1 - N_1(a_1 + a_2 T) \frac{B_1}{B + c_1} - a_3 B_1 \frac{T}{T + c_2} + N_2 \mu_{M_1} M_1 \\
\frac{dM_1}{dt} &= (a_1 + a_2 T) \frac{B}{B_1 + c_2} - a_4 M_1 \frac{T}{T + c_3 M} - \mu_{M_1} M_1 \\
\frac{dT}{dt} &= \alpha_T (M_1 + M_2) - \mu_T T \\
\frac{dB_2}{dt} &= \alpha_{B_2} B_2 \left(1 - \frac{B}{N_{max}}\right) - iB_2 - \hat{N}_1(a_1 + a_2 T) \frac{B_2}{B + \hat{c}_1} - a_3 B_2 \frac{T}{T + c_2} + \hat{N}_2 \mu_{M_2} M_2 \\
\frac{dM_2}{dt} &= (a_1 + a_2 T) \frac{B_2}{B_2 + \hat{c}_1} - a_4 M_2 \frac{T}{T + c_3 M} - \mu_{M_2} M_2
\end{aligned} \tag{3.4}$$

Analysis of 3.1 separates the parameters into the three disease outcome classes. As such when the second strain is introduced, the new parameters (7 in scenario 1 and 4 in scenario 2) can be drawn from any one of the three disease outcome classes. Drawing strain parameter values from the three classes separates the strains into different ‘fitness’ levels. For example,

a strain that is drawn from the clearance class would generally be cleared in a host and not be as fit as a strain with parameters drawn from the other two classes. In summary, the virtual experiments involve simulating the one-strain model (equation 3.1) with parameter values that lead to latency, and introducing a second strain at equilibrium when latency is achieved. When the second strain is introduced, the system of equations becomes equation 3.2 for scenario 1 or equation 3.4 for scenario 2.

At the start of the experiment we have the strain 1 in the host i.e the set $\{h, B_1\}$. When the second strain is introduced, we have both strains in the host i.e. the set $\{h, B_1, B_2\}$. The outcome of disease can be determined for the set $\{h, B_1, B_2\}$ using total bacterial load of both strains. We track the bacterial load of each strain (i.e. B_1 and B_2) in the host to determine if they fall into any of the three outcome classes. We also consider a situation where the host is infected with only the second strain i.e. the set $\{h, B_2\}$. This would help in quantifying the effect that having a previous infection has on being infected with the second strain.

3.11.1 SCENARIO ONE

To explore the results of the virtual experiments under this scenario, 500 different sets of $\{h, B_1\}$ that lead to latency were used to simulate the one-strain model. Also from each disease outcome class, we draw 500 different sets of the 7 (4 strain parameters and 3 host parameters) parameters that are introduced in this scenario. Altogether we obtain 500 different sets of $\{h, B_1, B_2\}$. Figure 3.18 shows the proportion of the simulations that fall into the different disease outcome classes.

In this scenario, the immune response of a host is different for the two strains, as such the introduction of a new strain could easily shift the disease outcome to active disease, even when the new strain is not as 'fit' as the previous strain the host was infected with. For example when B_2 is drawn from the clearance class, the host progresses to active disease in 14% of the simulations (Figure 3.18). In all three cases, the host doesn't clear the disease as expected because an infection cannot be cleared by re-infection. When B_2 is drawn from the latent class, the host clears B_2 in more simulations when previously infected with B_1 than when infected with only B_2 (Figure 3.18). This shows

that the presence of B_1 in the host causes the growth of B_2 to be suppressed. This may be as a result of competition between the two strains, as both strains have similar fitness and competition between them is expected to be relatively strong.

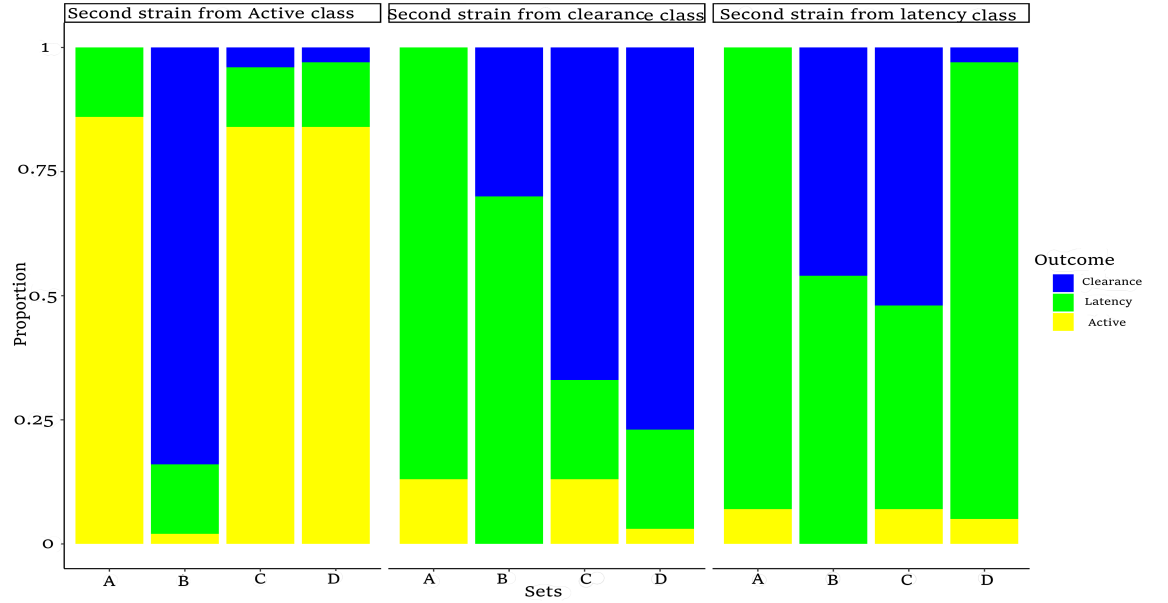


Figure 3.18: Proportions of simulations in scenario 1 that fall into the different outcome classes for the sets: A: the host infected with both strains $\{h, B_1, B_2\}$, B: the bacterial load of strain 1 after re-infection $\{B_1\}$, C: the bacterial load of strain 2 after re-infection $\{B_2\}$, D: the host infected with only strain 2 $\{h, B_2\}$.

3.11.2 SCENARIO TWO

Like in the first scenario, 500 different sets of $\{h, B_1\}$ that lead to latency were used to simulate the one-strain model. Also from each disease outcome class, we draw 500 different sets of the 4 (4 strain parameters) parameters that are introduced in this scenario. Altogether we obtain 500 different sets of $\{h, B_1, B_2\}$. Figure 3.19 shows the proportion of the simulations that fall into the different disease outcome classes.

In this scenario, since both strains elicit similar immunological host responses, when a strain from the clearance class is introduced into the latent

host, the second strain would be cleared more often than not. Figure 3.19 shows that when B_2 is from the clearance class, the host never progresses to active disease but continues in latency.

When B_2 is drawn from the active class, unlike scenario one, the host maintains latency in most of the simulations (Figures 3.19). Also when B_2 is drawn from all three classes, the presence of B_1 in the host suppresses the growth of B_2 , with the dually infected host being able to clear B_2 in more simulations than when infected with only B_2 .

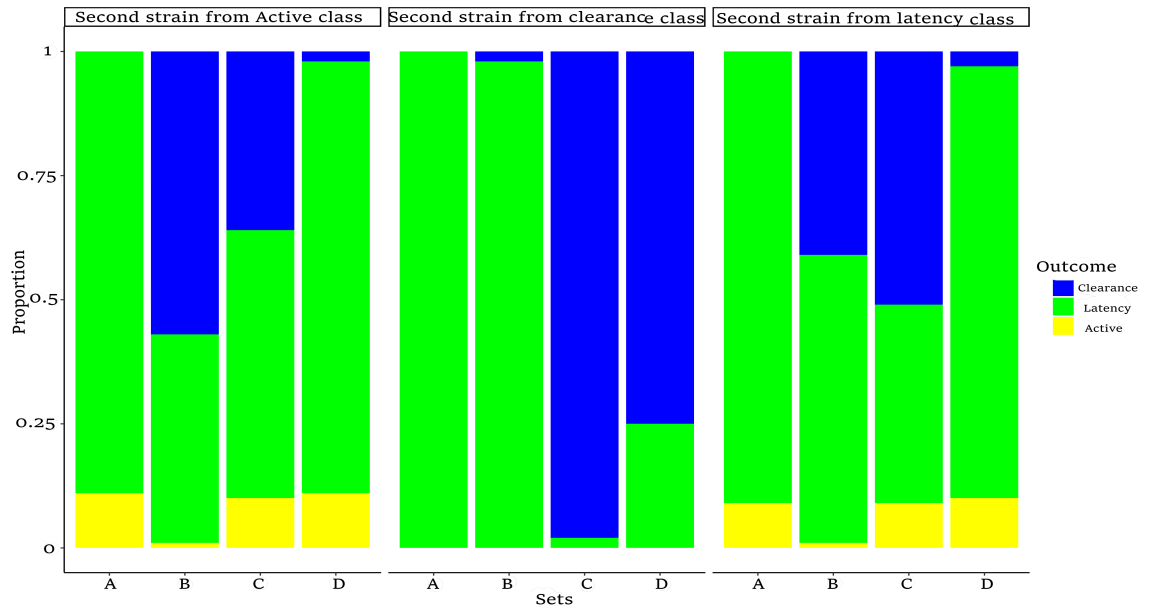


Figure 3.19: Proportions of simulations in scenario 2 that fall into the different outcome classes for the sets: A: the host infected with both strains $\{h, B_1, B_2\}$, B: the bacterial load of strain 1 after re-infection $\{B_1\}$, C: the bacterial load of strain 2 after re-infection $\{B_2\}$, D: the host infected with only strain 2 $\{h, B_2\}$.

3.12 CONCLUSION

The outcome of tuberculosis infection depends on complex host-pathogen interactions that occur across numerous scales. The models studied in this chapter are able to support the three disease states and the outcome of infection depends on the choice of parameter values. Traditional bifurcation

analysis would involve varying a parameter and studying how the disease outcome changes. However in cases where there are several parameters this becomes infeasible analytically [108, 109]. As such instead of adopting this approach, we use tools from data science to identify key parameters that discriminate between the three disease outcomes. The identified parameters are: the level of bacteria at which the infection of macrophages is half maximal (c_1), the rate at which T cells kill bacteria (a_3), the amount of bacteria engulfed by macrophages (N_1), the growth rate of the bacterial cell (α_B), and the constant source of macrophages (a_1). Infection can be controlled when the strength of the adaptive immune system (a_3) is sufficiently high. On the other hand, when the growth rate of the bacterial cells is sufficiently high, the bacteria are able to escape killing by the immune system and thus lead to active disease.

When a host is infected with more than one strain of tuberculosis, the interaction between the strains in the host can affect the outcome of the disease. Analysis of the two strain model (model 3.2) revealed that competition between strains in a host is stronger during latency. When competition between strains is strong, it affects their growth and invariably the immune system is able to contain the disease which leads to latency. On the other hand when there is less competition, the strains are able to co-exist and thrive well enough to progress to active disease. This is in agreement with the chemostat model in chapter two, where more diversity is maintained under fast dilution (which is usually assumed to depict the active phase of the disease).

Mathematical models have been used to perform virtual experiments in a variety of studies [145, 104]. The results of the virtual simulation of re-infection in section 3.11 show that when a host with latent tuberculosis is reinfected with a new strain, the outcome of the disease depends on the interaction between the strains, and on whether both strains elicit similar immunological host responses or not. When the immune response and strain type are not coupled and a host gets re-infected, the probability of maintaining latency is higher than when the immune response is different for the two strains.

The models analysed in this chapter are conceptual in nature and are only meant to capture certain key aspects of host-pathogen interactions. One possible direction for future improvement is to include the introduction of

anti-tuberculosis chemotherapy. Individuals with active tuberculosis have a very high mortality rate without medical intervention . It would be interesting to investigate the processes involved in drug uptake and see how they interact with the models discussed in this chapter. Also the effect of varying the immune response strength on the emergence of resistant bacteria can be studied.

Parameter	Description	Range	Reference
α_T	Rate of growth of T-cells	$10^{-4} - 1$	[168]
μ_T	Death rate of T-cells	$0.165 - 0.66$	[187]
i	Killing rate of innate immune response	$0.00025 - 10^{-3}$	[126]
a_1	Constant supply of macrophages	$5000 - 20000$	[126]
a_2	Recruitment rate of macrophages by T-cells	$10^{-5} - 0.1$	[126]
a_3	Maximum rate of killing of bacteria by T-cells	$10^{-4} - 1$	[126]
a_4	Maximum rate of killing of macrophages by T-cells	$10^{-4} - 1$	[187]
N_1	Amount of bacteria engulfed by macrophages	$25 - 100$	[201, 126]
N_2	Amount of bacteria released by bursting of infected macrophages	$12.5 - 50$	[187]
α_B	Maximum growth rate of bacteria	$0.1 - 0.4$	[126, 187]
c_1	Concentration bacteria at half maximal rate of infection of macrophages	$10^4 - 10^6$	[126]
c_2	Concentration of T-cells at half its maximum killing rate	$50 - 50000$	[126]
c_3	Ratio of $T : M$	$5 - 20$	[126]
N_{max}	Maximum carrying capacity of bacteria	10^9	[126]
μ_M	Death rate of macrophages	$0.0055 - 0.022$	[126]

Table 3.1: *Table of parameter definitions and ranges. Parameter ranges are specified with a – separating upper and lower ranges, and are obtained from [126]. Where a single value is given, the parameter was fixed.*

*Catch for us the little foxes, the little foxes that spoil
the vine.*

King Solomon

4

Detecting disruptive sites in tuberculosis genome

4.1 INTRODUCTION

Evolutionary tools are increasingly being used to study short-term variation in infectious pathogens, with the aim of improving our ability to understand pathogen adaptation and control infections. Determining the relationships among *M. tuberculosis* isolates is essential to understanding its ecology, evolutionary history, and transmission patterns. However, due to its low diversity, phylogenetic reconstructions have been challenging. As such it is necessary to carefully select genetic data which can be exploited for phylogenetic analysis [67]. Phylogenetic trees are central to our understanding of evolution, and are probably the most widely used tool to interpret data capturing variation in populations.

Not all regions of a genome evolve at precisely the same rate. Some parts of a multiple sequence alignment may be well conserved and therefore suitable for phylogenetic analysis, whereas others are divergent, fast evolving, hyper-variable or prone to homoplasies ([189], [24]). Because such regions do not fit the models used in phylogenetic inference, they may produce low likelihoods and adversely affect phylogenetic reconstruction. Identifying such regions is

therefore important for any phylogenetic analysis.

In this chapter, we describe some of the approaches used to filter bad regions from a multiple sequence alignment in particular the one proposed by Bar-Hein et al [11]. We also present a phylogeny-based method to detect disruptive sites. We test the performance of this method via simulations and on a tuberculosis dataset.

4.2 INFERENCE OF PHYLOGENETIC TREES

A phylogenetic tree is a graphical structure that describes the relationships between individuals or entities in a set, in a way that reflects their evolutionary histories. A phylogenetic tree is made up of branches, internal nodes and leaves. The leaves represent the extant species/individuals, the branches represent the amount of change between two nodes and descent from a common ancestor into distinct lineages, and the internal nodes represent the individuals/species ancestral to the leaves. Phylogenetic trees have been used to describe the histories of populations [54], and the evolutionary dynamics of pathogens ([76], [127]).

A simple example to aid understanding the idea of phylogenetic trees would be to consider a simple asexual organism as given in [140]. Each cell divides, yielding two cells with identical genomes except for those few genomic locations where a mutation occurred. Suppose this continues for say n generations after which a sample of size m of their genetic data is collected and compared to each other. Two cells that share a common ancestor would have accumulated fewer differences than a pair that diverged earlier, whose last common ancestor was many generations ago.

Phylogenetic trees are not observed directly but are rather inferred from sequence data or other kinds of data. In the past, phylogenetic trees were mainly constructed using morphological character traits [112]. However, with advances in sequencing technologies, there is an exponentially increasing volume of DNA and protein sequence data [200]. The first step in phylogenetics is to obtain sequences of the group of individuals to be studied. The next step is to align the sequences, usually to a reference genome, in order to generate a data matrix. Usually an alignment of N sequences is represented as an $n \times m$ matrix where the rows represent the individuals and the columns

are the sites (characters) of interest.

Theoretically, reliable or informative sites can be considered as those that have undergone only a few changes over time [45]. Multiple changes can lead to the occurrence of homoplasy which masks the genuine phylogenetic signal. Homoplasy occurs when there are identical site states (for example, the same nucleotide base in a DNA sequence) in two isolates, that are not the result of common ancestry (not homologous), but that arose independently in different ancestors by convergent mutations [45]. Homoplasy can also be seen as independently occurring mutations along the branches of a phylogenetic tree [114]. Since a phylogenetic tree is a representation of evolutionary history, it is needful to use sites that homologous in the inference procedure.

After a multiple sequence alignment is obtained, the next step is to pick an appropriate tree inference method. There are broadly two methods used to infer phylogenetic trees: distance based methods and character based methods. In distance based methods, the distance between every pair of sequences is calculated using a model of nucleotide substitution and this distance matrix is used for tree inference. An example of this is the neighbour joining method of tree inference which uses a clustering algorithm on the data matrix to infer the tree. On the other hand, character based methods involve a simultaneous comparison of all sequences in alignment, taking one site at a time to calculate a score for the tree [192]. Examples of this method are maximum parsimony (where the score of a tree is the sum of the minimum number of changes each site would undergo on the tree), maximum likelihood (where the score of a tree is log-likelihood value) and Bayesian inference methods (the score of a tree is its posterior probability). Crucially, each method has its own good and bad points, and the choice of one method over another for any particular analysis tends to depend primarily on a compromise between the desired complexity or accuracy of the analysis and the amount of time it will take to run. For the purpose of this chapter, all phylogenetic trees will be inferred using a maximum likelihood approach using using RAxML 8.2.4 [172].

sequence1	ACTCCCGGGA
sequence2	CCCCTAGGAT
sequence3	GCCTTTAGAG
sequence4	CCTTAGTGTT

Figure 4.1: *An example of multiple alignment of four sequences.*

4.3 *Mycobacterium tuberculosis* GENOME

The *M. tuberculosis* genome consists of about ~ 4000 genes, some of which play a distinct role in its pathogenesis [68]. Approximately a tenth of this is made up of two families of genes which constitute a highly repetitive class of antigenic genes and as such are poorly characterized ([150] [37]). These gene families (PE and PPE) were discovered over a decade ago but are not yet fully understood in terms of their function although they are stipulated to be important virulence factors involved with host-pathogen interactions ([62],[136]). It is usually difficult to characterize them using traditional mapping approaches, leading to their systematic exclusion from phylogenetic analysis.

Regions of a genome that are considered to be phylogenetically uninformative can be removed before tree inference [151]. Swofford et al [174] suggests that regions of a sequence alignment with substantial numbers of gaps, in which positional homology is too uncertain, should be omitted from the analysis. Although several authors ([157],[77],[121]) argue that such omissions are beneficial, others ([2], [111]) suggest that this could lead to loss of useful information. The benefit of removing disruptive sites must thus depend on the data being analysed. Removing problematic regions such as the PE and PPE genes from a multiple sequence alignment is often done in TB research in order to reconstruct high quality phylogenies. Pepperell et al [149] removed the PE/PPE gene regions and other regions prone to homoplasies before reconstructing their phylogeny. Merker et al [133] removed genes that are associated with drug resistance as these are likely to be homoplastic due

to selection, Bos et al [20] removed regions with missing data, Whitney et al [132] did not explicitly remove problematic regions but removed sites that did not meet certain mapping quality criterion and O’Niel et al [143] removed PE/PPE gene regions before their analysis. On the other hand, some authors ([62], [130]) have based their phylogenetic analysis on some of these problematic regions as they have been implicated in virulence and studying them may provide better understanding of their possible role in virulence [150].

Despite the fact that several studies remove these problematic regions, there is no consensus in literature as to what to exclude from phylogenetic analysis of *M. tuberculosis*. Also since the quality of tree reconstruction decreases with decreasing sequence length, it is important not to remove too many sites from an alignment if they may be informative. It remains unclear what sites or regions of a multiple sequence alignment of *M. tuberculosis* have an adverse affect on phylogenetic reconstruction, and it remains unclear under what circumstances which kinds of sites should be excluded.

4.4 CURRENT METHODS

There are a number of methods available in literature to filter a multiple sequence alignment in order to remove phylogenetically uninformative sites. These include Noisy [189] which detects homoplastic and noisy sites using the concept of circular orderings, Gblocks [176] which classifies each site as nonconserved, conserved, or highly conserved depending on the number of identical residues in the site, and trimAl [21], which computes scores for each site and removes sites beyond a certain threshold, amongst others. Ge et al [177] in their review of these methods conclude that that current alignment filtering methods do not generally lead to better trees because they removes sites too aggressively.

Bar-hen et al [11] adapt the concept of influence function to filter alignments in the context of maximum likelihood reconstructions. They calculate the influence of each site on the likelihood of the inferred tree and as such study sites one at a time. They do this by comparing the log likelihood of the tree T (that maximizes the likelihood for the entire data set) and the log likelihood of each tree T^h (that maximizes the likelihood of the data without

site h) with formula

$$IV_h = (m - 1)(L(T) - L(T^h)),$$

where m is the number of sites in the data set, $L(T)$ and $L(T^h)$ are the log likelihoods of tree T and T^h respectively.

A positive influence value shows support for topology T and vice versa. They assume that only a few sites would disrupt the robustness of an inferred topology and as such focus on and remove sites with very large negative influence values. The assumption that only a few sites, when removed, will result in reconstruction of a less robust or different phylogeny is appropriate where the taxa have plentiful informative diversity, but in many current sequencing studies on short-term evolution in tuberculosis, it may not be correct.

4.5 THE METHOD

The approach summarized here can be used to identify sites that cause a change in tree shape and possibly have biological importance. Also it may be used to identify sites that should be removed prior to phylogenetic inference. The steps are summarized in Figure 4.2. Using the entire alignment, a tree is constructed; we call this the ‘reference tree’. Other trees are built from the alignment with a single site removed at a time, these trees are called ‘jackknifed trees’. These trees are compared to the reference tree using a phylogenetic tree metric. For an initial screening and to minimize computational time, in step 2, a sliding window containing k sites is removed from the alignment at a time and the corresponding phylogenetic trees are constructed. For sliding windows whose trees are not identical to the reference tree, each site within the window is dropped and a tree constructed. This is done because when a window of sites is removed from an alignment and the inferred tree is identical to the reference tree (i.e. distance 0), removing the individual sites within this window usually gives a tree that is identical to the reference tree. Using the above steps, there is a distance d_i and tree t_i associated with every site i and/or the i^{th} window of sites.

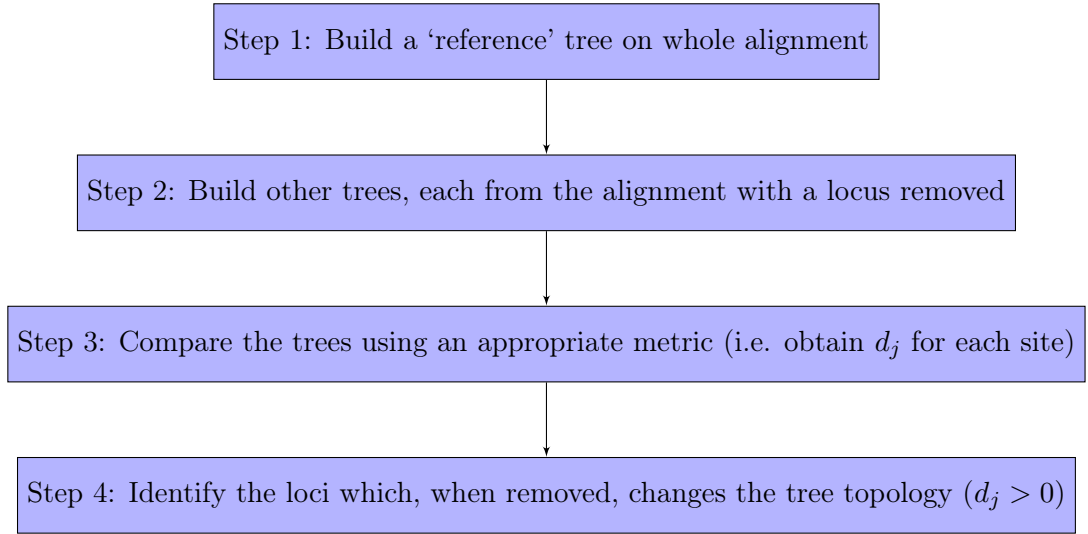


Figure 4.2: *Flow chart of the steps in the approach*

4.6 COMPARING PHYLOGENETIC TREES

A fundamental challenge in the study of phylogenetic trees is that different trees can be inferred from different combinations of input data and inference method [158]. Also when a Bayesian approach is used to infer phylogenetic trees, a posterior distribution of trees is produced. When trees are relatively few and consists of a few number of tips, direct qualitative comparison of trees may be possible but this becomes uninformative when the trees are large or differ considerably. A variety of tree comparison methods exist in literature. For the purpose of this chapter, the metric proposed by [101] will be used to compare and calculate the distance between trees. The metric calculates the distance between two trees by comparing tree topology (number of edges on the path between the root and the Most Recent Ancestor (MRCA) of every pair of tips) and/or branch lengths (the length of the path between the MRCA of every pair of tips). It has a parameter λ which allows the choice of what extent the branch lengths of a tree, vs its topology alone, contribute to the tree distance. For this analysis we use $\lambda = 0$ and therefore consider tree topology only. The metric is implemented using the function `refTreeDist` from R package `treespace` [97].

There are various techniques available to use distances between a set of objects (in this case, trees) to visualize, group, and compare them. One

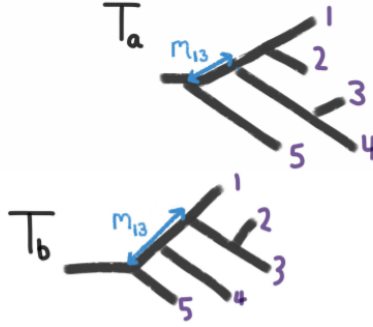


Figure 4.3: *Figure illustrating how the Kendall-Colijn metric works. The distance between tip 1 and 3 in T_a is 1, but in T_b , its 2.*

method is multidimensional scaling (MDS) which projects distances in such a way as to best capture them with just a few (e.g., two or three) dimensions. We visualized the distances between the trees using MDS (`dudi.pco` in the `ade4` package in [25]).

4.7 SIMULATION

We apply our approach to simulated data where the correct phylogenetic tree is known. We simulate sequences evolving down a tree, which we call the ‘true tree’. If we are able to re-construct a tree that is identical to the true tree, from the simulated sequences then, we can introduce disruptive sites/regions (i.e. homoplastic sites and hypervariable sites) and see how they affect the tree topology, then apply our method to identify these sites/regions.

Sequences were simulated using Seq-Gen [152]. Seq-Gen simulates the evolution of nucleotide or amino acid sequences along a phylogeny, using common models of the substitution process. We simulated a set of sequences evolving down a random tree on 20 tips, whose topology is shown in Figure 4.4. We considered datasets of varying sequence length (2000, 3000 and 4000 sites). However, with less than 4000 sites, it was not possible to reconstruct the true topology (i.e. obtain a tree of distance zero to the true tree).

Introducing disruptive sites: We introduce disruptive sites in the dataset

with 4000 sites. These sites are either hyper-variable or homoplastic. To introduce a hyper-variable site, we randomly select a site, copy the data in this site, randomly shuffle the rows, then append this shuffled site to the original dataset. In order to introduce homoplastic sites into the data, we pick a random edge of the tree inferred from the data and a random mutation is introduced at one of the nodes. To do this, the ancestral states of the given phylogeny and the simulated data are required. Ancestral state reconstruction is a method of inferring the sequences of the internal nodes of a given phylogeny. The ancestral state of the internal nodes are obtained using the function *ancestral.pml* in the phangorn R package. We added different number of homoplastic and hypervariable sites to the simulated data: 100,1000,2000 noisy sites. The tree reconstructed using the data with the noisy sites is taken as the reference tree.

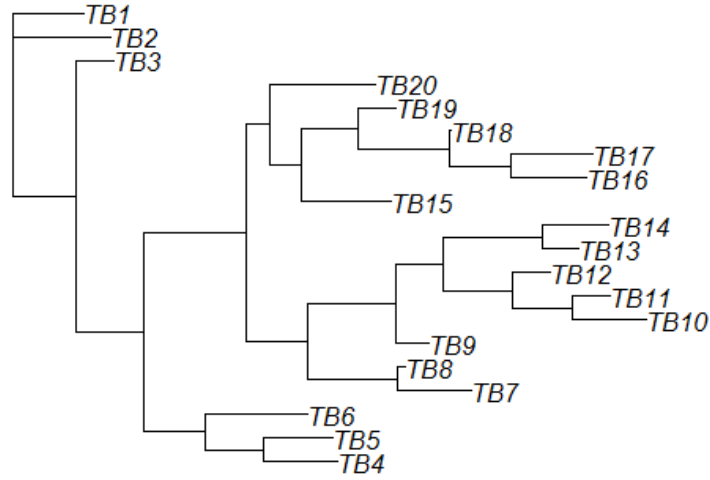


Figure 4.4: :The random tree used in the simulation

4.8 TUBERCULOSIS DATA

Whole genome sequences of 471 tuberculosis isolates each obtained from a different patient most of whom had multidrug resistant tuberculosis in the suburbs of Callao and South Lima between the periods of 2009-2013. A total of 20976 variable sites were identified in the multiple sequence alignment as described in [75]. This data was used to reconstruct the reference tree.

4.9 RESULTS

4.9.1 SIMULATED DATA

For each noisy data set, we reconstruct the reference tree and compare them to the true tree. Only the dataset with 1100 noisy sites (5100 sites altogether) gave a different tree from the true tree. Figure 4.5 shows the tree reconstructed from the noisy data i.e. the reference tree (A) beside the true tree (B). The reference tree is a distance of 3.16 from the true tree. We use our method (sliding window of size $k = 20$) on this dataset (with the 1110 noisy sites). Figure 4.6 shows that nine trees have a distance greater than zero from the reference tree. These trees are also distance zero from the true tree. This suggests that the sites dropped before reconstructing these nine trees are what contribute to the topological differences observed between the reference tree and the true tree. Although when we dropped each site in the nine windows, none of the inferred trees were different from the reference. Also only 20% of the sites in these nine windows are noisy sites. This suggests that the simulated data is so phylogenetically informative that we can add a lot of noise without altering the phylogenetic signal.

4.9.2 TUBERCULOSIS DATA

We found that about 50% of the windows of 10 sites, when removed, produced a tree with a different topology from the reference tree. This is quite different from the simulations where we ensured that we could get back the true tree. Figure 4.7 shows that most of the trees lie in one of three distinct bands which are at distances 0, 11.7 and 27.2 from the reference tree. Further investigation showed that each band consists of trees that are of distance 0

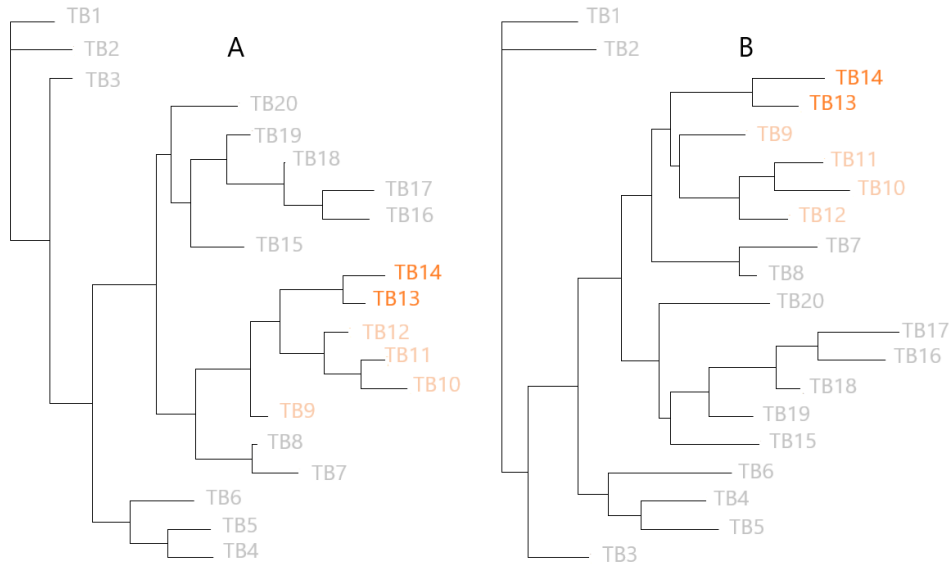


Figure 4.5: *A: The true tree and B: reference tree constructed with simulated data and noisy sites. This shows the topological differences between the two trees. For each tip, the function colours it grey if its ancestors in tree1 has the same partition of tip descendants in tree2, otherwise the tip gets coloured from pale orange to red to reflect the dissimilarity. Trees are plotted using `plotTreeDiff`, a function in the R package `treescap`.*

from each other. We then re-ran the method, removing a single site of the alignment at a time, each belonging to a window which caused a change in the first run. Approximately 10% (1043 sites) of the entire alignment were found to be disruptive. Figure 4.8 shows that most of the trees built on data with a single site dropped also line up in two distinct bands: one at distance 0 and the other at distance 27.2 from the reference tree.

The sites with $d > 0$ belong to 470 genes spread across the genome. Over 50% of these belong to the PE/PPE regions, 27% belong to intergenic regions, 2.8% belong to resistance genes while the remaining belong to repeat regions and other genes across the genome. The first and second most disruptive sites are in genes ‘Rv1578c’ and ‘RV0731c’ respectively. The function of these genes are unknown [115]. Table 4.1 shows the top 15 disruptive sites, the genes they are located in and if these genes are associated with resistance to any antimicrobial drugs.

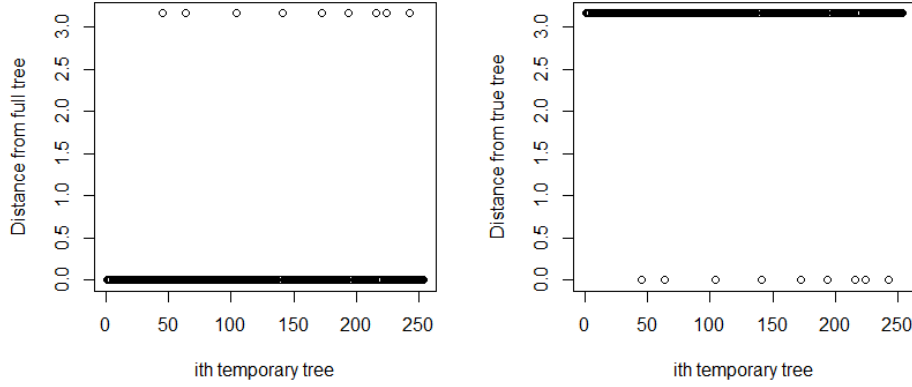


Figure 4.6: *Distance from reference tree and true tree of the simulated data. Nine trees have distance greater than zero from the reference tree. These trees also have distance zero from the true tree.*

We used the influence function proposed by Bar-hen et al [11] and explained in section 4.5 on the tuberculosis dataset to identify outlier sites. In their paper, Bar-hen et al [11] call sites with very negative influence values ‘outliers’ and thus recommend removing these sites from an analysis. Figure 4.9 shows the influence values of each site and the gene names of the top outliers. The strongest outlier in our data is site 4036 which has a distance 138.16 and is in the ‘rpoB’, gene which is associated with resistance to rifampicin [115, 134]. Over 70% of the sites have a positive influence value and as such have support for the reference tree which is reflected in the distances (as most of these sites have distance 0). Bar-Hen et al [11] suggest that there is a strong connection between changes in topology and likelihood of phylogenetic trees. This is similar to the results obtained here, except for the fact that half of the sites that have a negative influence value also have the same topology as the reference tree. This method does indeed identify disruptive sites, however, there may be differences in likelihood scores that are not necessarily associated with differences in tree topology [79] and as such it may be misleading.

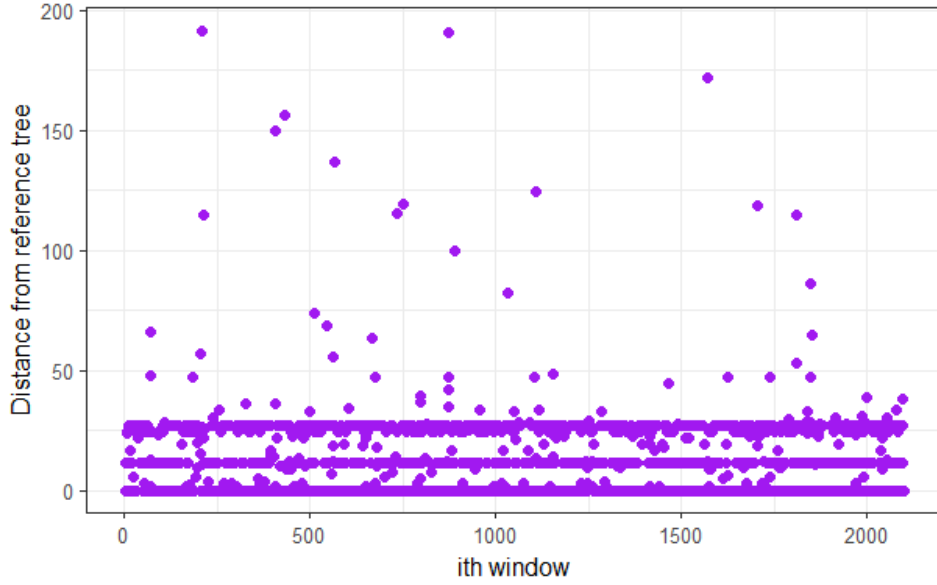


Figure 4.7: *Distances of trees with window of 10 sites removed from the reference tree*

4.10 OBTAINING A MORE ROBUST TREE

In order to obtain a more robust tree, we removed the sites with the greatest distances from the data before tree reconstruction. We hypothesize that after discarding these sites from the analysis, removing or adding a single site will not drastically change the inferred topology. To achieve this, we ranked the sites in order of decreasing distances, deleted the highest m sites with m ranging from 1 : 1043, and inferred the tree. Figure 4.11A shows the distances of these trees from the reference tree. We see that removing the first few sites ($m = 1 : 36$), leads to a steady increase in the distance from the reference tree. There is a bump in the distance after this with the trees having a distance between 700 and 1000 from the reference tree. After removing the top 990 sites, the distance reduces further and remains at this level. Figure 4.10A gives a visualization of a multi-dimensional scaling projection of the distances between these trees. In the plot each point represents a tree, and the distance between any pair of points approximates the distance between these trees. The points are coloured according to the number of disruptive sites removed. We see that removing the sites incrementally

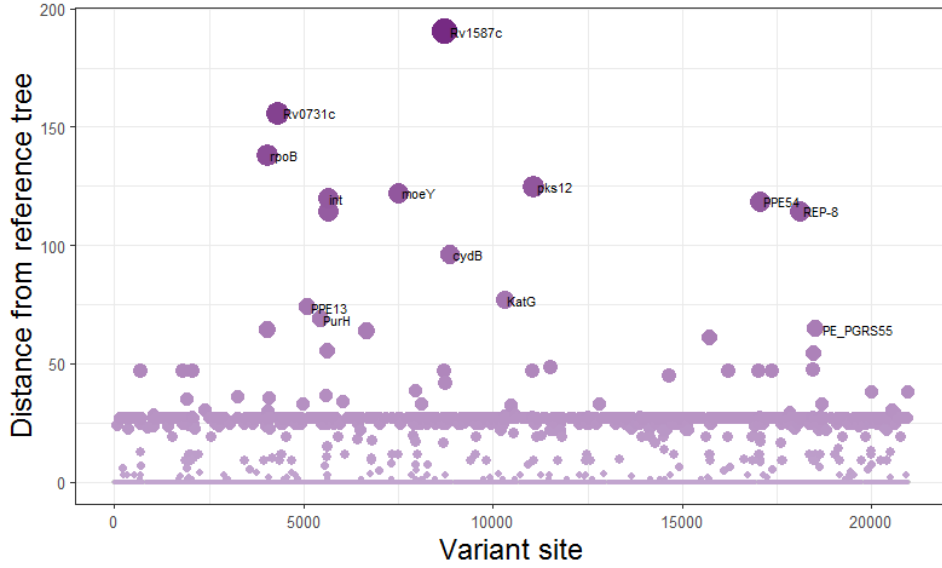


Figure 4.8: *Distance of each site along full alignment, with the genes of the top 15 sites (sites with the top 15 distances) labelled. Each dot represents a site and the size of the dots are mapped to the distance associated with each site.*

moves the inferred tree to different regions in the tree space. Using hierarchical clustering, we identified five distinct groups of trees, most of which are arranged in visually separated clusters (4.10B). We obtain the median tree of each group and carried out a bootstrap analysis on them.

Felsenstein [61] introduced the use of bootstrap in the estimation of phylogenetic trees. His technique, which has been widely used, provides assessments of ‘confidence’ for each clade of an observed tree, based on the proportion of bootstrap trees showing that same clade. It involves sampling the same number of sites from the original data with replacement. Thus each bootstrap sample is a new dataset with the same sets of isolates but with some of the original sites duplicated and others dropped. This process of sampling is repeated a large number of times, the corresponding trees are reconstructed and the proportions of bootstrap trees agreeing with the original tree are calculated [55]. This provides assessments of ‘confidence’ for each node of an observed tree. Figure 4.11B shows the mean bootstrap values of the nodes of the median trees of the groups and the reference tree (annotated by RT). The mean bootstrap value increases in the median tree of the first group

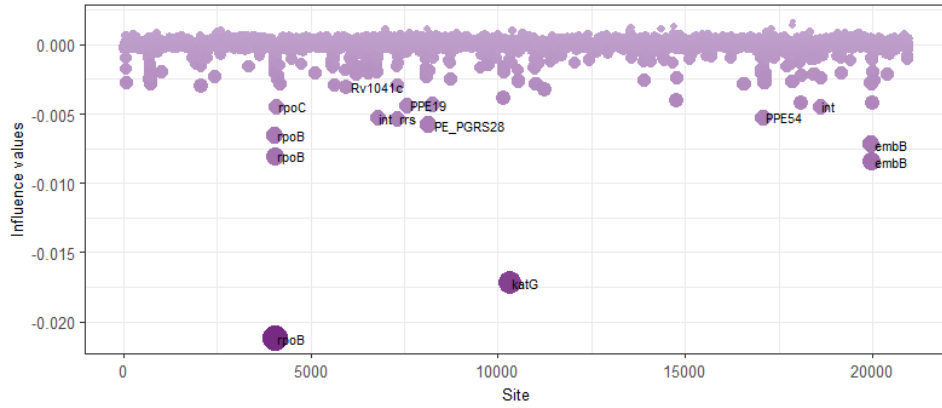


Figure 4.9: *Influence values of sites along the entire alignment with the genes of the top 15 outlier sites highlighted. Each dot represents a site and the size of the dots are mapped to the distance associated with each site.*

and then reduces in the others. This suggests that we obtain a more stable topology by removing the top disruptive sites.

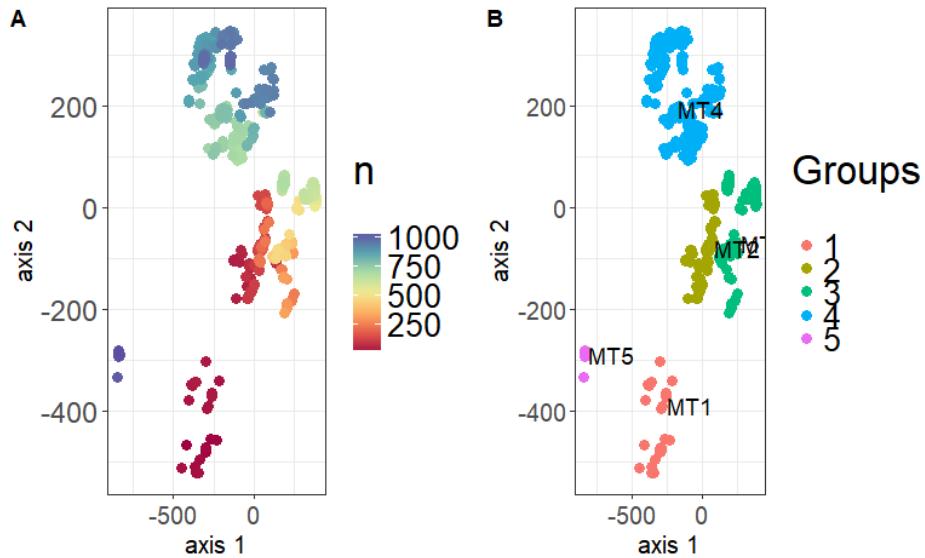


Figure 4.10: *A. MDS projection of distances between these trees coloured by their ranking, B. MDS projection of distances between these trees coloured by their groups (median tree of each group is annotated by MT'group number')*

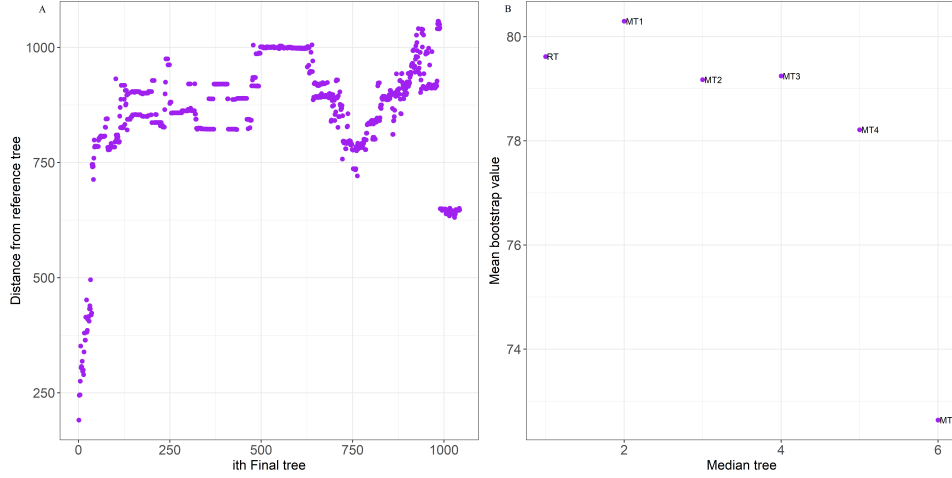


Figure 4.11: *A. Distance from reference tree of trees with top m disruptive sites removed ($m = 1 : 1043$), B. Mean bootstrap values of the nodes of median trees of the 5 groups*

4.11 CONCLUSION

We have introduced a method for detecting sites along a multiple sequence alignment that are phylogenetically disruptive. It reveals the extent to which a single site or group of sites shapes the reconstructed phylogenetic tree. It provides an easy way to identify sites that should be excluded before a dataset is used for phylogenetic inference. The method is able to detect sites that are attributed to drug resistance and homoplasies. There have been studies where trees from different partitions of the genome are compared in order to quantify phylogenetic incongruence or detect recombination. This is because trees based on single genes usually differ from each other [156], although not here, as TB is a highly clonal organism. However when a single site is removed from an alignment, it is not expected that a different tree topology would be inferred as the likelihood of the data at a site given the tree is usually a very small contribution to the estimation. In our dataset, however, approximately 5% of the sites individually affect the estimated tree topology. These sites may contain signals that mask the true evolutionary relationship between the isolates.

The influence value approach proposed by [11] aims to detect sites along the alignment that do not support the reference phylogeny and ultimately

delete them from the tree analysis. This is quite different from the approach described here as we look for sites that support the reference tree (i.e when they are removed from the alignment, a different tree is inferred) and term them disruptive. The major reason for the difference in approach is that in [11], the data analysed meets certain criteria that differs from a typical tuberculosis dataset. Firstly, It was stated that their data had enough variation that would clearly resolve the phylogenetic topology. As such the outlier sites are sites that do not strengthen the support for the topology of the reference tree and should hence be removed. Secondly, our data is based on the whole genome whereas theirs is based on a single gene and as such the hypothesis of different areas of the genome evolving at a different rates and masking tree signal does not hold for their data.

The metric used in this analysis detects differences near the root and as a result, trees that have distances much greater than zero from the reference tree may be as a result of sites that support certain branching deep in the tree that may not be reflective of the true evolutionary relationships. Branching events deep in the tree are typically challenging to infer [70] and the presence of a disruptive site in a whole sequence alignment indeed makes it more difficult.

It is always hoped that in a sequence alignment the true phylogenetic signal is present in at least some of the sites, other sites will represent, to varying degrees, phylogenetically uninformative noise or signals of selection (or recombination, where present). In the case of the simulated data, there is enough variation to infer the tree, thus when we remove certain sites, they change tree topology. For empirical data such as our tuberculosis dataset, there may not enough variation to infer the true tree and as such a single site may contain crucial piece of information that helps infer the true topology and removing such sites would be detrimental to phylogenetic inference.

Some of the top distance sites are in genes associated with resistance: these genes are generally thought to be diversifying more frequently [199] and as such may contain disruptive signals. Studies have found that more than half of new cases of MDR TB are among people who have never been treated for tuberculosis before pointing to the importance of transmission of resistance [161]. This suggests that there is enough transmitted resistance and as such sites located in resistance genes may actually be informative. The sampling

strategy plays a key role in the sense that if isolates in the data set are specifically chosen because they are resistant as is the case here, then one would expect to see more homoplasies and disruptive sites.

Site number	Distance	Gene	Resistance Gene
Site 8728	190.55	Rv1587c	No
Site 4314	156.08	Rv0731c	No
Site 4036	138.16	rpoB	Yes
Site 11079	124.67	pks12	No
Site 7506	122.32	moeY	No
Site 5634	119.93	intergenic	No
Site 17043	118.54	PPE54	No
Site 18097	114.51	REP-8	No
Site 5633	114.24	Intergenic	No
Site 8864	96.18	cydB	No
Site 10298	76.90	katG	Yes
Site 5090	73.76	PPE13	No
Site 5447	68.84	purH	No
Site 18529	64.69	PE-PGRS55	No
Site 4031	64.49	rpoB	Yes

Table 4.1: Table the genes, distances of top distance sites along alignment. The resistance column is compiled from the TB drug resistance mutation database [159]

*The secret of change is to focus all of one's energy not
on fighting the old but on building the new*

Socrates

5

Inferring transmission events using genomic and epidemiological data: is there ongoing transmission in Norway?

INTRODUCTION

Understanding the dynamics of transmission is key for developing effective policies and strategies that reduce the spread of infectious diseases. In particular, this can provide a mechanism to quantify factors associated with heightened transmissibility and susceptibility to carriage and infection which could help identify effective interventions to reduce transmission. All infectious disease outbreaks are characterized by a transmission network: a directed graph that shows who infected whom. A transmission tree consists of a set of nodes and edges where there is an edge from node a to node b if a infected b . A transmission network may be composed of multiple unconnected subtrees, each representing independent chains of transmission. Each transmission chain has an origin, representing a new introduction of the pathogen into the population. Outbreak reconstruction based on traditional epidemiology typically involves collection of epidemiological data such as details of exposure to infection, exhaustive contact tracing and locations where the disease may have spread. This is usually challenging especially

in cases of chronic infections where there is a considerable amount of time from when a patient is infected to when he/she is diagnosed. Genetic data of pathogens provides valuable additional information on inferring transmission events between cases in a disease outbreak, especially in situations where reliable contact tracing data is not available [94]. This data has the potential to inform inference on transmission with the central idea being that genetic change and disease transmission occur on a comparable timescale and as such comparing pathogen sequences of infected individuals would yield valuable insights to the inference of who infected whom.

In this chapter we give a summary of current methods used to investigate transmission of infectious diseases using sequence data with a focus on TransPhylo, a method developed by Didelot et al [48]. We use this method on tuberculosis data from Norway in order to determine if there is ongoing transmission of tuberculosis among migrants.

5.1 GENETIC DATA AND TRANSMISSION INFERENCE

The availability of next generation sequencing methods and the reduction in the cost of sequencing has made a huge amount of genomic data readily available. The basic question to answer is: given the genetic sequences of isolates collected from infected individuals, can we compare them in a statistical way to determine the infector/infectee relationships? Several statistical methods have been developed to answer this question.

5.1.1 THRESHOLD OF GENETIC SIMILARITY

A simple approach would be to assess pairwise sequence differences and set a threshold for the number of substitutions above which transmission is ruled out. The main idea is that the greater the similarity between samples taken from two different hosts, the more likely they are to be a transmission pair. Walker et al [185] ruled out transmission between pairs when they differed by more than 12 single nucleotide polymorphisms (SNPs) over 6 years. Other studies have used existing SNP thresholds to define transmission links [122, 78, 188]. This method also can be used to group individuals into clusters and then rule out transmission between individuals in different clusters.

Although this approach can rule out transmission in a quite simple way, it cannot generally determine the direction of transmission except with the use of field data which may not be complete.

5.1.2 GENETIC DISTANCES

This method makes use of genetic distances between sampled isolates. Genetic distance data can be used as an indicator of how many transmissions there might be along a transmission tree from an individual down to another individual whose distance is being measured [102]. This method involves weighting transmission links by a function of observed genetic distance. The distances can be used to create a network such as minimum spanning, neighbour joining or median joining networks to visualize transmission. Jombart et al [96] constructs a transmission tree by finding the spanning tree of the sampled sequences that minimizes/maximizes a set of edge weights. Spada et al [170] also used genetic distances to infer the transmission tree of five children with Hepatitis C virus by finding the spanning tree of the sequences that minimizes the sum of genetic distances across its edges.

5.1.3 GENETIC DISTANCES AND EPIDEMIOLOGICAL DATA

A third approach involves the combination of genetic and epidemiological data to weight transmission events using a pseudo likelihood approach [98]. This approach utilizes a likelihood component that describes the probability that a set of mutations occurs between two pathogen samples from different hosts given their epidemiological relationships. This is usually based on certain assumptions on transmission bottleneck (number pathogens transmitted from infector to infectee), pathogen evolution and so on. Jombart et al [94] developed a method that makes use of genetic distance data and sampling times to infer likely transmission events, dates of infections, and unobserved cases. They also allow for multiple introductions of infection thus distinguishing whether the cases are linked by a single connected transmission tree or a transmission network consisting of smaller transmission trees. Ypma et al [196] developed a method that combines genetic, geographical and temporal data to calculate the likelihood of a transmission tree and used this method to infer transmission events of an influenza out-

break among poultry farms in the Netherlands. Both methods assume that pathogen evolution occurs at the point of transmission and neglect in-host pathogen diversity. Morelli et al [138], on the other hand proposes a method that is a variation of both but considers in-host genetic diversity.

5.1.4 USING PHYLOGENETIC TREES

Another commonly used approach is use of phylogenetic trees. A phylogenetic tree reconstructed from genetic sequence data contains information of past transmission dynamics [184] and as such can inform transmission inference. This approach is based on the reconstruction of ancestries between hypothetical common ancestor (internal nodes) and sampled isolates (tips) [96]. Although the transmission tree and the phylogenetic tree of an outbreak may appear as two representations of the same tree, they are in fact different in meaning and properties for the following reasons :

- **Topology**

Phylogenetic trees are undirected, on the other hand, transmission trees are directed, therefore there is ambiguity around the direction of transmission even if the transmission tree is topologically identical to the phylogenetic tree. The internal nodes of a phylogenetic tree depict the hypothetical common ancestors of sampled pathogen sequences (tips), on the other hand, the internal nodes and tips of a transmission tree are the hosts in an outbreak (which may be sampled or not). Also in a phylogenetic tree, the timing of coalescence of lineages take place prior to transmission while in a transmission tree, this coalescence corresponds to a transmission event. This difference in branch length between the two trees depends on the generation time (time from when an individual gets infected to when the individual becomes infectious) and within-host dynamics [197].

- **Within host diversity**

Some methods [39, 38] identify transmission events as branching events in a phylogenetic tree, this depends on the assumption that no evolution occurs within a host from infection to onset of infectiousness and is well suited for pathogens with a small generation time. For

pathogens with long latent periods it is expected that the infecting agent would undergo evolutionary changes with several lineages being transmitted to secondary cases [49]. Also the pathogens that infect a host are not necessarily identical to the ones sampled from the host. Ypma et al [197] show that not considering the fact that several lineages of a pathogen can be present in a host can lead to biased estimates of transmission parameters.

- **Sampling density**

Typically not all cases of an outbreak are sampled and as such the possibility of linking two sampled cases via an unsampled case is not considered. This can introduce errors in the inference process. For example, in an outbreak, it seems intuitive to rule out transmission between two cases, a and b that are not paired in a phylogenetic tree (i.e. a pair of tips that are adjacent to a common ancestor node) as one would expect the evolutionary distance between transmission pairs to be relatively small. However even when a and b are not phylogenetically paired, there are scenarios when the transmission pair may be genuine. For instance, either of them may be a common source of infection for the other along side other sampled cases in a clade and ruling out transmission in this situation would be erroneous.

Despite these differences, phylogenetic trees can be used in a framework that ensures its consistency with the transmission tree. Several frameworks have been developed [197, 49, 48, 98].

5.1.5 BAYESIAN INFERENCE OF TRANSMISSION TREES

Bayesian inference provides a framework for better tree inference as it involves the use of likelihood functions that are consistent with models that appropriately represent the underlying processes [110]. The output of Bayesian inference is a posterior distribution which can be used to generate statistics of interest. This approach usually involves specifying a model that defines the underlying epidemiological process (which could be any of the generic SIR type models), and in some cases a model for sequence evolution. The joint posterior distribution of the model parameters and data is defined as

$$P(\beta, T|G) \propto P(G|\beta, T)P(\beta)$$

where T is the transmission tree, β are the model parameters and G is the observed data which could be genetic sequences or a phylogenetic tree. $P(G|\beta, T)$ is the likelihood function and $P(\beta)$ represents our prior knowledge of model parameters. There are several variations of this in literature depending on whether there is joint inference of the phylogenetic tree and transmission tree [197, 44], inclusion of location data [137, 138], whether within host diversity is considered [44, 44, 197, 84], whether unsampled cases are considered [44, 148, 48]. Table 5.6 shows some available Bayesian inference methods used to infer transmission trees with differences with respect to these variations .

5.2 COMPARING TRANSMISSION TREES

Bayesian inference of transmission trees leads to a posterior collection of transmission trees alongside posterior model parameters. In order to analyze these posterior collections of transmission trees, a means of comparison is needed. One of such is the transmission tree metric proposed by [99]. A metric such as this allows us to understand posterior uncertainty, sensitivity to priors and also identify a representative median tree from a collection of trees. This metric assumes that each individual is infected at most once, and that each transmission tree has a unique source (i.e. no multiple introductions of the disease). It also allows for unsampled cases to be amongst the nodes of the transmission tree. A vector containing the depth of every pair of nodes is created for each transmission tree. The depth of a pair of nodes is the number of edges from the source case to the most recent ancestor of the pair. The distance between two transmission trees is the Euclidean distance between their vectors. Since the number of unsampled cases may differ in trees to be compared, the vectors are constructed for the sampled cases only.

Due to the fact that the metric considers distance from the source case, differences in historic transmissions are given more emphasis than recent transmission differences. This could cause two trees that have a different source case but very similar transmission events to be a large distance apart.

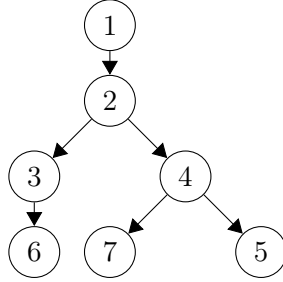


Figure 5.1: *simple transmission tree.* Here, $v_{5,7} = 2$ because the MRCI of cases 5 and 7 is case 4, which is 2 steps from the source case 1. But $v_{2,6} = v_{2,5} = v_{2,7} = 0$, and so on for pairs of cases whose MRCI is the source case, 1.

However, during investigation of an outbreak, importance is usually given to finding the source case as this could help with preventing onward transmission. The metric is implemented in the package `treespace` [99] in R.

5.3 TRANSPHYLO

In this section we describe a Bayesian method used to reconstruct transmission trees using phylogenetic trees that accounts for unsampled cases and within host diversity developed by [48]. This is done in a two-step process as specified in [49] :

- Construct a timed phylogeny $\mathcal{P}$ using any of the existing tree reconstruction methods.
- Infer the underlying transmission tree T given $\mathcal{P}$ by colouring the branches of phylogeny with a unique colour for each host so that a change in colour represents a transmission event between hosts.

Let θ be the transmission and sampling model parameters, and $N_e g$ the within host population evolutionary parameter i.e. the product of the in-host effective population size N_e (the size of an idealized (neutrally evolving, homogeneous) population that is otherwise equivalent to an observed population.) and the generation time g (time from between getting infected and onset of infectiousness). Then the Bayesian decomposition is

$$P(\theta, N_{eg}, T|\mathcal{P}) \propto P(\mathcal{P}|N_{eg}, T)P(T|\theta)P(\theta)P(N_{eg}) \quad (5.1)$$

The last two terms are prior distributions which are assumed to follow a gamma distribution .

MODEL FOR EVOLUTION WITHIN HOSTS

For the model of evolution within hosts, the following assumptions concerning the link between phylogenetic trees and a transmission trees are made:

- Each individual is infected at most once
- A single pathogen initiates each infection and a single genome of this pathogen is obtained at sampling (complete bottleneck).
- The pathogen is a random sample from the infector's pathogen population at the time of infection. When an individual is infected, evolution occurs and it is assumed that this is a neutral coalescent process with constant population size N_e i.e. any two lineages in a host coalesce at a constant rate N_{eg}

Using the above, the phylogeny $\mathcal{P}$ is simulated under a given transmission tree T and parameter N_{eg} by independently generating a subtree for each host and pasting them together to produce $\mathcal{P}$. The subtree in each host j has a single root at the time j is infected and a leaf when j is sampled and a leaf for each host that j infects. Each lineage within a host coalesce at the rate N_{eg} and rejection sampling is used to ensure that only a single lineage exists when a host is infected.

5.3.1 MODEL FOR EPIDEMIOLOGICAL PROCESS BETWEEN HOSTS

In this section we describe the epidemiological model as provided in [48]. The process of spread of the pathogen in a population is through contact transmission. The model for the spread of the pathogen in a population is considered to be a stochastic branching process where the number of individuals an infected individual transmits to is drawn from an offspring distribution $\alpha(k)$

which is assumed to be the negative binomial distribution with parameters r and p . The probability of having k offspring is thus $\alpha(k) = \binom{k+r-1}{k} p^k (1-p)^r$. This determines the reproduction number of the outbreak, which is the average number of individuals a single infectious individual can transmit to during an outbreak. It is given as the mean of the negative binomial distribution i.e. $\frac{rp}{1-p}$. The time between when an individual gets infected and when the individual becomes infectious (generation time) is drawn from a generation time distribution $\gamma(\tau)$ which is assumed to be a gamma distribution. The time from when an individual gets infected to when the individual is identified by the health sector is drawn from a sampling time distribution $\sigma(\tau)$ which also assumed to be a gamma distribution.

5.3.2 PROBABILITY OF A TRANSMISSION TREE $P(T|\theta)$

In order to calculate the probability of a transmission tree, TransPhylo considers two scenarios as specified in [48]: a finished outbreak (no more infected cases) and an ongoing outbreak (observation ends at some time $t = T$). TransPhylo accounts for the possibility of unsampled cases, as such an individual may be included in a transmission tree (either by being sampled or by leading through transmission to at most one sampled individual) or excluded i.e. the individual and all their descendants are not sampled. Let the probability of being excluded in a transmission be ω and the probability of being sampled be π .

Finished Outbreak: In this scenario, the outbreak is finished and there are no more infected individuals. We let the subscript $+$ denote all probabilities associated with this scenario. Since the outbreak is finished, all individuals are sampled with the same probability. In this scenario, the probability of being excluded is given by

$$\omega_+ = (1 - \pi) \sum_{k=0}^{\infty} \alpha(k) \omega_+^k \quad (5.2)$$

The probability generating function of the offspring distribution is given as $G(z) = \sum_{k=0}^{\infty} \alpha(k) z^k = (\frac{1-p}{1-pz})^r$. Applying this to 5.2, we have

$$\omega_+ = (1 - \pi) \left(\frac{1 - p}{1 - p\omega_+} \right)^r$$

The value for ω_+ is obtained by finding the root of the equation. The probability that an individual has k offspring of which d are included is

$$\sum_{k=d}^{\infty} \binom{k}{d} \alpha(k) \omega_+^{k-d} (1 - \omega_+)^d.$$

Each included individual in the transmission tree will have its own term and as such the modified offspring function is defined as:

$$\alpha_+(d) = \sum_{k=d}^{\infty} \binom{k}{d} \alpha(k) \omega_+^{k-d} = \alpha(d) + (d+1)\alpha(d+1)\omega_+ + \dots$$

We see that when there is full sampling $\pi = 1$ and $\omega_+ = 0$, then $\alpha_+(d) = \alpha(d)$. Each transmission tree would have n nodes which correspond to the included individuals of which some would be sampled and others not sampled. Let t_i^{sam} and t_i^{inf} be the sampling and infection times of individual i , with d_i representing the number of its offspring. To calculate the probability of a transmission tree, the root node ρ of the tree is considered. The node has d_ρ offspring and a subtree for each of its offspring. The probability of a transmission tree, T is thus :

$$p_+(T|\theta) = (1-\pi)^{1-s_p} (\pi \sigma(t_p^{sam} - t_p^{inf}))^{s_p} \sum_{k=d_p}^{\infty} \left(\binom{k}{d_p} \alpha(k) \omega_+^{k-d_p} \prod_{j=1}^{d_p} \left[p_+(T_j) \gamma(t_j^{inf} - t_p^{inf}) \right] \right)$$

where $s_i = 1$ if an individual is sampled and zero if unsampled and T_j are the subtrees of a node. Rearranging this and using the modified offspring function gives

$$p_+(T|\theta) = (1 - \pi)^{1-s_p} (\pi \sigma(t_p^{sam} - t_p^{inf}))^{s_p} \alpha_+(d_p) \prod_{j=1}^{d_p} \left[p_+(T_j) \gamma(t_j^{inf} - t_p^{inf}) \right]$$

Repeating this for all the nodes the probability of the transmission tree gives

$$p_+(T|\theta) = \prod_{i=1}^n \left[(1-\pi)^{1-s_p} (\pi \sigma(t_p^{sam} - t_p^{inf}))^{s_p} \alpha_+(d_p) \prod_{j=1}^{d_i} \gamma(t_j^{inf} - t_i^{inf}) \right] \quad (5.3)$$

Ongoing outbreak scenario: In this scenario, let π_t and ω_t be the probability of being sampled and excluded at time t respectively. The ongoing outbreak is observed up to a time $t = T$. As such individuals who become infected just before time T have a lower chance of being sampled. The probability of being excluded is also affected by stopping the observation before the end of the outbreak. Individuals who are infected at time $t \geq T$ are excluded with probability 1 and individuals infected close to the start of the outbreak are excluded with a probability that is close to the finished outbreak scenario i.e. as $t \rightarrow -\infty$, $\omega_t \rightarrow \omega_+$. Thus we have :

$$\pi_t = \int_0^{T-t} \sigma(\tau) d\tau,$$

and

$$\omega_t = (1-\pi_t) \sum_{k=0}^{\infty} \alpha(k) \left[\int_t^{\infty} \gamma(\tau) \omega_{t+\tau} d\tau \right]^k = (1-\pi_t) \sum_{k=0}^{\infty} \alpha(k) \prod_{j=1}^k \left[\int_t^{\infty} \gamma(\tau_j - t) \omega_{\tau_j} d\tau \right].$$

If we let $\hat{\omega}_t = \left[\int_t^{\infty} \gamma(\tau_j - t) \omega_{\tau_j} d\tau \right]$, then we have that

$$\omega_t = (1 - \pi_t) G(\hat{\omega}_t) = (1 - \pi_t) \left[\frac{1-p}{1-p\hat{\omega}_t} \right]^r \quad (5.4)$$

as in the finished outbreak scenario. It remains to compute the $\hat{\omega}_t$. Simplifying and applying the fact that $\omega_t = 1$ for $t \geq T$

$$\hat{\omega}_t = \int_t^T \gamma(\tau - t) \omega(\tau) d\tau + \int_T^{\infty} \gamma(\tau - t) d\tau$$

The second term of the sum can be calculated explicitly to give a function say $H(t)$. To compute the first term, the trapezoid method of integration is

used. The trapezoid method is a numerical method that approximates the value of a definite integral by using n trapezoids formed by using straight line segments between the points as shown in Figure 5.2.

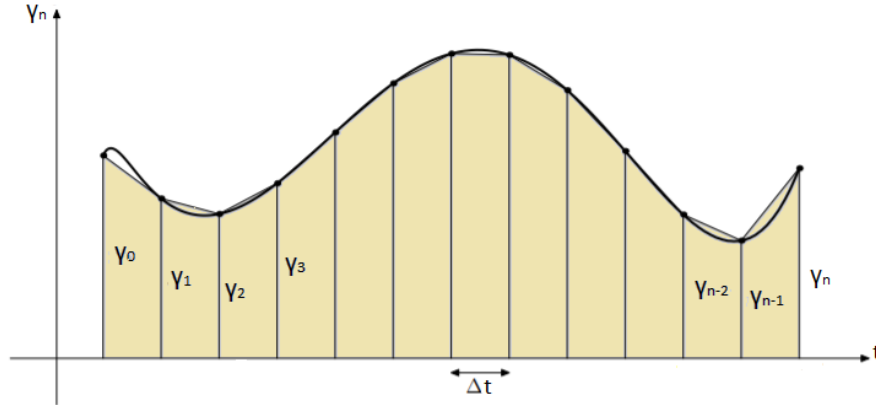


Figure 5.2: *Illustration of the trapezoid method. The area of a trapezoid is equal to the sum of the area of the triangle and the rectangle, the integral is the sum of the areas of the trapezoids.*

The trapezoid method gives that

$$\int_a^b \gamma(a) = \frac{\Delta t}{2} \left(\gamma(t_0) + 2\gamma(t_1) + \dots + 2\gamma(t_{n-1}) + \gamma(t_n) \right)$$

With k trapezoids we have $t_1, t_2, \dots, t_k$ time points. Let $t_i = T - i\Delta t$. Applying this we have

$$\int_t^T \gamma(\tau-t)\omega(\tau)d\tau = \frac{\Delta t}{2} \left(\gamma(T-t_0)\omega(t_0) + 2\gamma(T-t_1)\omega(t_1) + \dots + \gamma(T-t_{k-1})\omega(t_{k-1}) + \gamma(0)\omega(t_k) \right)$$

$\gamma(0) = 0$ thus we have

$$\int_t^T \gamma(\tau - t) \omega(\tau) d\tau = \sum_{i=0}^{k-1} c_i \gamma(T - t_i) \omega(t_i),$$

where

$$c_i = \begin{cases} \frac{\Delta t}{2}, & \text{if } i = 0 \\ \Delta t, & \text{otherwise} \end{cases}$$

Substituting these in 5.4, we have

$$\omega(t_k) = \left[\frac{1 - p}{1 - p(H(t) + \sum_{i=0}^{k-1} c_i \gamma(T - t_i) \omega(t_i))} \right]^r \quad (5.5)$$

5.3.3 PROBABILITY OF THE PHYLOGENETIC TREE GIVEN THE TRANSMISSION TREE AND WITHIN HOST PARAMETER $P(\mathcal{P}|N_{eg}, T)$

In this section, we calculate the probability of the phylogenetic tree as described in [49]. The phylogeny $\mathcal{P}$, is related to the transmission tree by associating each point on $\mathcal{P}$ to a host: each host is represented by a unique color cl_i , and a branch segment is coloured with cl_i if it represents evolution within host i . Thus a change in colour along the branches of $\mathcal{P}$ represent a transmission event. Therefore $\mathcal{P}$ can be seen as a collection of subtrees $\mathcal{P}_i$ for each included host, where $\mathcal{P}_i$ is the collection of branches coloured with cl_i . These subtrees are independent as $\mathcal{P}_i$ represents the evolutionary process in host i . Each $\mathcal{P}_i$ has n_i leaves, where n_i is the number of individuals infected by host i plus one (if i is sampled). Since $\mathcal{P}_i$ is conditional on T , the dates $d_{i,j}$ of the $j = 1 \dots n_i$ leaves are known. These dates correspond to the infection times of all i 's infectees and its sampling time (when i is sampled). The date r_i of the root of $\mathcal{P}_i$ is also known, as this is the infection time of host i . The probability is thus

$$P(\mathcal{P}|N_{eg}, T) = \prod_{i=1}^n P(\mathcal{P}_i|N_{eg}, d_{i,j} r_i).$$

$P(\mathcal{P}_i|N_{eg}, d_{i,j} r_i)$ is the probability of the subtree $\mathcal{P}_i$ under the coalescent given the dates of its leaves and that there is only one ancestor at the time r_i ,

when host i is infected. To calculate this probability, we consider arranging the leaves in order of increasing age, such if $k > j$, then $d_{i,k} > d_{i,j}$. For each leaf j , consider adding it to the phylogeny formed by the previous leaves $1 \dots j-1$. The first leaf is a linear genealogy with probability 1. The second leaf has to coalesce with the first leaf before the infection time r_i . Similarly the third leaf has to coalesce with the phylogeny formed by the previous leaves before r_i . Repeating this process for all the leaves make up $\mathcal{P}_i$. Let A_j depict the sum of the branches formed so far between the time of leaf j and the time where it coalesces with an ancestor of a previously added leaf. Also let B_j be the sum of the branch lengths between the time of leaf j and time r_i . A_j is exponentially distributed with rate N_{eg} , however there is a condition that $A_j < B_j$, therefore we have that:

$$P(\mathcal{P}_i | N_{eg}, d_{i,j} r_i) = \prod_{j=2}^{n_i} \frac{\exp(-(A_j/N_{eg}))}{N_{eg}(1 - \exp(-(B_j/N_{eg})))} \quad (5.6)$$

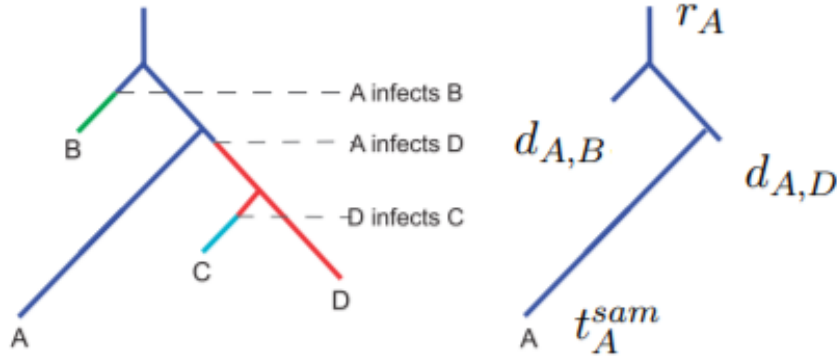


Figure 5.3: *Left: A coloured phylogenetic tree, each colour corresponds to a host and each lineage is coloured by the host it was in at the time. Right: the subtree in host A, with the $d_{A,j}$ shown.*

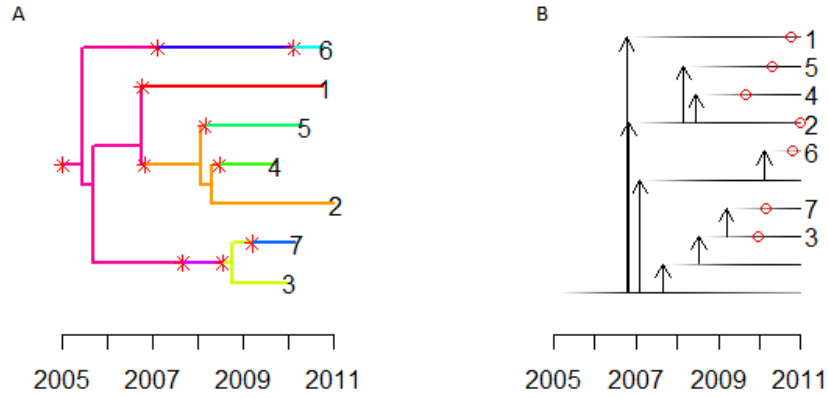


Figure 5.4: **A**: An example of a coloured phylogeny on 7 tips: each host has a unique colour to depict the evolution within each host, the red stars depict transmission events. **B**: Corresponding transmission tree: vertical arrows represent transmission between cases, and the sampled individuals have a red circle on them

5.4 PENALIZED TUBERCULOSIS TRANSMISSION

Smear-negative patients are not as infectious as smear-positive patients, and it is reasonable to assume that they transmit tuberculosis with less efficiency. We therefore modified the probability of a transmission tree in a way to account for this. We do this by penalizing transmission events which have infectors that are smear negative. We assume that each smear negative infector makes a transmission tree only 75% as likely as it would be if the infector was smear-positive. As such in the update process of the MCMC, the chances of accepting a transmission event that has such an infector is reduced. We also assume that patients with extra-pulmonary tuberculosis are only 1% as likely to transmit the disease as pulmonary tuberculosis patients, and apply this penalization to the probability of a transmission tree. These penalties imposes a mathematical preference for transmission trees that do not have infectors who have extra-pulmonary tuberculosis and/or are smear negative. As such the inferred transmission trees are pulled away

from estimates in the baseline model without smear status and/or type of tuberculosis, and pulled towards estimates that have fewer transmission events from smear negative and/or extra-pulmonary TB patients.

This approach is generally referred to as a penalized likelihood. Penalized likelihood methods augment the usual likelihood with a penalty function that encodes information about what parameter values are undesirable. A penalized log-likelihood is just the log-likelihood with a penalty subtracted from it that will pull or shrink the final estimates from regions in the posterior distribution that are undesirable [31]. The penalized likelihood, $\hat{L}(T)$ is therefore

$$\hat{L}(T) = L(T) - \lambda$$

where $L(T)$ is the log-likelihood function, and λ is the penalty term.

5.5 MIGRATION AND TUBERCULOSIS IN LOW INCIDENCE COUNTRIES

The ‘End TB strategy’ of the WHO has as a goal to reduce the global incidence of tuberculosis (TB) to 100 cases per million population by 2035. This means that the incidence rates across the globe must reduce to the what is obtainable in low incidence countries. Low incidence countries also must aim to attain lower incidence levels. Over the years, the incidence rates of tuberculosis (TB) have fallen in many high income countries but risen in many developing countries. International migration, which is one of the most notable characteristics of human society, has increased over the years as the world’s population becomes more mobile. More than three percent of the world’s population is currently living outside their country of birth as a result of industrialization [180]. As the number of people that migrate across wider ecological spaces, the probability of migration to affect public health, especially in terms of the spread of infectious diseases, increases. [22]. Immigrants who move from countries with high infectious disease incidence can pose a significant challenge for national disease control and elimination strategies for the countries they enter. This is because most of the tuberculosis burden is concentrated in high-burden settings of Africa and Asia (28% and 58%, respectively) where the cause of morbidity and mortality has been constantly attributed to tuberculosis.

International migration is the main challenge for TB prevention program in the low incidence countries [42], as tuberculosis particularly affects poor and vulnerable populations of which migrants make up a key proportion. Although tuberculosis incidence has fallen in these countries over the years, the overall changes hide an important disparity: while local-born cases have remained static or decreased, foreign-born cases have decreased more slowly or increased. The proportion of tuberculosis cases in low incidence countries that are immigrants can be anticipated to continue to increase and may lead to overall resurgence of the TB epidemic in western world [36]. Pareek et al [147] point out that this observed tuberculosis burden among migrants occurs as a result of either of the following:

- Migrants originating from countries with high tuberculosis burden have a high risk of acquiring tuberculosis before migration [198]. Also although most migrants are healthy, conditions surrounding the migration process can pose health risks and vulnerability to tuberculosis. As such they may arrive in the host country with active tuberculosis although this is relatively a small fraction [7].
- Migrants arrive host country with latent tuberculosis and difficulties such as inequalities in accessing health services, substandard quality of care, marginalization, and discrimination or the living condition of immigrants predispose them to a higher risk of reactivation.
- Migrants acquire tuberculosis after arrival through local transmission.

In low-incidence countries, preventing transmission within immigrant groups originating from high-incidence countries is of utmost importance if the overall TB incidence is to be reduced further. However, detecting transmission within immigrant populations remains a complicated task as it requires the ability to distinguish between repeated import and transmission post arrival in the host country.

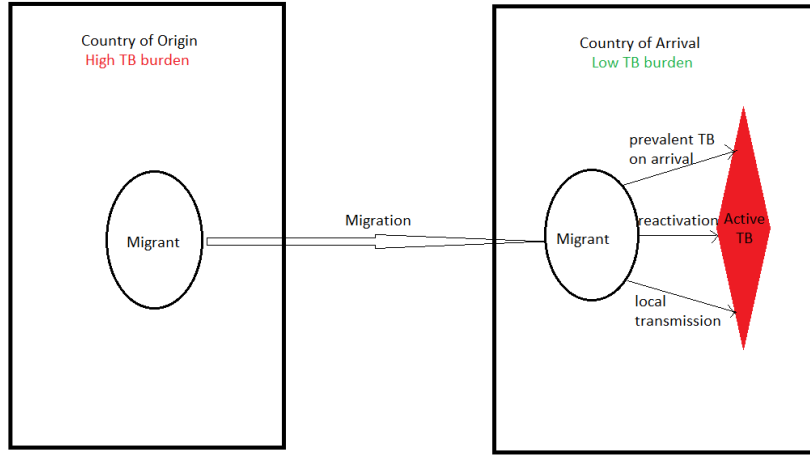


Figure 5.5: *Schematic diagram of migration and tuberculosis burden*

5.6 DETERMINING TRANSMISSION EVENTS AMONG IMMIGRANTS IN NORWAY

Although Norway is one of the few countries with low tuberculosis incidence, the incidence is not on the decline. It is generally believed that the TB level and incidence in Norway largely reflects the level of immigration from countries with a high incidence of the disease [22, 60]. During the last decade, Norway has received about 50,000 immigrants, the majority of these being from former Yugoslavia, Iraq, and countries in the horn of Africa [41]. With effective case-finding and case-management, TB transmission from immigrant populations to the Norwegian-born population has been found to be very limited [40], but does occasionally occur [57].

We apply TransPhylo to tackle a major public health conundrum, namely whether immigrants diagnosed with TB were infected before (imported cases) or after arrival in Norway (local transmission). [82]. Based on 24-loci Mycobacterial Interspersed Repetitive Unit (MIRU) typing, 129 clinical *M.tuberculosis* NAL3C isolates from 127 patients, collected between 1997 and 2015 were sequenced at the National Reference Laboratory for Mycobacteria (NRLM) Norway. A total of 1418 variable sites were identified, resulting in a mean pairwise-SNP distance of 43.22 separating the NAL3C isolates. We focus on this a large genotypic cluster for our analysis. One of the outputs

of TransPhylo is the posterior distribution of infection times, T_{inf} , of the cases. These are then compared with the time of arrival in Norway of each case to ascertain whether they became infected before or after their arrival in Norway.

A posterior distribution of timed phylogenetic trees were constructed using BEAST [51] with a GTR model and a relaxed clock model. A total of 2700 timed phylogenetic trees were obtained from BEAST. Analysing each tree separately would be computationally intensive. One commonly used approach to summarise a distribution of phylogenetic trees is the maximum credibility method. This method evaluates each of the sampled posterior trees and every clade within a tree is given a score based on the fraction of times it occurs in the sampled distribution. The score of a tree is then the product of these scores. The tree with the highest score is called the maximum credibility tree. Another method that would reduce computational cost involves analysing the distribution with a phylogenetic tree metric such as Kendall-Colijn metric discussed in section 4.6. Figure 5.7 gives a visualization using multidimensional scaling of the distances between these trees, with the median (MT) and median credibility clade (MCC) trees highlighted. In the plot each point represents a tree, and the distance between any pair of points approximates the distance between them given by the metric described in section 4.6.

5.6.1 TRANSMISSION RECONSTRUCTION

The MCC (figure 5.6) tree of the 129 isolates is characterized by long branch lengths with a few clades that have relatively shorter branch lengths (figure 5.6). As the branch lengths of a timed phylogenetic tree represents duration of evolution [5], it is intuitive to assume that these clades represent densely sampled clusters of cases whereas the longer branches depict cases with unsampled infectors. Indeed the overall structure of the phylogeny suggests that the clinical TB cases in Norway represented are samples drawn from a large unsampled population. Thus we reasoned that putative transmission-clusters in Norway would be represented by sub-clades of closely related isolates. Based on the the overall structure of the phylogeny, we selected clades with a minimum of four cases, and with a maximal mean pairwise SNP-distance of 5 or fewer SNPs within the clade for transmission inference.

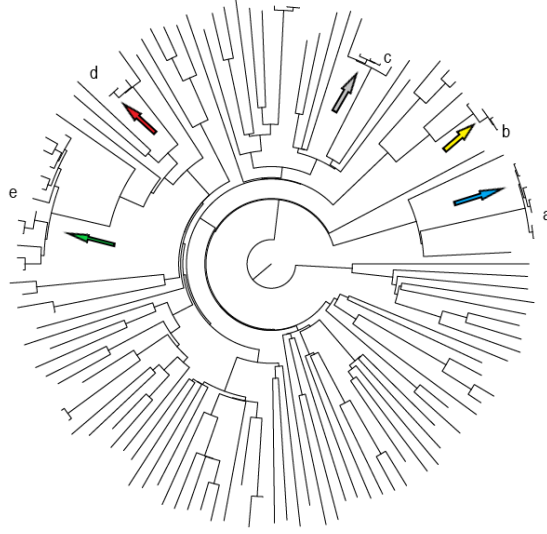


Figure 5.6: *Clinical M. tuberculosis isolates and phylogenetic reconstruction, with clades of interest highlighted.*

Five clades match this criteria (clades *a*, *b*, *c*, *d* and *e* in figure 5.8). Most of the cases in the selected clades come from countries in the horn of Africa, with one case each from Ghana, Gambia, Iran, Thailand and Norway as seen in figure 5.8.

The prior of the generation time distribution in TransPhylo, which is a gamma distribution can be seen to reflect the variable disease progression of tuberculosis which could either be rapid with short time interval from time of infection to onset of infectiousness or very long with infection leading to long latent periods before the onset of infectiousness. Estimating the parameters of the prior distribution is challenging, especially when the sampling density is unknown. However, for the five sub-clades identified above, we can assume a high sampling density based on the extremely limited observed diversity within each clade. We estimated the parameters of the sampling and generation time distributions from the subtree of the least diverse clade (clade *A*; mean pairwise SNP-distance = 0), assuming 95% sampling.

We ran 100,000 MCMC iterations with ten different choices for the priors for the sampling and generation times. We chose shape and scale parameters of the gamma distribution that give a mean of 4 – 5 years for the generation time distribution. This is uninformative as the variance is large and as such

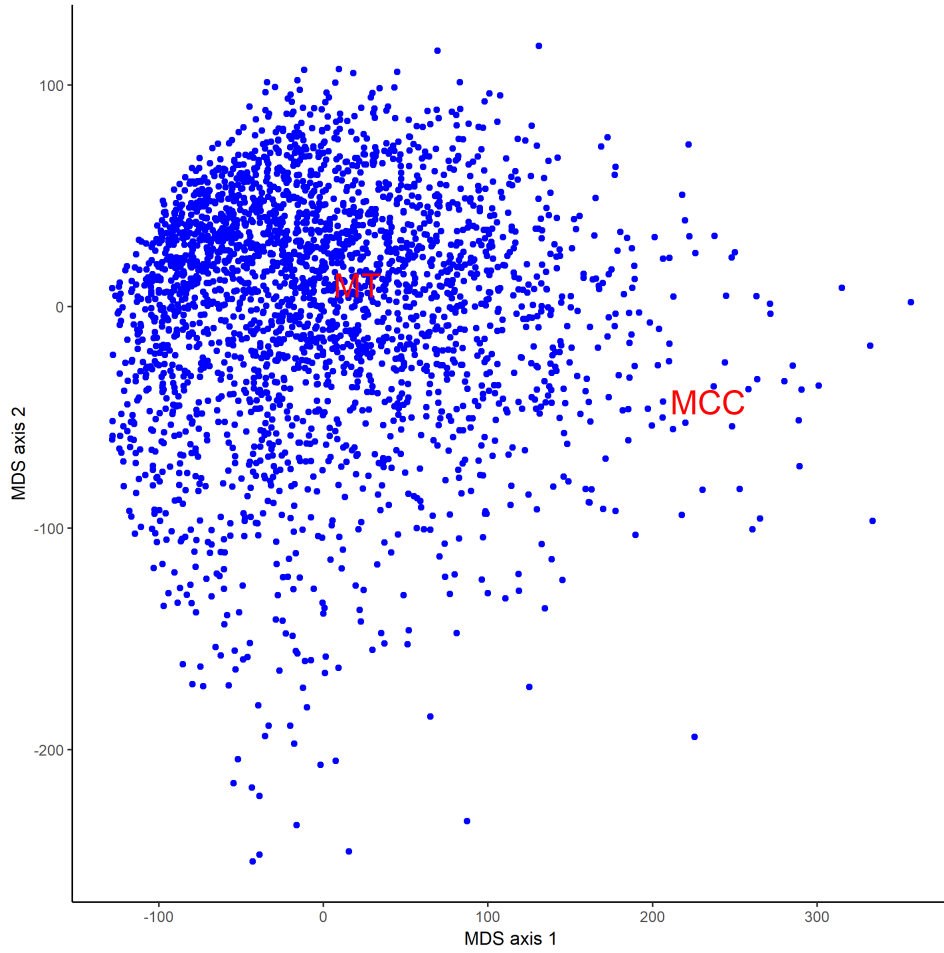


Figure 5.7: *Posterior distribution of timed phylogenetic trees with the maximum credibility clad tree (MCC) and median tree (MT) highlighted in red. The median tree is obtained using the Kendall-Colijn metric described in section 4.6*

is permissive to allow for very short and long generation times. Case finding and management of tuberculosis is quite effective in Norway, and as such, we chose a gamma sampling distribution with mean between 2.5 – 3 years. Altogether, we chose ten different shape and scale parameters that meet these criteria (Table 2a).

We start off with clade *A*, run the inference procedure using these different choices for the priors of the sampling and generation times, whilst assuming a very high sampling proportion. We thus obtain ten different posterior sam-

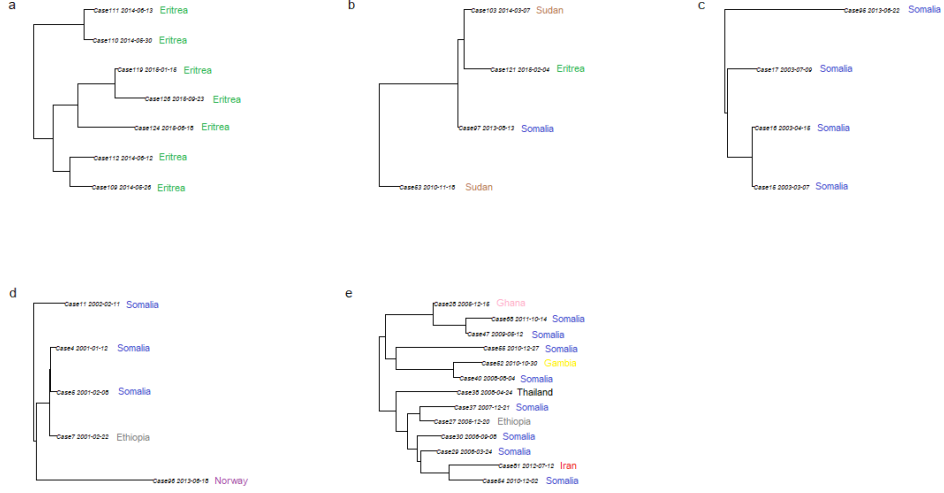


Figure 5.8: *The smaller clades with the cases, sampling dates and country of birth highlighted.*

pling and generation time distributions whose shape and scale parameters (obtained by fitting a gamma distribution to the posterior generation and sampling times respectively using the function *fitdistr* in the MASS package in R; these are shown in Table 2b) are then used as inputs for the inference of transmission events on the other clades.

5.6.2 PROBABILITY OF INFECTION PRIOR TO ARRIVAL

We can quantify the probability that an individual was exposed to their TB strain prior to their arrival in Norway, using the posterior times of infection. If we know the arrival time t_{arr}^i for case i , and we let the posterior time of infection density be called $L^i(\tau)$, then the probability that i was infected after arrival is just the portion of the posterior that lies above t_{arr}^i :

$$P(t_{inf}^i > t_{arr}^i) = \int_{t_{arr}^i}^{t_{max}} L^i(\tau) d\tau. \quad (5.7)$$

If the arrival is uncertain, and we only know that case j arrived between minimum time m_j and maximum time M_j , then we can integrate out the unknown time of arrival to find the marginal probability that j was infected

(a)	Run	G_{sh}	G_{sc}	S_{sh}	S_{sc}
	1	1.30	3.33	0.80	3.20
	2	2.20	1.98	1.00	2.50
	3	1.50	2.86	1.40	1.89
	4	3.33	1.30	1.50	1.74
	5	1.98	2.20	2.00	1.28
	6	3.33	1.50	2.50	1.20
	7	1.83	2.40	2.43	1.04
	8	2.00	2.10	2.79	1.07
	9	1.70	2.52	3.00	0.80
	10	3.50	1.23	4.20	1.07

(b)	Run	G_{sh}	G_{sc}	S_{sh}	S_{sc}
	1	1.22	0.54	0.71	0.70
	2	1.89	0.49	0.79	0.76
	3	1.38	0.56	1.02	0.64
	4	2.35	0.54	0.94	0.83
	5	2.04	0.42	1.30	0.58
	6	2.77	0.47	1.45	0.55
	7	1.85	0.47	1.32	0.69
	8	1.89	0.51	1.54	0.56
	9	1.87	0.44	1.68	0.49
	10	3.11	0.43	1.67	0.62

Table 5.1: Tables showing (a) prior and (b) posterior parameter values for the generation time distribution on clade a (shape is G_{sh} scale is G_{sc}) and sampling time distribution (shape is S_{sh} and scale is S_{sc})

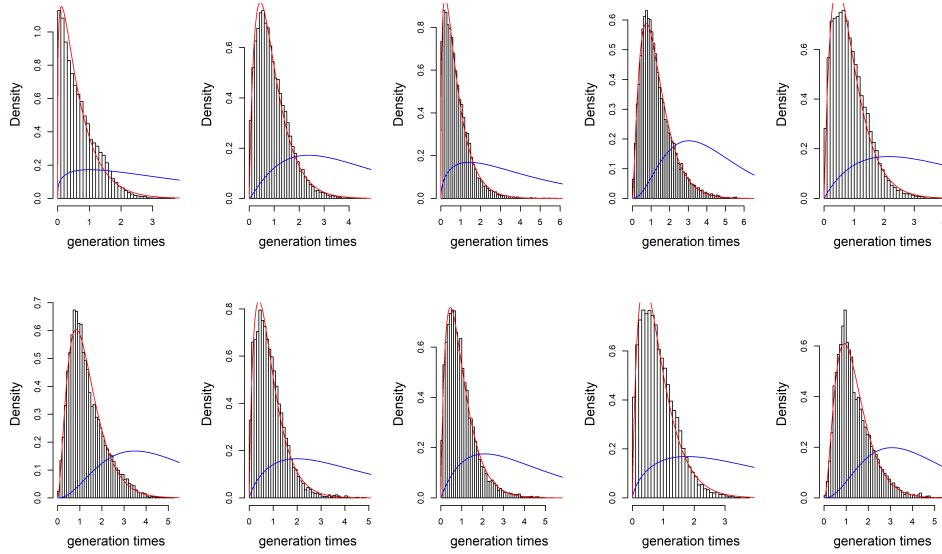


Figure 5.9: Posterior generation time distribution for Clade a. The prior distribution is highlighted in blue and the posterior fits are in red.

after arrival in Norway:

$$P(t_{inf}^j > t_{arr}^j) = \int_{m_j}^{M_j} P(t_{arr} = s)P(t_{inf}^j > s)ds$$

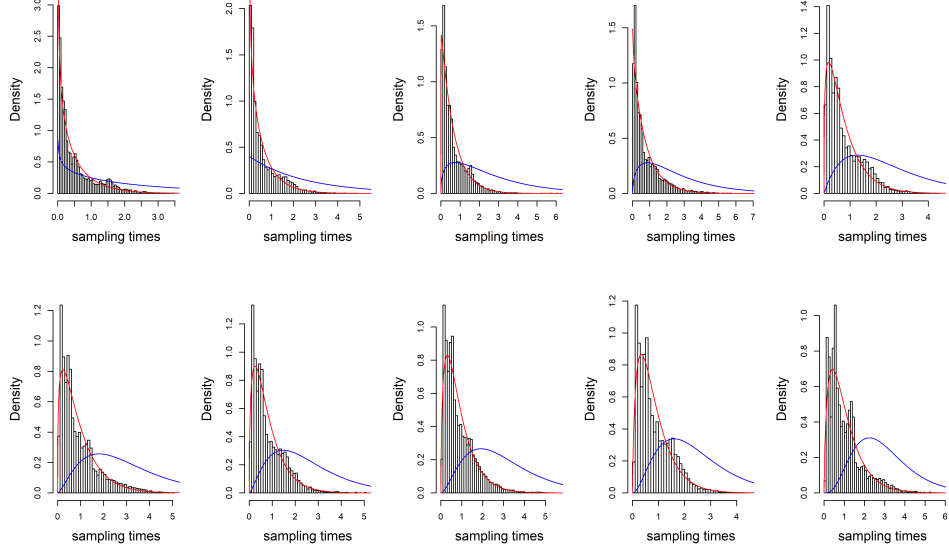


Figure 5.10: *Posterior sampling time distribution for Clade a. The prior distribution is highlighted in blue and the posterior fits are in red.*

and we use (5.7) to obtain $P(t_{inf}^i > s)$, and a uniform distribution (blue rectangles in the figures) for $P(t_{arr} = s)$. These probabilities are averaged over 10 inference procedures using different prior distribution parameters.

5.6.3 RESULTS

Genetic diversity: The high genetic diversity within the NAL3C cluster, combined with an overall phylogenetic structure characterized by multiple long terminal branches interspersed by a handful of tight clusters, suggest that the clinical TB cases in Norway represented samples drawn from a larger population of mainly unsampled cases presumably circulating in the horn of Africa.

Posterior distribution of infection times: Using the baseline model and the penalized likelihood approach outlined in section 5.4, we obtain posterior distributions of infection times for the cases in our clades of interest. Arrival time information is only available for some of the cases in our clades of interest. Figure 5.11 shows the arrival times of all cases from these clades for whom arrival times were retrievable (all in clades *A*, *B* and *E*), plotted

on top of the posterior infection time distribution for each individual (see methods section for details). It is quite clear that some of these patients (cases 30, 37, 40, 47, 54, 68 and 126) arrived in Norway before the estimated time T_{inf} . For all other cases, the estimated range for the time of infection has at least some overlap with the time of arrival.

Posterior infection times using different input trees: Clade a was extracted from the median tree and three other random trees from the posterior distribution of timed phylogenetic trees obtained from BEAST. Using the same prior parameters for the inference procedure, we reconstruct the transmission trees. Figure 5.12 shows the posterior distributions of the infection times of the five trees plotted alongside the arrival times of the cases in the clade. Although, there are differences in the distributions as expected, with some portion of the distributions falling in regions not similar to those of the MCC tree, the conclusion of the probability of infection post arrival does not change drastically across the different trees.

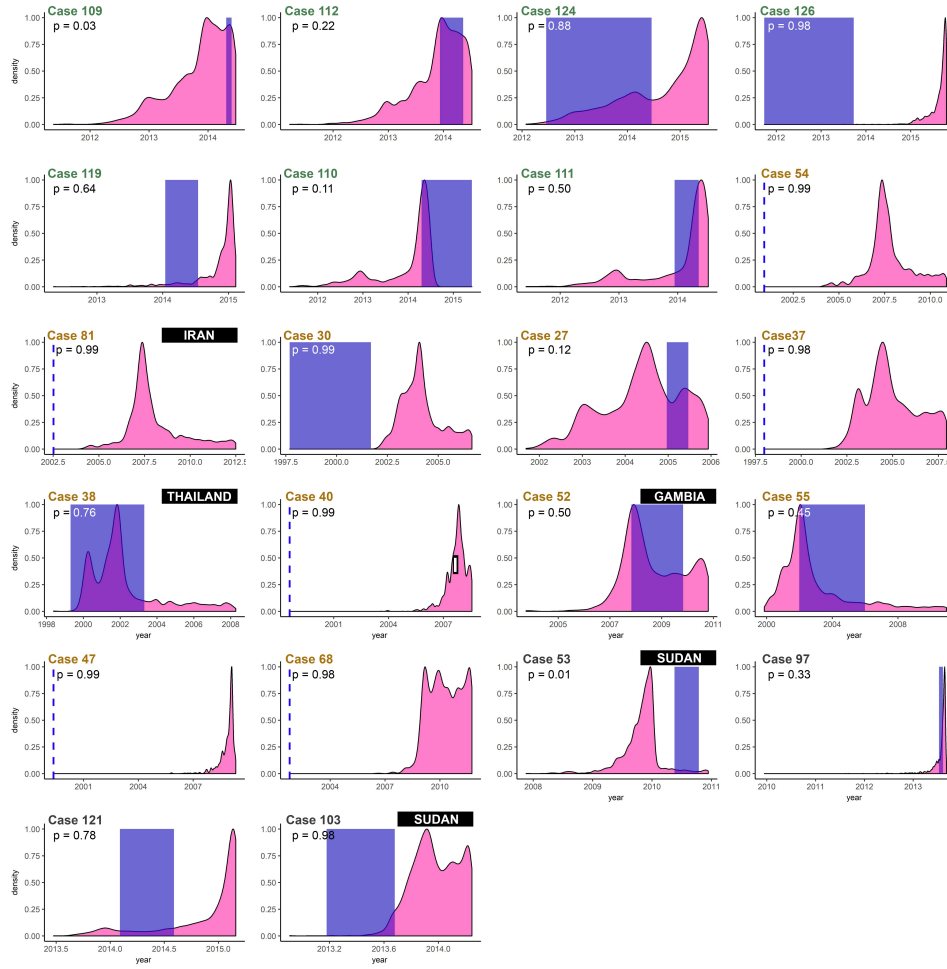


Figure 5.11: *Arrival times (in blue) plotted on top of estimated infection times for all cases of interest with available data. The case numbers are colored by clade assignment (clade a in green, clade e in orange and clade b in grey). The blue shaded area covers the time from earliest and latest possible arrival times, whereas a dotted single line indicates the latest possible arrival time. P-values indicate probability of infection after arrival in Norway, averaged over 10 different TransPhylo inference procedures. The country of origin of patients not originating from the Horn of Africa is annotated in black boxes.*

Posterior infection times using different priors: We compare the posterior infection times obtained using different prior parameter combinations to assess the sensitivity of inference results to parameter choices. Figures 5.13 show this comparison for clades *a*, *b* and *e* using six different prior pa-

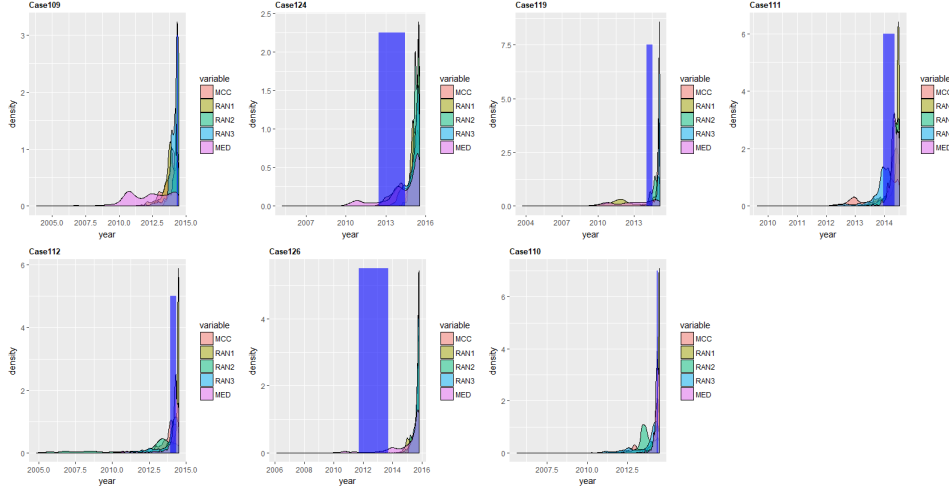


Figure 5.12: *Infection times of cases in clade a using the same priors but different input trees: MCC - maximum credibility clade tree, MED - medium tree and three random trees (RAN1, RAN2, RAN3)*

parameter values. The priors for clade *a* are taken from table 1a, while the priors for the other clades are from Table 1b. Although there are slight differences in the infection time distributions inferred using different priors, the conclusion generally remains the same when these distributions are compared with arrival times in Norway.

Probability of infection post arrival for two inference approaches:

Using the posterior densities of infection times alongside the arrival times of the cases, we obtain probabilities of infection post arrival in Norway ($P(t_{inf} > t_{arr})$) as outlined in section 5.6.2. We average these probabilities over the 10 different inference procedures using the different priors in Table 2. Table 3 shows the probabilities obtained for the cases using the two approaches (penalized and baseline inference). On average these probabilities obtained from the penalized inference are higher than those obtained using the baseline inference. The cumulative frequency plot of these probabilities (Figure 5.15) shows that there are 16 cases with $P(t_{inf} > t_{arr}) > 0.5$ and 12 cases with $P(t_{inf} > t_{arr}) > 0.9$ for the penalized inference.

Cases with no arrival information : In clades *a*, *b* and *e* there were two cases without arrival time information. For one Somali case we could conclude that the patient likely contracted TB in Norway, as the inferred

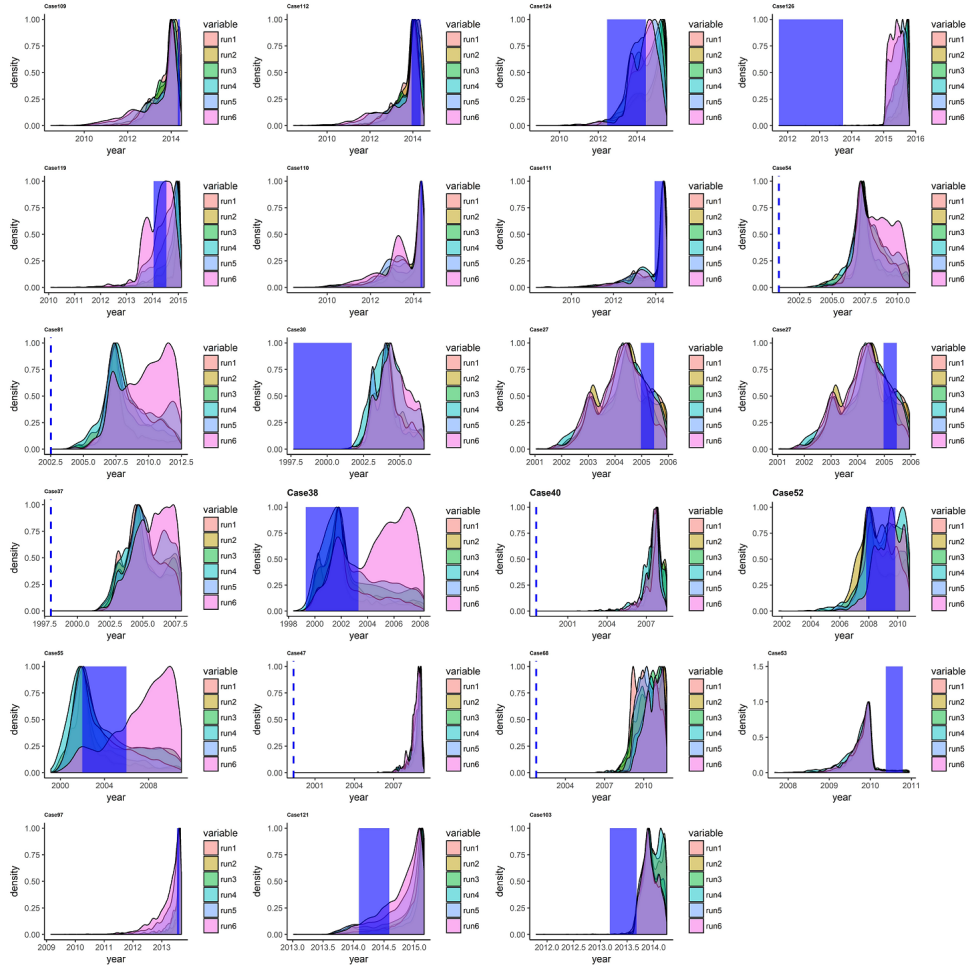


Figure 5.13: *Arrival times of cases in clades of interest plotted on their infection times obtained using six different prior parameter sets*

infector was also infected in Norway. The other case was an immigrant from Ghana, and hence also likely infected in Norway. Next, we looked into clades *c* and *d* for which arrival info was lacking for all isolates. For clade *d*, TransPhylo inferred that the same unsampled case had infected both a Norwegian, Ethiopian and two Somali patients, as well as a final Somali patient via another unsampled intermediate. This strongly suggest that all five patients contracted TB in Norway (Fig. 5.16). In clade *d*, all patients were Somali. This, combined with a lack of arrival information for these patients, makes it impossible to distinguish between transmission before or after arrival.

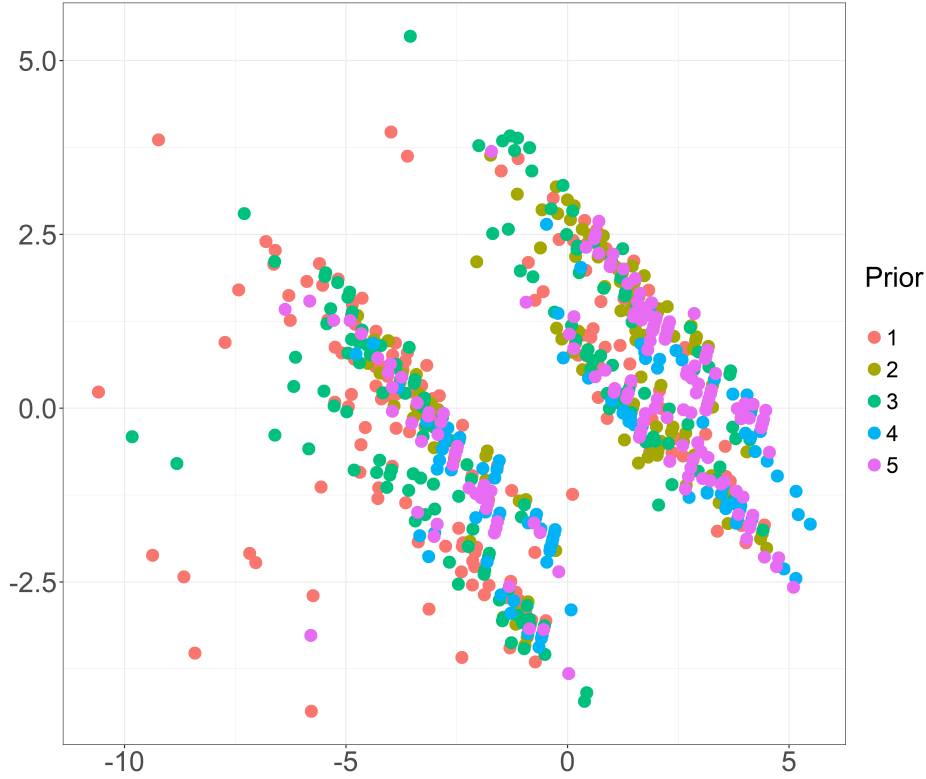


Figure 5.14: *Multidimensional scaling plot of 1,000 transmission trees obtained from using 5 different prior parameters 1 – 5 in Table 4.2, coloured according to prior parameter set.*

Cases not in selected clades: In order to obtain a more complete picture of transmission in Norway, beyond clades that were amenable to transmission inference using TransPhylo, the temporal phylogeny was further investigated manually for pairs and triplets of isolates for evidence of transmission in Norway. As the vast majority of immigrants from the Horn of Africa came to Norway after 1998 [59], we only included pairs and triplets with an estimated most recent common ancestor after 1995. Based on a combination of arrival times, disease manifestation and country of origin, we were able to identify another three instances of very probable transmission in Norway. For six cases it can be concluded that transmission in Norway was highly unlikely, whereas no conclusion could be drawn for 12 of the cases (see Tables 4, 5 and 6).

Case	Clade	Country	$P_{NP}(t_{inf} > t_{arr})$	$P_P(t_{inf} > t_{arr})$	LNorway
Case 109	A	Eritrea	0.03	0.08	No
Case 110	A	Eritrea	0.11	0.30	No
Case 111	A	Eritrea	0.50	0.98	Yes
Case 112	A	Eritrea	0.22	0.40	No
Case 119	A	Eritrea	0.64	0.57	Yes
Case 124	A	Eritrea	0.88	0.81	Yes
Case 126	A	Eritrea	0.98	0.98	Yes
Case 53	B	Sudan	0.01	0.30	Yes
Case 97	B	Somalia	0.33	0.35	No
Case 121	B	Eritrea	0.78	0.80	Yes
Case 103	B	Somalia	0.98	0.98	Yes
Case 54	E	Somalia	0.99	0.97	Yes
Case 81	E	Iran	0.99	0.97	Yes
Case 30	E	Somalia	0.99	0.99	Yes
Case 27	E	Ethiopia	0.12	0.27	No
Case 37	E	Somalia	0.98	0.97	Yes
Case 38	E	Somalia	0.76	0.96	Yes
Case 40	E	Somalia	0.99	0.99	Yes
Case 52	E	Gambia	0.50	0.88	Yes
Case 55	E	Somalia	0.45	0.96	Yes
Case 47	E	Somalia	0.99	0.99	Yes
Case 68	E	Somalia	0.98	0.97	Yes

Table 5.2: Table summarizing main results of study. Probabilities of infection after arrival in Norway (Not Penalized : $P_{NP}(t_{inf} > t_{arr})$ and Penalized: $P_P(t_{inf} > t_{arr})$) are shown where available. We conclude that the cases are likely infected in Norway (LNorway) based on these probabilities

Altogether, using the work-flow of contrasting TransPhylo inferred time of infection with time of arrival in Norway, 16 cases in clades *A*, *B* and *E* are identified that have a higher than 0.5 probability of having contracted TB in Norway. Two additional patients in these clades lacking arrival information were also determined to have contracted TB in Norway, as were all the five patients in clade *C* despite a lack of arrival times for all these patients. Finally, three additional instances of probable transmission in Norway, represented by two pairs of closely related isolates were identified. In total, it can be concluded that 26 out of 129 NAL3C cases were probably infected in Norway. For 16 cases it is impossible to conclude based on available data and/or inference, whereas the remaining 87 probably represented instances

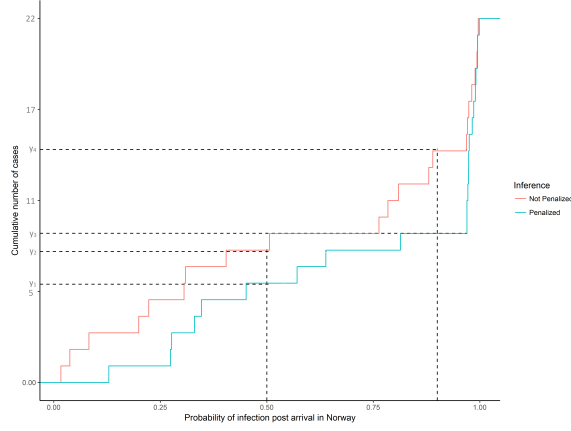


Figure 5.15: *Probability of infection after arrival in Norway for 22 cases included in TransPhylo analyses. The lines annotate the number of cases with probabilities equal to 0.5 and 0.9 in the cumulative distribution plot. With $y_1 = 16$, $y_2 = 16$, $y_3 = y_4 =$*

of imported TB.

5.7 CONCLUSION

In this chapter, approaches that involve the use of genomic data to infer transmission networks have been studied with a focus on TransPhylo. Patterns of genetic variation of pathogen isolates are informative about evolution. As such they can be used in the inference of transmission routes between epidemiologically related individuals.

TransPhylo models within host dynamics using a single parameter N_{eg} . The higher the value of this parameter, the greater the branching (diversity, as lineages coalesce at the rate $1/N_{eg}$) in a host which in turn leads to earlier infection times being inferred. This is because, at transmission, the donor has an established infection with some genetic variation in the bacterial load. From this population, a single lineage is transmitted to the recipient due to complete transmission bottleneck. The lineage that is transmitted already existed in the donor at the time of transmission and must have branched out from its ancestor in the donor prior to transmission. As such when there is more diversity in a donor, infection times are shifted to the past. Within host dynamics of pathogens is generally not well understood when sampling

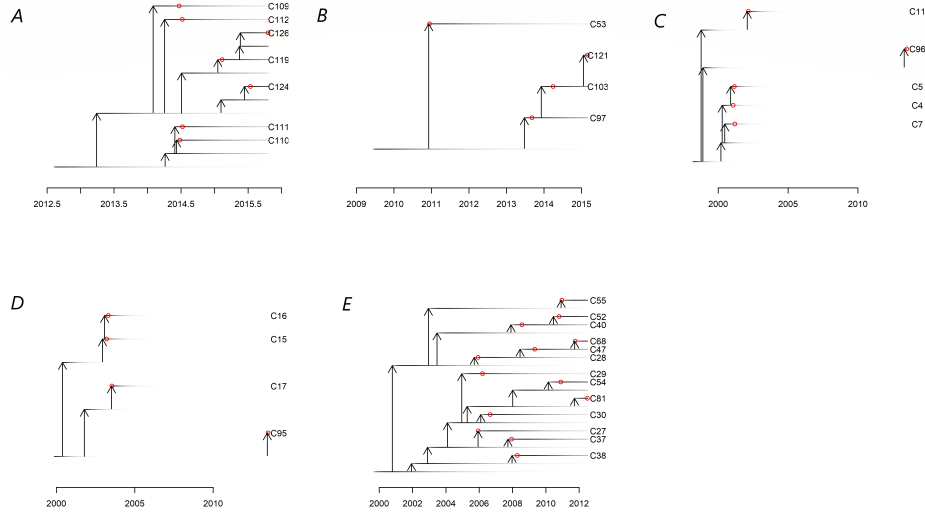


Figure 5.16: *Median transmission trees of the posterior distribution of transmission trees of the selected clades. Each horizontal line represents a case with the darkness of the line representing changing infectivity over time, the red circles indicate the sampled cases and time of sampling, the vertical arrows represent transmission from case to case. The cases are illustrated as Cn , where n is the case number.*

is not dense [191], however it can be estimated using serial sequence sampling of individual infected hosts over an extended period of time [46]. Using such estimates in the inference procedure would therefore improve the accuracy of the inferred infection times.

A key feature of TransPhylo is that it infers transmission events using a two-step procedure: obtaining a timed phylogenetic tree and inferring transmission events given this phylogenetic tree. This approach therefore makes it difficult to pass the uncertainty in the phylogenetic reconstruction to transmission inference especially when a posterior distribution of phylogenetic trees are produced in the first step. One way to account for this would be to apply the transmission inference to a sample of phylogenetic trees from step one. The two step procedure has the advantage of detecting separate transmission clusters over an approach that jointly infers phylogenetic and transmission trees [105]. These clusters can then be analysed separately, as we did on the TB data from Norway.

TransPhylo was used to analyse clusters of tuberculosis cases in Norway to determine whether there is ongoing transmission among migrants. Individual-level data on arrival time, TB smear status, time of sampling and whether the patients' tuberculosis disease was pulmonary or extra-pulmonary was used to refine the inference of the transmission process. The posterior estimates of case infection times were compared to the times when patients arrived in Norway and probabilities of infection post arrival in Norway were calculated. For many patients, the genetic structure of the isolates and/or a lack of epidemiological information left considerable uncertainty in their time of infection. However, for some patients, by taking all the available evidence into account, it was possible to infer with relative certainty whether they had contracted TB before or after arrival in Norway. Overall, it was found that there is substantial evidence of ongoing transmission in Norway. Our approach and findings highlight the importance of collecting and keeping good epidemiological records on individual patients.

In order to superimpose a transmission tree on a phylogenetic tree, there would have to be at least one transmission event on the path between every pair of tips in the phylogenetic tree, since each tip corresponds to a host. The recent timing of the clades of interest in the NAL3C, therefore suggest that there is recent transmission. Be that as it may, there is uncertainty with regards to the question of who infected who and when. Inference of times of infection using TransPhylo provides a statistical way to quantify these unknowns.

Immigrants from high-incidence countries typically make up a significant portion of TB cases in low-incidence countries such as Norway, but it is notoriously difficult to elucidate whether TB disease among them is a result of import or recent transmission in the receiving country. This difficulty has multiple roots, including the fact that cultural and ethnic identity play a role in forming social connections and low institutional trust in some immigrant groups, which can lead to an unwillingness to openly share information necessary for contact tracing. Furthermore, if public health authorities do not suspect local transmission (instead assuming that TB cases among immigrants are imported) then they may not perform outbreak control measures such as contact tracing. Indeed, national TB programs often apply a stricter evidence threshold for initiating active case finding around patients

diagnosed with TB when they are immigrants from high-incidence countries. This is in many ways rational, as such cases have an a priori higher probability of being the result of reactivation of infections acquired before arrival, but it precludes strong outbreak control when there is in fact ongoing transmission. As a result, a proper understanding of tuberculosis transmission among immigrant populations is lacking, and this hampers the design of improved interventions. The approach used in this chapter, combining epidemiological and genomic data, can assist public health authorities in understanding where and when patients are infected, and can aid in the design of appropriate TB control measures.

Case	IY	PIN	P/E	Country	MPD	Clade	Transmission in Norway	Interpretation
Case 14	2003	0.5 – 1	Pulm	Somalia	0	-	Probably one transmission event	Identical isolates, putative infector has pulmonary infection Different countries so one import event plus one transmission event
Case 107	2014	> 10	Pulm	Ethiopia		-		
Case 63	2011	1 – 3	E-pulm	Somalia	8	-	No	Putative infector not infectious
Case 108	2011	5 – 9	Pulm	Somalia		-		
Case 61	2011	0.5 – 1	Pulm	Somalia	0	-	?	Impossible to separate infection pre/post arrival due to timing of disease
Case 62	2011	0.5 – 1	Pulm	Somalia				
Case 71	2012	> 10	Pulm	Somalia	2	-	?	Impossible to separate infection pre/post arrival due to timing of disease
Case N1	1999	?	?	?		-		
Case 46	2009	1 – 3	E-Pulm	Ethiopia	14	-	Probably two transmission events	Case 98 probably infected in Norway due to country of origin. This implies an unsampled infector in Norway which also suggests Case 46 was infected in Norway
Case 98	2013	?	Pulm	Cote d’ivoir		-		
Case 57	2010	5 – 9	E-Pulm	Tanzania	11		?	N2 probably infected 57 in Norway based on arrival and diagnosis times. Tanzania is reservoir for strain as country of N2 not known, so can’t conclude
Case N2	2001	?	?	?				

Table 5.3: Table summarizing additional inference: **PIN** means period in Norway, **P/E** means Pulmonary or Extra-pulmonary TB, **MPD** means mean pairwise distance

Case	IY	PIN	P/E	Country	MPD	Clade	Transmission in Norway	Interpretation
Case 60	2012	3 – 5	E-Pulm	Somalia	10	-	No	Very likely import alternatively both infected by unsampled infector
Case 93	2012	1 – 3	E- Pulm	Somalia		-		
Case 77	2011	3 – 5	E-pulm	Somalia	6	-	No	Very likely import alternatively both infected by unsampled infector
Case 89	2011	5 – 9	E-Pulm	Somalia		-		
Case 48	2009	< 1	Pulm	Somalia	6	-	No	Both cases had E-pulmonary TB pre/post arrival due to timing of disease
Case 79	2012	< 1	E-Pulm	Somalia				
Case 101	2013	0.5 – 1	E-Pulm	Somalia	4	-	?	Temporal information compatible with 101 having been infected in Norway. But not information to conclude
Case N6	2012	?	?	?		-		
Case 75	2012	1 – 3	Pulm	Eritrea	0	-	?	Too much info lacking impossible to say much with certainty
Case N4	2012	?	?	?		-		
Case N5	2012	?	?	?		-		

Table 5.4: Table summarizing additional inference: **PIN** means period in Norway, **P/E** means Pulmonary or Extra-pulmonary TB, **MPD** means mean pairwise distance

Case	IY	PIN	P/E	Country	MPD	Clade	Transmission in Norway	Interpretation
Case 4	2012	?	Pulm	Somalia	1.2	C	Probably all infected in Norway	Based on inferred transmission chain (see main paper) and country of origin infected in Norway by unsampled host
Case 5	2012	?	Pulm	Somalia		C		
Case 7	2011	?	pulm	Ethiopia		C		
Case 11	2011	?	Pulm	Somalia		C		
Case 96	2009	?	E-Pulm	Norway		C		
Case 15	2003	?	E-Pulm	Somalia	2.5	D	?	Too much info lacking impossible to say much with certainty
Case 16	2003	?	Pulm	Somalia		D		
Case 17	2003	?	pulm	Somalia		D		
Case 95	2013	?	Pulm	Somalia		D		
Case 28	2012	?	E-Pulm	Ghana	2.5	E	Yes, both	Country of origin makes transmission before arrival extremely unlikely
Case 29	2012	?	Pulm	Somalia		E		Infector identified in Transphylo was infected in Norway

Table 5.5: Table summarizing additional inference: *PIN* means period in Norway, *P/E* means Pulmonary or Extra-pulmonary TB, *MPD* means mean pairwise distance

Name/Author	Ref	Multi-introduction (more than one source case)	Multiple sequences for each case	In-host diversity	Bottleneck > 1 (number of lineages transmitted)	Unsampled (Inferred cases not known to health authorities)	Phylogenetic Tree
Outbreaker	[94]	Yes	No	No	No	Yes	No
TransPhylo	[48]	No	Yes	Yes	No	Yes	Yes
SCOTTI	[44]	Yes	Yes	Yes	Yes	Yes	Est.
Kenah et al	[98]	No	Yes	Yes	No	No	Yes
Numinnen et al	[142]	No	No	Lim.	No	No	Est.
Mollentze et al	[137]	Yes	No	No	No	No	Yes
Morelli et al	[138]	No	No	No	No	Yes	No
Soubeyrand	[169]	No	No	Yes	Yes	No	Yes
Hall et al	[84]	No	Yes	Yes	No	Yes	Est
Phybreak	[105]	No	Yes	Yes	No	No	Est.
Trepar	[171]	No	No	Yes	No	Yes	Yes
Worby et al	[191]	No	Yes	Yes	Yes	No	No

Table 5.6: *Some available Bayesian methods for inferring transmission trees using genetic and epidemiological data. If ‘Yes’, then method uses or has the property and otherwise if ‘No’. If ‘Est’, then property is estimated by the method. If ‘Lim’, then the method includes a limited version of the property.*

*Better is the end of a matter than the beginning
thereof*

The book of Ecclesiastes

6

CONCLUSION

Significant amounts of data on the genetic diversity of *M. tuberculosis* has been generated over the years and this can be harnessed to make inference on its dynamics and interactions. Mathematical and statistical analysis of genetic data has made tremendous contribution to the study of *M. tuberculosis*. Also mathematical models of *M. tuberculosis* have been used to integrate various types of experimental data and suggest new hypotheses, mechanisms, and therapeutic approaches to tuberculosis [103].

The maintenance of diversity, be it at the species or genotypic level, is a fundamental problem in biology [12]. It has been suggested that the diversity of *M. tuberculosis* and the competition between its various strain types would affect the impact of public health intervention strategies [28]. In chapter two, we modelled *M. tuberculosis* growth and diversity in continuous culture under the action of isoniazid. We also obtained posterior parameter distributions that match experimental data and use these to explore the dynamics of diverse INH-resistant sub-populations in continuous culture. We find that the amount of resources available affects the level of competition between strains, which in turn affects the maintenance of diversity. The study did not include slow-growing persisters in response to antibiotic action but uses only simple mutation to be able to match experimental data. Even in a very simple, constrained and highly selective environment of nutrient-limited

continuous culture systems, *M. tuberculosis* can rapidly generate fit drug-resistant mutants that can establish long-term survival.

Over the years, models of within-host dynamics of TB infection have been developed and have provided insights to advance the study of this disease [103]. In chapter three, we study within host models of *M. tuberculosis*. Using tools from data science, the key parameters that discriminate between disease outcomes were identified. We find that when the strength of adaptive immunity is sufficiently high, infection can be controlled. Analysis of the two strain within host model showed that competition between strains is stronger during latency. During active disease, co-existence of the strains is more generic. This is in agreement with findings of chapter two, where more diversity is maintained under fast dilution (which is assumed to depict active disease [16]).

Individuals with latent tuberculosis can progress to active TB disease, in some cases decades after the initial infection [118]. Although the cause of this has been largely attributed to reactivation, exogenous re-infection together with primary infection have been shown to be the major causes of active tuberculosis [181, 23]. The results of the virtual experiments in chapter three show that reinfection can indeed lead to active disease especially in cases where the strain of the initial infection and the strain of the subsequent one do not elicit similar immunological host responses. The models in chapter three have the limitation of including only simple descriptions of the immune response. A natural future direction would be to include additional biological processes which provide a more realistic description of the host immune response to tuberculosis and allow better comparison with experimental data. In addition, these models could be modified to incorporate drug treatment, which would allow us investigate how competition of strains which differ in terms of drug susceptibility affect disease outcome.

For many datasets of infectious disease only a proportion of cases are observable and as such in the case of reconstructing a phylogenetic tree, there are prone to be errors. Also certain sites along a multiple sequence alignment may be well conserved and therefore suitable for phylogenetic analysis, whereas others are divergent, fast evolving, hyper-variable or prone to homoplasies and such sites can be removed before tree inference [151]. Our findings in chapter four suggest that when there is enough diversity in a

dataset to infer the evolutionary history of isolates, removing a single site does not affect the results in any substantial way. Also if the original multiple sequence alignment already yields phylogenetic trees with very high average bootstrap support, there is nothing to be gained from our method. In contrast to manual manipulation of alignments, reducing data sets using our method is transparent and easy to reproduce. Assuming that these sites are, at best, phylogenetically uninformative or, in the worst case, just misleading, we propose a new way of phylogenetic reconstruction that is based on minimizing the number of disruptive sites.

Sequence and epidemiological data contain information regarding the transmission process, as such analysing both kinds of data would yield more accurate estimates of epidemiological parameters than separate analyses [116]. In chapter five we apply this approach to elucidate local transmission from imported tuberculosis in Norway. This study was limited to the choice of prior model parameters. However, we carefully chose prior model parameters that typically depict a tuberculosis outbreak. Our findings show that there is substantial evidence of ongoing transmission in Norway. These findings imply that control efforts that focus on reducing the transmission of disease, especially in countries with low tuberculosis incidence, might have a more important role than previously appreciated. The work done in this chapter, presents a conceptually novel approach to tackle an important public health issue in countries with low tuberculosis incidence and aid the design of TB control measures.

The study of competition and diversity of *M. tuberculosis* offers the opportunity to increase our knowledge of the mechanisms and impact of strain interactions on disease dynamics. This would allow us to better project the trajectories of tuberculosis epidemics and the expected diversity of pathogen populations. This work provides some insight into the mechanisms that govern the outcome of infection, the impact of competition on the diversity of *M. tuberculosis* and contributes to our understanding of how genomic and epidemiological data can be combined to study infectious diseases.

REFERENCES

- [1] WHO | Global tuberculosis report 2017. *WHO*, 2017.
- [2] L. Aagesen. The information content of an ambiguously alignable region, a case study of the trnL intron from the Rhamnaceae. *Organisms Diversity & Evolution*, 4(1):35–49, 2004.
- [3] Z. Ahmad, L. G. Klinkenberg, M. L. Pinn, M. M. Fraig, C. A. Peloquin, W. R. Bishai, E. L. Nuermberger, J. H. Grosset, and P. C. Karakousis. Biphasic kill curve of isoniazid reveals the presence of drug tolerant, not drug resistant, *Mycobacterium tuberculosis* in the Guinea pig. *The Journal of Infectious Diseases*, 200(7):1136–1143, Oct 2009.
- [4] J.-W. Ai, Q.-L. Ruan, Q.-H. Liu, and W.-H. Zhang. Updates on the risk factors for latent tuberculosis reactivation and their managements. *Emerging microbes & infections*, 5(2):e10, feb 2016.
- [5] D. Alexei J. and B. Remco R. *Bayesian evolutionary analysis with BEAST*. Cambridge University Press, 2015.
- [6] P. Ankomah and R. B. Levin. Exploring the collaboration between antibiotics and the immune response in the treatment of acute self-limiting infections. *PNAS*, 111(23):8331–8338, 2014.
- [7] S. Arshad, L. Bavan, K. Gajari, S. N. J. Paget, and I. Baussano. Active screening at entry for tuberculosis among new immigrants: a systematic review and meta-analysis. *The European respiratory journal*, 35(6):1336–45, jun 2010.
- [8] D. Ayabina, C. Hendon-Dunn, J. Bacon, and C. Colijn. Diverse drug-resistant subpopulations of *Mycobacterium tuberculosis* are sustained in continuous culture. *Journal of the Royal Society, Interface*, 13(124):20160745, nov 2016.
- [9] J. Bacon, B. W. James, L. Wernisch, A. Williams, K. A. Morley, G. J. Hatch, J. A. Mangan, J. Hinds, N. G. Stoker, P. D. Butcher, and P. D.

- Marsh. The influence of reduced oxygen availability on pathogenicity and gene expression in *Mycobacterium tuberculosis*. *Tuberculosis (Edinburgh, Scotland)*, 84(3-4):205–17, 2004.
- [10] L. Baker, T. Brown, M. C. Maiden, and F. Drobniewski. Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis*. *Emerging infectious diseases*, 10(9):1568–77, sep 2004.
- [11] A. Bar-Hen, M. Mariadassou, M.-A. Poursat, and P. Vandenkoornhuyse. Influence Function for Robust Phylogenetic Reconstructions. *Molecular Biology and Evolution*, 25(5):869–873, feb 2008.
- [12] F. Bashey. Within-host competitive interactions as a mechanism for the maintenance of parasite diversity. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 370(1675), aug 2015.
- [13] E. D. Benavente, F. Coll, N. Furnham, R. McNerney, J. R. Glynn, S. Campino, A. Pain, F. R. Mohareb, T. G. Clark, N. Reiling, S. Homolka, K. Walter, J. Brandenburg, L. Niwinski, M. Ernst, F. Coll, R. McNerney, J. Guerra-Assuncao, J. Glynn, J. Perdigao, C. Ford, R. Shah, M. Maeda, S. Gagneux, M. Murray, T. Cohen, F. Coll, K. Mallard, M. Preston, S. Bentley, J. Parkhill, T. Clark, K. Mallard, F. Coll, M. Preston, J. Guerra-Assunção, R. Houben, A. Crampin, T. Mzembe, K. Mallard, F. Coll, P. Khan, L. Banda, A. Chiwaya, R. Pereira, R. McNerney, D. Harris, J. Parkhill, T. Clark, J. Glynn, A. Sandgren, M. Strong, P. Muthukrishnan, B. Weiner, G. Church, E. Chernyaeva, M. Shulgina, M. Rotkevich, P. Dobrynin, F. Coll, M. Preston, J. Guerra-Assuncao, J. Glynn, J. Perdigao, A. Wattam, D. Abraham, O. Dalay, T. Disz, M. Skinner, A. Uzilov, L. Stein, C. Mungall, I. Holmes, and H. Li. PhyTB: Phylogenetic tree visualisation and sample positioning for *M. tuberculosis*. *BMC Bioinformatics*, 16(1):155, dec 2015.
- [14] D. J. V. Beste, M. Espasa, B. Bonde, A. M. Kierzek, G. R. Stewart, and J. McFadden. The genetic requirements for fast and slow growth in *Mycobacteria*. *PloS One*, 4(4):e5349, Jan 2009.
- [15] D. J. V. Beste, J. Peters, T. Hooper, C. Avignone-Rossa, M. E. Bushell, and J. McFadden. Compiling a molecular inventory for *mycobacterium*

- bovis* bcg at two growth rates: Evidence for growth rate-mediated regulation of ribosome biosynthesis and lipid metabolism. *Journal of Bacteriology*, 187(5):1677–1684, 2004.
- [16] D. J. V. Beste, J. Peters, T. Hooper, C. Avignone-Rossa, M. E. Bushell, and J. McFadden. Compiling a Molecular Inventory for *Mycobacterium bovis* BCG at Two Growth Rates: Evidence for Growth Rate-Mediated Regulation of Ribosome Biosynthesis and Lipid Metabolism. *Journal of Bacteriology*, 187(5):1677–1684, feb 2005.
 - [17] J. C. Betts, P. T. Lukey, L. C. Robb, R. A. McAdam, and K. Duncan. Evaluation of a nutrient starvation model of *Mycobacterium tuberculosis* persistence by gene and protein expression profiling. *Molecular Microbiology*, 43(3):717–31, Feb 2002.
 - [18] K. Bhatt and P. Salgame. Host Innate Immune Response to *Mycobacterium tuberculosis*. *Journal of Clinical Immunology*, 27(4):347–362, jun 2007.
 - [19] M. Boots and A. Sasaki. ‘Small worlds’ and the evolution of virulence: infection occurs locally and at a distance. *Proceedings. Biological sciences*, 266(1432):1933–8, oct 1999.
 - [20] K. I. Bos, K. M. Harkins, A. Herbig, M. Coscolla, N. Weber, I. Comas, S. A. Forrest, J. M. Bryant, S. R. Harris, V. J. Schuenemann, T. J. Campbell, K. Majander, A. K. Wilbur, R. A. Guichon, D. L. Wolfe Steadman, D. C. Cook, S. Niemann, M. A. Behr, M. Zumarraga, R. Bastida, D. Huson, K. Nieselt, D. Young, J. Parkhill, J. E. Buikstra, S. Gagneux, A. C. Stone, and J. Krause. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature*, 514(7523):494–497, aug 2014.
 - [21] S. Capella-Gutierrez, J. M. Silla-Martinez, and T. Gabaldon. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–1973, aug 2009.
 - [22] M. Carballo and M. Mboup. International migration and health A paper prepared for the Policy Analysis and Research Programme of the Global Commission on International Migration. 2005.

- [23] P.-J. Cardona. Reactivation or reinfection in adult tuberculosis: Is that the question? *International Journal of Mycobacteriology*, 5(4):400–407, dec 2016.
- [24] J. Castresana. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution*, 17(4):540–52, apr 2000.
- [25] D. Chessel, A. B. Dufour, and J. Thioulouse. *R News*, 4, June 2004.
- [26] A. V. Chichurin and H. N. Shvychkina. Computer simulation of two chemostat models for one nutrient resource. *Mathematical Biosciences*, 278:30–36, aug 2016.
- [27] J. C. Cho and S. J. Kim. Increase in bacterial community diversity in subsurface aquifers receiving livestock wastewater input. *Applied and environmental microbiology*, 66(3):956–65, mar 2000.
- [28] T. Cohen, C. Colijn, and M. Murray. Modeling the effects of strain diversity and mechanisms of strain competition on the potential performance of new tuberculosis vaccines. *Proceedings of the National Academy of Sciences of the United States of America*, 105(42):16302–7, oct 2008.
- [29] T. Cohen, P. D. van Helden, D. Wilson, C. Colijn, M. M. McLaughlin, I. Abubakar, and R. M. Warren. Mixed-strain mycobacterium tuberculosis infections and the implications for tuberculosis treatment and control. *Clinical Microbiology Reviews*, 25(4):708–19, Oct 2012.
- [30] T. Cohen, D. Wilson, K. Wallengren, E. Y. Samuel, and M. Murray. Mixed-strain Mycobacterium tuberculosis infections among patients dying in a hospital in KwaZulu-Natal, South Africa. *Journal of clinical microbiology*, 49(1):385–8, jan 2011.
- [31] S. R. Cole, H. Chu, and S. Greenland. Maximum likelihood, profile likelihood, and penalized likelihood: a primer. *American journal of epidemiology*, 179(2):252–60, jan 2014.
- [32] C. Colijn and T. Cohen. How competition governs whether moderate or aggressive treatment minimizes antibiotic resistance. *eLife*, 4:e10559, sep 2015.

- [33] C. Colijn, T. Cohen, A. Ganesh, and M. Murray. Spontaneous emergence of multiple drug resistance in tuberculosis before and during therapy. *PLoS One*, 6(3):e18327, 30 Mar. 2011.
- [34] C. Colijn, T. Cohen, and M. Murray. Latent coinfection and the maintenance of strain diversity. *Bulletin of mathematical biology*, 71(1):247–63, jan 2009.
- [35] G. W. Comstock. Epidemiology of Tuberculosis 13.
- [36] C. Contini, M. Maritati, M. di Nuzzo, L. Massoli, S. Lomenzo, and A. Grilli. The impact of tuberculosis among immigrants: Epidemiology and strategies of control in high-income countries—current data and literature review. In *People’s Movements in the 21st Century - Risks, Challenges and Benefits*. InTech, 2017.
- [37] M. Coscolla and S. Gagneux. Does M. tuberculosis genomic diversity explain disease diversity? *Drug discovery today. Disease mechanisms*, 7(1):e43–e59, 2010.
- [38] E. M. Cottam, D. T. Haydon, D. J. Paton, J. Gloster, J. W. Wilesmith, N. P. Ferris, G. H. Hutchings, and D. P. King. Molecular epidemiology of the foot-and-mouth disease virus outbreak in the United Kingdom in 2001. *Journal of virology*, 80(22):11274–82, nov 2006.
- [39] E. M. Cottam, G. Thebaud, 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. *Proceedings of the Royal Society B: Biological Sciences*, 275(1637):887–895, apr 2008.
- [40] U. R. Dahle, V. Eldholm, B. A. Winje, T. Mannsåker, and E. Heldal. Impact of Immigration on the Molecular Epidemiology of *Mycobacterium tuberculosis* in a Low-Incidence Country. *American Journal of Respiratory and Critical Care Medicine*, 176(9):930–935, nov 2007.
- [41] U. R. Dahle, P. Sandven, E. Heldal, and D. A. Caugant. Continued low rates of transmission of *Mycobacterium tuberculosis* in Norway. *Journal of clinical microbiology*, 41(7):2968–73, jul 2003.

- [42] M. H. Dangisso, D. G. Datiko, and B. Lindtjørn. Trends of tuberculosis case notification and treatment outcomes in the Sidama Zone, southern Ethiopia: ten-year retrospective trend analysis in urban-rural settings. *PloS one*, 9(12):e114225, 2014.
- [43] H. L. David. Probability distribution of drug-resistant mutants in unselected populations of *Mycobacterium tuberculosis*. *Applied microbiology*, 20(5):810–4, nov 1970.
- [44] N. De Maio, C.-H. Wu, D. J. Wilson, S. Gaudieri, S. Pham, and A. Chopra. SCOTTI: Efficient Reconstruction of Transmission within Outbreaks with the Structured Coalescent. *PLOS Computational Biology*, 12(9):e1005130, sep 2016.
- [45] F. Delsuc, H. Brinkmann, and H. Philippe. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, 6(5):361–375, may 2005.
- [46] K. Dialdestoro, J. A. Sibbesen, L. Maretty, J. Raghwani, A. Gall, P. Kellam, O. G. Pybus, J. Hein, and P. A. Jenkins. Coalescent Inference Using Serially Sampled, High-Throughput Sequencing Data from Intrahost HIV Infection. *Genetics*, 202(4):1449–1472, apr 2016.
- [47] K. R. Dickman, L. Nabyonga, D. P. Kateete, F. A. Katabazi, B. B. Asiimwe, H. K. Mayanja, A. Okwera, C. Whalen, and M. L. Joloba. Detection of multiple strains of *Mycobacterium tuberculosis* using MIRU-VNTR in patients with pulmonary tuberculosis in Kampala, Uganda. *BMC Infectious Diseases*, 10(1):349, dec 2010.
- [48] X. Didelot, C. Fraser, J. Gardy, and C. Colijn. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular Biology and Evolution*, 195(4):msw075, jan 2017.
- [49] X. Didelot, J. Gardy, and C. Colijn. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Molecular biology and evolution*, 31(7):1869–79, jul 2014.
- [50] U. Dobrindt. Whole genome plasticity in pathogenic bacteria. *Current Opinion in Microbiology*, 4(5):550–557, oct 2001.

- [51] A. J. Drummond, M. A. Suchard, D. Xie, and A. Rambaut. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29(8):1969–1973, aug 2012.
- [52] J. Dunbar, S. Takala, S. M. Barns, J. A. Davis, and C. R. Kuske. Levels of bacterial community diversity in four arid soils compared by cultivation and 16S rRNA gene cloning. *Applied and environmental microbiology*, 65(4):1662–9, apr 1999.
- [53] D. E. Dykhuizen and D. L. Hartl. Selection in chemostats. *Microbiological Reviews*, 47(2):150–68, June 1983.
- [54] S. V. Edwards. IS A NEW AND GENERAL THEORY OF MOLECULAR SYSTEMATICS EMERGING? *Evolution*, 63(1):1–19, jan 2009.
- [55] B. Efron, E. Halloran, and S. Holmes. Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences*, 93(23):13429–13429, nov 1996.
- [56] M. El-Khairi and Muna. Mathematical Modelling of Tuberculosis Infection. 2013.
- [57] V. Eldholm, J. H.-O. Pettersson, O. B. Brynildsrud, A. Kitchen, E. M. Rasmussen, T. Lillebaek, J. O. Rønning, V. Crudu, A. T. Mengshoel, N. Debech, K. Alfsnes, J. Bohlin, C. S. Pepperell, and F. Balloux. Armed conflict and population displacement as drivers of the evolution and dispersal of Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences of the United States of America*, 113(48):13881–13886, nov 2016.
- [58] H. Esmail, C. E. Barry, D. B. Young, R. J. Wilkinson, and R. J. Wilkinson. The ongoing challenge of latent tuberculosis. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 369(1645):20130437, 2014.
- [59] K. Fangen. Humiliation Experienced by Somali Refugees in Norway. *Journal of Refugee Studies*, 19(1):69–93, mar 2006.
- [60] M. Farah, H. Meyer, R. Selmer, E. Heldal, and G. Bjune. Long-term risk of tuberculosis among immigrants in Norway. *International Journal of Epidemiology*, 34(5):1005–1011, jul 2005.

- [61] J. Felsenstein. CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP. *Evolution*, 39(4):783–791, jul 1985.
- [62] S. Fishbein, N. van Wyk, R. M. Warren, and S. L. Sampson. Phylogeny to function: PE/PPE protein evolution and impact on *Mycobacterium tuberculosis* pathogenicity. *Molecular Microbiology*, 96(5):901–916, jun 2015.
- [63] H. A. Fletcher and L. Schrager. TB vaccine development and the End TB Strategy: importance and current status. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, 110(4):212–218, apr 2016.
- [64] J. L. Flynn and J. Chan. IMMUNOLOGY OF TUBERCULOSIS. *Annual Review of Immunology*, 19(1):93–129, apr 2001.
- [65] C. B. Ford, P. L. Lin, M. R. Chase, R. R. Shah, O. Iartchouk, J. 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*, 43(5):482–6, May 2011.
- [66] S. Gagneux. Fitness cost of drug resistance in *Mycobacterium tuberculosis*. *Clinical Microbiology and Infection*, 15:66–68, jan 2009.
- [67] S. Gagneux and P. M. Small. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. *The Lancet. Infectious diseases*, 7(5):328–37, may 2007.
- [68] J. E. Galagan. Genomic insights into tuberculosis. *Nature Reviews Genetics*, 15(5):307–320, mar 2014.
- [69] D. Gammack, C. Doering, and D. Kirschner. Macrophage response to *Mycobacterium tuberculosis* infection. *Journal of Mathematical Biology*, 48(2):218–242, feb 2004.
- [70] L. Garamszegi. *Modern comparative and their application in evolutionary Biology*. Springer, 2014.

- [71] D. García de Viedma, M. Marín, S. Hernangómez, M. Díaz, M. J. Ruiz Serrano, L. Alcalá, and E. Bouza. Tuberculosis recurrences: reinfection plays a role in a population whose clinical/epidemiological characteristics do not favor reinfection. *Archives of internal medicine*, 162(16):1873–9, sep 2002.
- [72] S. H. Gillespie. Evolution of drug resistance in *Mycobacterium tuberculosis*: clinical and molecular perspective. *Antimicrobial Agents and Chemotherapy*, 46(2):267–274, Feb 2002.
- [73] A. Gold, T. Giraud, and M. E. Hood. Within-host competitive exclusion among species of the anther smut pathogen. *BMC ecology*, 9:11, may 2009.
- [74] M. G. M. Gomes, M. L. Barreto, P. Glaziou, G. F. Medley, L. C. Rodrigues, J. Wallinga, and S. B. Squire. End TB strategy: the need to reduce risk inequalities. *BMC Infectious Diseases*, 16(1):132, dec 2016.
- [75] L. Grandjean, R. H. Gilman, T. Iwamoto, C. U. Köser, J. Coronel, M. Zimic, M. E. Török, D. Ayabina, M. Kendall, C. Fraser, S. Harris, J. Parkhill, S. J. Peacock, D. A. J. Moore, and C. Colijn. Convergent evolution and topologically disruptive polymorphisms among multidrug-resistant tuberculosis in Peru. *PLOS ONE*, 12(12):e0189838, dec 2017.
- [76] B. T. Grenfell, O. G. Pybus, J. R. Gog, J. L. N. Wood, J. M. Daly, J. A. Mumford, and E. C. Holmes. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science*, 303(5656):327–332, jan 2004.
- [77] W. N. Grundy and G. J. Naylor. Phylogenetic inference from conserved sites alignments. *Journal of Experimental Zoology*, 285(2):128–139, aug 1999.
- [78] J. Guerra-Assunção, A. Crampin, R. Houben, T. Mzembe, K. Mallard, F. Coll, P. Khan, L. Banda, A. Chiwaya, R. Pereira, R. McNerney, P. Fine, J. Parkhill, T. Clark, and J. Glynn. Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *eLife*, 4, mar 2015.

- [79] S. Guindon, J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59(3):307–321, may 2010.
- [80] T. Gumbo, A. Louie, W. Liu, P. G. Ambrose, S. M. Bhavnani, D. Brown, and G. L. Drusano. Isoniazid’s bacterial activity ceases because of the emergence of resistance not depletion of mycobacterium tuberculosis in the log phase of growth. *Journal of Infectious Diseases*, 195(3):194–201, 2007.
- [81] S. Gupta and M. C. Maiden. Exploring the evolution of diversity in pathogen populations. *Trends in Microbiology*, 9(4):181–185, apr 2001.
- [82] B. R. Guzman Herrador, K. Rønning, K. Borgen, T. Mannsåker, and U. R. Dahle. Description of the largest cluster of tuberculosis notified in Norway 1997–2011: is the Norwegian tuberculosis control programme serving its purpose for high risk groups? *BMC Public Health*, 15(1):367, Dec 2015.
- [83] H. Haario, M. Laine, A. Mira, and E. Saksman. DRAM: Efficient adaptive MCMC. *Statistics and Computing*, 16(4):339–354, Dec 2006.
- [84] M. Hall, M. Woolhouse, and A. Rambaut. Epidemic reconstruction in a phylogenetics framework: Transmission trees as partitions of the node set. *PLoS Comput. Biol.*, 11(12):e1004613, Dec. 2015.
- [85] T. Hampshire, S. Soneji, J. Bacon, B. W. James, J. Hinds, K. Laing, R. A. Stabler, P. D. Marsh, and P. D. Butcher. Stationary phase gene expression of Mycobacterium tuberculosis following a progressive nutrient depletion: a model for persistent organisms? *Tuberculosis (Edinburgh, Scotland)*, 84(3-4):228–38, Jan 2004.
- [86] K. E. Hines, T. R. Middelndorf, and R. W. Aldrich. Determination of parameter identifiability in nonlinear biophysical models: A Bayesian approach. *The Journal of General Physiology*, 143(3):401–416, mar 2014.
- [87] S. B. Hsu, S. Hubbell, and P. Waltman. A mathematical theory for single-nutrient competition in continuous cultures of micro-organisms. *SIAM Journal on Applied Mathematics*, 32(2):366–383, Mar 1977.

- [88] S. B. Hsu, S. Hubbell, and P. Waltman. A Mathematical Theory for Single-Nutrient Competition in Continuous Cultures of Micro-Organisms. *SIAM Journal on Applied Mathematics*, jul 2006.
- [89] M. Imani and H. Ghassemian. Principal component discriminant analysis for feature extraction and classification of hyperspectral images. In *2014 Iranian Conference on Intelligent Systems (ICIS)*, pages 1–5. IEEE, feb 2014.
- [90] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning with applications in R*. Springer, New York, 2008.
- [91] R. E. Jeeves, A. A. N. Marriott, S. T. Pullan, K. A. Hatch, J. C. Allnut, I. Freire-Martin, C. L. Hendon-Dunn, R. Watson, A. A. Witney, R. H. Tyler, C. Arnold, P. D. Marsh, T. D. McHugh, and J. Bacon. Mycobacterium tuberculosis Is resistant to isoniazid at a slow growth rate by single nucleotide polymorphisms in katG codon Ser315. *PloS One*, 10(9):e0138253, 2015.
- [92] C. Jenkins, J. Bacon, and J. Allnut. Enhanced heterogeneity of rpob in *Mycobacterium tuberculosis* found at low ph. *Journal of Antimicrobial Chemotherapy*, 63:1118–20, 2009.
- [93] A. Jindani, V. R. Aber, E. A. Edwards, and D. A. Mitchison. The early bactericidal activity of drugs in patients with pulmonary tuberculosis. *Am Rev Respir Dis*, 121(6):939–49, 1980.
- [94] T. Jombart, A. Cori, X. Didelot, S. Cauchemez, C. Fraser, N. Ferguson, D. Haydon, M. Chase-Topping, D. Shaw, L. Matthews, J. Friar, S. Cauchemez, P. Boelle, C. Donnelly, N. Ferguson, G. Thomas, J. Wallinga, M. Lipsitch, S. Cauchemez, N. Ferguson, S. Cauchemez, A. Bhattarai, T. Marchbanks, R. Fagan, S. Ostroff, J. Heijne, M. Rondy, L. Verhoef, J. Wallinga, M. Kretzschmar, J. Wallinga, P. Teunis, N. Ferguson, C. Donnelly, R. Anderson, C. Fraser, C. Donnelly, S. Cauchemez, W. Hanage, M. V. Kerkhove, Y. Ruan, C. Wei, A. Ee, V. Vega, H. Thoreau, J. Liu, S. Lim, Y. Ruan, A. Ling, L. Ng, S. Harris, E. Feil, M. Holden, M. Quail, E. Nickerson, U. Nübel, J. Dordel, K. Kurt, B. Strommenger, H. Westh, A. Mutreja, D. Kim, N. Thomson, T. Connor, J. Lee, E. Snitkin, A. Zelazny,

P. Thomas, F. Stock, T. Walker, C. Ip, R. Harrell, J. Evans, G. Kapatai, O. Pybus, A. Rambaut, X. Didelot, R. Bowden, D. Wilson, T. Peto, D. Crook, C. Koser, M. Holden, M. Ellington, E. Cartwright, N. Brown, D. Eyre, T. Golubchik, N. Gordon, R. Bowden, P. Piazza, E. Cottam, G. Thebaud, J. Wadsworth, J. Gloster, L. Mansley, T. Jombart, R. Eggo, P. Dodd, F. Balloux, R. Ypma, A. Bataille, A. Stegeman, G. Koch, J. Wallinga, M. Morelli, G. Thebaud, J. Chadoeuf, D. King, D. Haydon, V. Vega, Y. Ruan, J. Liu, W. Lee, C. Wei, M. Lipsitch, T. Cohen, B. Cooper, J. Robins, S. Ma, S. Cauchemez, P. Boelle, G. Thomas, A. Valleron, A. Cori, N. Ferguson, C. Fraser, S. Cauchemez, O. Pybus, M. Suchard, P. Lemey, F. Bernardin, A. Rambaut, E. Volz, J. Koopman, M. Ward, A. Brown, S. Frost, D. Rasmussen, O. Ratmann, K. Koelle, B. Young, T. Golubchik, E. Batty, R. Fung, H. Larner-Svensson, S. Cauchemez, L. Temime, D. Guillemot, E. Varon, A. Valleron, S. Cauchemez, L. Temime, A. Valleron, E. Varon, G. Thomas, S. Harris, E. Cartwright, M. Torok, M. Holden, N. Brown, C. Varin, N. Reid, D. Firth, N. Hens, L. Calatayud, S. Kurkela, T. Tamm, J. Wallinga, and R. Edgar. Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLoS Computational Biology*, 10(1):e1003457, jan 2014.

- [95] T. Jombart, S. Devillard, and F. Balloux. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, 11(1):94, oct 2010.
- [96] T. Jombart, R. M. Eggo, P. J. Dodd, and F. Balloux. Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity*, 106(2):383–390, feb 2011.
- [97] T. Jombart, M. Kendall, J. Almagro-Garcia, and C. Colijn. treespace : Statistical exploration of landscapes of phylogenetic trees. *Molecular Ecology Resources*, 17(6):1385–1392, nov 2017.
- [98] E. Kenah, T. Britton, M. E. Halloran, I. M. Longini, T. Leitner, and V. de Jager. Molecular Infectious Disease Epidemiology: Survival Analysis and Algorithms Linking Phylogenies to Transmission Trees. *PLOS Computational Biology*, 12(4):e1004869, apr 2016.

- [99] M. Kendall, D. Ayabina, and C. Colijn. Estimating transmission from genetic and epidemiological data: a metric to compare transmission trees. sep 2016.
- [100] M. Kendall, D. Ayabina, Y. Xu, J. Stimson, and C. Colijn. Estimating Transmission from Genetic and Epidemiological Data: A Metric to Compare Transmission Trees. *Statistical Science*, 33(1):70–85, feb 2018.
- [101] M. Kendall and C. Colijn. Mapping Phylogenetic Trees to Reveal Distinct Patterns of Evolution. *Molecular biology and evolution*, 33(10):2735–43, oct 2016.
- [102] F. Képès. *Biological Networks*, volume 3 of *Complex Systems and Interdisciplinary Science*. WORLD SCIENTIFIC, dec 2007.
- [103] D. Kirschner, E. Pienaar, S. Marino, and J. J. Linderman. A review of computational and mathematical modeling contributions to our understanding of Mycobacterium tuberculosis within-host infection and treatment. *Current Opinion in Systems Biology*, 3:170–185, jun 2017.
- [104] D. E. Kirschner, D. Young, and J. L. Flynn. Tuberculosis: global approaches to a global disease. *Current Opinion in Biotechnology*, 21(4):524–531, aug 2010.
- [105] D. Klinkenberg, J. A. Backer, X. Didelot, C. Colijn, J. Wallinga, and D. Haydon. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLOS Computational Biology*, 13(5):e1005495, may 2017.
- [106] K. Kovárová-Kovar and T. Egli. Growth kinetics of suspended microbial cells: from single-substrate-controlled growth to mixed-substrate kinetics. *Microbiology and molecular biology reviews : MMBR*, 62(3):646–66, sep 1998.
- [107] A. Kuwata and T. Miyazaki. Effects of ammonium supply rates on competition between *Microcystis novacekii* (Cyanobacteria) and *Scenedesmus quadricauda* (Chlorophyta): simulation study. *Ecological Modelling*, 135(1):81–87, Nov 2000.

- [108] V. A. Kuznetsov, I. A. Makalkin, M. A. Taylor, and A. S. Perelson. Nonlinear dynamics of immunogenic tumors: Parameter estimation and global bifurcation analysis. *Bulletin of Mathematical Biology*, 56(2):295–321, mar 1994.
- [109] Y. A. Kuznetsov and C. Piccardi. Bifurcation analysis of periodic SEIR and SIR epidemic models. *Journal of Mathematical Biology*, 32(2):109–121, jan 1994.
- [110] M. S. Y. Lau, G. Marion, G. Streftaris, G. Gibson, M. Chase-Topping, and D. Haydon. A Systematic Bayesian Integration of Epidemiological and Genetic Data. *PLOS Computational Biology*, 11(11):e1004633, nov 2015.
- [111] M. S. Lee. Unalignable sequences and molecular evolution. *Trends in Ecology & Evolution*, 16(12):681–685, 2001.
- [112] M. S. Lee and A. Palci. Morphological Phylogenetics in the Genomic Age. *Current Biology*, 25(19):R922–R929, oct 2015.
- [113] P. D. Leenheer, B. Li, and L. S. Hal. Competition in the chemostat: some remarks.
- [114] L. Levin, I. Zhidkov, Y. Gurman, H. Hawlena, and D. Mishmar. Functional recurrent mutations in the human mitochondrial phylogeny: dual roles in evolution and disease. *Genome biology and evolution*, 5(5):876–90, 2013.
- [115] J. M. Lew, A. Kapopoulou, L. M. Jones, and S. T. Cole. Tuberculist - 10 years after. *Tuberculosis*, 91(1), 2011.
- [116] L. M. Li, N. C. Grassly, and C. Fraser. Genomic analysis of emerging pathogens: methods, application and future trends. *Genome Biology*, 15(11):541, nov 2014.
- [117] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.
- [118] T. Lillebaek, A. Dirksen, I. Baess, B. Strunge, V. Ø. Thomsen, and Å. B. Andersen. Molecular Evidence of Endogenous Reactivation of *Mycobacterium tuberculosis* after 33 Years of Latent Infection. *The Journal of Infectious Diseases*, 185(3):401–404, feb 2002.

- [119] M. Lipsitch, C. Colijn, T. Cohen, W. P. Hanage, and C. Fraser. No coexistence for free: neutral null models for multistrain pathogens. *Epidemics*, 1(1):2–13, mar 2009.
- [120] J. LOBRY, J. FLANDROIS, G. CARRET, and A. PAVE. Monod’s bacterial growth model revisited. *Bulletin of Mathematical Biology*, 54(1):117–122, 1992.
- [121] A. Löytynoja and M. C. Milinkovitch. SOAP, cleaning multiple alignments from unstable blocks. *Bioinformatics (Oxford, England)*, 17(6):573–4, jun 2001.
- [122] 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*, 94(4):434–440, jul 2014.
- [123] M. D. Madhukar Pai, Marcel A.Behr, David Dowdy, Keertan Dheda, M. S. Catharina C.Boehme, AnnGinsberg, Soumya Swaminathan7, D. M. Haileyesus Getahun, and MarioRaviglione. Tuberculosis. *Nature Reviews*, 2, 2016.
- [124] R. P. Maharjan, T. Ferenci, P. R. Reeves, Y. Li, B. Liu, and L. Wang. The multiplicity of divergence mechanisms in a single evolving population. *Genome Biology*, 13(6):R41, Jan. 2012.
- [125] C. Manca, M. B. Reed, S. Freeman, B. Mathema, B. Kreiswirth, C. E. Barry, and G. Kaplan. Differential Monocyte Activation Underlies Strain-Specific Mycobacterium tuberculosis Pathogenesis. *Infection and Immunity*, 72(9):5511–5514, sep 2004.
- [126] S. Marino and D. E. Kirschner. The human immune response to Mycobacterium tuberculosis in lung and lymph node. *Journal of Theoretical Biology*, 227(4):463–486, apr 2004.
- [127] M. A. Marra, S. J. M. Jones, C. R. Astell, R. A. Holt, A. Brooks-Wilson, Y. S. N. Butterfield, J. Khattra, J. K. Asano, S. A. Barber, S. Y. Chan, A. Cloutier, S. M. Coughlin, D. Freeman, N. Girn, O. L. Griffith, S. R. Leach, M. Mayo, H. McDonald, S. B. Montgomery, P. K.

- Pandoh, A. S. Petrescu, A. G. Robertson, J. E. Schein, A. Siddiqui, D. E. Smailus, J. M. Stott, G. S. Yang, F. Plummer, A. Andonov, H. Artsob, N. Bastien, K. Bernard, T. F. Booth, D. Bowness, M. Czub, M. Drebot, L. Fernando, R. Flick, M. Garbutt, M. Gray, A. Grolla, S. Jones, H. Feldmann, A. Meyers, A. Kabani, Y. Li, S. Normand, U. Stroher, G. A. Tipples, S. Tyler, R. Vogrig, D. Ward, B. Watson, R. C. Brunham, M. Krajden, M. Petric, D. M. Skowronski, C. Upton, and R. L. Roper. The Genome Sequence of the SARS-Associated Coronavirus. *Science*, 300(5624), 2003.
- [128] J. P. Mather. In vitro models. *STEM CELLS*, 30(2):95–99, feb 2012.
- [129] A. E. McCaig, L. A. Glover, and J. I. Prosser. Molecular analysis of bacterial community structure and diversity in unimproved and improved upland grass pastures. *Applied and environmental microbiology*, 65(4):1721–30, apr 1999.
- [130] C. R. E. McEvoy, R. Cloete, B. Müller, A. C. Schürch, P. D. van Helden, S. Gagneux, R. M. Warren, and N. C. Gey van Pittius. Comparative analysis of *Mycobacterium tuberculosis* *pe* and *ppe* genes reveals high sequence variation and an apparent absence of selective constraints. *PloS one*, 7(4):e30593, 2012.
- [131] J. McFadden, D. J. Beste, and A. M. Kierzek. Systems biology of tuberculosis. 2013.
- [132] J. Mears, I. Abubakar, T. Cohen, T. D. McHugh, P. Sonnenberg, N. G. Stoker, P. D. Butcher, and T. D. McHugh. Effect of study design and setting on tuberculosis clustering estimates using *Mycobacterium* Interspersed Repetitive Units-Variable Number Tandem Repeats (MIRU-VNTR): a systematic review. *BMJ Open*, 5(1):e005636–e005636, jan 2015.
- [133] M. Merker, C. Blin, S. Mona, N. Duforet-Frebourg, S. Lecher, E. Willery, M. G. B. Blum, S. Rüsch-Gerdes, I. Mokrousov, E. Alek-sic, C. Allix-Béguec, A. Antierens, E. Augustynowicz-Kopeć, M. Bal-lif, F. Barletta, H. P. Beck, C. E. Barry, M. Bonnet, E. Borroni, I. Campos-Herrero, D. Cirillo, H. Cox, S. Crowe, V. Crudu, R. Diel, F. Drobniewski, M. Fauville-Dufaux, S. Gagneux, S. Ghebremichael,

- M. Hanekom, S. Hoffner, W.-w. Jiao, S. Kalon, T. A. Kohl, I. Kontse-
vaya, T. Lillebæk, S. Maeda, V. Nikolayevskyy, M. Rasmussen, N. Ras-
togi, S. Samper, E. Sanchez-Padilla, B. Savic, I. C. Shamputa, A. Shen,
L.-H. Sng, P. Stakenas, K. Toit, F. Varaine, D. Vukovic, C. Wahl,
R. Warren, P. Supply, S. Niemann, and T. Wirth. Evolutionary history
and global spread of the *Mycobacterium tuberculosis* Beijing lineage.
Nature Genetics, 47(3):242–249, jan 2015.
- [134] L. P. Miller, J. T. Crawford, and T. M. Shinnick. The *rpoB* gene of
Mycobacterium tuberculosis. *Antimicrobial agents and chemotherapy*,
38(4):805–11, apr 1994.
- [135] S. R. Moaddab, S. Farajnia, D. Kardan, S. Zamanlou, and M. Y.
Alikhani. Isoniazid MIC and KatG Gene Mutations among *Mycobac-*
terium tuberculosis Isolates in Northwest of Iran. *Iranian journal of*
basic medical sciences, 14(6):540–5, nov 2011.
- [136] K. Mohareer, S. Tundup, and S. E. Hasnain. Transcriptional regulation
of *Mycobacterium tuberculosis* PE/PPE genes: a molecular switch to
virulence? *Journal of molecular microbiology and biotechnology*, 21(3-
4):97–109, 2011.
- [137] N. Mollentze, L. H. Nel, S. Townsend, K. le Roux, K. Hampson, D. T.
Haydon, and S. Soubeyrand. A Bayesian approach for inferring the
dynamics of partially observed endemic infectious diseases from space-
time-genetic data. *Proceedings of the Royal Society of London B: Bi-*
ological Sciences, 281(1782), 2014.
- [138] 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, 8(11):e1002768, nov 2012.
- [139] R. Müller, C. A. Roberts, and T. A. Brown. Genotyping of ancient
Mycobacterium tuberculosis strains reveals historic genetic diversity.
Proceedings. Biological sciences, 281(1781):20133236, apr 2014.
- [140] B. Nicholas H., B. Derek E.G., E. Jonathan A., G. David B., and
P. Nipam H. *Evolution*. Cold Spring Harbor, New York, New York,
USA, 2007.

- [141] S. Niemann, E. Richter, S. Rüscher-Gerdes, M. Schlaak, and U. Greinert. Double infection with a resistant and a multidrug-resistant strain of *Mycobacterium tuberculosis*. *Emerging infectious diseases*, 6(5):548–51, 2000.
- [142] E. Numminen, C. Chewapreecha, J. Sirén, C. Turner, P. Turner, S. D. Bentley, and J. Corander. Two-phase importance sampling for inference about transmission trees. *Proceedings of the Royal Society of London B: Biological Sciences*, 281(1794), 2014.
- [143] M. B. O’Neill, T. D. Mortimer, and C. S. Pepperell. Diversity of *Mycobacterium tuberculosis* across Evolutionary Scales. *PLoS pathogens*, 11(11):e1005257, 2015.
- [144] R. S. Ostfeld and F. Keesing. Effects of Host Diversity on Infectious Disease. *Annual Review of Ecology, Evolution, and Systematics*, 43(1):157–182, dec 2012.
- [145] L. P. Ling, K. Denise, and F. JoAnne L. Modeling pathogen and host: in vitro, in vivo and in silico models of latent *Mycobacterium tuberculosis* infection. *Drug Discovery Today: Disease Models*, 2(2):149–154, jun 2005.
- [146] G. S. Palanisamy, E. E. Smith, C. A. Shanley, D. J. Ordway, I. M. Orme, and R. J. Basaraba. Disseminated disease severity as a measure of virulence of *Mycobacterium tuberculosis* in the guinea pig model. *Tuberculosis (Edinburgh, Scotland)*, 88(4):295–306, jul 2008.
- [147] M. Pareek, C. Greenaway, T. Noori, J. Munoz, and D. Zenner. The impact of migration on tuberculosis epidemiology and control in high-income countries: a review. *BMC Medicine*, 14(1):48, dec 2016.
- [148] T. Parish and A. Brown. *Mycobacterium: Genomics and molecular Biology*. Caister Academic Press, Norfolk, UK, 2009.
- [149] C. S. Pepperell, A. M. Casto, A. Kitchen, J. M. Granka, O. E. Cornejo, E. C. Holmes, B. Birren, J. Galagan, M. W. Feldman, M. Achtman, R. Hershberg, M. Lipatov, P. Small, H. Sheffer, S. Niemann, Z. Udawadia, R. Amale, K. Ajbani, C. Rodrigues, S. Gagneux, K. DeRiemer, T. Van, M. Kato-Maeda, B. de Jong, A. Boyko, S. Williamson,

A. Indap, J. Degenhardt, R. Hernandez, B. Charlesworth, M. Morgan, D. Charlesworth, D. Charlesworth, B. Charlesworth, M. Morgan, Y. Fu, E. Santiago, A. Caballero, I. Gordo, A. Navarro, B. Charlesworth, S. Williamson, M. Orive, K. Zeng, Y. Fu, S. Shi, C. Wu, A. Hirsh, A. Tsolaki, K. DeRiemer, M. Feldman, P. Small, N. Smith, R. Hewinson, K. Kremer, R. Brosch, S. Gordon, P. Supply, R. Warren, A. Banuls, S. Lesjean, G. V. D. Spuy, T. Wirth, F. Hildebrand, C. Allix-Beguec, F. Wolbeling, T. Kubica, I. Comas, J. Chakravartti, P. Small, J. Galagan, S. Niemann, C. Sasseti, E. Rubin, C. Sasseti, D. Boyd, E. Rubin, K. Balbi, E. Rocha, E. Feil, C. Kuo, N. Moran, H. Ochman, E. Rocha, J. Smith, L. Hurst, M. Holden, J. Cooper, S. Kryazhimskiy, J. Plotkin, C. Pepperell, V. Hoepfner, M. Lipatov, W. Wobeser, G. Schoolnik, C. Pepperell, J. Granka, D. Alexander, M. Behr, L. Chui, L. Fenner, M. Egger, T. Bodmer, H. Furrer, M. Ballif, J. Tennesen, A. Bigham, T. O'Connor, W. Fu, E. Kenny, M. Nelson, D. Wegmann, M. Ehm, D. Kessner, P. S. Jean, A. Namouchi, X. Didelot, U. Schock, B. Gicquel, E. Rocha, P. Hadrill, L. Loewe, B. Charlesworth, O. Cornejo, T. Lefebvre, P. P. Bitar, P. Lang, V. Richards, N. Smith, S. Gordon, R. de la Rua-Domenech, R. Clifton-Hadley, R. Hewinson, I. Comas, S. Gagneux, M. Gutierrez, S. Brisse, R. Brosch, M. Fabre, B. Omais, J. Becq, M. Gutierrez, V. Rosas-Magallanes, J. Rauzier, B. Gicquel, F. Veyrier, D. Pletzer, C. Turenne, M. Behr, X. Liu, M. Gutacker, J. Musser, Y. Fu, G. McVean, B. Charlesworth, B. Charlesworth, D. Stucki, S. Gagneux, P. Supply, M. Marceau, S. Mangenot, D. Roche, C. Rouanet, E. Elyashiv, K. Bullaughey, S. Sattath, Y. Rinott, M. Przeworski, P. A. D. Silva, J. Palomino, F. Wang, R. Langley, G. Gulten, L. Dover, G. Besra, T. Reddy, R. Riley, F. Wymore, P. Montgomery, D. DeCaprio, M. Pasca, P. Guglielame, F. Arcesi, M. Bellinzoni, E. D. Rossi, B. Choudhuri, S. Bhakta, R. Barik, J. Basu, M. Kundu, G. Rodriguez, I. Smith, A. Farhana, S. Kumar, S. Rathore, P. Ghosh, N. Ehtesham, G. Stewart, J. Patel, B. Robertson, A. Rae, D. Young, L. Camacho, P. Constant, C. Raynaud, M. Laneele, J. Triccas, I. Cheeseman, B. Miller, S. Nair, S. Nkhoma, A. Tan, T. Mixson-Hayden, V. Jain, A. McCollum, A. Poe, A. Nagpal, N. Osorio, F. Rodrigues, S. Gagneux, J. Pedrosa, M. Pinto-Carbo, K. Adams,

K. Takaki, L. Connolly, H. Wiedenhof, K. Winglee, L. Kennemann, X. Didelot, T. Aebischer, S. Kuhn, B. Drescher, T. Andrews, T. Gojobori, S. Chen, C. Hung, J. Xu, C. Reigstad, V. Magrini, D. Fitzpatrick, J. McInerney, F. Jiggins, L. Petersen, J. Bollback, M. Dimmic, M. Hubisz, R. Nielsen, R. Urwin, E. Holmes, A. Fox, J. Derrick, M. Maiden, C. Tarr, T. Whittam, M. Touchon, C. Hoede, O. Tenaillon, V. Barbe, S. Baeriswyl, M. May, D. Brown, B. Nystedt, A. Frank, M. Thollessen, S. Andersson, S. Uplekar, B. Heym, V. Friocourt, J. Rougemont, S. Cole, M. Lipsitch, J. O'Hagan, G. Aguilera, J. Lengelle, S. Marthey, H. Chiapello, F. Rodolphe, H. van der Mei, H. Busscher, A. McGuire, B. Weiner, S. Park, I. Wapinski, S. Raman, T. Stinear, T. Seemann, S. Pidot, W. Frigui, G. Reyssset, W. Qi, M. Kaser, K. Roltgen, D. Yeboah-Manu, G. Pluschke, T. Stinear, T. Seemann, P. Harrison, G. Jenkin, J. Davies, J. Felsenstein, J. McDonald, M. Kreitman, D. Huson, D. Bryant, D. Paraskevis, K. Deforche, P. Lemey, G. Magiorkinis, A. Hatzakis, F. Ronquist, J. Huelsenbeck, G. Kubica, T. Kim, F. Dunbar, A. Drummond, A. Rambaut, D. Ayers, A. Darling, D. Zwickl, P. Beerli, M. Holder, C. Ford, P. Lin, M. Chase, R. Shah, O. Iartchouk, S. Ho, M. Phillips, A. Cooper, A. Drummond, K. Bos, V. Schuenemann, G. Golding, H. Burbano, N. Waglechner, A. Mutreja, D. Kim, N. Thomson, T. Connor, J. Lee, G. Watterson, R. Hernandez, P. Underhill, T. Kivisild, L. Cavalli-Sforza, and M. Feldman. The Role of Selection in Shaping Diversity of Natural *M. tuberculosis* Populations. *PLoS Pathogens*, 9(8):e1003543, aug 2013.

- [150] J. E. Phelan, F. Coll, I. Bergval, R. M. Anthony, R. Warren, S. L. Sampson, N. C. Gey van Pittius, J. R. Glynn, A. C. Crampin, A. Alves, T. B. Bessa, S. Campino, K. Dheda, L. Grandjean, R. Hasan, Z. Hasan, A. Miranda, D. Moore, S. Panaiotov, J. Perdigao, I. Portugal, P. Sheen, E. de Oliveira Sousa, E. M. Streicher, P. D. van Helden, M. Viveiros, M. L. Hibberd, A. Pain, R. McNerney, T. G. Clark, M. Wilkie, H. McShane, S. Gagneux, P. Small, M. Caws, G. Thwaites, S. Dunstan, T. Hawn, N. Lan, N. Thuong, K. Stepniewska, M. Huyen, N. Bang, T. Loc, S. Gagneux, D. Soolingen, K. Kremer, M. Sande, P. Small, P. Anh, N. Chinh, H. Quy, N. Duyen, D. Tho, N. Hieu, E. Torok, T. Hien, N. Dung, N. Nhu, P. Duy, C. Vinh, J. Farrar, S. Niemann, P. Supply, Y. Akhter, M. Ehebauer, S. Mukhopadhyay, S. Hasnain,

- G. P. NC, S. Sampson, H. Lee, Y. Kim, P. Helden, R. Warren, L. Majlessi, R. Prados-Rosales, A. Casadevall, R. Brosch, G. Delogu, S. Cole, R. Brosch, K. Mohareer, S. Tundup, S. Hasnain, H. Wang, D. Dong, S. Tang, X. Chen, Q. Gao, K. Singh, X. Zhang, A. Patibandla, P. Chien, S. Laal, J. Galagan, J. Musser, A. Amin, S. Ramaswamy, X. Liu, M. Gutacker, J. Musser, Y.-X. Fu, T. Bruen, H. Philippe, D. Bryant, R. Tatusov, M. Galperin, D. Natale, E. Koonin, Z. Yang, W. Wong, R. Nielsen, H. Målen, F. Berven, K. Fladmark, H. Wiker, S. Sampson, M. Chaitra, M. Shaila, R. Nayak, C. Mehaffy, J. Guthrie, D. Alexander, R. Stuart, E. Rea, F. Jamieson, A. Zimin, G. Marçais, D. Puiu, M. Roberts, S. Salzberg, J. Yorke, J. Simpson, R. Durbin, D. Zerbino, E. Birney, A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, I. Tsai, T. Otto, M. Berriman, M. Price, P. Dehal, A. Arkin, A. Vilella, A. Blanco-Garcia, S. Hutter, J. Rozas, Z. Yang, J. Guerra-Assunção, R. Houben, A. Crampin, T. Mzembe, K. Mallard, F. Coll, I. Bergval, F. Coll, A. Schuitema, H. Ronde, K. Mallard, A. Pain, Z. Hasan, A. Ali, R. McNerney, K. Mallard, G. Hill-Cawthorne, F. Coll, J. Perdigão, H. Silva, D. Machado, R. Macedo, F. Maltez, and C. Silva. Recombination in *pe/ppe* genes contributes to genetic variation in *Mycobacterium tuberculosis* lineages. *BMC Genomics*, 17(1):151, dec 2016.
- [151] V. Rajan. A Method of Alignment Masking for Refining the Phylogenetic Signal of Multiple Sequence Alignments. *Molecular Biology and Evolution*, 30(3):689–712, mar 2013.
- [152] A. Rambaut and N. C. Grassly. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer applications in the biosciences : CABIOS*, 13(3):235–8, jun 1997.
- [153] D. A. Rasmussen, O. Ratmann, and K. Koelle. Inference for Nonlinear Epidemiological Models Using Genealogies and Time Series. *PLoS Computational Biology*, 7(8):e1002136, aug 2011.
- [154] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, aug 2009.

- [155] N. reviews. Nature Reviews Immunology Focus on Monocytes and macrophages.
- [156] D. A. Robinson, E. J. Feil, and D. Falush. *Bacterial Population Genetics in Infectious Disease*. John Wiley and sons, 2010.
- [157] A. G. Rodrigo, P. R. Bergquist, and P. L. Bergquist. Inadequate Support for an Evolutionary Link between the Metazoa and the Fungi. *Systematic Biology*, 43(4):578–584, dec 1994.
- [158] A. Rokas, B. L. Williams, N. King, and S. B. Carroll. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425(6960):798–804, oct 2003.
- [159] A. Sandgren, M. Strong, P. Muthukrishnan, B. K. Weiner, G. M. Church, and M. B. Murray. Tuberculosis Drug Resistance Mutation Database. *PLoS Medicine*, 6(2):e1000002, feb 2009.
- [160] N. SCHLUGER and W. ROM. The Host Immune Response to Tuberculosis. *American Journal of Respiratory and Critical Care Medicine*, 157(3):679–691, mar 1998.
- [161] J. A. Seddon, R. M. Warren, D. A. Enarson, N. Beyers, and H. S. Schaaf. Drug-resistant tuberculosis transmission and resistance amplification within families. *Emerging infectious diseases*, 18(8):1342–5, aug 2012.
- [162] J. L. Segovia-Juarez, S. Ganguli, and D. Kirschner. Identifying control mechanisms of granuloma formation during M. tuberculosis infection using an agent-based model. *Journal of Theoretical Biology*, 231(3):357–376, dec 2004.
- [163] R. Sergeev, C. Colijn, and T. Cohen. Models to understand the population-level impact of mixed strain M. tuberculosis infections. *Journal of Theoretical Biology*, 280(1):88–100, jul 2011.
- [164] C. E. Shannon and W. Weaver. *A mathematical theory of communication*. Univ. Illinois Press Urbana, 1963.
- [165] P. Sheen, D. Couvin, L. Grandjean, M. Zimic, M. Dominguez, G. Luna, R. H. Gilman, N. Rastogi, and D. A. J. Moore. Genetic Diversity of

- Mycobacterium tuberculosis in Peru and Exploration of Phylogenetic Associations with Drug Resistance. *PLoS ONE*, 8(6):e65873, jun 2013.
- [166] H. L. Smith and H. R. Thieme. Chemostats and epidemics: competition for nutrients/hosts. *Mathematical Biosciences and Engineering : MBE*, 10(5-6):1635–50, Jan.
 - [167] D. Sorensen and D. Gianola. *Likelihood, Bayesian and MCMC methods in quantitative genetics*.
 - [168] T. Sornasse, P. V. Larenas, K. A. Davis, J. E. de Vries, and H. Yssel. Differentiation and stability of T helper 1 and 2 cells derived from naive human neonatal CD4+ T cells, analyzed at the single-cell level. *The Journal of experimental medicine*, 184(2):473–83, aug 1996.
 - [169] S. Soubeyrand. Construction of semi-markov genetic-space-time SEIR models and inference. *Journal de la Société Française de Statistique*, 157(1):129–152, 4 Dec. 2016.
 - [170] E. Spada, L. Sagliocca, J. Sourdis, A. R. Garbuglia, V. Poggi, C. De Fusco, and A. Mele. Use of the Minimum Spanning Tree Model for Molecular Epidemiological Investigation of a Nosocomial Outbreak of Hepatitis C Virus Infection. *Journal of Clinical Microbiology*, 42(9):4230–4236, sep 2004.
 - [171] T. Stadler and S. Bonhoeffer. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1614):20120198, mar 2013.
 - [172] A. Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, may 2014.
 - [173] D. Sun and D. Zhang. A New Discriminant Principal Component Analysis Method with Partial Supervision. *Neural Processing Letters*, 30(2):103–112, oct 2009.
 - [174] D. L. Swofford, G. J. Olsen, P. J. Waddell, and D. M. Hillis. *Phylogenetic Inference, In Molecular Systematics*. Sinauer Associates, Inc., Sunderland, MA., 1996.

- [175] K. Takeya, A. Kuwata, M. Yoshida, and T. Miyazaki. Effect of dilution rate on competitive interactions between the cyanobacterium *Microcystis novacekii* and the green alga *Scenedesmus quadricauda* in mixed chemostat cultures. *Journal of Plankton Research*, 26(1):29–35, Jan 2004.
- [176] G. Talavera and J. Castresana. Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. *Systematic Biology*, 56(4):564–577, aug 2007.
- [177] G. Tan, M. Muffato, C. Ledergerber, J. Herrero, N. Goldman, M. Gil, and C. Dessimoz. Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference. *Systematic biology*, 64(5):778–91, sep 2015.
- [178] M. Tanveer, Z. Hasan, A. Kanji, R. Hussain, and R. Hasan. Reduced TNF- α and IFN- γ responses to Central Asian strain 1 and Beijing isolates of *Mycobacterium tuberculosis* in comparison with H37Rv strain. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 103(6):581–587, jun 2009.
- [179] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, 6(31):187–202, Feb 2009.
- [180] United Nations. International Migration Report 2015.
- [181] P. W. Uys, P. D. van Helden, and J. W. Hargrove. Tuberculosis reinfection rate as a proportion of total infection rate correlates with the logarithm of the incidence rate: a mathematical model. *Journal of the Royal Society, Interface*, 6(30):11–5, jan 2009.
- [182] H. Veldkamp and H. W. Jannasch. Mixed culture studies with the chemostat. *Journal of Applied Chemistry and Biotechnology*, 22(1):105–123, April 2007.
- [183] J. Viridi, P. Gulati, and M. Pai. Genetic diversity of pathogenic microorganisms and its medical and public health significance. *Indian Journal of Medical Microbiology*, 25(1):2, jan 2007.

- [184] E. M. Volz, S. D. W. Frost, T. Stadler, R. Kouyos, and S. Yerly. Inferring the Source of Transmission with Phylogenetic Data. *PLoS Computational Biology*, 9(12):e1003397, dec 2013.
- [185] T. M. Walker, M. K. Lalor, A. Broda, L. S. Ortega, M. Morgan, L. Parker, S. Churchill, K. Bennett, T. Golubchik, A. P. Giess, C. Del Ojo Elias, K. J. Jeffery, I. C. J. W. Bowler, I. F. 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, T. E. A. Peto, and C. P. Conlon. Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007 - 12, with whole pathogen genome sequences: an observational study. *The Lancet Respiratory Medicine*, 2(4):285–292, apr 2014.
- [186] J. Weiner and S. H. E. Kaufmann. Recent advances towards tuberculosis control: vaccines and biomarkers. *Journal of Internal Medicine*, 275(5):467–480, may 2014.
- [187] J. E. Wigginton and D. Kirschner. A model to predict cell-mediated immune regulatory mechanisms during human infection with Mycobacterium tuberculosis. *Journal of immunology (Baltimore, Md. : 1950)*, 166(3):1951–67, feb 2001.
- [188] A. A. Witney, K. A. Gould, A. Arnold, D. Coleman, R. Delgado, J. Dhillon, M. J. Pond, C. F. Pope, T. D. 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*, 53(5):1473–1483, may 2015.
- [189] A. Wm Dress, C. Flamm, G. Fritzsche, S. Grünwald, M. Kruspe, S. J. Prohaska, and P. F. Stadler. Algorithms for Molecular Biology Noisy: Identification of problematic columns in multiple sequence alignments. *Algorithms for Molecular Biology*, 3(3), 2008.
- [190] S. Wohl, S. F. Schaffner, and P. C. Sabeti. Genomic Analysis of Viral Outbreaks. *Annual review of virology*, 3(1):173–195, sep 2016.
- [191] C. J. Worby, M. Lipsitch, and W. P. Hanage. Shared genomic vari-

- ants: identification of transmission routes using pathogen deep sequence data. *American Journal of Epidemiology*, jun 2017.
- [192] V. d. P. Y. *Phylogeny inference based on distance methods*. In: Salemmi M, Vandamme AM (eds) *The phylogenetic handbook, a practical approach to DNA and protein phylogeny*. Cambridge University Press, New York, USA, 2009.
 - [193] Z. Yang, M. Rosenthal, N. A. Rosenberg, S. Talarico, L. Zhang, C. Marrs, V. Ø. Thomsen, T. Lillebaek, and A. B. Andersen. How dormant is *Mycobacterium tuberculosis* during latency? A study integrating genomics and molecular epidemiology. *Infection, Genetics and Evolution*, 11(5):1164–1167, jul 2011.
 - [194] S. Yazdani, J. Shanbehzadeh, and M. T. Manzuri Shalmani. RPCA: A Novel Preprocessing Method for PCA. *Advances in Artificial Intelligence*, 2012:1–7, dec 2012.
 - [195] S. A. Yimer, G. Norheim, A. Namouchi, E. D. Zegeye, W. Kinander, T. Tønjum, S. Bekele, T. Mammsåker, G. Bjune, A. Aseffa, and C. Holm-Hansen. *Mycobacterium tuberculosis* lineage 7 strains are associated with prolonged patient delay in seeking treatment for pulmonary tuberculosis in Amhara Region, Ethiopia. *Journal of clinical microbiology*, 53(4):1301–9, apr 2015.
 - [196] R. J. F. Ypma, A. M. A. Bataille, A. Stegeman, G. Koch, J. Wallinga, and W. M. van Ballegooijen. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings. Biological sciences / The Royal Society*, 279(1728):444–50, feb 2012.
 - [197] R. J. F. Ypma, W. M. van Ballegooijen, and J. Wallinga. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics*, 195(3):1055–62, nov 2013.
 - [198] L. Zammarchi, F. Bartalesi, and A. Bartoloni. Tuberculosis in tropical areas and immigrants. *Mediterranean journal of hematology and infectious diseases*, 6(1):e2014043, 2014.
 - [199] H. Zhang, D. Li, L. Zhao, J. Fleming, N. Lin, T. Wang, Z. Liu, C. Li, N. Galwey, J. Deng, Y. Zhou, Y. Zhu, Y. Gao, T. Wang, S. Wang,

- Y. Huang, M. Wang, Q. Zhong, L. Zhou, T. Chen, J. Zhou, R. Yang, G. Zhu, H. Hang, J. Zhang, F. Li, K. Wan, J. Wang, X.-E. Zhang, and L. Bi. Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance. *Nature Genetics*, 45(10):1255–1260, sep 2013.
- [200] J. Zhang, R. Chiodini, A. Badr, and G. Zhang. The impact of next-generation sequencing on genomics. *Journal of genetics and genomics* = *Yi chuan xue bao*, 38(3):95–109, mar 2011.
- [201] M. Zhang, J. Gong, Y. Lin, and P. F. Barnes. Growth of virulent and avirulent *Mycobacterium tuberculosis* strains in human macrophages. *Infection and immunity*, 66(2):794–9, feb 1998.
- [202] Y. Zhang. Persisters, persistent infections and the yin-yang model. *Emerging Microbes and Infections*, 3, 2014.
- [203] A. Zumla, P. Nahid, and S. T. Cole. Advances in the development of new tuberculosis drugs and treatment regimens. *Nature Reviews Drug Discovery*, 12(5):388–404, apr 2013.