Mathematical Models for the Epidemiology and Evolution of *Mycobacterium tuberculosis*

15

Jūlija Pečerska, James Wood, Mark M. Tanaka, and Tanja Stadler

Abstract

This chapter reviews the use of mathematical and computational models to facilitate understanding of the epidemiology and evolution of *Mycobacterium tuberculosis*. First, we introduce general epidemiological models, and describe their use with respect to epidemiological dynamics of a single strain and of multiple strains of *M. tuberculosis*. In particular, we discuss multi-strain models that include drug sensitivity and drug resistance. Second, we describe models for the evolution of *M. tuberculosis* within and between hosts, and how the resulting diversity of strains can be assessed by considering the evolutionary relationships among different strains. Third, we discuss developments in integrating evolutionary and epidemiological models to analyse *M. tuberculosis* genetic sequencing data. We conclude the chapter with a discussion of the practical implications of modelling – particularly modelling strain diversity – for controlling the spread of tuberculosis, and future directions for research in this area.

Keywords

Strain variation • Heterogeneity • Population biology • Phylogeny • Phylodynamics • Molecular epidemiology • Compartmental model

J. Pečerska • T. Stadler
Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland
e-mail: julija.pecerska@bsse.ethz.ch; tanja.stadler@bsse.ethz.ch

J. Wood
School of Public Health and Community Medicine, UNSW Sydney, Sydney, NSW, Australia
e-mail: james.wood@unsw.edu.au

M.M. Tanaka (✉)
School of Biotechnology & Biomolecular Sciences, and Evolution & Ecology Research Centre, UNSW Sydney, Sydney, NSW, Australia
e-mail: m.tanaka@unsw.edu.au

S. Gagneux (ed.), *Strain Variation in the* Mycobacterium tuberculosis *Complex: Its Role in Biology, Epidemiology and Control*, Advances in Experimental Medicine and Biology 1019,
DOI 10.1007/978-3-319-64371-7_15

15.1 Introduction

The causative agent of tuberculosis, *Mycobacterium tuberculosis*, emerged as a human pathogen around 70,000 years ago (Comas et al. 2013; Gagneux 2012), although conflicting estimates point to much later dates of around 5,000 years ago (Bos et al. 2014). Forms of tuberculosis such as *M. bovis* that infect non-human animals evolved from human tuberculosis, indicating that the disease first appeared in humans before adapting to other animals. Mounting genetic evidence indicates that strain-to-strain variation in *M. tuberculosis* is more extensive than previously thought (Borrell and Gagneux 2011; Gagneux and Small 2007). Seven major lineages of modern-day *M. tuberculosis* have been identified (Galagan 2014) and specific strains are highly associated with geographic location (see Chap. 1) (Gagneux and Small 2007; Hershberg et al. 2008). Molecular methods have helped identify finer scale variation within lineages, which we discuss in more detail in Sect. 15.3.

The increased availability of data both at the epidemiological and molecular level allows us to start raising complex questions about data interpretation and analysis. For instance, how do we understand and predict tuberculosis epidemics on the population level? How do we best use molecular data to shed light on the transmission dynamics of different *M. tuberculosis* lineages? These questions typically require collated data analysis under specific assumptions on the properties of *M. tuberculosis*, such as, the mechanisms of mutational change. Mathematical, or computational, modelling is a methodology that enables the precise description of assumptions in order to investigate model behaviour, qualitatively or quantitatively. Defined models can be combined with data and thus provide answers to scientific questions concerning a given dataset. This approach has been instrumental in the understanding of physical sciences, and it has become more widely used in biology as biological data have become increasingly refined and quantitative in nature.

Mathematical models that are applied in biology range from being extremely simple and generic to being complex and specific. Simple models often enable a qualitative understanding of complex phenomena, while complex models have the advantage of being more realistic and detailed and thus may offer detailed quantitative insight. In the words of statistician George Box, however, "all models are wrong but some are useful". The aim of modelling is to shed light on a phenomenon rather than to create a maximally realistic description of it.

In the study of infectious diseases models can extend our understanding of an epidemic by allowing us to predict population dynamics from basic knowledge of the natural history of a disease. Models can help evaluate the effects of any potential or actual interventions at the population level. By providing precise quantitative predictions mathematical models also play a role in drawing inferences from observational data, for example, by producing estimates of parameters relating to disease transmission.

In this chapter we consider how mathematical and computational models can be used to understand the variation in *M. tuberculosis* that has been revealed using molecular techniques. Two different modelling traditions are pertinent to this topic. First, epidemiological models address the dynamics of infectious diseases at the population level and enable researchers to consider possible outcomes including the effects of intervention strategies. Second, models of molecular evolution and population genetics concern the processes by which genomes undergo change. These models are generic in that they have not been developed for any particular species, and can be applied to *M. tuberculosis* to understand its variation and to reconstruct its evolutionary history. We will describe both of these approaches and their applications to *M. tuberculosis*. We will further discuss progress made in combining epidemiological and evolutionary elements within the same framework to analyse the diversity of *M. tuberculosis*.

15.2 Epidemiological Modelling and Analysis

In this section we will focus on epidemiological models and their application to *M. tuberculosis*. We will primarily consider models that assume that the host population is homogeneous, ignoring possible effects of heterogeneity in host behaviour on the dynamics of the epidemic. We will begin with generic models that describe epidemiological dynamics of a single disease variant and then describe models of TB epidemiology with heterogeneity in the pathogen population, e.g. due to the occurrence of drug-resistant strains.

15.2.1 Epidemiological Modelling of *M. tuberculosis*

Epidemiological models traditionally separate host populations into distinct compartments according to their infection status. In the simplest scenario, an individual is either susceptible to infection, infectious, or recovered and therefore immune to reinfection. The numbers of individuals in each compartment are tracked by $S(t)$, $I(t)$ and $R(t)$ respectively, where t stands for time. In what follows we will drop the "(t)" from the notation except where it is clearer to retain it. Typically, a susceptible host is assumed to transition to the infected state at a rate proportional to the number of infected individuals I (say βI), and an infected individual transitions to the recovered state at a constant rate (say γ). Without host birth or host death events, the total number of individuals in the three compartments is constant $N = S+I+R$ where N is the total population size. The structure of this model is illustrated in the left panel of Fig. 15.1. This model, which is known as the Susceptible-Infectious-Recovered (SIR) model, was initially studied in depth by Kermack and McKendrick

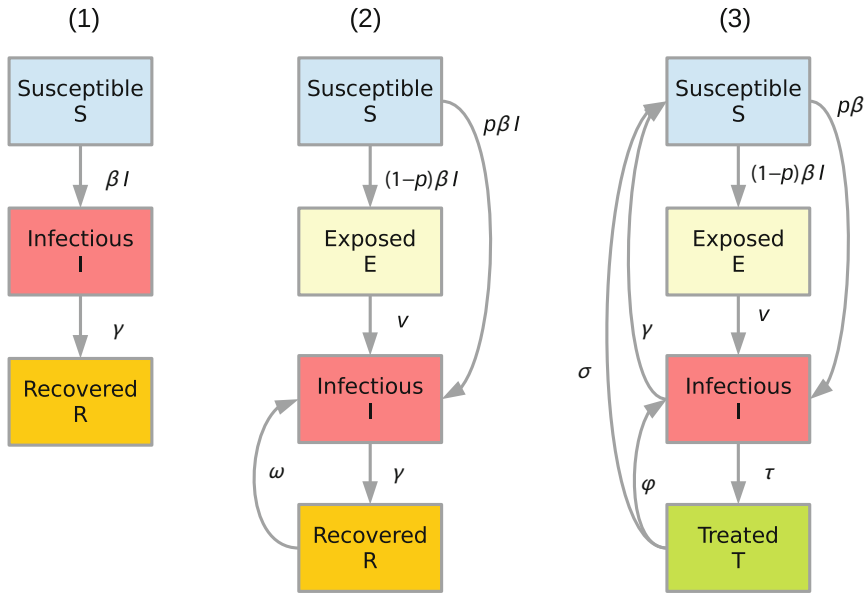

Fig. 15.1 Examples of compartmental models of infectious disease dynamics. (*1*) Classic SIR model with transmission rate β and recovery rate γ. (*2*) A more complex model with an exposed class E. In the case of tuberculosis, a proportion $(1-p)$ of new cases enter a state of latent infection modelled with E while the remainder (p) progress to active disease I. Latent infections reactivate at rate v, active cases recover at rate γ and recovered individuals regress to the disease state at rate ω. (*3*) When modelling antibiotic use and drug resistance it is useful to modify the model to include a state for infected treated individuals T. In this model, active cases are detected and treated at rate τ, stop treatment at rate ϕ and treated individuals return to the uninfected class S at rate σ. Not shown are death from each compartment or birth into S. We note that published models of TB dynamics are varied and while they are in general similar to the structures shown in (*2*) and (*3*) there are differences that reflect differences in the questions being addressed

(1927) and has since been elaborated upon in many ways (Diekmann et al. 1990; Keeling and Rohani 2008).

The SIR dynamics can be modelled deterministically or stochastically. In the deterministic version, the change in compartment sizes follows ordinary differential equations

$$\frac{d}{dt}S(t) = -\beta I(t)S(t) \tag{15.1}$$

$$\frac{d}{dt}I(t) = \beta I(t)S(t) - \gamma I(t) \tag{15.2}$$

and $R(t) = N - S(t) - I(t)$. If the epidemic starts with a single introduction of the infection into the population, the initial conditions are set to $S(0) = N - 1$ and $I(0) = 1$.

In the stochastic formulation of the SIR model, S, I, and R are integer-valued rather than real-valued, and when an infection occurs I increases by one and S decreases by one. Given a very small time interval Δt, the probability for infection to happen is assumed to be $\beta SI\Delta t + o(\Delta t)$. The term $\beta SI\Delta t$ is the probability for precisely one infection event to happen in time interval Δt. The term $o(\Delta t)$ summarises the probability for more than one infection event to happen, with the term $o(\Delta t)/\Delta t$ approaching zero as Δt approaches zero. This means that the waiting time until an infection event where I increases by one and S decreases by one is exponentially distributed with parameter βSI. Similarly, upon recovery I decreases by one and R increases by one, and this event occurs with probability $\gamma I\Delta t + o(\Delta t)$.

The dynamics of the SIR model are well understood and are described in multiple sources such as the text by Keeling and Rohani (2008). Unfortunately, the assumptions of the SIR model are clearly too narrow to be directly applicable to *M. tuberculosis*. In particular, *M. tuberculosis* infection is characterised by a long and highly variable incubation period known as latent infection. Furthermore, hosts generally do not have strong immune protection against further infection and asymptomatic hosts often relapse to disease years after an acute infection. However, the basic methodology of SIR modelling can be modified to reflect the natural history of *M. tuberculosis*. Adaptation of the SIR model to *M. tuberculosis* began with the work of Waaler et al. (1962), with the inclusion of long-term latency in the form of non-symptomatic cases as a key feature. The model divides individuals into noninfected, noncase, actual disease case, and recovered compartments. The noncase individuals are the latently infected individuals that do not show symptoms immediately upon infection but can potentially progress (with some rate ν). The actual disease case individuals show symptoms and are thus infectious. Individuals move into the recovered compartment after an active *M. tuberculosis* infection.

Figure 15.1 (middle) shows a simple version of a model structure that captures key features, in particular latency and relapse, of *M. tuberculosis* natural history. Note that the R compartment here represents "recovery" from active tuberculosis, but individuals in this compartment have not necessarily cleared the *M. tuberculosis* pathogen which is why they may relapse with some rate ω.

Different approaches have been used to incorporate latency due to the lack of detailed quantitative information about the long-term dynamics of infection and the immune response within humans. Blower et al. (1995) popularised the use of a dichotomous short-term/long-term characterisation of latency based around the rule of thumb that around 5% of infections progress quickly to active disease and another 5% progress slowly over the remainder of a person's lifespan. This would be captured by $p = 0.05$ in Fig. 15.1 (middle). In these models, a fraction of infected individuals progress immediately to active disease while the remaining fraction enter a latent state and progress to disease at a low rate. More recently, a modified form of this dichotomous transition has been introduced that more accurately captures the timing of active disease in relation to infection through stratification of latency into 2-stages (see for instance Dowdy et al. 2013).

Models of tuberculosis epidemiology have been used to characterise the decline of *M. tuberculosis* epidemics in the US and UK (Blower et al. 1995) and to determine the

contribution of endogenous reactivation and exogenous reinfection to the overall risk of disease (Vynnycky and Fine 1997). Blower et al. (1995) used a model that allowed for infection, reactivation, and relapse, and showed that the apparent decline might be explained as a temporary effect following a large epidemic. In the work by Vynnycky and Fine (1997), a more data-driven approach to *M. tuberculosis* infection risk estimation was taken, without modelling the underlying transmission process directly. The model was designed to evaluate the impact of new infections compared to reinfection and reactivation of the disease. The results suggest that in the UK reinfection made a strong epidemiological impact during the first half of the twentieth century, but had negligible effects by 1980, by which time the incidence reached its lowest point. This approach is less relevant to more recent epidemiological history in countries such as the UK, where cases have been increasingly driven by migration from high-incidence settings.

The observed variation in the effect of the BCG vaccine has also been investigated, for instance through the work by Gomes et al. (2004), drawing on the earlier model-based analysis by Fine and Vynnycky (1998). The latter work was aimed at explaining the differential success of the BCG vaccine in different settings, ranging from high efficacy observed in the UK trials as opposed to no efficacy in the large Chingleput trial in India. This paper noted how exposure to *M. tuberculosis* and/or related mycobacteria could have a confounding effect on the estimates of vaccine efficacy. Animal studies have shown that the BCG vaccine did not provide a further benefit over the immunity derived from tuberculosis infection. The study by Fine and Vynnycky showed how variation in this background level of immunity affected estimates of vaccine efficacy, with these estimates becoming negligibly small when natural infection rates were very high and provided similar levels of protection as vaccination. Gomes et al. (2004) explored these issues in a population dynamic model and linked the analysis with the concept of reinfection threshold, which is defined as the reproduction number (see Sect. 15.2.2) becoming greater than 1 in a population of previously exposed individuals.

Further developments in tuberculosis modelling have focused on potential effects of variation in treatment strategies. Treatment strongly affects incidence, prevalence and mortality from TB and reduces the average duration of infectiousness and thus reduces the possibility of transmission. Treatment is typically incorporated in models through a modification of the rate of cure (as in Fig. 15.1, right), although in models seeking to capture treatment programs more precisely, this may be described through multiple model states and transitions. Such models have been used to examine the traditional World Health Organisation DOTS (directly observed treatment, short-course) approach (Dye et al. 1998) and various means of improvement, for example, through active case finding, changes in diagnosis and treatment regimens and wider application of preventive treatment of latent tuberculosis infection (Dye and Williams 2010). Of particular interest within the context of this book are the models that have looked at interactions between treatment programs and the development of resistance, which we cover in more detail in Sect. 15.5.

Modelling studies have also aimed to understand the origins of tuberculosis (Chisholm et al. 2016) and the longer term evolution of *M. tuberculosis* and its characteristics, such as latency (Chisholm and Tanaka 2016; Zheng et al. 2014) and virulence (Basu and Galvani 2009).

15.2.2 Transmission Parameters and Their Estimation

Epidemiological modelling has identified a critical quantity in infectious disease dynamics known as the basic reproduction number or R_0 (Anderson and May 1979; Diekmann et al. 1990). This quantity is defined as the average number of new infectious cases arising from a typical case in a completely susceptible population. One of the principal reasons for interest in R_0 is that it constitutes a threshold quantity for a large class of infectious disease

models; namely, $R_0 > 1$ implies that the disease can persist whereas $R_0 < 1$ leads to disease elimination. Despite this simplicity, the mathematical definition of R_0 is a function of model structure and reflects details of disease epidemiology.

Under the SIR model, $R_0 = \beta S(0)/\gamma$. However, even quite simple models of TB such as that defined in Blower et al. (1995) lead to more complex expressions, with R_0 defined as the sum

$$R_0 = R_0^{\text{FAST}} + R_0^{\text{SLOW}} + R_0^{\text{RELAPSE}}.$$

Each component here is a function of multiple parameters. The SIR model formulation of R_0 resembles only the R_0^{FAST} component whereas the transmission potential for tuberculosis is complicated by the processes of reactivation (R_0^{SLOW}) and relapse (R_0^{RELAPSE}). Other studies have additionally included the process of reinfection (Gomes et al. 2004). For a general derivation and discussion of the basic reproduction number we refer readers to the text by Diekmann and Heesterbeek (2000).

In practice, populations are not usually fully susceptible and so the production of new cases is slower than indicated by R_0. New cases are better understood through the related quantity known as the *effective* reproduction number, usually labelled R_e. This quantity is defined as the average number of new infectious cases produced by a typical case regardless of the susceptible proportion. At the start of a local epidemic $R_e = R_0$, and over time it decreases to a value < 1 if the epidemic ends and is unable to persist, or remains close to 1 if the disease persists endemically. Knowledge of reproduction numbers is important for informing strategies for controlling infectious diseases but estimating these quantities poses some challenges. In particular, direct observation of the average number of secondary infections produced by infectious individuals is not feasible with conventional epidemiological methods. Instead reproduction numbersare typically estimated implicitly through comparisons of model outcomes with infection history data or incidence and prevalence time-series. Earlier approaches are summarised in Dietz (1993), while examples of approaches for stochastic models and structured communities are provided in the text by Becker (2015).

Empirical data can be used to estimate parameters which in turn allow investigation of the long term characteristics of epidemics using epidemiological models. The dynamics can also be studied through computer simulations of stochastic formulations of models, which can in turn be compared with data that directly reflect transmission such as contact tracing studies or analysis of *M. tuberculosis* infection within households. Brooks-Pollock et al. (2011), for example, used incidence data across multiple-person households to estimate the relative contributions of community and households to *M. tuberculosis* transmission.

For tuberculosis, the differing components of the reproduction number as described above make this more challenging. Estimation of the basic reproduction number is also complicated by the absence of good diagnostic markers for immunity. As such, in contrast to infections such as influenza and measles, practical epidemiological models of tuberculosis have focused more on projections of disease incidence than analysis of the reproduction number and the potential for elimination (Dye and Williams 2010). However, the growing availability of molecular data provides opportunities to overcome some of the issues outlined above. In particular, molecular data provide information that can potentially allow models to separate between recent infection, reactivation and relapse (Small et al. 1994). Stochastic formulations of the epidemiological models in combination with molecular data, rather than deterministic formulations in combination with the traditional epidemiological data discussed here, will be the focus of Sect. 15.4.

Fig. 15.2 Examples of compartmental models with pathogen variation and host variation. (*1*) When pathogen variation is taken into account the infected classes are partitioned into multiple states corresponding to the multiple pathogen types. The variation can be with respect to resistance or other genetic characters. Here we do not show transitions between types, which depend on the details of the model. (*2*) When host variation is taken into account the hosts are partitioned into susceptibility classes. For simplicity the figure does not show births and deaths for either model

15.2.3 Modelling Heterogeneous Epidemics

Thus far we have discussed models in which infections are homogenous with all prevalent infectious cases represented by the single I compartment. However, pathogen populations may be variable, for example, in levels of drug resistance. Such variation could cause hosts infected with different strains to transmit with different rates. One can model a simple case by subdividing the I class into I_S and I_R classes: the hosts infected with a drug sensitive strain, and the hosts infected with a drug resistant strain, respectively. Correspondingly, one needs to divide the E class into E_S, E_R. Hosts may move from I_S to I_R through resistance evolution, and potentially move from I_R to I_S through loss of resistance.

Similarly, there may be variation in the host population both in terms of transmission risks (e.g. geographic variation in incidence) and in risks of developing active TB for example through HIV co-infection. Again, we can divide the compartments according to this variation, for example into I_1 and I_2. As done above, we subdivide the S class as well as the E class as the difference in risk behaviour does not depend on the TB infection status. Figure 15.2 illustrates an example of how variation in pathogens (left) and variation in hosts (right) could be modelled, extending the homogeneous epidemic model displayed on the right of Fig. 15.1.

Such models have been used to represent different types of variation in viral epidemics, such as geographical variation, drug resistance levels and super spreader behaviour. For *M. tuberculosis*, outside of settings with high HIV prevalence, the main source of variation that has been considered to date is in relation to drug-resistance and its interaction with treatment programs, which we now discuss in more detail.

As with other infectious diseases, the rise of antimicrobial drug resistance is a problem for the control of *M. tuberculosis*. While multi-drug resistant (MDR) or extremely-drug resistant (XDR) tuberculosis is likely to be the result of treatment

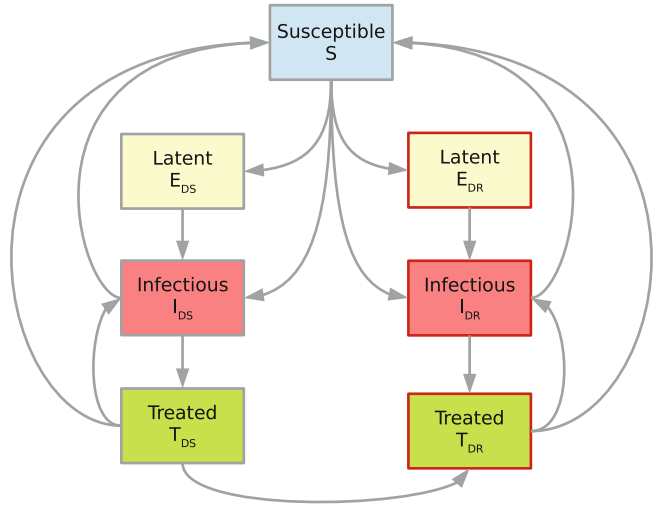

Fig. 15.3 An example of a compartmental model of tuberculosis with drug resistance evolution and transmission. This model is an extension of the treated TB model shown in Fig. 15.1 (right) in which drug sensitive *M. tuberculosis* can evolve resistance de novo and subsequently transmit to new hosts. The subscripts indicate drug sensitive (DS) and drug resistant (DR) infection states. For simplicity, rates are omitted from this diagram

failure, they also occur with lower frequency in new cases. These issues are discussed in more detail in Chaps. 11 and 14 but here we briefly describe key features of epidemiological models of resistance.

In the context of population dynamics, drug resistant strains have two critical properties. First, they can persist longer in patients because standard treatment is less effective or not effective at all. Second, they may come with a fitness cost that lowers their rate of transmission. The cost may be due to a trade-off between the original function of the gene and the resistance phenotype (Chap. 11). The fitness cost has been measured in vitro (Gagneux et al. 2006) and it has been observed to be variable across *M. tuberculosis* strains. Such in vitro measurements assess the replicative capacity of a strain which differs from the ability of the strain to infect new hosts, i.e. the transmission fitness, but the two fitness concepts are assumed to be linked. In vitro, strains have shown a replication fitness disadvantage when compared to their rifampicin-susceptible ancestors, while some of the clinically-derived strains show no fitness costs. One reason for this variability is the possibility that further mutations occur that lower or compensate the cost of resistance. Thus, it is unlikely that drug resistant *M. tuberculosis* will easily revert to sensitivity (Cohen and Murray 2004).

The dynamics of drug resistance in *M. tuberculosis* have been studied using compartmental models as introduced above and shown in Fig. 15.2 (left). The first such models were introduced by Blower and colleagues (Blower and Gerberding 1998; Blower et al. 1996) modelling two types of infection: drug sensitive and drug resistant. Models have since been extended and refined, maintaining a structure similar to that shown in Fig. 15.3 (see also Ozcaglar et al. 2012). This figure is a special case of models with pathogen heterogeneity as shown on the left in Fig. 15.2, with the addition of transition between sensitive and resistant treatment classes.

Such models of the evolution and spread of drug resistance have been studied to address a number of epidemiological problems, in particular to clarify the importance and variability of the reproductive and transmission fitness of drug resistant strains. Such insight allows us to quantify the future burden of drug resistant strains on an epidemiological scale. The replicative fitness of resistant strains is variable (Dye and Williams 2009; Gagneux et al. 2006), but there is some evidence for lower rates of transmission of resistant strains (Dye and Williams 2009; Luciani et al. 2009). On the epidemiological scale this cost is balanced by the advantage resistant strains have in patients under treatment (Luciani et al. 2009). Furthermore, since the cost of resistance can be

lowered by compensatory mutations, the resulting variation in *M. tuberculosis* fitness means that without adequate control strategies, resistant strains will likely dominate in the long run (Blower and Gerberding 1998; Cohen and Murray 2004; Luciani et al. 2009; Shrestha et al. 2014; Trauer et al. 2014). Models can be used to explore control scenarios – e.g. rates of detection and cure success – that could lead to the control of drug-resistant tuberculosis (Dye and Espinal 2001; Dye and Williams 2000). Modelling also allows quantification of the relative importance of the two factors contributing to the spread of drug resistant strains, namely de novo evolution of resistance versus transmitted resistance. Modelling studies strongly suggest that in most settings a large majority of drug resistant cases are due to the transmission of resistant strains rather than the de novo acquisition of resistance (Cohen and Murray 2004; Kendall et al. 2015; Luciani et al. 2009; Trauer et al. 2014).

15.3 Molecular Evolution of *M. tuberculosis*

The fields of population genetics and molecular evolution are dedicated to analysing and understanding genetic variation within and among species. Many models and methods have been developed for these purposes, most of which are general and not specific to pathogens. In this section we discuss applications of this theory to *M. tuberculosis*. We start with standard models and methods in molecular evolution and phylogenetics, and then discuss some of the specific issues that arise when applying these methods to genetic data obtained from *M. tuberculosis* isolates.

All organisms may accumulate mutations through replication, including *M. tuberculosis*. Many of these changes are never observed as the resulting mutants may suffer a large fitness disadvantage so that they are eliminated from the population. Other mutations, however, rise in frequency and may become fixed in a population. A mutation which is fixed in a population is called a *substitution*. To define this term more precisely, it is necessary to clarify what population is being considered. The evolution of *M. tuberculosis* occurs on at least two different population scales. First, the global population of *M. tuberculosis* may undergo evolutionary substitution. This process of substitution is important to consider when comparing *M. tuberculosis* to other species. Second, evolutionary substitution occurs at a local level within hosts. In this case, mutants arise within each host and may reach fixation in the host. In this section we are primarily concerned with substitution at the local host scale rather than the global population scale. Genetic changes accumulate within hosts so that bacterial populations in different hosts generally diverge. This genetic divergence provides information about the evolutionary history of the bacterium, which we can represent with a phylogenetic tree. A similar but distinct tree concept used in population genetics studies and implemented in some software is the coalescent, which describes genealogical history viewing time going backwards (for further details see Wakeley 2009). Coalescents are also known as gene trees.

Phylogenies are appropriate when there is no horizontal gene transfer or recombination and no convergent evolution. In contrast to many other bacteria, *M. tuberculosis* exhibits remarkably little recombination (Liu et al. 2006) and there is currently no evidence it carries plasmids (Zainuddin and Dale 1990). This makes the analysis of genomes relatively straightforward as only substitutions on a phylogenetic tree need to be modelled. In particular, phylogenies of *M. tuberculosis* based on different genetic loci have tree-like structures rather than being reticulate or networks. On the other hand, some genes in the *M. tuberculosis* genome may be under strong natural selection and can therefore exhibit convergent evolution. In particular, sites conferring drug resistance are known to undergo convergence and are therefore typically excluded from any phylogenetic analysis.

In the next subsections we will first discuss evolutionary models for markers of *M. tuberculosis* and then evolutionary models for all nucleotides within the *M. tuberculosis* genome.

We will then illustrate how these models are used in population genetics to assess the within- and between-host variation of *M. tuberculosis* strains.

15.3.1 Evolutionary Models for Markers of *M. tuberculosis*

In an effort to characterise and understand the transmission of tuberculosis, isolates of *M. tuberculosis* have been genotyped for many years using a variety of molecular methods (Chap. 3). Early methods included typing based on the mobile gene IS6110 (Alland et al. 1994; Cave et al. 1991; Small et al. 1994), spacer-oligonucleotide typing or spoligotyping (Kamerbeek et al. 1997) and MIRU-VNTR typing (Supply et al. 2006). In recent years, the declining cost of DNA sequencing has enabled the use of whole-genome sequencing (WGS) as a strategy for studying *M. tuberculosis* transmission (Gardy et al. 2011; Walker et al. 2013). WGS allows a fine-scale genetic characterisation of *M. tuberculosis* strains within a population, and has the advantage over previous technologies of minimising the impact of parallel evolution (Chap. 3).

The genetic variation observed at marker loci ultimately arises from mutation processes. To exploit this variation for studying *M. tuberculosis* epidemiology it is useful to know how and at what rate the underlying genetic loci mutate and undergo within-host substitution (Tanaka and Francis 2005). The simplest model to apply is a Poisson process in which mutations appear randomly at a constant rate per unit time and reach fixation instantly. This model and elaborations of it have been used to estimate mutation rates. An alternative approach is to compare the extent of variation using different markers to estimate rates against known mutation rates (Reyes and Tanaka 2010).

Genomic variation in *M. tuberculosis* is partly due to the movement of the mobile gene IS6110. The rate of movement of IS6110 has been estimated to be around 0.1 to 0.3 changes per year using serial isolates of *M. tuberculosis* (de Boer et al. 1999; Rosenberg et al. 2003; Warren et al. 2002). Spoligotype variation is due to deletion of repeats at a CRISPR locus (Aranaz et al. 2004). The evolutionary process at this locus, estimated to be around 0.02 to 0.09 per year (Reyes and Tanaka 2010), is slower than changes due to IS6110 movement. VNTR loci mutate by expanding or contracting in the number of repeats. The stepwise mutation models developed for microsatellites in humans (Di Rienzo et al. 1994) are likely to apply well to bacteria (Vogler et al. 2006). Under the simplest version of such models, repeats increase or decrease by a single step and with equal probability. Estimates of the mutation rate of VNTR loci have varied considerably with low rates around 10^{-5} per locus per year (Grant et al. 2008) and $10^{-3.9}$ (Wirth et al. 2008) and high rates of 10^{-3} to 10^{-2} (Aandahl et al. 2012; Ragheb et al. 2013; Reyes and Tanaka 2010). The variation in these estimates may reflect the diversity of models, methods and data used to obtain them.

Single nucleotide polymorphisms (SNPs) occur through point mutation which can appear throughout the entire genome. Whole genome sequences analysed with phylogenetic and similar methods have yielded conflicting estimates of mutation rates varying from 3×10^{-9} (Comas et al. 2013) to 10^{-7} (Bos et al. 2014; Walker et al. 2013). The higher rates are supported by in vitro studies of mutation rates (Ford et al. 2013) but it should be noted that the "long-term" rate of evolution is likely to be lower because estimates based on recent variation includes polymorphisms that ultimately will not be fixed in the population (e.g. deleterious mutations) (Ho et al. 2005). Mutation rates also vary substantially depending on genetic background (Ford et al. 2013). Whether mutation rates during latent infection are equal to (Ford et al. 2011; Lillebaek et al. 2016) or lower than (Colangeli et al. 2014) rates during active infection is not settled. We suggest that the uncertainty in mutation rates can be further addressed in the future through modelling and careful consideration of assumptions underlying models and data.

One of the challenges in using genetic markers for phylogenetic analysis is the occurrence of parallel evolution (or homoplasy), by which iden-

tical states are reached by mutation in independent infections. Such events can potentially undermine the phylogenetic analysis of *M. tuberculosis* (Comas et al. 2009), although a simulation study has shown that the impact of homoplasy is not necessarily large on the epidemiological scale (Reyes et al. 2012). The arrival of low-cost genome sequencing has removed the obstacle of homoplasy since it does not strongly affect genome-wide SNPs (after removing SNPs implicated in drug resistance).

15.3.2 Evolutionary Models for Whole Genomes of *M. tuberculosis*

A wide array of molecular evolution models have been developed for nucleotide changes along a genome, covering a range of complexities. These models all assume that each nucleotide evolves independently of the other nucleotides. The simplest is the Jukes-Cantor model (JC69) which assumes that all substitutions among nucleotide bases occur at an equal rate and that base frequencies are all equal (Jukes and Cantor 1969). Kimura's two parameter model (K80) allows different transition and transversion rates, while keeping equal base frequencies (Kimura 1980), whereas Felsenstein's model (F81) keeps equal rates while allowing varying nucleotide base frequencies (Felsenstein 1981). More complex models include the HKY85 model which does not assume equal base frequencies and accounts for the difference between transitions and transversions (Hasegawa et al. 1985), the TN93 model which distinguishes not only transitions/transversions, but also differentiates between purine and pyrimidine transitions (Tamura and Nei 1993), and the generalised time-reversible (GTR) model, which is the least restrictive time-reversible model possible (Tavaré 1986).

Generally, the rate of substitution has been inferred to vary across sites. In order to account for such variation, it is common to assume either a continuous gamma distribution or a discrete approximation with a suitable number of rate classes (Yang 1994a, 1996) for variation across sites in the substitution rate. These substitution models can then be used in combination with genetic data to compute likelihoods or genetic distances for phylogenetic reconstruction, as we next discuss.

15.3.3 Phylogenetic Reconstruction

A number of approaches have been developed for reconstruction of phylogenetic trees from genetic data. Distance-based methods use a measure of genetic distance – an estimate of the degree of similarity two sequences share – to reconstruct evolutionary history. Other methods optimise a criterion measuring how well a tree explains the genetic sequence alignment over the space of possible phylogenetic trees.

In distance-based methods, similar sets of taxa are grouped together whereas more distant ones are placed further apart on a tree according to entries in a pairwise distance matrix. The branch lengths of the inferred phylogeny are a close, but in general imperfect, representation of the inter-sequence distance matrix. An example of a distance-based method is the neighbour-joining method which is a bottom-up clustering algorithm that joins nodes of the tree according to the shortest distance between two existing nodes (Saitou and Nei 1987). It is statistically consistent,[1] but does not maximise a criterion for measuring how well a tree explains the data.

Alternative methods of tree reconstruction from genetic data include maximum parsimony, maximum likelihood and Bayesian tree reconstruction, which all search tree space while optimising a criterion. The maximum parsimony method minimises the number of substitutions required to explain the inferred phylogeny. This method is quick and easy but it has been shown to be statistically inconsistent (Felsenstein 1978).

[1] Statistical consistency of a phylogenetic method means that when given infinitely long genetic sequences, the method –employing the model under which the sequences evolved– will recover the true underlying phylogeny.

The maximum likelihood method searches the tree space to maximise the probability of the data given a particular tree structure (the likelihood of the tree given the data). Bayesian methods also search the tree space, yielding a posterior distribution of trees. As such, Bayesian methods produce multiple trees; they also accommodate prior distributions on trees and parameters. In the maximum likelihood and Bayesian frameworks, the substitution model is used to evaluate the likelihood of trees and model parameters when considering the data. Thus, the methods allow multiple substitutions, parallel substitution, convergent evolution and back substitution along branches, and the complexity of the substitution model can be adjusted to improve the fit to the data. Both frameworks are statistically consistent (Steel 2013; Yang 1994b), and are currently the most widely used phylogenetic reconstruction methods.

Phylogenetic reconstruction methods produce trees with branch lengths measured in numbers of substitutions. Exceptions are the Bayesian methods assuming a clock model (Drummond et al. 2006; Zuckerkandl and Pauling 1965) which leads to branch lengths in units of calendar time. Branch lengths in calendar time are important for quantifying the timing of epidemic spread. Recently, a method to infer branch lengths in calendar time based on a tree with branch lengths on number of substitutions was introduced (LSD (least-squares dating) software (To et al. 2016)). For more details on phylogenetic methods, we refer readers to the texts by Yang (2014) or Felsenstein (2004).

15.3.4 Classification of TB Using Genetic Data

The application of molecular technologies and phylogenetic methods also enables the classification of pathogen isolates into broad classes. In relation to *M. tuberculosis* extensive genetic research has identified the structure of the complex of closely related *M. tuberculosis* species and the seven major extant lineages (see Chap. 1) (Galagan 2014). The classification of *M. tuberculosis* can be further refined by considering relationships within lineages using fast-evolving molecular markers. Attempts have been made to date the first introduction of *M. tuberculosis* into the human population using phylogenetic methods and to describe the patterns of world-wide *M. tuberculosis* distribution (Gagneux 2012). The suitability of alternative genetic approaches can be evaluated by comparing phylogenies reconstructed from different types of markers to a "gold standard" phylogeny in order to identify flaws in commonly used methods and to provide a means of quickly typing isolates (Filliol et al. 2006).

When the isolates in a data set are closely related to each other – such as isolates from a single outbreak – an alternative approach to phylogenies is to show the direct relationships among the genotypes in graphs such as minimal spanning trees. The underlying assumption is that all substitutions that occurred are observed, so that complex substitution models to account for hidden ancestral substitutions are not needed. This approach to visualisation and classification is aided by specifically modelling the processes of substitution underlying the genetic markers. Models of spoligotype evolution have been used to show relationships among isolates (Brudey et al. 2006; Ozcaglar et al. 2011; Reyes et al. 2008; Shabbeer et al. 2012a,b). Relationships among isolates based on MIRU-VNTR can also be visualised using graphs (Weniger et al. 2012). Large international databases based on multiple genotyping methods including MIRU-VNTR and spoligotypes aid in the classification of isolates (see Chap. 3) (Demay et al. 2012; Weniger et al. 2012).

SNPs obtained from whole genome sequencing can also be visualised through graphs within outbreaks (Walker et al. 2013) or through phylogenies when analysing highly divergent isolates with ancient origins (Bos et al. 2014; Comas et al. 2013; Wirth et al. 2008). It is anticipated that future epidemiological studies will increasingly make use of whole genome sequencing.

15.3.5 Population Genetics of TB

So far we have considered models of substitution (i.e. fixation of a mutation in a population), and how the variation between isolates can be represented in a tree. Population genetics "zooms into" the process of substitution by modelling the origins and dynamics of genetic variation including the process of fixation. A natural null model in population genetics is the neutral model in which all genetic variants are selectively equivalent. In this model the process of mutation generates new variants (alleles) while randomness – genetic drift – leads to loss of variation or to fixation (i.e. substitution). The dynamic balance between these two processes – mutation and drift – has been characterised along with properties of samples from a population in such balance (Ewens 1972). The theory of this balance between mutation and drift is generally applicable to bacteria including *M. tuberculosis*.

Because most viable mutations are expected to have zero or negligible effect on bacterial phenotype, variation at marker loci can be considered selectively neutral, as a first approximation. An important exception is antibiotic resistance genes which are removed for the purposes of phylogenetic analysis because they are known to be under selection (Comas et al. 2013). Exceptions to strict neutrality can also occur at marker loci: for instance, movement of IS6110 can lead to deleterious or even advantageous effects. Moreover, in the absence of recombination – as is the case for *M. tuberculosis* – any mutation under positive selection will be linked to neutral variation throughout the genome which may hitchhike to fixation in a selective sweep (Smith et al. 2009). Nevertheless, a first step to analysing most molecular variation is to treat it as selectively neutral to understand broad patterns in data using theory from population genetics (Ewens 2004).

Neutral models have been found to often adequately describe the distribution of cluster sizes under IS6110-RFLP typing and spoligotyping, and the fit can be improved compared to the standard Wright-Fisher population genetic model by allowing the infected population to expand according to a birth-death process (Luciani et al. 2008). These simple population models focus on genetic variation without considering details of the population dynamics or the natural history of disease. For example, at the epidemiological level the population of interest is the set of infected hosts, and the replication process is the transmission process. However, standard population genetic models do not explicitly account for the process of transmission nor other processes such as host recovery, death or latent infection. Thus there is a need to integrate population genetic models with epidemiological dynamics such as those described in Sect. 15.2. Section 15.4 will describe progress towards this goal.

15.3.6 Models of Within-Host Variation and Mixed Infections

At epidemiological scales, it is convenient to assume that each infected case corresponds to a single strain and that mutation leads to the instantaneous replacement of the ancestor, but in reality more than one strain can exist within an infection (Sergeev et al. 2011; Warren et al. 2004). Such infections are called *mixed* or *complex infections*. This bacterial variation may be due to mutation within the host or reinfection of the host by another strain. In order to understand the source and consequences of variation, models of bacterial dynamics at the within-host level are needed. Such modelling has led, for example, to methods to detect mixed infection (Plazzotta et al. 2015) and to classify whether the variation is due to reinfection or mutation (Chindelevitch et al. 2016) using genetic data. In cases where within-host variation is due to mutation, serial isolates of *M. tuberculosis* from an infection can be used to estimate mutation rates (Ragheb et al. 2013; Rosenberg et al. 2003; Tanaka et al. 2004). A benefit of these estimates is that unlike "snapshot" data they make use of temporal information. Within-host data can also be used to study population genetic statistics to quantify the action of natural selection (O'Neill et al. 2015).

We note that the dynamics of *M. tuberculosis* within patients are highly complex and involve a large number of interactions between the pathogen and the host. The roles of the immune system, inter-cellular signals and spatial effects have been modelled (Gammack et al. 2004; Wigginton and Kirschner 2001). In those models, the variation in disease dynamics is due to the complex interactions between *M. tuberculosis* and the host response. To be able to draw conclusions for population level dynamics, models must suppress some of the complexities of the intra-host dynamics while focusing on competition between strains. For example, different *M. tuberculosis* strains can be modelled by imposing structure on the pathogen population (Fig. 15.2, left), and different immune responses can be modelled by imposing structure on host population (Fig. 15.2, right).

Just as drug resistance is an important source of variation on the epidemiological scale, it is also important to consider on the within-host scale. Models that combine the population dynamics of *M. tuberculosis* with the pharmacokinetics of drugs provide a quantitative description of the emergence of resistance within a patient, which can then be used to optimise treatment regimens to minimise drug resistance (Lipsitch and Levin 1997). The effects of nonadherence and drug synergies can be considered under such models (Lipsitch and Levin 1997). Within-host modelling of multidrug resistant *M. tuberculosis* has led to predictions that such strains can arise at an unexpectedly high rate from apparently pansensitive within-host populations because of standing genetic variation in those populations (Colijn et al. 2011; Lipsitch and Levin 1997).

15.4 Integrating Epidemiological and Evolutionary Models

In Sect. 15.2.1 we described models of the epidemiology of disease and of *M. tuberculosis* in particular. In Sect. 15.3 we described how models from the fields of molecular evolution and population genetics can be applied to genetic data from *M. tuberculosis* isolates. We suggest that to fully understand genetic data, it is necessary to combine both kinds of approaches. We will describe developments in this area, including the integration of epidemiological and evolutionary models that involve the analysis of phylogenies – a young field dubbed *phylodynamics* (Grenfell et al. 2004).

15.4.1 Between-Host Variation, Clustering and Transmission

A central goal of molecular epidemiology is to draw inferences about disease transmission using genetic information (Glynn et al. 1999b; Mathema et al. 2006; Murray 2002a; Tanaka and Francis 2005). Above we introduced phylogenetic trees to investigate the diversity of *M. tuberculosis* strains, namely to infer the past history (or state) of the epidemic. In this section, we go a step further and in fact assess the transmission dynamics in addition to the state.

The degree to which isolates cluster into identical genotypes carries information about the extent of recent transmission. The underlying assumption is that genotypes evolve on the same timescale as the process of disease transmission so that each cluster of isolates represents a set of cases that arose recently through transmission, but isolates that are not connected via recent transmission are different through accumulated mutations. Simple clustering statistics have therefore been used to quantify recent transmission of tuberculosis (Alland et al. 1994; Small et al. 1994). Widely used are the "n" and "n-1" statistics ((Glynn et al. 1999b), also known as RTI_n and RTI_{n-1} respectively (Tanaka and Francis 2005)). These are defined as:

$$RTI_n = n_c/n \tag{15.3}$$

$$RTI_{n-1} = (n_c - c)/n \tag{15.4}$$

where n_c is the number of isolates in clusters (of 2 or more isolates), c is the number of clusters and n is the total number of isolates. These can be alternatively written as

$$RTI_n = 1 - u/n \quad (15.5)$$

$$RTI_{n-1} = 1 - g/n \quad (15.6)$$

where g is the number of distinct genotypes in the sample and u is the number of unique genotypes in the sample (also called non-clustered isolates or singletons).

Mathematical modelling has contributed to the goal of improving inferences from these data. First, modelling and simulation studies have shown that incomplete sampling leads to an underestimation of the extent of recent transmission which also interferes with assessing risk factors for recent transmission (Borgdorff et al. 2011; Glynn et al. 1999a,b; Murray 2002b; Murray and Alland 2002). Second, the degree of clustering cannot easily be compared across different types of markers because different markers mutate at different rates, a fact not accounted for in these simple clustering statistics (Tanaka and Francis 2005). Thus, in order to interpret patterns of clustering in terms of disease transmission it is important to know the mutation rates of different markers. Furthermore, separate clusters may not be completely different strains – a single mutation event could split a cluster into two clusters, which should be treated as a single epidemiological cluster of cases. Ideally then, methods of inference from genetic data would both incorporate the speed of evolution and account for sampling.

An approach to analysing genetic data is to use mathematical models that account for disease transmission, marker mutation and sampling. The extent of transmission can be quantified by estimating the effective reproduction number of the pathogen. Because models can be complex and difficult to work with directly, computational methods that approximate the likelihood have been applied to analyse data using models (Aandahl et al. 2014; Luciani et al. 2009; Tanaka et al. 2006). These models do not consider the phylogenetic history of the genetic data. An alternative approach is to augment the data by using trees which permits exact calculation of likelihoods conditional on trees (Stadler 2011). This approach requires a clear definition of trees that represent the evolutionary and epidemiological history of a sample of bacterial isolates. The next section introduces the concept of the transmission tree as a step towards integrating epidemiology and phylogeny.

15.4.2 Transmission Trees

In order to characterise the connection between epidemiological and evolutionary relationships it is necessary to introduce the concept of transmission trees to be studied in light of phylogenetic trees. If all of the infected individuals and the times and sources of every new infection were known, we could represent the spread of an infection as a bifurcating tree. In such a tree, which we will refer to as the transmission tree, each branch represents an infectious case, each bifurcation event represents a secondary case, and the root branch represents the initial introduction of the infection to the population. One of the very few cases of perfect transmission tree reconstruction via complete contact tracing was done in a 1980 study on a quarantined US naval vessel described by Houk (1980).

Unfortunately, in most epidemics it is difficult to achieve complete sampling and to record the precise timing of events; nor can we perform contact tracing, as the infection event might have taken place years prior to the onset of symptoms. Instead, genetic sequencing data are used to estimate the transmission tree relying on the following observation. Upon each infection a subset of the genotypes occurring in the donor is transferred to the recipient, producing a new case. From the time of that transmission event, the pathogen populations in the two distinct infections evolve independently of each other within the two hosts. Thus patients close in the transmission chain have *M. tuberculosis* genomes closer to each other compared to patients distant in the transmission chain. The phylogenetic tree of the pathogen sequences puts similar sequences close together and serves as a proxy for the transmission chain. Note that upon a bifurcation,

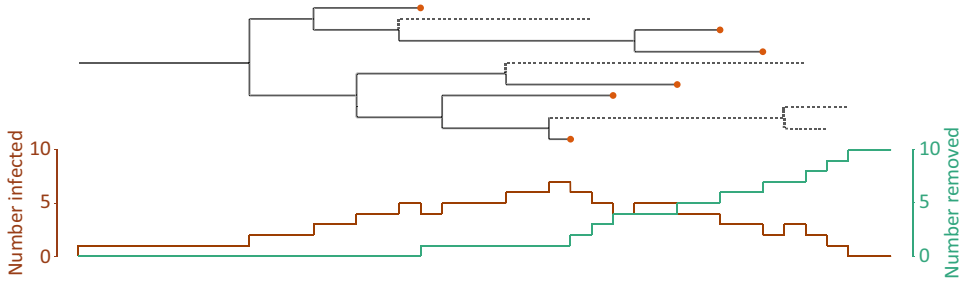

Fig. 15.4 An example of a transmission tree depicting SIR dynamics. The tree shows sampled branches with solid lines and unsampled ones with dotted lines, with samples shown as orange circles. The curves underneath the tree show the changes in the number of infected individuals and the recovered individuals

we do not know which patient is the donor or the recipient in this reconstructed tree.

Genetic sequencing data represent only a subset of active tuberculosis infection because of incomplete sampling due to financial or other constraints. Reconstructed transmission chains are therefore also incomplete. Figure 15.4 shows an example of a transmission tree in an epidemic with SIR dynamics, with incomplete sampling. By observing the reconstructed phylogenetic trees and interpreting them as a proxy for the transmission chain, one can make conclusions about multiple characteristics of the epidemic, such as possible hotspots of infection (e.g. households where multiple family members were infected) and identify whether a specific patient could have been a source of infection in a cluster. Methods to infer the exact transmission tree of *M. tuberculosis* from a phylogenetic tree, including the direction of infection have been recently proposed in Didelot et al. (2014, 2017). Essentially, these methods assign a corresponding sampled or unsampled infected host to lineages.

15.4.3 Phylodynamic Methods

As sequencing of genomes becomes more cost-effective, fast and reliable, increasing amounts of sequence data are sampled from ongoing epidemics, and phylogenetic trees as well as transmission trees are thus being reconstructed. This increase in sequence data has stimulated the development of phylodynamic methods that combine evolutionary and epidemiological analyses to quantify the parameters of the epidemiological models. In fact, the structure and branch length of the reconstructed phylogenetic tree contains information on the different compartments and rates of movement (i.e. dynamics) between the compartments in the underlying epidemiological model. For example, in the case of an SIR model with all patients being sampled upon recovery, the waiting times for a new branching event will be exponentially distributed with mean rate $\lambda = \beta I$ and the branches will terminate (the individual will stop being infectious) with mean rate γ. Incomplete sampling requires the development of sophisticated statistical tools integrating over all non-sampled patients (Stadler et al. 2013; Volz et al. 2009). The dependency will be more complex for a model including a latent or other non-infectious class, or allowing for heterogeneous pathogen or host population as in Fig. 15.2 (Kühnert et al. 2016; Stadler and Bonhoeffer 2013; Volz 2012). In such a case the reconstructed phylogenetic tree is extended by labelling each tip in the tree with the compartment the corresponding host belongs to: if tips from the same compartment cluster in the tree we conclude that there is transmission within the compartment, while if tips from the same compartment are spread over the tree we conclude independent migration into that compartment. The phylodynamic methods allow us to quantify the rates of transmission and migration in the epidemiological models based on the phylogenetic trees with the tip labels.

There are two main approaches to infer epidemiological parameters from pathogen sequencing data, which we call the two-step and the one-step approaches.

In the two-step approach, one first produces a phylogenetic tree with branch lengths in calendar units, using a tree reconstruction method as discussed in Sect. 15.3.3. The reconstructed tree is used as fixed input in the second analysis step to infer epidemiological parameters (see e.g. Kühnert et al. 2014, 2016; Stadler and Bonhoeffer 2013; Stadler et al. 2013; Volz 2012; Volz et al. 2009). Most of these methods are available within the software package BEAST2 (Bouckaert et al. 2014) and stand-alone implementations are mentioned in the individual papers. This parameter inference based on a fixed tree can be done using maximum likelihood (ML) or Bayesian inference. ML inference focuses on finding the combination of parameters that was the most likely to have produced the phylogenetic tree that is being studied under the given epidemiological model. The ML framework does not make use of prior knowledge of the parameters of the underlying models. In contrast, Bayesian methods can incorporate prior distributions of parameters and yield posterior distributions over the parameter space. Posterior distributions are a natural way to interpret the uncertainty in the resulting estimates. The incorporation of prior distributions allows for the better use of all the information available but requires great care in prior specification, as inappropriate priors can significantly and incorrectly influence the results of the analysis.

One-step approaches simultaneously estimate trees and parameters from the genetic sequences, typically in a Bayesian framework. In the one-step Bayesian approach the uncertainty in the phylogenetic trees is naturally integrated out. In other words, the posterior distributions of our epidemiological parameters take into account tree uncertainty. This will become particularly useful for *M. tuberculosis* analyses as the low diversity in whole genome *M. tuberculosis* strains leads to high uncertainty in trees. For this one-step approach we jointly assume an epidemiological model such as the SIR model, which gives rise to a probability distribution over the tree space, and an evolutionary model such as GTR, which defines the probability distribution of the alignment of sequences. The output is a posterior distribution of trees, and the epidemiological and evolutionary parameters. Software packages BEAST2 (Bouckaert et al. 2014) and BEAST1 (Drummond et al. 2012) can simultaneously infer the evolutionary history and the epidemiological parameters under some simple epidemiological models.[2]

This one-step approach had been used for viral datasets such as HCV (Pybus et al. 2003), HIV (Stadler et al. 2012), or influenza (Kühnert et al. 2016). Phylodynamics on viruses can be done based on single genes as the substitution rates are high enough to see differences in single genes of the virus in donor compared to recipient. The slower-evolving *M. tuberculosis* requires whole genomes to reconstruct the phylogenetic trees and the epidemiological parameters. With next-generation sequencing technologies, such whole genomes become increasingly available, opening the door for phylodynamic analysis.

Such analyses have been done for assessing the rate of transmitted drug resistance. In Casali et al. (2014) the tips in the reconstructed phylogenetic tree were labelled according to the drug resistance status. Short inter-sequence distance was used to infer transmission links and to assess the transmission fitness costs in drug-resistant strains.

15.5 Practical Implications

Mathematical modelling helps us explain and predict the dynamics of tuberculosis, including the origins and future of strain diversity. Models aid in estimating the rates of transmission and reactivation, which in turn can influence the design of population interventions, and therefore models that incorporate at least the conclusions

[2]BEAST2 started out as a re-design of BEAST1, but over the course of time the two platforms continued to evolve independently with new features being implemented in both.

from strain diversity studies are of importance in targeting interventions to achieve WHO goals of eliminating *M. tuberculosis*.

15.5.1 Classification and Outbreaks

In a practical sense, model-based analysis of genetic diversity data for *M. tuberculosis* can be useful both for reactive purposes such as outbreak and contact investigations and longer-term policy definitions for addressing problems such as drug-resistance.

In relation to *M. tuberculosis* cases and contact interventions, modelling has most potential to be useful in high-resource environments where relevant data collection on cases and contacts provide nearly complete strain information including genotyping and sequencing of samples. A recent demonstration study suggested that use of WGS-based testing and identification of resistance was a more cost-effective solution for resistance testing and case follow-up than existing methods (Pankhurst et al. 2016).

Application of modelling tools such as BEAST2 to analyse genotype and whole genome sequencing data (Gardy et al. 2011; Outhred et al. 2016) for *M. tuberculosis* isolates can inform outbreak investigations by more precisely identifying links, which in turn has implications for epidemiological analyses of risk-factors for transmission and disease. In addition, modelling can be used to better understand measurement processes and biases in collection of samples, for instance the simulations conducted by Plazzotta et al. (2015) show how the prevalence of mixed-infection can be corrected for based on modelled properties of the infection and sample collection process.

A related question that can be informed by diversity data is the estimation of the proportions of recent transmission, relapse and reactivation. Determining this balance is important as it can help decide which aspects of treatment and prevention programs require attention. For instance a high rate of recent transmission would suggest prioritising case finding and treatment success rates, while high rates of reactivation might suggest the need for increased use of preventive therapy. Studies using WGS to investigate transmission were recently reviewed by Hatherell et al. (2016) who suggest that while these approaches are very helpful, improvements are still needed not only to data fidelity but also to the models of transmission trees and to the development of model-based thresholds for genetic distance to distinguish linked and unlinked cases.

15.5.2 Correlation Between Pathogen Genetics and Host Outcomes

A separate question that can be addressed with genetic data is whether differences in infection and disease outcomes are due to pathogen diversity or due to factors unrelated to the pathogen characteristics (see Chap. 5). For example, differences in progression to disease from infections with *M. africanum* as opposed to *M. tuberculosis* have been observed in a large epidemiological cohort study in the Gambia (de Jong et al. 2008).

While such differences in disease natural history between related species might be expected, it does raise the question of whether epidemiological differences in infection, progression or disease outcomes differ between strain groupings and whether this might need attention in terms of disease control. Evidence for this variation, discussed in Chap. 5 and in the comprehensive review by Coscolla and Gagneux (2014), demonstrates a range of differences between *M. tuberculosis* lineages at molecular, in vivo and in vitro levels. However, epidemiological evidence for special properties of, for example, the Beijing strain, is less conclusive and has not yet been a major focus in modelling studies. We note that Comas and Gagneux (2011) have argued for a "systems epidemiology" approach using computational models to address such questions and expect the increased use of the mathematical techniques such as those summarised here.

15.5.3 Dynamics of Resistance

A major practical focus for modelling strain diversity is the phenomena of multi-drug and extremely drug resistant tuberculosis. While poor individual outcomes and the high cost of such treatment have been influential in altering WHO policy for detecting and treating MDR-TB, models have played a key part in showing the potential for expansion of resistant strains (Blower and Gerberding 1998; Cohen and Murray 2004; Dye and Espinal 2001; Dye and Williams 2000) and identifying the need for drug-sensitivity testing as part of approaches to mitigate these effects. Implementation of treatment programs for MDR-TB in lower income settings has been facilitated by the development of molecular genetic tests such as *geneXpert* (Boehme et al. 2010) and more detailed modelling studies that directly assess and compare the cost-effectiveness of relevant population treatment strategies.

These issues have been explored primarily through deterministic epidemiological models, with strain heterogeneity simplified to sensitive and resistant classes, including the potential for latent infection with both resistant and sensitive TB (Colijn et al. 2009). These approaches are valuable for assessing the impacts of relative fitness under strong treatment related selection, either for active TB (Cohen and Murray 2004) or preventive therapy (Cohen et al. 2006) in latently infected populations.

Extensions of such models have also been applied to describe trends in MDR+ resistance through an expanded classification of resistance properties from mono-resistance through to MDR and XDR-TB (Menzies et al. 2012). Models with this additional detail on resistance allow prediction of the effects of molecular testing for resistance and appropriate on epidemic trajectories and the prevalence of MDR+ resistance. Modelling studies have in general found that tests such as *geneXpert* offer acceptable value for money even in lower resource settings and these studies have helped facilitate a rapid scale-up in testing and treatment for MDR-TB since 2013 (WHO 2016). There has been a recent proliferation of studies assessing the effectiveness and cost-effectiveness of such strategies, as summarised by Zwerling et al. (2016) who note that molecular tests have generally had positive cost-effectiveness findings but that future models need to feature increased use of setting-specific parameters in relation to treatment and diagnostic programs.

In 2015 WHO established its end-TB strategy which sets ambitious goals for reducing new cases by 90% by 2035. This target does not seem achievable without reducing the burden of reactivated TB cases through treatment of latent tuberculosis (Rangaka et al. 2015) and suggests that the prevalence of resistance in latent infections will have substantial impact on the success of preventative treatment strategies. It will be important to detect and know the extent of mixed infection involving sensitive and resistant bacteria (Cohen et al. 2012; Mills et al. 2013), through both data collection and use of models to correct biases in observations.

Models have also considered that increases in the prevalence of resistance relate to the variance of reproductive fitness among resistant strains. In accordance with expectations from evolutionary theory, variance in fitness enhances the success of resistance even if the mean fitness is relatively low (Knight et al. 2015). In particular these results indicate that while resistance that emerges under treatment will in general be poorly transmissible, transmission will nevertheless become dominated by resistant strains with the highest fitness (Knight et al. 2015) in the absence of rapid identification and appropriate treatment for MDR-TB. Such models have the advantage of illustrating how high-fitness resistant strains can gain an increase in prevalence in poorly controlled epidemics.

Dynamic models that consider strain diversity more directly have been less common and have been concerned with scientific questions such as estimation of fitness costs of resistance. For instance stochastic epidemiological models with genotype evolution have been used to estimate the relative fitness of drug resistant strains, and to estimate the relative importance of transmission of resistant strains versus acquired resistance (Luciani et al. 2009). The practical impli-

cation of such studies is to emphasise the value in reducing the average period of infectiousness for individuals with resistant *M. tuberculosis* infections. Models can then be used as decision tools to help guide us to the most effective means to achieve this goal, for example through increasing overall case-finding, reducing the time between identification of a case and tests for resistance or improving treatment compliance and outcomes. These findings can have both policy and research implications, for instance in terms of suggesting the characteristics of potential diagnostics (Dowdy et al. 2014) or treatment regimens that would be most beneficial (Zwerling et al. 2016).

15.5.4 Future Directions

Tuberculosis modelling is a rapidly growing field, with a number of key directions in which modelling research is progressing. In relation to models of *M. tuberculosis* variation, it is particularly important to refine models for whole genome sequencing (WGS) data analysis because much future data will be of this kind. We discussed some of the recent developments in this broad area in Sect. 15.4 but further work can be done. For example, more realistic models could be developed to link dynamics at within- and between-host levels.

Models of the within-host *M. tuberculosis* infection should ideally feature a more fine-grained characterisation of the natural history of disease, including interactions between the immune system and the pathogen population and understanding features of TB infection such as granulomas. Research in this area has moved from more theoretical explorations (Chang et al. 2005) to studies of potential biomarkers (Marino et al. 2016) and enhancements to therapy (Linderman and Kirschner 2015). One other open topic of research is the quantification of substitution rates for *M. tuberculosis* in the latent and in the acute stage of the disease.

Another aspect of tuberculosis epidemiology that has been relatively underexplored is the variation in host susceptibility to infection and disease. We briefly described approaches to describing host variation through models in Sect. 15.2.3. Future work may extend such models to consider coevolution between the host immune system and *M. tuberculosis*. We note that variation in hosts can be due to genetic or non-genetic factors and that although genetic susceptibility to tuberculosis infection has been studied (Bellamy et al. 2000) the important sources of host variation are arguably non-genetic factors such as age, HIV status and other immunosuppression, diabetes, BCG vaccination and living conditions including nutrition, crowding and smoking behaviour. While host factors such as these are commonly included in risk-prediction models for non-communicable diseases (e.g. coronary artery disease), their adoption in transmission models for *M. tuberculosis* has been slow. However, there is increasing recognition that social determinants of transmission and disease are important in characterizing *M. tuberculosis* epidemics and in setting priorities for control (Andrews et al. 2015). As host factors are often closely linked to key characteristics of treatment programs (e.g. compliance with treatment), there can be flow on effects to pathogen variation. As such, it is likely that there is value in integrated approaches that take both host and pathogen variation into account. We see these approaches as valuable not only in terms of explaining existing epidemiology but in more local prediction of epidemic outcomes under changes to control measures.

References

Aandahl RZ, Reyes JF, Sisson SA, Tanaka MM (2012) A model-based Bayesian estimation of the rate of evolution of VNTR loci in *Mycobacterium tuberculosis*. PLoS Comput Biol 8(6):e1002573. doi:10.1371/journal.pcbi.1002573, http://dx.doi.org/10.1371/journal.pcbi.1002573

Aandahl RZ, Stadler T, Sisson SA, Tanaka MM (2014) Exact vs. approximate computation: reconciling different estimates of *Mycobacterium tuberculosis* epidemiological parameters. Genetics 196(4):1227–1230. doi:10.1534/genetics.113.158808

Alland D, Kalkut G, Moss A, McAdam R, Hahn J, Bosworth W, Drucker E, Bloom B (1994) Transmission of tuberculosis in New York City. An analysis by DNA fingerprinting and conventional epidemiologic methods. N Engl J Med 330(24):1710–1716

Anderson RM, May RM (1979) Population biology of infectious diseases: Part I. Nature 280:361–367

Andrews JR, Basu S, Dowdy DW, Murray MB (2015) The epidemiological advantage of preferential targeting of tuberculosis control at the poor. Int J Tuberc Lung Dis 19(4):375–380

Aranaz A, Romero B, Montero N, Alvarez J, Bezos J, de Juan L, Mateos A, Domínguez L (2004) Spoligotyping profile change caused by deletion of a direct variable repeat in a *Mycobacterium tuberculosis* isogenic laboratory strain. J Clin Microbiol 42(11):5388–5391. doi:10.1128/JCM.42.11.5388-5391.2004, http://dx.doi.org/10.1128/JCM.42.11.5388-5391.2004

Basu S, Galvani AP (2009) The evolution of tuberculosis virulence. Bull Math Biol 71(5):1073–1088. doi:10.1007/s11538-009-9394-x, http://dx.doi.org/10.1007/s11538-009-9394-x

Becker N (2015) Modeling to inform infectious disease control. Chapman & Hall/CRC biostatistics series. CRC Press/Taylor & Francis, Boca Raton. https://books.google.com.au/books?id=F_MQrgEACAAJ

Bellamy R, Beyers N, McAdam KP, Ruwende C, Gie R, Samaai P, Bester D, Meyer M, Corrah T, Collin M, Camidge DR, Wilkinson D, Hoal-Van Helden E, Whittle HC, Amos W, van Helden P, Hill AV (2000) Genetic susceptibility to tuberculosis in Africans: a genome-wide scan. Proc Natl Acad Sci U S A 97(14):8005–8009. doi:10.1073/pnas.140201897, http://dx.doi.org/10.1073/pnas.140201897

Blower SM, Gerberding JL (1998) Understanding, predicting and controlling the emergence of drug-resistant tuberculosis: a theoretical framework. J Mol Med (Berl) 76(9):624–636

Blower SM, McLean AR, Porco TC, Small PM, Hopewell PC, Sanchez MA, Moss AR (1995) The intrinsic transmission dynamics of tuberculosis epidemics. Nature Med 1:815–821

Blower SM, Small PM, Hopewell PC (1996) Control strategies for tuberculosis epidemics: new models for old problems. Science 273:497–500

Boehme CC, Nabeta P, Hillemann D, Nicol MP, Shenai S, Krapp F, Allen J, Tahirli R, Blakemore R, Rustomjee R, Milovic A, Jones M, O'Brien SM, Persing DH, Ruesch-Gerdes S, Gotuzzo E, Rodrigues C, Alland D, Perkins MD (2010) Rapid molecular detection of tuberculosis and rifampin resistance. N Engl J Med 363(11):1005–1015

Borgdorff MW, van den Hof S, Kalisvaart N, Kremer K, van Soolingen D (2011) Influence of sampling on clustering and associations with risk factors in the molecular epidemiology of tuberculosis. Am J Epidemiol 174(2):243–251. doi:10.1093/aje/kwr061, http://dx.doi.org/10.1093/aje/kwr061

Borrell S, Gagneux S (2011) Strain diversity, epistasis and the evolution of drug resistance in *Mycobacterium tuberculosis*. Clin Microbiol Infect 17(6):815–820. doi:10.1111/j.1469-0691.2011.03556.x, http://www.ncbi.nlm.nih.gov/pubmed/21682802

Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, Forrest SA, Bryant JM, Harris SR, Schuenemann VJ et al (2014) Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature 514(7523):494–497

Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol 10(4):e1003537. doi:10.1371/journal.pcbi.1003537, http://www.ncbi.nlm.nih.gov/pubmed/24722319

Brooks-Pollock E, Becerra MC, Goldstein E, Cohen T, Murray MB (2011) Epidemiologic inference from the distribution of tuberculosis cases in households in Lima, Peru. J Infect Dis 203(11):1582–1589. doi:10.1093/infdis/jir162, http://www.ncbi.nlm.nih.gov/pubmed/21592987

Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, Al-Hajoj SA, Allix C, Aristimuño L, Arora J, Baumanis V et al (2006) *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol 6(1):23

Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, Kontsevaya I, Corander J, Bryant J, Parkhill J, Nejentsev S, Horstmann RD, Brown T, Drobniewski F (2014) Evolution and transmission of drug-resistant tuberculosis in a Russian population. Nat Genet 46(3):279–286. doi:10.1038/ng.2878, http://www.ncbi.nlm.nih.gov/pubmed/24464101

Cave MD, Eisenach KD, McDermott PF, Bates JH, Crawford JT (1991) IS *6110*: conservation of sequence in the *Mycobacterium tuberculosis* complex and its utilization in DNA fingerprinting. Mol Cell Probes 5(1):73–80

Chang ST, Linderman JJ, Kirschner DE (2005) Multiple mechanisms allow *Mycobacterium tuberculosis* to continuously inhibit MHC class II-mediated antigen presentation by macrophages. Proc Natl Acad Sci U S A 102(12):4530–4535

Chindelevitch L, Colijn C, Moodley P, Wilson D, Cohen T (2016) ClassTR: classifying within-host heterogeneity based on tandem repeats with application to *Mycobacterium tuberculosis* infections. PLoS Comput Biol 12(2):e1004475. doi:10.1371/journal.pcbi.1004475, http://dx.doi.org/10.1371/journal.pcbi.1004475

Chisholm RH, Tanaka MM (2016) The emergence of latent infection in the early evolution of *Mycobacterium tuberculosis*. Proc Biol Sci 283(1831). doi:10.1098/rspb.2016.0499, http://dx.doi.org/10.1098/rspb.2016.0499

Chisholm RH, Trauer JM, Curnoe D, Tanaka MM (2016) Controlled fire use in early humans might have triggered the evolutionary emergence of tuberculosis. Proc Natl Acad Sci U S A 113(32):9051–9056. doi:10.1073/pnas.1603224113

Cohen T, Murray M (2004) Modeling epidemics of multidrug-resistant M. tuberculosis of heterogeneous fitness. Nat Med 10(10):1117–1121. doi:10.1038/nm1110, http://dx.doi.org/10.1038/nm1110

Cohen T, Lipsitch M, Walensky RP, Murray M (2006) Beneficial and perverse effects of isoniazid preventive therapy for latent tuberculosis infection in HIV-tuberculosis coinfected populations. Proc Natl Acad Sci U S A 103(18):7042–7047. doi:10.1073/pnas.0600349103, https://www.scopus.com/inward/record.uri?eid=2-s2.0-33646493800&partnerID=40&md5=affd24e21a2915b31375d88cb57caf88

Cohen T, van Helden PD, Wilson D, Colijn C, McLaughlin MM, Abubakar I, Warren RM (2012) Mixed-strain *Mycobacterium tuberculosis* infections and the implications for tuberculosis treatment and control. Clin Microbiol Rev 25(4):708–719. doi:10.1128/CMR.00021-12, https://www.scopus.com/inward/record.uri?eid=2-s2.0-84867198547&partnerID=40&md5=bb6952b52ec058f2821621f8f8b3c52d

Colangeli R, Arcus VL, Cursons RT, Ruthe A, Karalus N, Coley K, Manning SD, Kim S, Marchiano E, Alland D (2014) Whole genome sequencing of *Mycobacterium tuberculosis* reveals slow growth and low mutation rates during latent infections in humans. PLoS One 9(3):e91024. doi:10.1371/journal.pone.0091024, http://dx.doi.org/10.1371/journal.pone.0091024

Colijn C, Cohen T, Murray M (2009) Latent coinfection and the maintenance of strain diversity. Bull Math Biol 71(1):247–263. doi:10.1007/s11538-008-9361-y, http://dx.doi.org/10.1007/s11538-008-9361-y

Colijn C, Cohen T, Ganesh A, Murray M (2011) Spontaneous emergence of multiple drug resistance in tuberculosis before and during therapy. PLoS One 6(3):e18327. doi:10.1371/journal.pone.0018327, http://dx.doi.org/10.1371/journal.pone.0018327

Comas I, Gagneux S (2011) A role for systems epidemiology in tuberculosis research. Trends Microbiol 19(10):492–500

Comas I, Homolka S, Niemann S, Gagneux S (2009) Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. PLoS One 4(11):e7815. doi:10.1371/journal.pone.0007815, http://dx.doi.org/10.1371/journal.pone.0007815

Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G et al (2013) Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat Genet 45(10):1176–1182

Coscolla M, Gagneux S (2014) Consequences of genomic diversity in *Mycobacterium tuberculosis*. Semin Immunol 26(6):431–444. doi:10.1016/j.smim.2014.09.012, https://www.ncbi.nlm.nih.gov/pubmed/25453224

de Boer AS, Borgdorff MW, de Haas PE, Nagelkerke NJ, van Embden JD, van Soolingen D (1999) Analysis of rate of change of IS6110 RFLP patterns of *Mycobacterium tuberculosis* based on serial patient isolates. J Infect Dis 180(4):1238–1244. doi:10.1086/314979, http://dx.doi.org/10.1086/314979

Demay C, Liens B, Burguière T, Hill V, Couvin D, Millet J, Mokrousov I, Sola C, Zozio T, Rastogi N (2012) SITVITWEB – a publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology. Infect Genet Evol 12(4):755–766

Di Rienzo A, Peterson AC, Garza JC, Valdes AM, Slatkin M, Freimer NB (1994) Mutational processes of simple-sequence repeat loci in human populations. Proc Natl Acad Sci U S A 91(8):3166–3170

Didelot X, Gardy J, Colijn C (2014) Bayesian inference of infectious disease transmission from whole-genome sequence data. Mol Biol Evol 31(7):1869–1879. doi:10.1093/molbev/msu121, http://dx.doi.org/10.1093/molbev/msu121

Didelot X, Fraser C, Gardy J, Colijn C (2017) Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. Mol Biol Evol 34(4):997–1007.

Diekmann O, Heesterbeek JAP (2000) Mathematical epidemiology of infectious diseases: model building, analysis and interpretation. Wiley, Chichester

Diekmann O, Heesterbeek JAP, Metz JAJ (1990) On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. J Math Biol 28(4). doi:10.1007/bf00178324

Dietz K (1993) The estimation of the basic reproduction number for infectious diseases. Stat Methods Med Res 2(1):23–41

Dowdy DW, Dye C, Cohen T (2013) Data needs for evidence-based decisions: a tuberculosis modeler's 'wish list'. Int J Tuberc Lung Dis 17(7):866–877. doi:10.5588/ijtld.12.0573, http://www.ncbi.nlm.nih.gov/pubmed/23743307

Dowdy DW, Houben R, Cohen T, Pai M, Cobelens F, Vassall A, Menzies NA, Gomez GB, Langley I, Squire SB, White R (2014) Impact and cost-effectiveness of current and future tuberculosis diagnostics: the contribution of modelling. Int J Tuberc Lung Dis 18(9):1012–1018

Drummond AJ, Ho SY, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. PLoS Biol 4(5):e88. doi:10.1371/journal.pbio.0040088, https://www.ncbi.nlm.nih.gov/pubmed/16683862

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution 29(8):1969–1973. doi:10.1093/molbev/mss075

Dye C, Espinal MA (2001) Will tuberculosis become resistant to all antibiotics? Proc R Soc Lond B Biol Sci 268(1462):45–52

Dye C, Williams BG (2000) Criteria for the control of drug-resistant tuberculosis. Proc Natl Acad Sci U S A 97(14):8180–8185

Dye C, Williams BG (2009) Slow elimination of multidrug-resistant tuberculosis. Sci Transl Med 1(3):3ra8. doi:10.1126/scitranslmed.3000346, http://dx.doi.org/10.1126/scitranslmed.3000346

Dye C, Williams BG (2010) The population dynamics and control of tuberculosis. Science 328(5980):856–861

Dye C, Garnett GP, Sleeman K, Williams BG (1998) Prospects for worldwide tuberculosis control under the WHO DOTS strategy. The Lancet 352(9144):1886–1891. http://dx.doi.org/10.1016/S0140-6736(98)03199-7, http://www.sciencedirect.com/science/article/pii/S0140673698031997

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor Popul Biol 3(1):87–112

Ewens WJ (2004) Mathematical population genetics 1: theoretical introduction, vol 27, 2nd edn. Springer, New York

Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27(4):401–410. doi:10.2307/2412923, <GotoISI>://WOS:A1978GH36300002

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17(6):368–376. doi:10.1007/bf01734359, http://www.ncbi.nlm.nih.gov/pubmed/7288891

Felsenstein J (2004) Inferring phylogenies, vol 2. Sinauer Associates Sunderland

Filliol I, Motiwala AS, Cavatore M, Qi W, Hazbon MH, Bobadilla del Valle M, Fyfe J, Garcia-Garcia L, Rastogi N, Sola C, Zozio T, Guerrero MI, Leon CI, Crabtree J, Angiuoli S, Eisenach KD, Durmaz R, Joloba ML, Rendon A, Sifuentes-Osornio J, Ponce de Leon A, Cave MD, Fleischmann R, Whittam TS, Alland D (2006) Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. J Bacteriol 188(2):759–772. doi:10.1128/JB.188.2.759-772.2006, http://www.ncbi.nlm.nih.gov/pubmed/16385065

Fine PE, Vynnycky E (1998) The effect of heterologous immunity upon the apparent efficacy of (e.g. bcg) vaccines. Vaccine 16(20):1923–1928. http://dx.doi.org/10.1016/S0264-410X(98)00124-8, http://www.sciencedirect.com/science/article/pii/S0264410X98001248

Ford CB, Lin PL, Chase MR, Shah RR, Iartchouk O, Galagan J, Mohaideen N, Ioerger TR, Sacchettini JC, Lipsitch M, Flynn JL, Fortune SM (2011) Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. Nat Genet 43(5):482–486. doi:10.1038/ng.811, http://dx.doi.org/10.1038/ng.811

Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, Johnston JC, Gardy J, Lipsitch M, Fortune SM (2013) *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. Nat Genet 45(7):784–790

Gagneux S (2012) Host-pathogen coevolution in human tuberculosis. Philos Trans R Soc B 367(1590):850–859

Gagneux S, Small PM (2007) Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. Lancet Infect Dis 7(5):328–337. doi:10.1016/S1473-3099(07)70108-1, http://dx.doi.org/10.1016/S1473-3099(07)70108-1

Gagneux S, Long CD, Small PM, Van T, Schoolnik GK, Bohannan BJM (2006) The competitive cost of antibiotic resistance in *Mycobacterium tuberculosis*. Science 312(5782):1944–1946. doi:10.1126/science.1124410, http://dx.doi.org/10.1126/science.1124410

Galagan JE (2014) Genomic insights into tuberculosis. Nat Rev Genet 15(5):307–320. doi:10.1038/nrg3664, http://dx.doi.org/10.1038/nrg3664

Gammack D, Doering CR, Kirschner DE (2004) Macrophage response to *Mycobacterium tuberculosis* infection. J Math Biol 48(2):218–242

Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJM, Brinkman FSL, Brunham RC, Tang P (2011) Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med 364(8):730–739. doi:10.1056/NEJMoa1003176, http://dx.doi.org/10.1056/NEJMoa1003176

Glynn JR, Bauer J, de Boer AS, Borgdorff MW, Fine PE, Godfrey-Faussett P, Vynnycky E (1999a) Interpreting DNA fingerprint clusters of *Mycobacterium tuberculosis*. European Concerted Action on Molecular Epidemiology and Control of Tuberculosis. Int J Tuberc Lung Dis 3(12):1055–1060

Glynn JR, Vynnycky E, Fine PE (1999b) Influence of sampling on estimates of clustering and recent transmission of *Mycobacterium tuberculosis* derived from DNA fingerprinting techniques. Am J Epidemiol 149(4):366–371

Gomes MGM, Franco AO, Gomes MC, Medley GF (2004) The reinfection threshold promotes variability in tuberculosis epidemiology and vaccine efficacy. Proc Biol Sci 271(1539):617–623. doi:10.1098/rspb.2003.2606, http://dx.doi.org/10.1098/rspb.2003.2606

Grant A, Arnold C, Thorne N, Gharbia S, Underwood A (2008) Mathematical modelling of *Mycobacterium tuberculosis* VNTR loci estimates a very slow mutation rate for the repeats. J Mol Evol 66(6):565–574. doi:10.1007/s00239-008-9104-6, http://dx.doi.org/10.1007/s00239-008-9104-6

Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, Holmes EC (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303(5656):327–332. doi:10.1126/science.1090727, http://dx.doi.org/10.1126/science.1090727

Hasegawa M, Kishino H, Yano Ta (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22(2): 160–174

Hatherell HA, Colijn C, Stagg HR, Jackson C, Winter JR, Abubakar I (2016) Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. BMC Med 14:21. doi:10.1186/s12916-016-0566-x, https://www.ncbi.nlm.nih.gov/pubmed/27005433

Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, Roach JC, Kremer K, Petrov DA, Feldman MW, Gagneux S (2008) High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. PLoS Biol 6(12):e311. doi:10.1371/journal.pbio.0060311, http://dx.doi.org/10.1371/journal.pbio.0060311

Ho SYW, Phillips MJ, Cooper A, Drummond AJ (2005) Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Mol Biol Evol 22(7):1561–1568. doi:10.1093/molbev/msi145, http://dx.doi.org/10.1093/molbev/msi145

Houk VN (1980) Spread of tuberculosis via recirculated air in a naval vessel: the Byrd study. Ann N Y Acad Sci 353:10–24. http://www.ncbi.nlm.nih.gov/pubmed/6939378

de Jong BC, Hill PC, Aiken A, Awine T, Antonio M, Adetifa IM, Jackson-Sillah DJ, Fox A, Deriemer K, Gagneux S, Borgdorff MW, McAdam KP, Corrah T, Small PM, Adegbola RA (2008) Progression to active tuberculosis, but not transmission, varies by *Mycobacterium tuberculosis* lineage in The Gambia. J Infect Dis 198(7):1037–1043. doi:10.1086/591504, https://www.ncbi.nlm.nih.gov/pubmed/18702608

Jukes T, Cantor C (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism, vol 3. Academic, New York, pp 21–132

Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M, van Embden J (1997) Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J Clin Microbiol 35(4):907–914

Keeling MJ, Rohani P (2008) Modeling infectious diseases in humans and animals. Princeton University Press, Princeton

Kendall EA, Fofana MO, Dowdy DW (2015) Burden of transmitted multidrug resistance in epidemics of tuberculosis: a transmission modelling analysis. Lancet Respir Med 3(12):963–972. doi:10.1016/s2213-2600(15)00458-0

Kermack WO, McKendrick AG (1927) Contributions to the mathematical theory of epidemics. Proc R Soc Ser 115:700–721

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16(2):111–120

Knight GM, Colijn C, Shrestha S, Fofana M, Cobelens F, White RG, Dowdy DW, Cohen T (2015) The distribution of fitness costs of resistance-conferring mutations is a key determinant for the future burden of drug-resistant tuberculosis: a model-based analysis. Clin Infect Dis 61(Suppl 3):S147–S154. doi:10.1093/cid/civ579, http://dx.doi.org/10.1093/cid/civ579

Kühnert D, Stadler T, Vaughan TG, Drummond AJ (2014) Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death SIR model. J R Soc Interface 11(94):20131106. doi:10.1098/rsif.2013.1106, http://dx.doi.org/10.1098/rsif.2013.1106

Kühnert D, Stadler T, Vaughan TG, Drummond AJ (2016) Phylodynamics with migration: a computational framework to quantify population structure from genomic data. Mol Biol Evol. doi:10.1093/molbev/msw064, http://dx.doi.org/10.1093/molbev/msw064

Lillebaek T, Norman A, Rasmussen EM, Marvig RL, Folkvardsen DB, Andersen ÅB, Jelsbak L (2016) Substantial molecular evolution and mutation rates in prolonged latent *Mycobacterium tuberculosis* infection in humans. Int J Med Microbiol. doi:10.1016/j.ijmm.2016.05.017, http://dx.doi.org/10.1016/j.ijmm.2016.05.017

Linderman JJ, Kirschner DE (2015) In silico models of m. tuberculosis infection provide a route to new therapies. Drug Discov Today Dis Models 15:37–41. http://dx.doi.org/10.1016/j.ddmod.2014.02.006, http://www.sciencedirect.com/science/article/pii/S1740675714000115. Computational models of lung diseases

Lipsitch M, Levin BR (1997) The within-host population dynamics of antibacterial chemotherapy: conditions for the evolution of resistance. Ciba Found Symp 207:112–127; discussion 127–130

Liu X, Gutacker MM, Musser JM, Fu YX (2006) Evidence for recombination in *Mycobacterium tuberculosis*. J Bacteriol 188(23):8169–8177. doi:10.1128/JB.01062-06, http://www.ncbi.nlm.nih.gov/pubmed/16997954

Luciani F, Francis AR, Tanaka MM (2008) Interpreting genotype cluster sizes of *Mycobacterium tuberculosis* isolates typed with IS6110 and spoligotyping. Infect Genet Evol 8(2):182–190. doi:10.1016/j.meegid.2007.12.004, http://dx.doi.org/10.1016/j.meegid.2007.12.004

Luciani F, Sisson SA, Jiang H, Francis AR, Tanaka MM (2009) The epidemiological fitness cost of drug resistance in *Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A 106(34):14711–14715. doi:10.1073/pnas.0902437106, http://dx.doi.org/10.1073/pnas.0902437106

Marino S, Gideon HP, Gong C, Mankad S, McCrone JT, Lin PL, Linderman JJ, Flynn JL, Kirschner DE (2016) Computational and empirical studies predict *Mycobacterium tuberculosis*-specific T cells as a biomarker for infection outcome. PLoS Comput Biol 12(4):e1004804

Mathema B, Kurepina NE, Bifani PJ, Kreiswirth BN (2006) Molecular epidemiology of tuberculosis: current insights. Clin Microbiol Rev 19(4):658–685. doi:10.1128/CMR.00061-05, http://dx.doi.org/10.1128/CMR.00061-05

Menzies NA, Cohen T, Lin HH, Murray M, Salomon JA (2012) Population health impact and cost-effectiveness of tuberculosis diagnosis with Xpert MTB/RIF: a dynamic simulation and economic evaluation. PLoS Med. 9(11). doi:10.1371/journal.pmed.1001347, https://www.scopus.com/inward/record.uri?eid=2-s2.0-84870266092&partnerID=40&md5=557a31f6d3f2341c03f44cb526513bdf

Mills HL, Cohen T, Colijn C (2013) Community-wide isoniazid preventive therapy drives drug-resistant tuberculosis: a model-based analysis. Sci Transl Med 5(180). doi:10.1126/scitranslmed.3005260, https://www.scopus.com/inward/record.uri?eid=2-s2.0-84877765955&partnerID=40&md5=feec00c0015ae17a30d3181030db7f0c

Murray M (2002a) Determinants of cluster distribution in the molecular epidemiology of tuberculosis. Proc Natl Acad Sci U S A 99(3):1538–1543

Murray M (2002b) Sampling bias in the molecular epidemiology of tuberculosis. Emerg Infect Dis 8(4):363–369

Murray M, Alland D (2002) Methodological problems in the molecular epidemiology of tuberculosis. Am J Epidemiol 155(6):565–571

O'Neill MB, Mortimer TD, Pepperell CS (2015) Diversity of *Mycobacterium tuberculosis* across evolutionary scales. PLoS Pathog 11(11):e1005257. doi:10.1371/journal.ppat.1005257, http://dx.doi.org/10.1371/journal.ppat.1005257

Outhred AC, Holmes N, Sadsad R, Martinez E, Jelfs P, Hill-Cawthorne GA, Gilbert GL, Marais BJ, Sintchenko V (2016) Identifying likely transmission pathways within a 10-Year community outbreak of tuberculosis by high-depth whole genome sequencing. PLoS One 11(3):e0150550. doi:10.1371/journal.pone.0150550, https://www.ncbi.nlm.nih.gov/pubmed/26938641

Ozcaglar C, Shabbeer A, Vandenberg S, Yener B, Bennett KP (2011) Sublineage structure analysis of *Mycobacterium tuberculosis* complex strains using multiple-biomarker tensors. BMC Genomics 12(Suppl 2):S1. doi:10.1186/1471-2164-12-S2-S1, http://dx.doi.org/10.1186/1471-2164-12-S2-S1

Ozcaglar C, Shabbeer A, Vandenberg SL, Yener B, Bennett KP (2012) Epidemiological models of *Mycobacterium tuberculosis* complex infections. Math Biosci 236(2):77–96. doi:10.1016/j.mbs.2012.02.003, http://dx.doi.org/10.1016/j.mbs.2012.02.003

Pankhurst LJ, Del Ojo Elias C, Votintseva AA, Walker TM, Cole K, Davies J, Fermont JM, Gascoyne-Binzi DM, Kohl TA, Kong C, Lemaitre N, Niemann S, Paul J, Rogers TR, Roycroft E, Smith EG, Supply P, Tang P, Wilcox MH, Wordsworth S, Wyllie D, Xu L, Crook DW, Group CTS (2016) Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. Lancet Respir Med 4(1):49–58. doi:10.1016/S2213-2600(15)00466-X, https://www.ncbi.nlm.nih.gov/pubmed/26669893

Plazzotta G, Cohen T, Colijn C (2015) Magnitude and sources of bias in the detection of mixed strain *M. tuberculosis* infection. J Theor Biol 368:67–73

Pybus OG, Drummond AJ, Nakano T, Robertson BH, Rambaut A (2003) The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. Mol Biol Evol 20(3):381–387. doi:10.1093/molbev/msg043, https://www.ncbi.nlm.nih.gov/pubmed/12644558

Ragheb MN, Ford CB, Chase MR, Lin PL, Flynn JL, Fortune SM (2013) The mutation rate of mycobacterial repetitive unit loci in strains of M. tuberculosis from cynomolgus macaque infection. BMC Genomics 14:145. doi:10.1186/1471-2164-14-145, http://dx.doi.org/10.1186/1471-2164-14-145

Rangaka MX, Cavalcante SC, Marais BJ, Thim S, Martinson NA, Swaminathan S, Chaisson RE (2015) Controlling the seedbeds of tuberculosis: diagnosis and treatment of tuberculosis infection. Lancet 386(10010):2344–2353. doi:10.1016/S0140-6736(15)00323-2, https://www.ncbi.nlm.nih.gov/pubmed/26515679

Reyes JF, Tanaka MM (2010) Mutation rates of spoligotypes and variable numbers of tandem repeat loci in *Mycobacterium tuberculosis*. Infect Genet Evol 10(7):1046–1051. doi:10.1016/j.meegid.2010.06.016, http://dx.doi.org/10.1016/j.meegid.2010.06.016

Reyes JF, Francis AR, Tanaka MM (2008) Models of deletion for visualizing bacterial variation: an application to tuberculosis spoligotypes. BMC Bioinformatics 9:496. doi:10.1186/1471-2105-9-496, http://dx.doi.org/10.1186/1471-2105-9-496

Reyes JF, Chan CHS, Tanaka MM (2012) Impact of homoplasy on variable numbers of tandem repeats and spoligotypes in *Mycobacterium tuberculosis*. Infect Genet Evol 12(4):811–818. doi:10.1016/j.meegid.2011.05.018, http://dx.doi.org/10.1016/j.meegid.2011.05.018

Rosenberg NA, Tsolaki AG, Tanaka MM (2003) Estimating change rates of genetic markers using serial samples: applications to the transposon IS6110 in *Mycobacterium tuberculosis*. Theor Popul Biol 63(4):347–363

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4(4):406–425. http://www.ncbi.nlm.nih.gov/pubmed/3447015

Sergeev R, Colijn C, Cohen T (2011) Models to understand the population-level impact of mixed strain M. tuberculosis infections. J Theor Biol 280(1):88–100. doi:10.1016/j.jtbi.2011.04.011, http://dx.doi.org/10.1016/j.jtbi.2011.04.011

Shabbeer A, Cowan LS, Ozcaglar C, Rastogi N, Vandenberg SL, Yener B, Bennett KP (2012a) TB-Lineage: an online tool for classification and analysis of strains of *Mycobacterium tuberculosis* complex. Infect Genet Evol 12(4):789–797.

doi:10.1016/j.meegid.2012.02.010, http://dx.doi.org/10.1016/j.meegid.2012.02.010

Shabbeer A, Ozcaglar C, Yener B, Bennett KP (2012b) Web tools for molecular epidemiology of tuberculosis. Infect Genet Evol 12(4):767–781. doi:10.1016/j.meegid.2011.08.019, http://dx.doi.org/10.1016/j.meegid.2011.08.019

Shrestha S, Knight GM, Fofana M, Cohen T, White RG, Cobelens F, Dowdy DW (2014) Drivers and trajectories of resistance to new first-line drug regimens for tuberculosis. Open Forum Infect Dis 1(2):ofu073. doi:10.1093/ofid/ofu073, http://dx.doi.org/10.1093/ofid/ofu073

Small PM, Hopewell PC, Singh SP, Paz A, Parsonnet J, Ruston DC, Schecter GF, Daley CL, Schoolnik GK (1994) The epidemiology of tuberculosis in San Francisco: a population-based study using conventional and molecular methods. N Engl J Med 330:1703–1709

Smith NH, Hewinson RG, Kremer K, Brosch R, Gordon SV (2009) Myths and misconceptions: the origin and evolution of *Mycobacterium tuberculosis*. Nat Rev Microbiol 7(7):537–544. doi:10.1038/nrmicro2165, http://dx.doi.org/10.1038/nrmicro2165

Stadler T (2011) Inferring epidemiological parameters on the basis of allele frequencies. Genetics 188(3):663–672. doi:10.1534/genetics.111.126466, http://dx.doi.org/10.1534/genetics.111.126466

Stadler T, Bonhoeffer S (2013) Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. Philos Trans R Soc B: Biolog Sci 368(1614):20120198

Stadler T, Kouyos RD, von Wyl V, Yerly S, Böni J, Bürgisser P, Klimkait T, Joos B, Rieder P, Xie D, Günthard HF, Drummond A, Bonhoeffer S, the Swiss HIV Cohort Study (2012) Estimating the basic reproductive number from viral sequence data. Mol Biol Evol 29:347–357

Stadler T, Kuhnert D, Bonhoeffer S, Drummond AJ (2013) Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proc Natl Acad Sci U S A 110(1):228–233. doi:10.1073/pnas.1207965110, http://www.ncbi.nlm.nih.gov/pubmed/23248286

Steel M (2013) Consistency of bayesian inference of resolved phylogenetic trees. J Theor Biol 336:246–249

Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rüsch-Gerdes S, Willery E, Savine E, de Haas P, van Deutekom H, Roring S et al (2006) Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. J Clin Microbiol 44(12):4498–4510

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10(3):512–526

Tanaka MM, Francis AR (2005) Methods of quantifying and visualising outbreaks of tuberculosis using genotypic information. Infect Genet Evol 5(1):35–43. doi:10.1016/j.meegid.2004.06.001, http://dx.doi.org/10.1016/j.meegid.2004.06.001

Tanaka MM, Rosenberg NA, Small PM (2004) The control of copy number of IS6110 in *Mycobacterium tuberculosis*. Mol Biol Evol 21(12):2195–2201. doi:10.1093/molbev/msh234, http://dx.doi.org/10.1093/molbev/msh234

Tanaka MM, Francis AR, Luciani F, Sisson SA (2006) Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. Genetics 173(3):1511–1520. doi:10.1534/genetics.106.055574, http://dx.doi.org/10.1534/genetics.106.055574

Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. Lect Math Life Sci 17:57–86

To TH, Jung M, Lycett S, Gascuel O (2016) Fast Dating Using Least-Squares Criteria and Algorithms. Syst Biol 65(1):82–97. doi:10.1093/sysbio/syv068, http://www.ncbi.nlm.nih.gov/pubmed/26424727

Trauer JM, Denholm JT, McBryde ES (2014) Construction of a mathematical model for tuberculosis transmission in highly endemic regions of the Asia-Pacific. J Theor Biol 358:74–84. doi:10.1016/j.jtbi.2014.05.023, http://dx.doi.org/10.1016/j.jtbi.2014.05.023

Vogler AJ, Keys C, Nemoto Y, Colman RE, Jay Z, Keim P (2006) Effect of repeat copy number on variable-number tandem repeat mutations in Escherichia coli O157:H7. J Bacteriol 188(12):4253–4263. doi:10.1128/JB.00001-06, http://dx.doi.org/10.1128/JB.00001-06

Volz EM (2012) Complex population dynamics and the coalescent under neutrality. Genetics 190(1):187–201. doi:10.1534/genetics.111.134627, http://dx.doi.org/10.1534/genetics.111.134627

Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SDW (2009) Phylodynamics of infectious disease epidemics. Genetics 183(4):1421–1430. doi:10.1534/genetics.109.106021, http://dx.doi.org/10.1534/genetics.109.106021

Vynnycky E, Fine PEM (1997) The natural history of tuberculosis: the implications of age-dependent risks of disease and the role of reinfection. Epidemiol Infect 119(2):183–201. doi:10.1017/S0950268897007917, <GotoISI>://WOS:A1997YD95900010

Waaler H, Geser A, Andersen S (1962) The use of mathematical models in the study of the epidemiology of tuberculosis. Am J Public Health Nat Health 52(6):1002–1013

Wakeley J (2009) Coalescent theory: an introduction. Roberts and Co., Greenwood Village

Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW et al (2013) Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. Lancet Infect Dis 13(2):137–146

Warren RM, van der Spuy GD, Richardson M, Beyers N, Borgdorff MW, Behr MA, van Helden PD (2002) Calculation of the stability of the IS6110 banding pattern in patients with persistent *Mycobacterium tuberculosis* disease. J Clin Microbiol 40(5):1705–1708

Warren RM, Victor TC, Streicher EM, Richardson M, Beyers N, Gey van Pittius NC, van Helden PD (2004) Patients with active tuberculosis often have different strains in the same sputum specimen. Am J Respir Crit Care Med 169(5):610–614. doi:10.1164/rccm.200305-714OC, http://dx.doi.org/10.1164/rccm.200305-714OC

Weniger T, Krawczyk J, Supply P, Harmsen D, Niemann S (2012) Online tools for polyphasic analysis of *Mycobacterium tuberculosis* complex genotyping data: now and next. Infect Genet Evol 12(4):748–754

WHO (2016) Global tuberculosis report 2016. Tech. rep., World Health Organization

Wigginton JE, Kirschner D (2001) A model to predict cell-mediated immune regulatory mechanisms during human infection with *Mycobacterium tuberculosis*. J Immunol 166(3):1951–1967

Wirth T, Hildebrand F, Allix-Béguec C, Wölbeling F, Kubica T, Kremer K, van Soolingen D, Rüsch-Gerdes S, Locht C, Brisse S, Meyer A, Supply P, Niemann S (2008) Origin, spread and demography of the *Mycobacterium tuberculosis* complex. PLoS Pathog 4(9):e1000160. doi:10.1371/journal.ppat.1000160, http://dx.doi.org/10.1371/journal.ppat.1000160

Yang Z (1994a) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39(3):306–314. doi:10.1007/BF00160154, http://www.ncbi.nlm.nih.gov/pubmed/7932792

Yang Z (1994b) Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. Syst Biol 43(3):329–342

Yang Z (1996) Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol 11(9):367–372. doi:10.1016/0169-5347(96)10041-0

Yang Z (2014) Molecular evolution: a statistical approach. Oxford University Press, Oxford

Zainuddin Z, Dale J (1990) Does *Mycobacterium tuberculosis* have plasmids? Tubercle 71:43–49. doi:10.1016/0041-3879(90)90060-L, http://www.sciencedirect.com/science/article/pii/004138799090060L

Zheng N, Whalen CC, Handel A (2014) Modeling the potential impact of host population survival on the evolution of M. tuberculosis latency. PLoS One 9(8):e105721. doi:10.1371/journal.pone.0105721, http://dx.doi.org/10.1371/journal.pone.0105721

Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. Evol Genes Proteins 97:97–166

Zwerling A, Gomez GB, Pennington J, Cobelens F, Vassall A, Dowdy DW (2016) A simplified cost-effectiveness model to guide decision-making for shortened anti-tuberculosis treatment regimens. Int J Tuberc Lung Dis 20(2):257–260