Models and methods to integrate epidemiological and whole genome sequence data for effectively analysing infectious disease outbreak data

Joseph Marsh

A Thesis presented for the degree of

Doctor of Philosophy

School of Mathematical Sciences

University of Nottingham

23rd September 2023

Abstract

Advances in sequencing technology and the reduction in associated costs have enabled scientists to obtain highly detailed genomic data on disease-causing pathogens on a scale never seen before. Combining genomic data with traditional epidemiological data (e.g. incidence data) provides a unique opportunity to determine the actual transmission pathway of the pathogen through a population. Despite recent advances, existing approaches have their own limitations, such as simplifications to the underlying biological processes, arbitrary phenomenological models or approximations to the likelihood function, to name a few.

We present a novel modelling framework for integrating epidemiological and whole genome sequence data to overcome the above limitations where (i) we use the matrix of pairwise horizontal distances between sequences as a summary statistic for the genetic data and (ii) explicitly derive joint probability distribution of pairwise genetic distances under the assumption of microevolution mutation models. We develop bespoke and computationally efficient data-augmentation MCMC algorithms to infer the transmission network, infection times and unobserved genetic distances from pathogen sequences at the time of transmission.

The framework presented is general and applicable to a variety of outbreak scenarios. For example, we explicitly consider a discrete time transmission model for healthcare associated infections and demonstrate the performance of our framework on simulated data and also analyse an outbreak of *S. aureus* in an intensive care unit in Brighton during 2011-2012. Our approach integrates healthcare worker data at an individual level and considers the possibility of multiple distinct genetic subtypes.

Finally we also consider integrating genetic data with a continuous time SEIR model and analyse an outbreak of foot-and-mouth disease in Darlington, a town in the north west of the UK in 2001. We validated our inferred transmission network with previous modelling studies and demonstrate that pairwise genetic distance is an informative summary of the raw sequence data.

Acknowledgements

Firstly I would like to express my gratitude to my supervisors, Theo for his always positive attitude and encouragement, and Phil for his attention to detail and reminding me to always be precise. I am eternally grateful for their patience and support over the years and for everything they have taught me. I would also like to thank my examiners Karthik and Simon for taking the time out of their busy schedules to read and assess my work.

I would like to thank all of the friends I have made along the way, in particular Rowland and Haris. I would also like to thank my wonderful wife Elisabeth for putting up with my over the years, without her this journey would have been much more difficult and I am lucky to have someone to complain to. Finally, I would like to thank my family who have provided constant love and support throughout not just the PhD but my entire life, even though they have no idea what I actually do, I know they are very proud!

I am also thankful to James Price and the rest of the team working with the Brighton and Sussex Medical School for kindly sharing their data and expertise. This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/N50970X/1.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Motivation

In the past decade, advances in genomic sequencing technology have provided scientists with an abundance of detailed and high dimensional data on a scale never seen before. These developments allow researchers to collect and process sequence data at unprecedented levels of resolution, at affordable prices and turnaround times (Gaiarsa et al., 2015). Whole-genome sequencing (WGS) has become an essential tool for public health surveillance and molecular epidemiology by unveiling the transmission dynamics of infection and microbial resistance allowing for effective and targetted interventions (Revez et al., 2017).

Not only does sequence data inform individual level pathogen phenotypic characteristics such as virulence and vaccine resistance (Neher and Bedford, 2018); there is also the opportunity to perform microbial source attribution, i.e. determining the transmission pathway of the pathogen through the population (Cottam et al., 2008). Isolates with high genetic similarity are likely to be closely related, as such this provides us with a starting point to develop models that readily incorporate the wealth of WGS data (Vasylyeva et al., 2016).

Integrating genetic data with traditional epidemiological data may provide greater insight into probable transmission links, exploiting the idea that high genetic similarity between isolates is likely indicative of a closer link in the transmission chain (Klinkenberg et al., 2019). Reconstruction of transmission trees provides answers to questions such as ‘who infected whom and when?’, and consequently can inform decision making regarding intervention strategies, such as vaccination programmes, deep cleaning procedures, treatment programs and quarantine measures (e.g. Chapman et al. (2020)).

Furthermore, the genetic diversity of sampled isolates allows researchers and clinicians to identify unexpected modes of transmission (Snitkin et al., 2012) and distinguish between within-hospital transmission and imported cases from the community. For healthcare associated infections (HCAIs), identification of clusters early in an outbreak allows the use of infection prevention and control (IPC) measures to minimize pathogen transmission and ensure adequate care and patient safety (Mutters et al., 2017). A detailed picture of emerging outbreaks for HCAIs refined by utilising the high resolution WGS data has the potential to revolutionize infection control practice on local, national and international scales (Price et al., 2013).

1.2 Literature review

Traditional analysis of data from outbreaks of infection disease typically involves host data such as the dates of symptom onset and recovery, age, sex, geographical location and dates of control measures. These can then be used to infer individual level parameters, such as the rate of infection, infectious period, risk factors, and population level parameters such as the basic reproduction ratio and the effectiveness of any control measures (Klinkenberg et al., 2019).

The application of molecular techniques to the study of heterogeneous organisms enhances epidemiological studies by improving our ability to subclassify these organisms into meaningful groups (Foxman, 2001). In the past few decades the accuracy, speed and cost effectiveness of genetic sequencing tools has improved at an unprecedented rate (van Dijk et al., 2018). Through genotyping, a molecular fingerprint of each pathogen isolate can be generated and compared with others in the same suspected cluster. Isolates with the same or highly similar genotypes, and linked by epidemiologic data, are likely to represent related cases of the infectious disease within an outbreak (Tang et al., 2017).

Epidemiological analysis based on genetic data often contains sequences from a limited subset of all cases in an outbreak, commonly referred to as a *sparsely sampled outbreak*. Conversely, epidemics with most genetic sequences observed are called *densely sampled outbreaks* (Klinkenberg et al., 2019).

Frameworks for analysing epidemiological and genetic data can be separated into two distinct paradigms, phylogenetic based inference which is based on the reconstruction of ancestries between hypothetical common ancestors and sampled isolates (Jombart et al., 2010), and non-phylogeny based methods which avoid the use of phylogenies by constructing transmission trees weighted by functions of genetic distance (Worby et al., 2016).

The heart of phylogenetic inference relies on the construction of *phylogenetic trees*. A phylogenetic tree is a a diagrammatic representation of the evolutionary relationship between various organisms (Choudhuri, 2014). The external nodes, also referred to as *leaves* or *clades* in the literature, represent the organisms in question, whereas the internal nodes represent ancestral states or evolutionary events (Nakhleh, 2013).

Although phylogenetic models are useful to describe the spread of pathogens, they do not aim to directly reconstruct transmission trees. Instead, phylogenetic models can be used to infer most recent common ancestors between pairs of isolates, which may not be suitable for densely sampled outbreaks that contain ancestral and descendant isolates (Jombart et al., 2010).

1.2.1 Phylogenetic based inference

Cottam et al. (2006, 2008) provide some of the earliest attempts to integrate epidemiological and genetic data to infer transmission pathways, specifically for the 2001 foot-and-mouth (FMD) outbreak in the UK. The authors construct a phylogenetic tree from sequence data using *maximum parsimony* and then enumerate all possible transmission trees consistent with the phylogenetic tree. Maximum parsimony seeks to construct a tree that exhibits the minimum amount of genetic evolution (Edwards, 2009). The transmission model consists of a discrete form of the gamma distribution to describe the incubation period and a discrete form of the beta distribution to describe the probability a particular farm is infected. The authors then compute the transmission likelihood for the epidemiological data for each plausible transmission tree and they found four trees which account for more than 95% of the total likelihood. This approach does not integrate sequence data directly in the model, rather the sequence data is used to narrow down plausible transmission trees.

In Hall et al. (2015) the authors assume a mutation model for nucleotide evolution, a coalescent process to model within-host diversity and an SEIR model with a spatial kernel to describe the transmission process, however making the assumption of a single individual initially infectious and complete observational data. A Markov chain Monte Carlo (MCMC) algorithm is used to simultaneously sample both the transmission tree and phylogeny with a focus on accurate phylogenetic inference. The model was used to analyse the 2003 H7N7 avian influenza outbreak in the Netherlands.

Klinkenberg et al. (2017) provides a MCMC algorithm to infer within a Bayesian framework the transmission tree and phylogeny simultaneously, using ideas that were first described by Ypma et al. (2013). The authors explicitly model the within-host dynamics using a coalescent process, and the mutation using the Jukes Cantor model of evolution. The limitation of this approach is the assumption of complete observation of cases, which is unlikely for most outbreaks.

1.2.2 Non-phylogeny based methods

Morelli et al. (2012) provides an integration of genetic and spatio-temporal epidemiological data in a Bayesian framework which is able to reconstruct transmission trees and estimate infection dates. The authors' describe a stochastic SEIR transmission model with gamma distributed exposure and infectious periods with a spatial transmission kernel. To model the genetic data the authors consider the probability distribution of the number of substitutions (mutations) between sequences and their ancestor under the assumption of the Jukes Cantor microevolution model. The conditional distribution of the observed sequences depends explicitly depends on sequences transmitted at the time of infection, which are often unobserved.

Consequently the authors consider a pseudo-distribution of observed sequences which calculates the likelihood of ancestor genetic sequences assuming that sequence ancestor pairs are independent of one another. The main drawbacks are that the model assumes a single initial infective and only one sequence per host. This method is applied to the 2007 and a subset of the 2001 foot-and-mouth outbreaks in the United Kingdom.

Mollentze et al. (2014) is an extension of the work in Morelli et al. (2012) which allows for multiple introductions of the pathogen. The authors' considered an alternative genetic microevolution model with the Kimura three-parameter model (Kimura, 1981) and include a sound post-processing algorithm to detect imported cases from the genetic data. This method is applied to an outbreak of Rabies in South Africa from March 2010 to June 2011.

Jombart et al. (2014) describes a fully integrated Bayesian reconstruction of transmission trees using pathogen genetic sequences and their collection dates. The authors use a discrete time Binomial model to describe the epidemiological data and assume a mutation rate per generation to model the genetic data. Furthermore the model assumes that mutations take place during the transmission event and only one genetic sequence per host which therefore may not be suitable for epidemics with multiple sequences from infected hosts. The authors examine the performance of the model on simulated data and also provide an analysis to the 2003 outbreak of Severe Acute Respiratory Syndrome (SARS) in Singapore.

Lau et al. (2015) proposes a stochastic SEIR model with a spatial kernel to model the epidemic process and a Kimura model to describe the process for genetic evolution. The authors integrate these components in a Bayesian framework which simultaneously infers the transmission tree and unobserved infection times and sequences using data augmentation MCMC techniques. Their model makes the assumption of a single *master sequence* to allow multiple introductions of the pathogen, and assumes a single dominant strain which does not consider the within-host diversity of the pathogen. A disadvantage to the model lies with needing to keep track of entire genetic sequences, which may be computationally demanding for large data sets. The authors apply the algorithm to the 2001 outbreak of foot-and-mouth disease, demonstrating both agreement with previous findings and also improvements on inference.

Worby et al. (2016) attempts to integrate epidemiological and genetic data by introducing phenomenological models that calculate the probability of pairwise genetic distances between isolates (SNPs), rather than entire sequences. The methods are applied to an outbreak out MRSA in 2011 and unobserved transmission dynamics are imputed using a data augmentation MCMC scheme. This approach is flexible and allows for multiple importations of the pathogen and multiple sequences per host, which is useful for modelling healthcare associated infections. A disadvantage is that the model assumes that genetic distances between each pair of isolates are independent, and that the distributions for pairwise genetic data are arbitrary.

Cassidy et al. (2020) builds upon the work of Worby et al. (2016) by relaxing the assumption of independent genetic distances and proposes flexible models to calculate the probability of pairwise genetic distances. The models assume that the number of mutations between sequences of length N over a fixed time period follows a Binomial distribution, which is then approximated using Poisson distributions. The author addresses the issue of independence by considering the relatedness of host sequences in the transmission chain and applies these models to data from a MRSA outbreak in Thailand and the avian influenza in the Netherlands.

We review the models of Cassidy (2019) and Worby et al. (2016) in more detail in Section 2.3 and a more comprehensive and systematic review of existing methods to integrate genomic and epidemiological data can be found in Duault et al. (2022).

The aim of this thesis is build upon the models introduced in Worby et al. (2016) and Cassidy (2019). The main benefit of these approaches, compared to the others, is that we wish to work with the pairwise genetic distances rather than the full sequences themselves. We seek to develop a framework similar to that in Lau et al. (2015) where we explicitly model the genetic evolutionary process, however we wish to use (and impute) genetic distances, rather than whole sequences.

Another important feature of these non-phylogeny based models is that these approaches have generative models for the epidemiological and genetic processes. This is an important feature as it allows one to simulate forward in time which can be useful many reasons, e.g. model assessment using posterior predictive distributions. This is advantageous to phylogeny based models that work backwards in time by inferring past events.

1.3 Epidemic modelling

Mathematical models for infectious disease have proven essential for understanding the transmission dynamics of pathogens in a population. Models can be useful at the theoretical level to understand how the system behaves for certain parameter values and initial conditions, for example the scenarios at which epidemics 'take off' or final size estimation. However, relating these models to real world data allows researchers to investigate a variety of practical situations, for example evaluation of scientific hypotheses, estimation of biological parameters and what-if scenarios (O'Neill, 2010).

Firstly, evaluation of scientific hypotheses usually involves some kind of model comparison and has been used to assess the role of isolation in transmission of MRSA (Kypraios et al., 2010), evaluate transmission routes for the plague (Whittles and Didelot, 2016) and analysing the impact of control measures for SARS (Wallinga, 2004).

Secondly, estimation is primarily concerned with inferring model parameters that may be of biological interest. Examples include estimating quantities that relate to the latent and infectious periods for Ebola (Lekone and Finkenstädt, 2006) and the basic reproduction number for smallpox (Clancy and O'Neill, 2008).

Finally, prediction and what-if analyses are useful to evaluate model predictions when comparing control or intervention strategies. Previous examples in the literature include evaluating the impact of different intervention strategies for lymphatic filariasis (Touloupou et al., 2022), analysing retrospective control strategies for foot-and-mouth disease (Keeling et al., 2001) and assessing the impact of non-pharmaceutical interventions for COVID-19 (Ferguson et al., 2020), to name a few. The latter of the three has been largely been attributed as one of the factors which influenced lockdown measures introduced in the UK during the COVID-19 pandemic.

Epidemic models can be divided into two distinct groups, *stochastic* and *deterministic* models. Deterministic models usually involve a system of differential equations which describe how individuals in different states evolve through time, e.g. the original SIR model described in Kermack and McKendrick (1927). On the other hand, stochastic models involve probabilistic arguments for how individuals change state (e.g. susceptible to infectious). In this thesis we focus on stochastic epidemic models, and to our knowledge there are no attempts in the literature to combine epidemiological and genetic data using deterministic models. For a more comprehensive overview of the mathematical theory of infectious diseases we direct the reader to Bailey (1975).

1.4 Whole genome sequence data

The genome is a store of biological information that consists of DNA, a polymer comprised of two helical polynucleotide chains each coiled round the same axis (Watson and Crick, 1953). Each DNA strand is a string of four different units called nucleotide bases and are adenine (A), guanine (G), cytosine (C) and thymine (T). Furthermore the two DNA strands are joined by bonds between nucleotide bases known as base pairs. Each nucleotide may bond with one other base, e.g. A may only bond with T and C may only bond with G, therefore we only need to observe one strand as the other is uniquely determined.

Whole-genome sequencing is a tool to analyse the entire genome which provide a unique snapshot of a particular organism at a point in time. A whole genome sequence is the chain of nucleotides of length N which may be represented as $X = (B_1, B_2, ..., B_N)$ where $B_i \in \{A, G, C, T\}$ denotes the nucleotide base in the ith position. Genome length varies by species, for example the average size of *S. aureus* is 280kb (Shukla et al., 2012) and the average size of foot-and-mouth disease virus is 8.5kb, where 1kb is equal to 1000 base pairs (Domingo et al., 2002).

Genomes are dynamic entities that change over time as a result of small-scale sequences alterations caused by *mutation* (Brown, 2018, Chapter 16). Mutations are caused by physical changes to the genetic material and can occur in a variety of ways, however we focus on *single nucleotide polymorphisms* (SNPs) which are substitutions of a single nucleotide at a specific position, also known as point mutations. Typically mutations arise as a result of errors in the DNA replication process or from mutagens (e.g. radiation).

Furthermore we can classify the type of mutation in two groups: *transitions* and *transversions*. In molecular biology, nucleotide bases can be classified as pyrimidines (T, C) and purines (A, G). A transition is a substitution between the two pyrimidines $(T \leftrightarrow C)$ or the two purines $(A \leftrightarrow G)$, and a transversion is a substitution between a pyrimidine and a purine $(T, C \leftrightarrow A, G)$.

Furthermore transitions occur more often than transversions (Lyons and Lauring, 2017), which should be taken into account when developing a realistic model that is motivated by the underlying biological processes. Genetic distance between isolates can be measured by counting the number of horizontal differences, that is the number of SNPs between the two sequences. Sequence pairs with small genetic distances indicate that they have many genes in common and are biologically related where they have both descended from a common ancestral species (Higgs and Derrida, 1992).

1.5 Continuous time Markov chains

The mutation models that are considered in this thesis use continuous time Markov chains to describe nucleotide substitution, therefore we briefly present the required preliminaries. Let $\{X(t), t \geq 0\}$ be a continuous time stochastic process on a discrete state space S.

Definition 1.5.1. The process $\{X(t) : t \geq 0\}$ is said to be a homogeneous continuous time Markov chain on the discrete state space S with $Q-$matrix $Q = (q_{ij})_{i,j \in S}$ if for all $t, h \geq 0$, $0 \leq s < t$ and $i, j, k \in S$,

1. $\Pr(X(t+s) = j | X(t) = i, X(s) = k) = \Pr(X(t+s) = j | X(t) = i)$
2. $\Pr(X(t+s) = j | X(t) = i) = \Pr(X(s) = j | X(0) = i)$
3. $\Pr(X(t+s) = j | X(t) = i) = \begin{cases} q_{ij}s + o(s) & i \neq j \\ 1 + q_{ii}s + o(s) & i = j \end{cases}.$

Note that (1) is the Markov property, (2) is time-homogeneity and (3) defines q_{ij} as the rate of jumping from i to j.

A consequence of the Markov property is that the waiting time until transition is exponentially distributed. If $X(t) = i$ then the chain remains in state i for some time $T_i \sim \text{Exp}(-q_{ii})$ and then transitions to state $j \neq i$ with probability $-q_{ij}/q_{ii}$.

Let $p_{ij}(t) = \Pr(X(t+s) = j | X(s) = 0)$ denote the probability that the process is in state i and will be in state j after t time units. These quantities are the transition probabilities of the CTMC define the transition matrix $P(t) = (p_{i,j}(t))_{i,j \in S}$. From the Chapman-Kolmogorov equations we have

$$p_{i,j}(t+s) = \sum_{k \in S} p_{i,k}(t) p_{k,j}(u),$$

which can be used to derive analytical transition probabilities for simple processes.

1.6 Bayesian inference

Often when relating epidemic models to outbreaks, the data are only partially observed in the sense that we may have information on the time of symptom onset, however we do not know the precise time of infection. As a consequence the likelihood for many epidemic models may be intractable as it involves integrating over all possible unobserved times which is typically a difficult task due to the high dimensional nature of the problem.

A natural approach to overcome the intractability of the likelihood is to use data imputation methods where the unknown quantities are treated as model parameters. Inference may then be performed from a classical viewpoint through the use of the EM algorithm (Becker, 1993) or from a Bayesian perspective using Markov chain Monte Carlo (MCMC) methods (O'Neill and Roberts, 1999). However, in O'Neill (2002) the author argues that the Bayesian approach provides information concerning parameter uncertainty that is otherwise difficult to obtain through the classical approach. Furthermore evaluation of the expectation step for the EM algorithm may be complicated compared to the MCMC methods.

For these reasons we shall use the Bayesian approach throughout this thesis and now we describe the fundamentals of Bayesian inference and computation. Let x denote the data that we assume are generated by some process parametrised by θ. In the classical framework, it is assumed that the parameters θ are fixed non-random quantities that we wish to infer, primarily through the likelihood function $L(\theta)$ which more precisely is the density of the data conditional on the model parameters, $f(x|\theta)$, regarded as a function of θ alone.

On the other hand, the Bayesian approach now considers the parameters θ to be random quantities and is assigned a probability distribution with density $\pi(\theta)$ known as the *prior distribution.* Then, according to Bayes' theorem we obtain

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\pi(x)}, \tag{1.1}$$

where $\pi(x)$ is the marginal likelihood and $\pi(\theta|x)$ is the *posterior distribution* which is the main object of interest. We may then estimate quantities such as the mean, median and credible intervals from the posterior distribution.

The main challenge with the Bayesian approach is that the marginal likelihood, or normalising constant, in Equation 1.1 is often difficult to compute analytically. Fortunately it is possible to simulate from a target density that is only known up to a proportionality using Markov chain Monte Carlo techniques.

1.6.1 Markov chain Monte Carlo

There is an vast and extensive literature of MCMC methods and applications and thus we do not go into great detail here, instead we give an overview and briefly outline two important methods, however the reader should consult Gilks (1998); Robert and Changye (2020) and Brooks et al. (2011) for a more broad overview.

Markov chain Monte Carlo (MCMC) methods are a class of algorithms which aim to generate a sequence of samples $X_1, ..., X_n$ from a target density $\pi(x)$. More precisely, the Markov chain is constructed in such a way that the equilibrium distribution is proportion to the target density. An important feature is that we only need know the target density up to a proportionality, which is useful for the Bayesian approach if the marginal likelihood $\pi(X)$ in Equation 1.1 is difficult to compute.

For the Bayesian approach the target density is the posterior distribution $\pi(\theta|X)$. We now briefly outline the two simple approaches which are used throughout this thesis, that is the Metropolis-Hastings algorithm (Hastings, 1970) and Gibbs sampler (Geman and Geman, 1984).

1.6.1.1 Metropolis-Hastings algorithm

Suppose we wish to draw approximate samples from a target density $\pi(x)$, which for our purposes is the posterior distribution and the general procedure can be found in Algorithm 1. The main idea is to propose a new value y from a distribution q and then accept the proposal with probability $\min\left(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right)$.

The algorithm is general in the sense that there is no restriction on the choice of proposal distribution q, however a popular choice are proposal distributions of the form

$$q(x, y) = q(|x - y|),$$

i.e. the proposal is symmetric in x and y. If the proposal distribution is symmetric, then the acceptance probability simplifies to $\min\left(1, \frac{\pi(y)}{\pi(x)}\right)$ since $q(x, y) = q(y, x)$ and note that this quantity depends only on the ratio of the density evaluated at candidate and current values.

In order to verify that the algorithm is sampling from the target density, it is sufficient to verify that the chain produced by the algorithm does have the stationary distribution $\pi(x)$. In other words, we require π and Q to satisfy the detailed balance equations

$$\pi(x)Q(x,y) = \pi(y)Q(y,x) \quad \forall x, y,$$

where $Q(x,y)$ is the transition kernel of the chain and is written as

$$Q(x,y) = \begin{cases} q(x,y)\alpha(x,y) & x \neq y \\ \int (1-\alpha(x,y))q(x,y)dy & x = y. \end{cases}$$

Algorithm 1 Structure of the Metropolis-Hastings algorithm

```
1: Initialise the chain with initial values $X_0 = x_0$
2: Suppose at time $t$ we have $X_t = x$. Propose a candidate value $y$ from a proposal
   distribution $q(x,y)$.
3: Calculate the Metropolis-Hastings acceptance probability written as
       $\alpha(x,y) = \min\left(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right)$
4: Accept the move with probability $\alpha(x,y)$, that is we set $X_{t+1} = y$. If the move is
   rejected we set $X_{t+1} = x$.
5: Repeat steps 2-4.
```

1.6.1.2 Gibbs sampler

The Gibbs sampler can be viewed as a special case of the Metropolis-Hastings algorithm which utilises the full conditional distributions to simulate values that are always accepted, that is the Metropolis-Hastings acceptance probability is always one.

Suppose we wish to sample from $\pi(\boldsymbol{x})$ where $\boldsymbol{x} = (x_1, ..., x_d)$ is a d-dimensional vector. We sample from the full-conditional distributions which is defined by $\pi(x_i|\boldsymbol{x}_{-(i)})$ where $\boldsymbol{x}_{-(i)} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_d)$ is $\boldsymbol{x}$ with the ith element removed.

Note that we can write the full-conditional distribution as

$$\pi(x_i|\boldsymbol{x}_{(i)}) = \frac{\pi(x_i, \boldsymbol{x}_{-(i)})}{\pi(\boldsymbol{x}_{-(i)})} = \frac{\pi(\boldsymbol{x})}{\int \pi(\boldsymbol{x})dx_i},$$

from which it is clear that $\pi(x_i|\boldsymbol{x}_{-(i)}) \propto \pi(\boldsymbol{x})$, that is the full conditional distribution is proportional to the target density. We repeat this process for each of the available full-conditional distributions that are known and can be sampled from. If the full conditional distribution is unknown, then we may update the parameter using other techniques such as a Metropolis-Hastings step. The general procedure for the Gibbs sampler can be found in Algorithm 2.

Algorithm 2 Structure of the Metropolis-Hastings

```
Initialise the chain with initial values $\boldsymbol{X}_0 = \boldsymbol{x}_0$
Given that we have $\boldsymbol{x}^{(j)} = (x_1^{(j)}, ..., x_d^{(j)})$, we sample $\boldsymbol{x}^{(j+1)}$ as follows:
    • Sample $x_1^{(j+1)}$ from $\pi(x_1|\boldsymbol{x}_{-(1)}^{(j)})$
    • Sample $x_2^{(j+1)}$ from $\pi(x_2|\boldsymbol{x}_{-(2)}^{(j)})$
    ⋮
    • Sample $x_d^{(j+1)}$ from $\pi(x_d|\boldsymbol{x}_{-(d)}^{(j)})$
Now we have $\boldsymbol{x}^{(j+1)} = (x_1^{(j+1)}, ..., x_d^{(j+1)})$, set $j = j + 1$ and repeat step 2.
```

1.6.2 Data augmentation

As discussed in Section 1.6, real world data are often partially observed and in most cases the likelihood is intractable, consequently we resort to *data augmentation* methods. The first examples of data augmentation MCMC techniques for analysis of outbreak data first appeared in Gibson and Renshaw (1998) and O'Neill and Roberts (1999) and for a more broader overview of data augmentation see van Dyk and Meng (2001).

Let x denote the observed data and θ the model parameters, and $\pi(x|\theta)$ the likelihood that is intractable. The main idea is to introduce missing data z such that the likelihood $\pi(x, z|\theta)$ is tractable and these quantities are related by

$$\pi(x|\theta) = \int_z \pi(x, z|\theta) dz.$$

In the Bayesian framework we augment the parameter space to include the unobserved quantities z and the resulting posterior is $\pi(\theta, z|x) \propto \pi(x, z|\theta)\pi(\theta, z)$. We wish to develop MCMC algorithms to sample both the model parameters θ and missing data z. Often direct sampling of the unobserved data z is problematic due to the complicated nature of the density $\pi(z|x, \theta)$, however we may use standard Metropolis-Hastings steps to explore the parameter space.

Let $q(z^*|z)$ and $q(z|z^*)$ denote the proposal densities for the forward and reverse moves for the unobserved data where z^* is the proposed value and z is the current state of the chain. After sampling z^* from $q(z^*|z)$, this move is accepted with probability

$$\min\left(1, \frac{\pi(\theta, z^*|x)q(z|z^*)}{\pi(\theta, z|x)q(z^*|z)}\right).$$

A key challenge with developing data-augmented MCMC algorithms is designing efficient proposal distributions that are sophisticated enough to explore the parameter space but also simple in the sense that the proposed moves get accepted often enough. This trade off between complexity and efficiency is more art than science however there are a list of tools available to assess the quality of MCMC outputs which we describe in the following section.

1.6.3 MCMC diagnostics

Markov chain Monte Carlo (MCMC) algorithms are essential tools in Bayesian statistics for complex high-dimensional problems, however there are several implementation issues that must be considered, hence it is natural to check whether these algorithms 'work'. That is, are the samples generated by the algorithm truly representative of the underlying stationary distribution of the Markov chain (Cowles and Carlin, 1996)?

The two primary areas of interest to assess the performance of MCMC algorithms are convergence and mixing. The notion of convergence refers to convergence of the underlying Markov chain to stationarity and convergence of Monte Carlo estimators to population quantities (Roy, 2020). On the other hand, *mixing* refers to efficient exploration of the parameter space, which is a concern for high-dimensional problems since conventional MCMC algorithms often scale poorly in problem size and complexity (Duan et al., 2018).

It is important ensure that the Markov chain we simulate from converges to a stationary distribution that sufficiently explores the parameter space. There is no definitive rule or method to assess the convergence of a Markov chain, however we shall outline some techniques below.

Starting point determination Unless prior knowledge is known about the target density, the choice of starting point may not be obvious. One option is to choose various points at random from the parameter space, which in principle can also be used as a measure of convergence to check visually that the univariate Markov chains converge to the same posterior mode.

Burn in Often MCMC algorithms take some number of iterations, M say, such that the underlying Markov chain converges to the stationary distribution which translates to sampling from the target distribution. Hence we wish to discard the first M iterations and keep the latter $M+1, ..., N$ samples which have been simulated from a stationary Markov chain and therefore the samples are from the distribution of interest. The first M values is typically known as the burn in period, the choice of M is not obvious however can be determined by several methods, one of which being visual inspection of trace plots.

Correlated observations Due to the Markov nature of these methods, consecutive samples are clearly correlated. A technique to reduce the dependent nature of observations and autocorrelation is *thinning*, which aims to store every k sample. Suppose we have N samples denoted by

$$\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, ..., \theta^{(N)},$$

then by thinning the output we end up with N/k observations which correspond to

$$\theta^{(1)}, \theta^{(k+1)}, \theta^{(2k+1)}, ..., \theta^{(N)},$$

assuming that k divides N. The practice of thinning MCMC outputs has been used often in the literature, however in Link and Eaton (2011) the authors claim that thinning is often unnecessary and always inefficient.

Visual Assessment Visual assessment of the Markov chains are usually achieved by the use of trace plots, which are a time series of the sampler iterations. Trace plots give a rough idea of the when a distribution reaches stationarity and also indicate how well the chain is mixing.

Effective sample size Effective sample size (ESS) is a metric that aims to measure the amount of information loss due to autocorrelation in samples. Given a sample of N dependent samples, the effective sample size N_{ESS} can be considered as the number of independent samples with the same estimation power (Gamerman and Lopes, 2006).

Parallel chains and the Gelman-Rubin diagnostic Originally proposed in Gelman and Rubin (1992), the authors use multiple chains and an analysis of variance technique to assess whether or not each of the chains has the same distribution. The method is implemented as follows.

Run J chains in parallel from an overdispersed distribution, i.e. widely different starting positions, for $2n$ iterations and discard the first n iterations as a burn in. For each scalar parameter of interest θ_i: label the draws from the J chains as ψ_{jk}, $j = 1, ..., J$ and $k = 1, ..., n$. Then the between-chain variance is defined as:

$$\frac{B}{n} = \frac{1}{J-1} \sum_{j=1}^{J} (\bar{\psi}_{j\cdot} - \bar{\psi}_{\cdot\cdot})^2,$$

where

$$\bar{\psi}_{j\cdot} = \frac{1}{n}\sum_{k=1}^{n}\psi_{jk} \qquad \bar{\psi}_{\cdot\cdot} = \frac{1}{J}\sum_{j=1}^{J}\bar{\psi}_{j\cdot}$$

The quantity $\bar{\psi}_{\cdot j}$ is the mean of the jth chain and $\bar{\psi}_{\cdot\cdot}$ is the overall mean. Next we defined the within-chain variance as:

$$W = \frac{1}{J}\sum_{j=1}^{J} s_j^2,$$

where

$$s_j^2 = \frac{1}{n-1}\sum_{k=1}^{n}(\psi_{jk} - \bar{\psi}_{j\cdot})^2.$$

Then calculate the weighted average

$$\hat{\sigma}^2 = \frac{n-1}{n}W + \frac{B}{n}$$

and monitor the ratio

$$R = \frac{\hat{\sigma}^2}{W} = \frac{n-1}{n} + \frac{B}{nW} \approx 1 + \frac{B}{nW},$$

which should decrease to 1 upon convergence to the stationary distribution. A general rule of thumb is to monitor the chains and stop the simulations when each parameter has $R < 1.1$ (Gelman et al., 2013).

A multivariate version of the Gelman-Rubin diagnostic was later proposed in Brooks and Gelman (1998) which is known as the *multivariate potential scale factor* which aims to estimate the posterior variance-covariance matrix. All of the above tools can be readily implemented in the R package *coda* (Plummer et al., 2006).

1.7 Thesis structure

This thesis is composed of three main contributions to the literature of epidemic modelling. This chapter aims to provide an overview of the existing approaches for integrating epidemiological and genetic data and some preliminary work that will be useful for later in this thesis.

In Chapter 2 we summarise the models described in Worby et al. (2016) and Cassidy et al. (2020) and outline how we aim to build upon this work. We then explore some mutation models used in the literature and consider the joint distribution of pairwise genetic distances under the assumption of a nucleotide substitution model. We derive the distributions for sequences of length $N = 1$ and then extend this to sequences of arbitrary length by utilising the independence assumption of nucleotides. The main strategy for deriving these distributions is to start with the pairwise genetic distances and consider the set of possible sequences that give rise to the distances.

We conclude that inference for the mutation parameters, conditional on an underlying genetic network, can be performed using the pairwise distance matrix due to sufficiency. Finally we outline how to relate these models to data, that is how one may construct the genetic network that is consistent with epidemiological data (i.e. the transmission network). We also consider a simple model that describes the contribution of genetic sequences that are of the same type but are unrelated, which is common for situations where an individual imports the pathogen from the community, such as nosocomial infections.

In Chapter 3 we provide an application for the models developed in Chapter 2 where we analyse a data set containing information on transmission of MRSA in a hospital in Brighton. Motivated by the exploratory analysis we extend the transmission and genetic model to include (i) integration healthcare worker data at an individual level and (ii) multiple distinct genetic subtypes.

We develop an MCMC algorithm to sample the model parameters and unobserved data, that is the transmission tree and unobserved genetic distances. We demonstrate the performance of the algorithm on simulated data and then provide a detailed analysis of the Brighton data set.

Next in Chapter 4 we analyse a well studied data set of foot-and-mouth disease known as the Darlington cluster. This application is an example of integrating the genetic data in a continuous time framework and assuming that the mutations can be described by the Kimura mutation model. We also demonstrate the performance of the algorithm on simulated data and provide a rigorous analysis of the data set and compare our inference to existing approaches in the literature.

Finally in Chapter 5 we conclude by summarising our findings and discuss how the work developed in this thesis contributes to the field of epidemic modelling and inference. We also discuss the limitations of our approach and directions for future research.

CHAPTER 2

Models to analyse whole genome sequence data

2.1 Introduction

This chapter aims to develop statistical methods to analyse genetic and epidemiological data simultaneously. Genome length varies by species, and as such raw sequence data are often high dimensional. A technique that is commonly used to reduce the dimensionality is to measure the genetic distance between isolates by counting the number of horizontal differences of aligned sequences, referred to as single nucleotide polymorphisms (SNP).

Sequences that are collected during an outbreak may be epidemiologically related, therefore interest lies in modelling the distribution of SNPs conditional on the outbreak. We aim to derive a joint distribution of pairwise genetic distances (SNPs) under the assumption of an underlying mutation model. Mutation models provide a framework to describe evolution of the genetic sequences by assuming that nucleotide bases change over time according to some random process.

We aim to build upon the work of Worby et al. (2016) and Cassidy (2019) and focus on analysis of genetic data using genetic distance matrices instead of raw sequence data. The work in Worby et al. (2016) models the genetic distance between pairs of isolates using arbitrary geometric distributions, and in a similar fashion the work in Cassidy (2019) uses arbitrary Poisson distributions. Our approach is to assume an underlying mutation model and derive the joint distribution of the observed pairwise distances, rather than using arbitrary phenomenological models such as those described later in Section 2.3.

In this chapter we do not focus on any pathogen in particular and describe the general methodology to integrate genetic data in any outbreak scenario. In Chapters 3 and 4 we apply the models develop here to an outbreak of methicillin-resistant *S. aureus* (MRSA) and foot-and-mouth-disease virus (FMDV) respectively.

This chapter is organised as follows. In Section 2.3 we review the models developed in Worby et al. (2016) and Cassidy et al. (2020). Next, in Section 2.4 we present the two simplest nucleotide Markov mutation models. Then, in Sections 2.5 and 2.6 we present some graph theoretic definitions of two important objects of interest, specifically transmission and genetic networks. Transmission networks are useful for describing the spread of the pathogen through the population and genetic networks for describing the evolution of the pathogen.

Next, in Section 2.7 we begin by considering a single nucleotide evolving through time and derive the probability distribution implied by pairwise genetic distances assuming that the underlying mutation process can be modelled by the Jukes-Cantor (Jukes and Cantor, 1969) and Kimura models (Kimura, 1980). Then, we consider the probability of observing independent and identically distributed distance matrices for each nucleotide in sequences of arbitrary length.

Next, in Section 2.8 we discuss the assumptions and general procedure for constructing genetic networks conditional on the transmission network and later in Section 2.9 we outline a simple approach to model multiple introductions of the pathogen. Finally in Section 2.10 we discuss the strengths and limitations of the models and distributions developed in this chapter.

2.2 Motivation

In this thesis we analyse two data sets, specifically the Brighton data set in Chapter 3 and the Darlington data set in Chapter 4, where both contain information on traditional epidemiological features such as removal/discharge times, symptom times and swab testing data. In addition, the data sets are scenarios where pathogenic genetic information have been collected, that is we have genetic sequences that have been isolated and sampled from individuals or farms.

Raw sequence data are often high dimensional depending on the type of organism under consideration. For example, the human genome is approximately 3.2 billion nucleotides in length (Brown, 2018, Chapter 1), whereas MRSA has been found to be approximately 2.8 million nucleotides (Shukla et al., 2012) and foot-and-mouth has been estimated to be 8500 nucleotides (Domingo et al., 2002).

Suppose we have $n_s \geq 1$ sequences observed that each have been assumed to be aligned and are of size N, where N is sufficiently large, the genetic data are of dimension $n_s \times N$ which may be computationally intractable (Bang, 2010, Chapter 14). In our analysis we propose to represent the genetic data by the matrix of pairwise distances which is of dimension $n_s \times n_s$.

A important feature of the Brighton data set analysed later in Chapter 3 is that we do not have access to the raw sequence data, as such the genetic data are the pairwise distances, times of sampling and the individuals that the sequences were sampled from. As a consequence, we wish to use the matrix of pairwise distances as a starting point for the analysis, similar to the approaches in Worby et al. (2016) and Cassidy et al. (2020). More precisely, we aim to derive the joint distribution of pairwise genetic distances under the assumption of microevolution mutation models.

Furthermore, many existing approaches for the joint analysis of genetic and epidemiological data (Lau et al. (2015); Morelli et al. (2012); Mollentze et al. (2014), to name a few) make the assumption of a *single dominant strain* in the population at any point in time. Such assumptions are reasonable for data sets that do not exhibit a large amount of genetic diversity (e.g. the Darlington data set in Chapter 4), however these may impractical for highly diverse pathogenic colonies. For example, there are individuals in the Brighton data set with multiple sequences with genetic distances greater than 10,000 SNPs, indicating that there are multiple distinct lineages or strains in the population.

In light of those features highlighted in the Brighton data set (Section 3.2), we seek to develop a flexible and efficient model using pairwise genetic distances which allows for the possibility of multiple genetic strains competing within hosts which are a result of coinfection.

2.3 Pairwise distance models

In this section we outline the models in Worby et al. (2016) and Cassidy et al. (2020) before discussing how we aim to build upon this work.

In Worby et al. (2016) the authors introduce two models for the pairwise genetic distances that are the *transmission diversity* and *importation structure* models. For both of these let Ψ denote the matrix of pairwise genetic distances that measures the number of SNPs (horizontal differences) between each pair where $\psi_{x,y}$ is the distance between sequences x and y for $x, y = 1, ..., n_s$.

2.3.1 Transmission diversity model

Define $t(x, y)$ to be the number of transmission events separating sequences x and y. If sequences x and y originate from the same person then $t(x, y) = 0$, and if sequences x and y originate from individuals in distinct transmission chains then $t(x, y) = \infty$. For $d = 0, 1, ...$, the distribution of the distance between sequences x and y is written as

$$\pi(\psi_{x,y} = d) = \begin{cases} \gamma k^{t(x,y)}(1 - \gamma k^{t(x,y)})^d & t(x, y) < \infty \\ \gamma_G(1 - \gamma_G)^d & t(x, y) = \infty, \end{cases}$$

where $\gamma k^{t(x,y)} \in [0, 1]$ and γ_G is the genetic diversity parameter between samples in distinct transmission chains and γ is the diversity parameter for distances in the same chain.

2.3.2 Importation structure model

In this model each sequence belongs to a group which contains genetically similar sequences. For $d = 0, 1, ...$ the distribution of the distance between sequences x and y is written as

$$\pi(\psi_{x,y} = d) = \begin{cases} \gamma(1 - \gamma)^d & x \text{ and } y \text{ are in the same group} \\ \gamma_G(1 - \gamma_G)^d & \text{otherwise}, \end{cases}$$

where γ and γ_G are the genetic diversity parameters for sequences in the same group and different groups respectively.

For both models, it is assumed that pairwise genetic distances are independent of one another, hence the probability of the pairwise distance matrix is written as

$$\pi(\Psi|X, T, \theta) = \prod_{y=2}^{n_s} \prod_{x=1}^{y} \pi(\psi_{x,y} = d),$$

where X is the screening data, T is the transmission tree and θ are the model parameters.

For both models, geometric distributions are used to describe the genetic distance between isolates which is reasonable as the probability mass function is decreasing. Furthermore, geometric distributions have been previously used to model the accumulation of SNPs (Sainudiin et al., 2007).

The work in Cassidy et al. (2020) naturally builds upon these models by relaxing the assumption of independence between isolates and utilising Poisson distributions to model the accumulation of SNPs. More precisely, the authors describe the *Poisson error dependence* and *Poisson chain dependence* models which are defined as follows.

2.3.3 Poisson chain dependence model

The Poisson error model assumes that the genetic distance between sequences x and y follows a Poisson distribution with parameter θ_G, θ_I or θ depending on the number of transmission events between x and y. Recall that $t(x, y)$ denotes the number of transmission events between isolates x and y, then the probability of a distance d between sequences x and y for $x = 0, 1, ...$ is written as

$$P(\psi_{x,y} = d) \begin{cases} \left(\theta_G^d/d!\right)\exp\left(-\theta_G\right) & t(x,y) = \infty \\ \left(\theta_I^d/d!\right)\exp\left(-\theta_I\right) & t(x,y) = 0 \\ \left(\theta^d/d!\right)\exp\left(-\theta\right) & t(x,y) = 1 \\ \left(D(x,y)^d/d!\right)\exp(-D(x,y)) & t(x,y) > 1, \end{cases}$$

where $D(x, y)$ is the sum of genetic distances between sequences x and y along the transmission chain.

2.3.4 Poisson error dependence model

The Poisson error model is identical to the chain dependence model apart from the contribution for isolates sampled from individuals separated by more than one transmission event. In this formulation, if $t(x, y) > 1$ then the genetic distance is defined as $D(x,y)+\xi W$ where $\Pr(\xi = 1) = \Pr(\xi = -1) = 0.5$, W is a Poisson random variable with parameter $t(x,y)\gamma$ truncated at $D(x,y)$ and ξ and W are independent. The authors then define the probability of a distance d between sequences x and y for $x = 0, 1, ...$ as

$$P(\psi_{x,y} = d) \begin{cases} \left(\theta_G^d/d!\right) \exp\left(-\theta_G\right) & t(x,y) = \infty \\ \left(\theta_I^d/d!\right) \exp\left(-\theta_I\right) & t(x,y) = 0 \\ \left(\theta^d/d!\right) \exp\left(-\theta\right) & t(x,y) = 1 \\ \frac{(t(x,y)\gamma)^{|d-D(x,y)|}}{|d-D(x,y)|!C_d} \left(\frac{1}{2}\right)^{\mathbb{1}_{\{d \neq D(x,y)\}}} & t(x,y) > 1, \end{cases}$$

where $D(x,y)$ the sum of genetic distances between sequences x and y along the transmission chain, $\mathbb{1}_A$ is the indicator function of the event A and $C_D = \sum_{l=0}^{D(x,y)} (t(x,y)\gamma)^l/l!$.

We wish to build upon these models in the sense that we want a probability distribution for the pairwise distance matrix conditional on the transmission tree, however we wish to consider the probability distribution implied by assuming an underlying mutation model, rather than simply assigning geometric or Poisson distributions to genetic distances. This now motivations the work in the rest of this chapter where we derive the joint distribution of genetic distances under the assumption of a nucleotide substitution model.

2.4 Models of nucleotide substitution

In this section we describe a class of probabilistic models used to describe the evolutionary changes over time, specially *Markov nucleotide substitution* models.

The evolutionary process of the pathogen is modelled at the level of nucleotide substitutions. To model DNA sequence evolution, probabilistic models were introduced to describe the evolutionary changes between nucleotides over time, and continuous-time Markov chains are commonly used for this purpose.

In previous literature it is assumed that substitutions at any particular site occur independently of one another (Yang, 2014). Jukes and Cantor (1969) originally proposed the simplest model where the mutation rate for any transition is constant, also commonly referred to as the JC69 model.

We also consider the Kimura 2 parameter mutation (K80) model described in Kimura (1980) which aims at a more realistic representation of the underlying biological processes and also the General Time Reversible (GTR) model (Tavaré, 1986).

The general setup for the following sections is as follows. Define $\{X(t) : t \geq 0\}$ to be a homogeneous continuous-time Markov chain where $X(t)$ is a nucleotide base at time t on the state space $\mathcal{E} = \{A, G, C, T\}$. For distinct $i, j \in \mathcal{E}$, let $\tilde{q}_{ij}$ be the transition rate from state i to state j, and $\tilde{q}_{ii} = -\sum_{j \neq i} \tilde{q}_{ij} < 0$ by definition.

2.4.1 Jukes Cantor (JC69) model

The JC69 model (Jukes and Cantor, 1969) assumes that nucleotides have the same instantaneous rate λ of transitioning to another nucleotide. Suppose that the rate of mutation for a single nucleotide base is constant, denoted by λ, so that $\tilde{q}_{ij} = \lambda$, for all $i \neq j$. The corresponding rate matrix is then given by

$$\tilde{\mathbf{Q}} = \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{pmatrix}.$$

Let $p_{ij}(t) = \Pr(X(t) = j \mid X(0) = i)$. It is straightforward to show that the probability that a nucleotide base $i \in \mathcal{E}$ mutates to $j \in \mathcal{E}$ in t time units is given by

$$p_{ij}(t) = \begin{cases} \frac{1}{4}(1 + 3e^{-4\lambda t}) & i = j \\ \frac{1}{4}(1 - e^{-4\lambda t}) & i \neq j. \end{cases}$$

2.4.2 Kimura (K80) model

The K80 model (Kimura, 1980) now relaxes the assumption of equal mutation rates and differentiates between the type of substitution, which may either be a *transition* or *transversion*. It is important to note that the transition described here is different to the transition for a Markov chain and the specific definition should be clear from the context.

Recall that a transition is a substitution between two pyrimidines ($T \leftrightarrow C$) or two purines ($A \leftrightarrow G$), and a transversion is a substitution between a pyrimidine and a purine ($T, C \leftrightarrow A, G$).

Let λ_T and λ_V denote the transition and transversion rate respectively, then the corresponding rate matrix is given by

$$\tilde{\mathbf{Q}} = \begin{pmatrix} -(2\lambda_V + \lambda_T) & \lambda_T & \lambda_V & \lambda_V \\ \lambda_T & -(2\lambda_V + \lambda_T) & \lambda_V & \lambda_V \\ \lambda_V & \lambda_V & -(2\lambda_V + \lambda_T) & \lambda_T \\ \lambda_V & \lambda_V & \lambda_T & -(2\lambda_V + \lambda_T) \end{pmatrix}.$$

We again define $p_{ij}(t) = \Pr(X(t) = j \mid X(0) = i)$ which is the probability that a nucleotide base $i \in \mathcal{E}$ mutates to $j \in \mathcal{E}$ in t time units. With some algebra it can be shown that

$$p_{ij}(t) = \begin{cases} \frac{1}{4}(1 + e^{-4\lambda_V t} + 2e^{-2(\lambda_T + \lambda_V)t}) & i = j \\ \frac{1}{4}(1 + e^{-4\lambda_V t} - 2e^{-2(\lambda_T + \lambda_V)t}) & i \neq j \text{ (transition)} \\ \frac{1}{4}(1 - e^{-4\lambda_V t}) & i \neq j \text{ (transversion).} \end{cases} \tag{2.1}$$

2.4.3 General time reversible (GTR) model

For completeness we present the most general of the Markov substitution models, first proposed in (Tavaré, 1986) and is referred to as the general time reversible (GTR) model.

An important property of the previous two models is *time-reversibility*, which loosely states that the random process of nucleotide substitution is identical if the direction of time is reversed. It is important to note that there is no biological reason to suggest this process is reversible, rather the notion of reversibility is a mathematical convenience (Yang, 2014). A Markov chain is said to be time-reversible if and only if

$$\pi_i p_{ij}(t) = \pi_j p_{ji}(t), \quad \text{for all } i, j, t \tag{2.2}$$

where Equations (2.2) are the *detailed balance* equations (Grimmett, 2001, Chapter 6).

The most general of models that satisfies the time reversibility conditions in Equation (2.2) has nine free parameters: the rates a-f and three frequency parameters and the rate matrix is given by

$$\tilde{\mathbf{Q}} = \begin{pmatrix} \cdot & a\pi_G & b\pi_C & c\pi_T \\ a\pi_A & \cdot & d\pi_C & e\pi_T \\ b\pi_A & d\pi_G & \cdot & f\pi_T \\ c\pi_A & e\pi_G & f\pi_C & \cdot \end{pmatrix},$$

where the diagonals are given by $\tilde{q}_{ii} = -\sum_{i \neq j} \tilde{q}_{ij}$ and π_y are the equilibrium frequencies for $y \in \{A, G, C, T\}$ and details of how to calculate the transition matrix is given in Lanave et al. (1984).

We now have a probabilistic frameworks for a nucleotide observed at two distinct time points, and we now wish to consider the joint distribution of these pairwise distances for sequences of length one.

Before that we first define some key concepts that will be used throughout this thesis, specifically some graph theoretic definitions that describe the evolutionary relationship between the observed genetic sequences.

2.5 Transmission networks

In the context of epidemic modelling, a commonly used definition for a transmission network usually has nodes which describe individuals or groups of individuals and links between the represent transmission events (e.g. Pellis et al. (2015) and references therein).

While this serves as a simple and intuitive way to describe the transmission through a population, we wish to extend this notion to explicitly consider the temporal progression of the disease, for reasons discussed later in Section 2.8.

We wish to represent the transmission network by the individual life histories of each infected individual. In the models we are considering, each individual ever to have been infected will have a time at which they contracted the pathogen (typically an infection time for SIR models, or exposure time for SEIR models, etc.) and also a time at which the individual is no longer infectious, i.e. a removal time. To avoid confusion between the various types of terminology we shall refer to the time at which the individual contracted the pathogen as the infection time.

In this formulation, rather than representing an individual with a single node, we now represent an individual by a collection of connected nodes between the infection and removal time to denote the period which an individual is present in the population with the pathogen. Furthermore, if an individual i infects another individual j we connect the two with a directed edge, however here the origin of the edge is at the precise time of infection.

More precisely, let V_i^I and V_i^R denote the nodes which correspond to the time of infection and removal respectively for individual i. Furthermore, suppose that a particular individual i infects $k \geq 0$ susceptibles and let $C_i = \{C_i^1, ..., C_i^k\}$ denote the set of nodes corresponding to these infections, ordered by the time of infection such that the first element corresponds to the first infection and so on. Note that if i does not infect anyone then $k = 0$ and $C_i = \emptyset$.

Define $V_i = \{V_i^I, C_i, V_i^R\}$ to be to complete set of nodes associated with i which are the infection time node, all infections and the removal time node. Finally we 'join' these nodes by adding an edge between sequential nodes and also between the infection node V_j^I with the corresponding node in the infector C_i^j where i infects j. In the event that a particular individual i has imported the pathogen from the community and hence has no source of infection, then the infection node V_j^I is the root of the tree. A simple example transmission network with two individuals can be found in Figure 2.1.

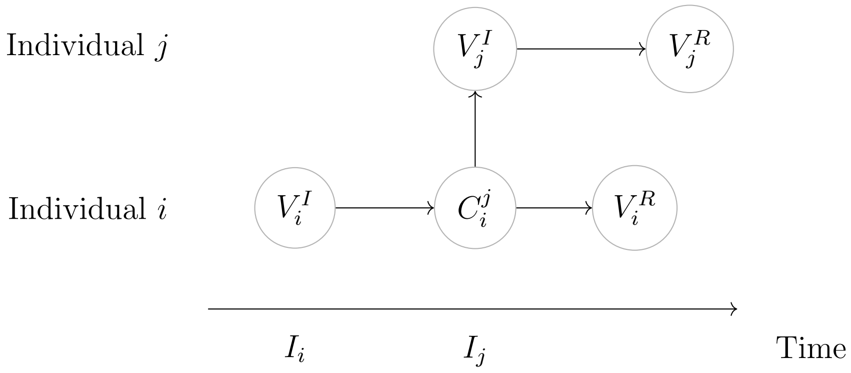


Figure 2.1: An example transmission network consisting of two individuals where i is initially infectious at time I_i and later infects j at time I_j where $I_i < I_j < \infty$. Here V_i^I and V_j^I are the nodes that correspond to the time of infection for individuals i and j and similarly V_i^R and V_j^R are the nodes that correspond to the time of removal. Since i infects j there also is C_i^j in the node set for i at the time of infection.

The nodes need not be aligned by time as in Figure 2.1, however this presentation is intuitive and straightforward to understand. The benefit of viewing the transmission network temporally is useful to visually describe the evolution of the pathogen, in the sense that an individual contracts the pathogen at the time of infection and then the pathogen will evolve over time along the network.

2.6 Genetic networks

In this section we define the two primary objects of interest, a *sub genetic network* and *genetic network*. The models of nucleotide substitution introduced in Section 2.4 provide us with a framework to evaluate the probability of various substitution types between a single nucleotide observed at distinct time points, however we need an idea of how each of the nucleotides are related. Throughout this section we shall use the terms nucleotides and genetic sequences interchangeably, since a nucleotide is a specific case of a sequence of length one and here we are interested in describing how these objects are related.

Suppose we observe multiple nucleotides, typically from a phylogenetic viewpoint the goal would be to reconstruct a phylogenetic tree which describes the evolutionary relationship between the observed sequences by inferring most recent common ancestors. In our work we wish to construct a genetic network, which plays the same role as a phylogenetic tree in the sense that it provides a graphical representation of the evolutionary relationship between sequences, however we do not explicitly aim to reconstruct unobserved ancestral states. To put it another way, in order to evaluate the probability of nucleotide substitution, we need to have knowledge of which sequences are directly linked.

Furthermore, motivated by real world data, there are often significant genetic diversity between organisms sampled from the same species which may be referred to as a specific ‘strain’ or ‘subtype’. In the literature there is not a universally accepted definition for the terms ‘strain’, ‘variant’ or ‘subtype’ (Kuhn et al., 2012), however throughout the rest of the thesis we shall use the term ‘subtype’ to refer to genetically similar organisms from the same species.

In light of the various genetic subtypes that are found in real world data, we wish to explicitly describe the structure of the genetic isolates within each subtype and also the genetic structure of all organisms as a whole.

Suppose there are n_s genetic sequences of length N sampled from a population which are assumed to belong to one of M distinct genetic subtypes. First we describe a *sub genetic network* which is defined as follows.

Definition 2.6.1 (Sub genetic network)**.** A *sub genetic network*, denoted by $\mathcal{G}_i = (V_i, E_i)$ for $i = 1, ..., M$ provides a graphical representation of the evolution of the sequences corresponding to the ith subtype.

In a more precise mathematical framework, a sub genetic network $\mathcal{G}_i$ for the ith subtype is a rooted directed acyclic graph with m_i nodes (or vertices) corresponding to each of the genetic sequences, with arbitrary labels $V_i = \{S_{i,1}, ..., S_{i,m_i}\}$, and edges E_i which represent an evolutionary transition from an ancestral genetic sequence. For any $(a, b) \in E_i$ with a specified direction $a \to b$, the genetic sequence b is assumed to have evolved from the ancestral genetic sequence a. We further note that the sub genetic network $\mathcal{G}_i$ is in fact weighted by an edge function $w : E_i \to \mathbb{R}^+$ which assigns the difference in time between genetic isolates. Assuming that we are modelling the evolution of a single genetic sequence, if the weight between two nodes is zero, then the sequences are sampled at the same time and therefore must be identical.

We wish to consider the possibility of multiple introductions of the pathogen and consequently the sub genetic network may consist of multiple disjoint trees. Each node that is not a root of a tree has exactly one source, which is referred to as the ancestor sequence or genetic source. Details of how to construct the genetic network given epidemiological data are introduced later in Section 2.8. An example sub genetic network with three sequences is given in Figure 2.2.

We introduce the term *genetic network* which is the graphical object (also known as a forest) that consists of all sub genetic networks and is defined as follows.

Definition 2.6.2 (Genetic network)**.** A *genetic network*, denoted by $\mathcal{G} = (V, E)$, is defined as the disjoint union of all sub genetic networks $\mathcal{G}_i$ for $i = 1, ..., M$. In line with the notation used in (Gross et al., 2018, Chapter 2), the disjoint union of graphs

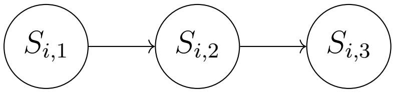


Figure 2.2: An example sub genetic network $\mathcal{G}_i = (V_i, E_i)$ for genetic subtype i consisting of three genetic sequences. The origin or root is sequence $S_{i,1}$, which evolves through time and is later observed, denoted by sequence $S_{i,2}$. The genetic sequence is then observed at some time later, denoted by sequence $S_{i,3}$ which is assumed to have evolved from sequence $S_{i,2}$. The genetic sequences are observed at times t_i for $i = \{1, 2, 3\}$ and the edges are weighted by the difference in time between the genetic sequences. The vertices are defined as $V_i = \{S_{i,1}, S_{i,2}, S_{i,3}\}$ with edges $E_i = \{\{S_{i,1}, S_{i,2}\}, \{S_{i,2}, S_{i,3}\}\}$

is defined as

$$\begin{aligned} \mathcal{G} &= \cup_{i=1}^{M} \mathcal{G}_i \\ &= \cup_{i=1}^{M} (V_i, E_i) \\ &= (\cup_{i=1}^{M} V_i, \cup_{i=1}^{M} E_i) \\ &= (V, E). \end{aligned}$$

The genetic network is therefore a forest that contains the M sub genetic networks, where $V = \cup_{i=1}^{M} V_i$ is the disjoint union of nodes in each sub network and $E = \cup_{i=1}^{M} E_i$ is the disjoint union of edges. An example of a genetic network can be found in Figure 2.3.

The main idea here is that we have multiple genetic sequences that are assumed to belong to one of M distinct subtypes. The graphical representation of the ith subtype is referred to as a sub genetic network, denoted by $\mathcal{G}_i$. The collection (or disjoint union) of the M sub genetic networks is then called the genetic network, which will be a convenient mathematical object that is necessary evaluate the probability of pairwise genetic distances used throughout the rest of this thesis.

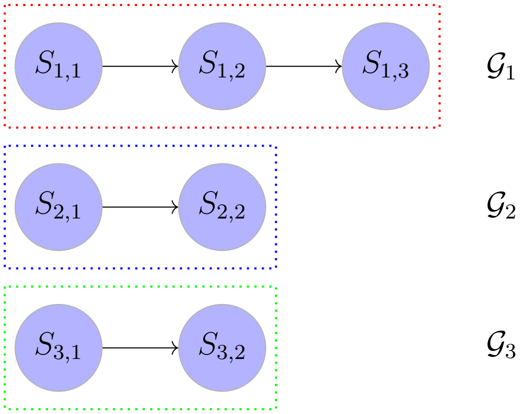


Figure 2.3: An example genetic network $\mathcal{G}$ consisting of observed sequences assumed to belong to one of three subtypes. The genetic network $\mathcal{G}$ is a union of three disjoint sub networks $\mathcal{G} = \mathcal{G}_1 \cup \mathcal{G}_2 \cup \mathcal{G}_3$ where each sub graph $\mathcal{G}_i$ describes the genetic network of the ith nucleotide for $i = 1, 2, 3$.

2.7 Distributions of pairwise genetic distances

In this section we consider the joint distribution of pairwise genetic distances under the assumption of two nucleotide substitution models. The assumed substitution models are the Jukes-Cantor (JC69) and Kimura (K80) which are described in Section 2.4. Furthermore, we would like to derive probability distributions in the general scenario where there are $M \geq 1$ competing organisms evolving through time.

The setup for each of the following sections is identical and is as follows. Consider nucleotides evolving through time evolving through time under an arbitrary genetic network $\mathcal{G} = (V, E)$ which is comprised of M sub genetic networks where the ith network describes the ith genetic subtype for $i = 1, ..., M$.

With each of these we first derive the distribution of the distances for a sequence of length $N = 1$. Then, by assuming that each nucleotide evolves independently of one another we derive expressions for sequences of arbitrary length.

Consider $M \geq 1$ distinct nucleotides evolving independently of one another through time denoted by $\{X(t)\} = \{X_1(t), X_2(t), ..., X_K(t)\}$ where $\{X_n(t)\}$ describes the process of the nth nucleotide where each $\{X_n(t)\}$ is a homogeneous continuous-time Markov chain for $n = 1, ..., M$ on a finite state space $\mathcal{E} = \{A, G, C, T\}$. For distinct $i, j \in \mathcal{E}$, let $q_{ij}^{[n]}$ be the transition rate for the nth nucleotide from state i to state j, and $q_{ii}^{[n]} = -\sum_{j \neq i} q_{ij}^{[n]} < 0$ by definition.

2.7.1 JC69 model

Before deriving the distribution of pairwise genetic distances under the JC69 mutation model we first outline properties of the distance matrix calculated from sequences of length one. Let (X, d_H) denote a metric space with $\{x_1, \ldots, x_n\}$ to be a set of points in X and let $\mathbf{d}(x_1, \ldots, x_n)$ denote its distance matrix which is a symmetric non-negative matrix with zeros along the diagonal and elements $(d_H(x_i, x_j))_{i,j=1}^n$. Define the distance function $d_H(x_i, x_j) = 1$ if $x_i = x_j$ otherwise $d_H(x_i, x_j) = 0$ and let $d_H(x_i, x_j) = d_{i,j}$ for shorthand.

Each x_i is a genetic sequence of length one, in other words each $x_i \in \mathcal{E}$ is a nucleotide base. We wish to determine properties of the distance matrix $\mathbf{d}$ in order for it to have been computed from the n nucleotide bases $x_1, \ldots, x_n$.

We define $\mathcal{D}_n$ to be the space of all valid distance matrices computed from n nucleotides. For example, consider nucleotides (x_1, x_2, x_3) and suppose the pairwise distances are $\boldsymbol{d} = (d_{1,2}, d_{1,3}, d_{2,3}) = (0, 1, 0)$. We note that $d_{1,2} = d_{2,3} = 0$ which implies that $x_1 = x_2$ and $x_2 = x_3$, however $d_{1,3} = 1$ and hence $x_1 \neq x_3$, which is a contradiction and therefore $\boldsymbol{d} \notin \mathcal{D}_3$.

A distance matrix $\mathbf{d}$ is considered valid if and only if there is complete agreement among all pairwise distances. An algorithmic approach to determine whether or not an arbitrary distance matrix has been computed from the metric space (X, d_H) and therefore is valid is described as follows.

Consider a graph $G = (V, E)$ with an adjacency matrix A which is defined as $A_{ij} = 1$

if and only if $d_{ij} = 0$ and $A_{ij} = 0$ otherwise. In other words, we construct a simple undirected graph such that an edge is connected between nodes i and j if the distance is zero. A graph G is said to be disconnected if and only if it can be expressed as a disjoint union of two or more subgraphs and we define the subgraphs to be *components* of G (Bondy, 1976).

We wish to find the number of components of G induced by the adjacency matrix A and assign the nodes $v \in V$ to one of these components. Note that the number of components of G must not be greater than four, which is the size of the set of nucleotide bases $\mathcal{E}$. Equivalently, we are determining a partition of the nodes which is later formalised by equivalence relations. Once we have assigned a group for each $v \in V$, all that remains is to check is that the subgraphs are complete - each node should have an edge with all other nodes in the group. This is straight forward to compute in linear time by verifying that the degree of nodes within the same group is equal.

In the example above with $\boldsymbol{d} = (d_{1,2}, d_{1,3}, d_{2,3}) = (0, 1, 0)$ we calculate the adjacency matrix of $G = (V, E)$ as

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

which implies that there is a singular component consisting of the nodes $\{1, 2, 3\}$. However, the graph is not complete since nodes $\{1, 3\}$ are of degree one but node $\{2\}$ is of degree two and hence $\boldsymbol{d} \notin \mathcal{D}_3$. Now we have an algorithm approach to determine whether or not the pairwise distances have been computed from nucleotides $x_1, \ldots, x_n$, we now outline a procedure for assigning nucleotide bases from valid distance matrices $\mathbf{d} \in \mathcal{D}_n$.

Recall that under the assumption of the JC69 mutation model, the probability that the nth nucleotide is in state j after t time units given the nucleotide started in state

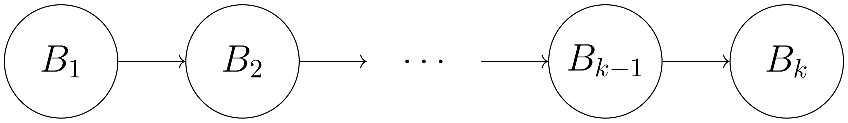


Figure 2.4: A single nucleotide evolving through time, observed at time points t_i with each B_i denoting the observed nucleotide base at that time.

i is given by

$$p_{ij}^{[n]}(t) = \begin{cases} \frac{1}{4}(1 + 3e^{-4\lambda_n t}) & i = j \\ \frac{1}{4}(1 - e^{-4\lambda_n t}) & i \neq j. \end{cases}$$

Since we are observing k processes assumed to have evolved independently from one another, the joint probability mass function of $X(t)$ is the product of the probability mass functions for $X_1(t), ..., X_k(t)$.

For each of the k nucleotides evolving independently of one another there are distinct mutation rates denoted by λ_n, however throughout the rest of the chapter we shall assume that these processes are identical and there is a common mutation rate, $\lambda_1 = \lambda_2 = ...\lambda_k = \lambda$ say.

Suppose there are k nodes in the genetic network that correspond to observations of the nucleotide at time points t_i for $i = 1, ..., k$ and label this set of vertices $V = \{1, ..., k\}$. Recall that a directed edge $(i, j) \in E$ indicates that nucleotide j has evolved from i and the edge is weighted by the difference in time, i.e. $t_j - t_i$.

Recall that $\mathcal{E} = \{A, G, C, T\}$ denotes the set of DNA nucleotide bases and let $(x_1, ..., x_k)$ denote the vector of observed nucleotide bases where $x_i \in \mathcal{E}$ for $i = 1, ..., k$. The general procedure is to use the pairwise genetic distances as the starting point of analysis, that is we wish to derive the probability distributions for the distances without observing the original sequence. The strategy we use is to consider all possible nucleotide bases that give rise to the observed genetic distances.

It is straightforward to calculate the matrix of pairwise distances $\boldsymbol{d}$ if we observe the genetic sequences directly, however we wish to categorize the reverse and determine the genetic sequences from the matrix of pairwise genetic distances. Mathematically, we are trying to characterize the mapping from the set of possible $k \times k$ distance matrices into the set of sequences $(x_1, ..., x_k)$.

Lemma 2.7.1. *Let* $\boldsymbol{d} \in \mathcal{D}_k$ *be a distance matrix for sequences of length one for a known genetic network* $\mathcal{G}$ *with* $k > 1$ *connected nodes. Then there exists a unique corresponding sequence of nucleotides, up to a permutation of the bases* $\mathcal{E} = \{A, G, C, T\}$.

Proof. We wish to assign each node in the genetic network to a nucleotide base that agrees with the distance matrix $\boldsymbol{d}$. Furthermore we wish to prove that this assignment is unique (up to permutation of bases), and that the maximum number of assignments is four. We begin by labelling the set of nodes $S = \{1, ..., k\}$ and define a binary relation on the set S by $i \sim j \iff d_{ij} = 0$. We note that this relation has the following properties:

1. Reflexive – Suppose $i \in S$, then $d_{ii} = 0$ by definition.
2. Symmetric – Suppose $i, j \in S$, and $i \sim j$. Then $d_{ij} = d_{ji} = 0$, and so $i \sim j$.
3. Transitive – Suppose $i, j \in S, i \sim j$ and $j \sim k$. Then $d_{ij} = 0$ and $d_{jk} = 0$. Then i has the same base as j, and k has the same base as j, hence all three are the same, and so $i \sim k$.

Therefore $\sim$ is an equivalence relation on S. This equivalence relation induces a partition on S; every element of S belongs to only one equivalence class, where each equivalence class is the set of nodes with a distance of zero, i.e. they share a common nucleotide base.

Suppose we wish to assign nucleotide bases to the nodes in S. We first label the equivalence classes as $1, ..., p$ where p is the number of partitions induced on S. Then we can assign all of the nodes in class 1 to A, all of the nodes in class 2 to G, all of the nodes in class 3 to C and finally all of the nodes in class 4 to T.

If we have more than four equivalence classes, then this corresponds to an invalid biological sequence as there are at most four nucleotide bases to assign. Furthermore, it does not matter which order the nucleotide bases are assigned since we are primarily concerned with the structure of the sequence, rather than the actual sequence itself. □

It has been shown that given a $k \times k$ pairwise distance matrix for genetic sequences of length $N = 1$, there exists a unique corresponding sequence (up to a permutation of bases). Before we introduce our main result, we first require a way to count the number of partitions of nodes in $\mathcal{G}$ induced by the pairwise distance matrix $\boldsymbol{d}$. Define $h(\boldsymbol{d})$ to be the number of distinct nucleotide bases in the genetic sequence.

Lemma 2.7.2. *Let $\boldsymbol{d} \in \mathcal{D}_k$ be a distance matrix for a sequence of length one for a known genetic tree $\mathcal{G}$ with $k > 1$ connected nodes. Then the number of distinct bases in the genetic network is given by*

$$h(\boldsymbol{d}) = 1 + \sum_{j=2}^{k} \left[\prod_{i=1}^{j-1} d_{i,j} \right], \tag{2.3}$$

where the number of distinct bases is the number of distinct elements from $\mathcal{E} = \{A, G, C, T\}$.

Proof. Begin by labelling the nodes $1, ..., k$ and the first node is the first base we are counting. Clearly there is always at least one distinct base in a given sequence. The next distinct base can be determined by inspecting each node j in order and observing the associated pairwise distances. For any $j = 2, ..., k$, suppose there exists a node $i < j$ such that $d_{ij} = 0$, then by definition this implies that $x_i = x_j$ and therefore node j is not distinct. However suppose $d_{ij} = 1$ for all $i = 1, ..., j-1$, then clearly $x_i \neq x_j$ and j is a new nucleotide base, and therefore the product term in Equation (2.3) will be equal to 1. The product term will be zero otherwise, hence the product term is equal to 1 if and only if node j is a different base to nodes $1, ..., j-1$. Furthermore by Lemma 2.7.1, the maximum value $h(\boldsymbol{d})$ can take is 4, therefore $h(\boldsymbol{d}) \in \{1, 2, 3, 4\}$. □

The result from Lemma 2.7.1 induces a partition of the nodes in $\mathcal{G}$ by the distance matrix $\boldsymbol{d}$, and Lemma 2.7.2 provides us with an expression to calculate the number of partitions.

Now we consider the probability of observing a distance matrix for a sequence of length one under the assumption that nucleotides mutate according to the JC69 model. Recall from Section 2.4.1 that the probability to observe a mutation between two connected nodes in t time units is given $q(t) = 1 - p_{ii}(t) = \frac{3}{4}(1 - e^{-4\lambda t})$. Note that q also depends on λ, however we shall consider q as a function of t for a fixed λ for all subsequent notation.

Theorem 2.7.3. *Let $\boldsymbol{d}$ be the pairwise distance matrix for a sequence of length one under a genetic network $\mathcal{G} = (V, E)$ with $k > 1$ connected nodes, $q(t)$ be the probability of observing a mutation in t time units under the JC69 model and $t_{ij} = t_j - t_i$ be the difference in time between nodes i and j. The probability mass function for the distance matrix is given by*

$$f(\boldsymbol{d}|\lambda, \mathcal{G}) = \begin{cases} 3^{\mathbb{1}_{\{h(\boldsymbol{d})>1\}}} 2^{\mathbb{1}_{\{h(\boldsymbol{d})>2\}}} \prod_{(i,j)\in E} (1 - q(t_{ij}))^{1-d_{ij}} \left(\frac{1}{3} q(t_{ij})\right)^{d_{ij}} & \boldsymbol{d} \in \mathcal{D}_k \\ 0 & \textit{otherwise} \end{cases}$$

where E is the set of edges in the genetic network, d_{ij} refers to the distance between nodes i and j and λ is the mutation rate.

Proof. Consider a partition of nodes in $\mathcal{G}$ induced by the distance matrix $\boldsymbol{d}$ and call the first node the reference base. It is clear that if $h(\boldsymbol{d}) > 1$, the second distinct base can be chosen from the three remaining bases in $\{A, G, C, T\}$, and if $h(\boldsymbol{d}) > 2$, the third distinct node can be chosen from the two remaining bases in $\{A, G, C, T\}$.

With any assignment we choose, the probability to observe $\boldsymbol{d}$ is obtained by simply evaluating the probability of mutation (or lack of mutation) in the links (or edges) of the genetic network. Every link in the genetic tree is independent of all others from the Markov assumption of the JC69 mutation model, and the probability of observing a specific mutation is given by $q(t)/3$, since there are three remaining nucleotide bases to mutate to, hence the 1/3 term. Finally, the probability of not observing a mutation is given by the $1 - q(t)$ term. □

We have verified computationally that $\sum_{\mathbf{d}\in\mathcal{D}_k} f(\mathbf{d}|\lambda, \mathcal{G}) = 1$ for genetic networks with $k \leq 8$ nodes and believe that this result is true for all $k > 1$, however we no formal proof for this claim.

It is important to note that this result is valid for any number of genetic subtypes, i.e. $M \geq 1$. We had made no explicit reference the subtypes since all information on the connectivity of the nodes is contained within the genetic network $\mathcal{G}$ which is assumed to be known. Details on how to calculate genetic networks conditional on an underlying transmission network is discussed in Section 2.8.

2.7.1.1 Sequences of length N

Up to this point we have only considered a sequences of length one evolving through time, however our primary interest lies with whole genome sequences which vary in length depending on the organism. We will assume that each nucleotide site evolves independently of one another and we wish to consider the joint distribution of genetic distances for each nucleotide.

Suppose each of the N nucleotides evolve through time under the genetic network $\mathcal{G}$, then let $\boldsymbol{d}^{[i]} \in \mathcal{D}_k$ denote the observed pairwise distance matrix for the ith nucleotide for $i = 1, ..., N$. Recall that the genetic tree $\mathcal{G} = (V, E)$ describes the structure of the genetic isolates.

Let $\boldsymbol{d}^{[1]}, \ldots, \boldsymbol{d}^{[N]}$ be independent and identically distributed random distance matrices from the distribution in Theorem 2.7.3. Then the joint probability mass function is given by

$$\begin{aligned}
f_{\boldsymbol{d}^{[1]},\ldots,\boldsymbol{d}^{[N]}}(\boldsymbol{d}^{[1]},\ldots,\boldsymbol{d}^{[N]}|\lambda,\mathcal{G}) &= \Pr(\boldsymbol{d}^{[1]}=\boldsymbol{d}^{[1]},\ldots,\boldsymbol{d}^{[N]}=\boldsymbol{d}^{[N]}) \\
&= \prod_{i=1}^{N} f(\boldsymbol{d}^{[i]}|\lambda,\mathcal{G}) \\
&= \prod_{i=1}^{N}\left[3^{\mathbb{1}_{\{h(\boldsymbol{d}^{[i]})>1\}}}2^{\mathbb{1}_{\{h(\boldsymbol{d}^{[i]})>2\}}}\prod_{(j,k)\in E}(1-q(t_{jk}))^{1-d_{jk}^{[i]}}\left(\frac{1}{3}q(t_{jk})\right)^{d_{jk}^{[i]}}\right] \\
&= 3^{a}\times 2^{b}\times\prod_{i=1}^{N}\prod_{(j,k)\in E}(1-q(t_{jk}))^{1-d_{jk}^{[i]}}\left(\frac{1}{3}q(t_{jk})\right)^{d_{jk}^{[i]}} \\
&= 3^{a}\times 2^{b}\times\prod_{(j,k)\in E}(1-q(t_{jk}))^{N-D_{jk}}\left(\frac{1}{3}q(t_{jk})\right)^{D_{jk}} \\
&= H_1(\boldsymbol{d}^{[1]},...,\boldsymbol{d}^{[N]})g_1(\lambda,\mathcal{G},T_1(\boldsymbol{d}^{[1]},...,\boldsymbol{d}^{[N]})), \qquad (2.4)
\end{aligned}$$

where $a=\sum_{i=1}^{N}\mathbb{1}_{\{h(\boldsymbol{d}^{[i]})>1\}}$ and $b=\sum_{i=1}^{N}\mathbb{1}_{\{h(\boldsymbol{d}^{[i]})>2\}}$ which are the number of distance matrices that contain more than one and two distinct nodes respectively. Furthermore we have $D_{jk}=\sum_{i=1}^{N}d_{jk}^{[i]}$ which is the total number of differences (SNPs) between sequence j and the sequence k, where $(j,k)\in E$, and the functions H_1, g_1 and T_1 are defined as

$$\begin{aligned}
H_1(\boldsymbol{d}^{[1]},...,\boldsymbol{d}^{[N]}) &= 3^{\sum_{i=1}^{N}\mathbb{1}_{\{h(\boldsymbol{d}^{[i]})>1\}}}\times 2^{\sum_{i=1}^{N}\mathbb{1}_{\{h(\boldsymbol{d}^{[i]})>2\}}} \\
g_1(\lambda,\mathcal{G},T_1(\boldsymbol{d}^{[1]},...,\boldsymbol{d}^{[N]})) &= \prod_{(i,j)\in E}(1-q(t_{jk}))^{N-D_{jk}}\left(\frac{1}{3}q(t_{jk})\right)^{D_{jk}} \\
T_1(\boldsymbol{d}^{[1]},...,\boldsymbol{d}^{[N]}) &= \sum_{i=1}^{N}\boldsymbol{d}^{[i]}
\end{aligned}$$

It is important to note that the quantities D_{ij} for $(i,j)\in E$ are straightforward to obtain by calculating a matrix of pairwise distances for each observed sequence. For all subsequent analysis, we shall refer to D as the pairwise distance matrix for sequences of length N. It follows from the Fisher-Neyman factorisation theorem that the (matrix) statistic T, which is the defined as the sum of distance matrices $\boldsymbol{d}^{[i]}$ for $i=1,...,N$, is sufficient for the underlying parameter λ.

We are unable to calculate H_1 explicitly as we require knowledge of the actual sequences rather than the matrix of pairwise distances, however since these are functions of the data alone these terms vanish in the posterior distribution when performing Bayesian inference.

To see this clearly, consider some data X with model parameters θ. Let $\pi(X \mid \theta)$ denote the likelihood and suppose that the likelihood can be decomposed into two functions H and g such that $\pi(X \mid \theta) = H(X)g(X \mid \theta)$ where $H(X)$ depends on the data alone and $g(X \mid \theta)$ depends on the data conditional on the model parameters. The posterior distribution is then given by

$$\begin{aligned}
\pi(\theta \mid X) &= \frac{\pi(X \mid \theta)\pi(\theta)}{\int_\theta \pi(X \mid \theta)\pi(\theta)d\theta} \\
&= \frac{H(X)g(X \mid \theta)\pi(\theta)}{\int_\theta H(X)g(X \mid \theta)\pi(\theta)d\theta} \\
&= \frac{H(X)g(X \mid \theta)\pi(\theta)}{H(X)\int_\theta g(X \mid \theta)\pi(\theta)d\theta} \\
&= \frac{g(X \mid \theta)\pi(\theta)}{\int_\theta g(X \mid \theta)\pi(\theta)d\theta}
\end{aligned}$$

From the above formulation, it is clear the function H_1 which involves the data alone cancel in the posterior distribution and the matrix of pairwise distances is sufficient to perform inference with a known genetic network $\mathcal{G}$.

For a genetic sequence of length one, we have found a unique mapping (up to a permutation) from the distance matrix to the sequence of nucleotide bases. Given that we are able to recover the structure of the nucleotides, we can then evaluate the probability of observing pairwise genetic distances under the JC69 model and provide the explicit probability mass function in Theorem 2.7.3.

Since each nucleotide site is assumed to evolve independently of all other sites, it has been shown that the joint probability mass function can be factored such that the sum of the distance matrices is sufficient to estimate the mutation rate. This formulation is convenient, since the aim of this chapter is to provide inference from the full matrix of pairwise distances alone. To show this, consider the following example.

Example 2.7.1. Suppose we observe a genetic sequence of length $N = 6$ at times

Nucleotide	(d_{12}, d_{13}, d_{23})	$h(\boldsymbol{d})$	Probability
1	$(0,0,0)$	1	$(1-q(t_{12}))\,(1-q(t_{23}))$
2	$(1,0,1)$	2	$\frac{1}{3}q(t_{12})q(t_{23})$
3	$(0,1,1)$	2	$(1-q(t_{12}))\,q(t_{23})$
4	$(0,0,0)$	1	$(1-q(t_{12}))\,(1-q(t_{23}))$
5	$(1,1,1)$	3	$\frac{2}{3}q(t_{12})q(t_{23})$
6	$(1,1,0)$	2	$q(t_{12})\,(1-q(t_{23}))$

Table 2.1: The distance matrix for each nucleotide and corresponding probability from Theorem 2.7.3 for the nucleotides $i = 1, ..., 6$ which are shown in Figure 2.5.

t_1, t_2 and t_3 such that $t_1 < t_2 < t_3$ with the setup in Figure 2.5. In this scenario the pairwise distance matrix can be summarised as $D = (d_{12}, d_{13}, d_{23}) = (3, 3, 3)$. Recall that the objective is to perform inference on the matrix of pairwise distances alone, without knowledge of the exact composition of the sequences themselves.

Suppose we do in fact observe the sequences directly, then the corresponding probability to observe each nucleotide site is calculated by Theorem 2.7.3 and is summarised in Table 2.1.

By considering the joint probability mass function described by Equation (2.4), we obtain

$$\begin{aligned} f_{\boldsymbol{d}^{[1]},\ldots,\boldsymbol{d}^{[6]}}(\boldsymbol{d}^{[1]}, \ldots, \boldsymbol{d}^{[6]}|\lambda, \mathcal{G}) &= \prod_{i=1}^{6} f(\boldsymbol{d}^{[i]}|\lambda, \mathcal{G}) \\ &= H_1(\boldsymbol{d}^{[1]}, ..., \boldsymbol{d}^{[6]})\,(1-q(t_{12}))^3 \left(\frac{1}{3}q(t_{12})\right)^3 \\ &\times (1-q(t_{23}))^3 \left(\frac{1}{3}q(t_{23})\right)^3 \end{aligned}$$

where H_1 is function that depends on the data alone and is explicitly given by $H_1 = 3^{\sum_{i=1}^{6} \mathbb{1}_{\{h(\boldsymbol{d}^{[i]})>1\}}} \times 2^{\sum_{i=1}^{6} \mathbb{1}_{\{h(\boldsymbol{d}^{[i]})>2\}}} = 3^4 \times 2$. It is clear that the statistic $T_1(\boldsymbol{d}^{[1]}, ..., \boldsymbol{d}^{[6]}) = \sum_{i=1}^{6} \boldsymbol{d}^{[i]} = (3, 3, 3)$ is sufficient to estimate the mutation rate λ which occurs in the joint mass function through the function $q(\cdot)$ alone.

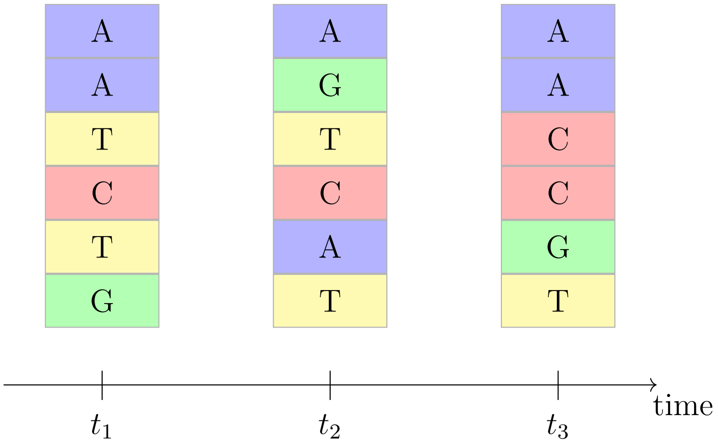


Figure 2.5: A genetic sequence of length $N = 6$ evolving through time observed at $t = t_1, t_2$ and t_3. The total pairwise differences between the sequences is given by $(d_{12}, d_{13}, d_{23}) = (3, 3, 3)$.

2.7.2 K80 model

Under the assumption of the K80 mutation model, we now differentiate between the type of substitution and hence require more detailed summary statistics for the genetic data, rather than simply pairwise distances introduced in Section 2.7.1. Define d_{ij}^T, d_{ij}^V to be the transition and transversion indicator functions which are equal to one if there is a transition or transversion mutation between nucleotides i and j respectively. Furthermore it follows that we have $d_{ij}^T + d_{ij}^V = d_{ij}$ where d_{ij} is defined as in Section 2.7.1 as equal to one if and only if the bases between nucleotides i and j differ.

Let $\mathcal{P}_u = \{A, G\}$ and $\mathcal{P}_y = \{C, T\}$ denote the purines and pyrimidines respectively and consider a nucleotide at distinct time points denoted by x_i and x_j. The two nucleotides are said to have a transversion mutation if and only if

$$x_i \in \mathcal{P}_u, x_j \in \mathcal{P}_y \quad \text{or} \quad x_j \in \mathcal{P}_y, x_i \in \mathcal{P}_u,$$

and in this case we set $d_{ij}^V = 1$ and $d_{ij}^T = 0$. On the other hand, the two nucleotides are said to have a transition mutation if and only if $x_i \neq x_j$ and

$$x_i, x_j \in \mathcal{P}_u \quad \text{or} \quad x_i, x_j \in \mathcal{P}_y,$$

and we set $d_{ij}^V = 0$ and $d_{ij}^T = 0$. Finally we have the case where the nucleotides are equal, i.e. $x_i = x_j$, in which case no mutation is said to have occurred and we set $d_{ij}^V = 0$ and $d_{ij}^T = 0$.

Finally, let $\boldsymbol{d}^T, \boldsymbol{d}^V$ denote the matrices of pairwise transition and transversions and define $\mathcal{D}^K$ to be the space of transition and transversion matrices that correspond to a valid biological sequence.

For example, consider the simplest scenario with two nucleotide bases (labelled 1 and 2) in which case we have two random variables denoted by (d_{12}^T, d_{12}^V). Suppose we have $(d_{12}^T, d_{12}^V) = (1, 1)$, i.e. both a transition and transversion has occurred. Notice $d_{12}^T = 1$ implies that both nucleotides are either purines or pyrimidines, however $d_{12}^V = 1$ implies that the nucleotides are not of the same group, hence we have a contradiction and $(1, 1) \notin \mathcal{D}^K$. Put another way, there are no possible sequences which give rise to the distances $(1, 1)$ and this particular realisation has probability zero.

The general strategy to derive the joint distribution of distances is to consider a mapping from genetic distances (transitions and transversions in this case) to a nucleotide bases. In other words, we trying to determine all possible genetic sequences that give rise to a particular set of distances. Mathematically we are trying to characterize a mapping from the set of possible transition and transversion matrices to the set of nucleotide bases, written as $\phi(\boldsymbol{d}^T, \boldsymbol{d}^V) \mapsto S \in \mathcal{E}^k$ where $\mathcal{E} = \{A, G, C, T\}$ is the set of DNA nucleotide bases.

Define the term ‘nucleotide type’ to denote the type of organic group the particular base belongs to, which may either be a purine or pyrimidine.

Theorem 2.7.4. *Let $\boldsymbol{d}^T, \boldsymbol{d}^V \in \mathcal{D}^K$ be the pairwise transitions and transversions for a sequence of length one under a known genetic network $\mathcal{G} = (V, E)$ with $k > 1$ connected nodes. Define $p_T(t)$ and $p_V(t)$ to be the probability of a transition and transversion respectively in t time units under the K80 model. Let $t_{ij} = t_j - t_i$ be the difference in time between nodes i and j. The probability mass function for matrices* *of pairwise transition and transversions is given by*

$$f(\boldsymbol{d}^T, \boldsymbol{d}^V | \lambda_T, \lambda_V, \mathcal{G}) = \prod_{(i,j)\in E} (1 - p_T(t_{ij}) - p_V(t_{ij}))^{1-d_{ij}^T - d_{ij}^V} p_T(t_{ij})^{d_{ij}^T} \left(\frac{1}{2} p_V(t_{ij})\right)^{d_{ij}^V} \times 2^{\mathbb{1}_{\{\sum_{i,j} d_{ij}^V > 0\}}} \tag{2.5}$$

if $(\boldsymbol{d}^T, \boldsymbol{d}^V) \in \mathcal{D}^K$ *and* $f(\boldsymbol{d}^T, \boldsymbol{d}^V | \lambda_T, \lambda_V, \mathcal{G}) = 0$ *otherwise. Recall that* d_{ij}^T*,* d_{ij}^V *are the transitions and transversions between sequences* i *and* j*, where* $(i, j) \in E$ *is an edge in the genetic network* $\mathcal{G}$ *and* λ_T *and* λ_V *are the mutation rate parameters.*

Proof. Given there are k nodes in the genetic network which can be written as $S = (B_1, ..., B_k)$, we wish to assign a nucleotide base to each x_i for $i = 1, ..., k$.

Without loss of generality set $x_1 = A$, then we wish to consider the pairwise transitions and transversions to deduce the nucleotide base for all other nodes. Let $\mathcal{T} = \{j : d_{1j}^T = 1, j > 1\}$ denote the set of nodes that have a transversion and $\mathcal{A} = \{j : d_{1j}^T = d_{1j}^T = 0, j > 1\}$ denote the set of nodes that have no difference between nodes 1 and j for $j > 1$.

If there exists a transversion substitution between nodes 1 and j for some $j > 1$, then the only way this may occur is if $x_j = G$ for all $j \in \mathcal{T}$. Similarly it is straightforward to deduce that all nodes that have no difference between node 1 must be of the same base, i.e. $x_j = A$ for $j \in \mathcal{A}$.

Now we wish to assign the remaining nodes a nucleotide base to either C or T. Let $\mathcal{X} = \{j : V_{1j} = 1, j > 1\}$ denote the set of nodes that have a transversion difference with node 1 and pick an element from this set at random and observe there are exactly two ways to assign the nucleotide base. Suppose we assign $x_y = C$ for any $y \in \mathcal{X}$, then the remaining nucleotide bases are easily assigned where $\{j : d_{yj}^T = d_{yj}^V = 0\}$ is the set of nodes that are the same as y and $\{j : d_{yj}^T = 1\}$ is the set of nodes with a transition difference to y which must be the last available nucleotide to assign T.

If there are any nodes not assigned by this point then we have distances that do not correspond to a valid biological sequence, i.e. $(\boldsymbol{d}^T, \boldsymbol{d}^V) \notin \mathcal{D}^K$.

Now that we have assigned a nucleotide base to each node, given an underlying genetic network $\mathcal{G}$, we may now look at the edges in the network and evaluate the probabilities of the different types of substitution given by Equation 2.1, which are product terms in Equation 2.5.

Finally, if there is a transversion substitution, there are exactly two ways to assign this base. Therefore we multiply by a factor of 2 given in the second line of Equation 2.5 if and only if there is a transversion substitution which is given by the event $\{\sum_{i,j} V_{ij} > 0\}$.

We remark that we assigned to first node to be base A, however this need not be the case. We may start with any base and by following the same logic the nodes will get assigned in a similar fashion. Furthermore, the labels assigned to the nodes are invariant to permutation, that is to say it does not matter which labels are assigned. □

We provide a sketch proof of Theorem 2.7.4 in Figure 2.6 for a genetic network with $k = 5$ nodes which illustrates the procedure of assigning nucleotide bases to the nodes, which can be seen as equivalent to colouring the nodes on a graph.

2.7.2.1 Sequences of length N

Similar to Section 2.7.1.1, we consider N nucleotides evolving independently of each other and derive the joint probability mass function.

Suppose we have N nucleotides evolving through time under the genetic network $\mathcal{G}$ and let $\boldsymbol{x}^{[i]} = (\boldsymbol{d}^{T[i]}, \boldsymbol{d}^{V[i]}) \in \mathcal{D}^K$ denote the observed pairwise transition and transversion matrices for the ith nucleotide for $i = 1, ..., N$.

Let $\mathbf{X}^{[1]}, ..., \mathbf{X}^{[N]}$ be independent and identically distribution random distance matrices from the probability mass function in Theorem 2.7.4. The joint probability distribution

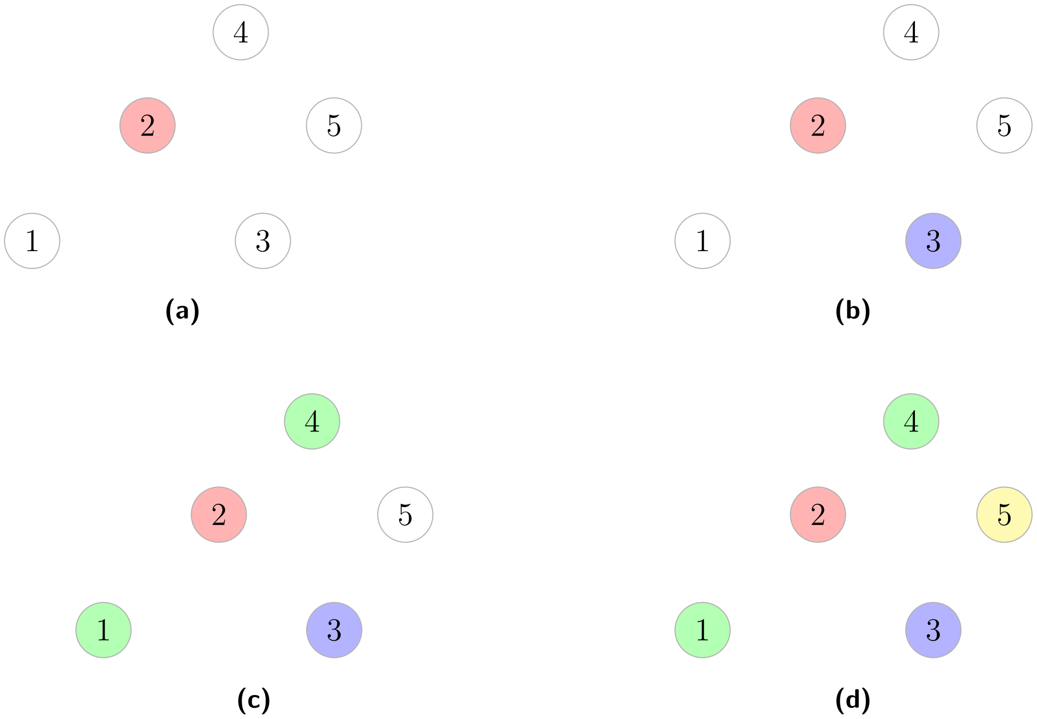


Figure 2.6: Sketch proof to illustrate the base assignment procedure described in Theorem 2.7.4. (a) Select an initial base at random and set this to be A indicated by red, and assign all other nodes with distance zero to also be red. (b) Find all nodes which have a transversion between the initial node and set these to G, indicated by blue. (c) Select one node uniformly at random from the set of nodes that have a transversion difference and find all other nodes which are identical and set these to C, indicated by green. (d) Fill in all remaining nodes by the final nucleotide base T, indicated by yellow.

is given by

$$\begin{aligned}
f_{\mathbf{X}^{[1]},...,\mathbf{X}^{[N]}}(\boldsymbol{x}^{[1]}, ..., \boldsymbol{x}^{[N]}|\lambda_T, \lambda_V, \mathcal{G}) &= \Pr(\mathbf{X}^{[1]} = \boldsymbol{x}^{[1]}, ..., \mathbf{X}^{[N]} = \boldsymbol{x}^{[N]}) \\
&= \prod_{i=1}^{N} f(\boldsymbol{d}^{T[i]}, \boldsymbol{d}^{V[i]}|\lambda_T, \lambda_V, \mathcal{G}) \\
&= \prod_{i=1}^{N} \Big[\prod_{(j,k)\in E} (1 - p_T(t_{jk}) - p_V(t_{jk}))^{1-d_{jk}^{T[i]}-d_{jk}^{V[i]}} p_T(t_{jk})^{d_{jk}^{T[i]}} \\
&\qquad \times \left(\frac{1}{2}p_V(t_{jk})\right)^{d_{jk}^{V[i]}} 2^{\mathbb{1}_{\{\sum_{j,k} d_{jk}^{V[i]}>0\}}} \Big] \\
&= \prod_{(j,k)\in E} (1 - p_T(t_{jk}) - p_V(t_{jk}))^{\sum_{i=1}^{N} 1-d_{jk}^{T[i]}-d_{jk}^{V[i]}} p_T(t_{jk})^{\sum_{i=1}^{N} d_{jk}^{T[i]}} \\
&\qquad \times \left(\frac{1}{2}p_V(t_{jk})\right)^{\sum_{i=1}^{N} d_{jk}^{V[i]}} 2^{\sum_{i=1}^{N} \mathbb{1}_{\{\sum_{j,k} d_{jk}^{V[i]}>0\}}} \\
&= \prod_{(j,k)\in E} (1 - p_T(t_{jk}) - p_V(t_{jk}))^{1-D_{jk}^T-D_{jk}^V} p_T(t_{jk})^{D_{jk}^T} \\
&\qquad \times \left(\frac{1}{2}p_V(t_{jk})\right)^{D_{jk}^V} 2^{\sum_{i=1}^{N} \mathbb{1}_{\{\sum_{j,k} d_{jk}^{V[i]}>0\}}} \\
&= H_2(\boldsymbol{x}^{[1]}, ..., \boldsymbol{x}^{[N]}) g_2(\lambda_T, \lambda_V, \mathcal{G}, T_2(\boldsymbol{x}^{[1]}, ..., \boldsymbol{x}^{[N]})),
\end{aligned} \tag{2.6}$$

where $D_{jk}^T = \sum_{i=1}^{N} d_{jk}^{T[i]}$ and $D_{jk}^V = \sum_{i=1}^{N} d_{jk}^{V[i]}$ are pairwise transition and transversion matrices for sequences of length N and the functions H_2, g_2 and T_2 are defined as

$$\begin{aligned}
H_2(\boldsymbol{x}^{[1]}, ..., \boldsymbol{x}^{[N]}) &= 2^{\sum_{i=1}^{N} \mathbb{1}_{\{\sum_{j,k} d_{jk}^{V[i]}>0\}}} \\
g_2(\lambda_T, \lambda_V, \mathcal{G}, T_2(\boldsymbol{x}^{[1]}, ..., \boldsymbol{x}^{[N]})) &= \prod_{(j,k)\in E} (1 - p_T(t_{jk}) - p_V(t_{jk}))^{1-D_{jk}^T-D_{jk}^V} p_T(t_{jk})^{D_{jk}^T} \\
&\qquad \times \left(\frac{1}{2}p_V(t_{jk})\right)^{D_{jk}^V} \\
T_2(\boldsymbol{x}^{[1]}, ..., \boldsymbol{x}^{[N]}) &= \sum_{i=1}^{N} \boldsymbol{x}^{[i]} \\
&= \sum_{i=1}^{N} (\boldsymbol{d}^{T[i]}, \boldsymbol{d}^{V[i]}).
\end{aligned}$$

It can be seen that the joint probability mass function takes the form required by the Fisher-Neyman factorization theorem, therefore it follows that the matrices of pairwise transition and transversions for sequences of length N, $T_2(\boldsymbol{x}^{[1]}, ..., \boldsymbol{x}^{[N]}) = \sum_{i=1}^{N}(\boldsymbol{d}^{T[i]}, \boldsymbol{d}^{V[i]})$ is a sufficient statistic for (λ_T, λ_V).

Furthermore, we are unable to calculate H_2 explicitly as we require knowledge of the actual sequences rather than pairwise distance data, however for reasons discussed in Section 2.7.1.1 this function vanishes in the posterior in a Bayesian setting and therefore is suitable for practical inference.

2.7.3 Determining genetic groups

In order to evaluate the probability of a pairwise distance matrix assumed to contain observations from multiple sequences described in Section 2.7, we must construct a genetic network $\mathcal{G} = \cup_{i=1}^{K}\mathcal{G}_i$. In order to do this we cluster the genetic sequences such that sequences within clusters are similar and sequences in different clusters are dissimilar. The process to classify and cluster genetic sequences to groups is typically a non trivial task and is an active area of research in bioinformatics, a review of some methods can be found in Zou et al. (2018), however we now describe some simple approaches.

It will be convenient to introduce two equivalent ways to describe an arbitrary allocation of each genetic sequence to one of the K groups. Suppose there are n_s genetic sequences denoted by $G_1, ..., G_{n_s}$, then we write $\eta = (\eta_1, ..., \eta_{n_s})$ to denote the genetic group. Consider the equivalence

$$\eta_i = j \iff G_i \in \mathcal{G}_j, \qquad j = 1, ..., K,$$

where $\mathcal{G}_j$ denotes the sub genetic network associated with group j for all $i = 1, ..., n_s$ and

$$\mathcal{G} = \cup_{j=1}^{K}\mathcal{G}_j, \quad \text{and} \quad \mathcal{G}_i \cap \mathcal{G}_j = \emptyset, \quad i \neq j,$$

that is to say the genetic groups induce a partition in the set of nodes of genetic sequences.

The process of determining the genetic groups is largely an arbitrary choice and is likely to be a modelling specific decision. In this section we introduce some methods to calculate the genetic groups η from the matrix of pairwise distances D. The aim here is to partition the sequences such that each partition contains similar sequences which can then be used to construct the sub genetic networks and hence the genetic network.

2.7.3.1 Group by threshold

Here we choose an arbitrary threshold, ζ say, and group similar genetic sequences that have pairwise distances less than the given threshold. The idea is that genetic distances with large distances are clearly unrelated, therefore intuitively it makes sense to group distances that are 'similar enough'. For any graph G consisting of k nodes (or vertices) denoted by v_i for $i = 1, ..., k$, we can construct a $k \times k$ adjacency matrix A such that A_{ij} is the number of edges joining nodes v_i and v_j (Bondy, 1976). First define the adjacency matrix A such that

$$A_{ij} = \begin{cases} 1 & D_{ij} < \zeta, \\ 0 & \text{otherwise.} \end{cases}$$

The adjacency matrix induces a partition of the nodes in the complete genetic network, with each partition assigned an arbitrary label.

It is important to note that the graph represented by the adjacency matrix in the above formula need not be strongly connected, that is there may be some sequences within the same group that have a distance greater than the threshold ζ. For example, consider three sequences labelled A, B and C with the following pairwise distances

$$\begin{aligned} d_{AB} &= 2 \\ d_{BC} &= 2 \\ d_{AC} &= 4, \end{aligned}$$

and suppose that the threshold is given as $\zeta = 3$. In this example there is an edge between nodes A,B and B, C but no edge connecting nodes A, C, however here A and C are considered to be in the same group due to the mutual connection with B.

Theorem 2.7.5. *Let G be the graph induced by the adjacency matrix A. Then G will be a disjoint union of complete graphs with an edge between nodes if the distance is smaller than the threshold value ζ. Furthermore, any graph induced by A is invariant to a permutation of labels.*

Proof. Suppose G can be decomposed into a disjoint union of sub graphs such that $G = G_1 \cup ... \cup G_K$ where K is the number of partitions induced by A and $G_n = (V_n, E_n)$ for $n = 1, ..., K$. Then for any fixed n we have $G_n = (V_n, E_n)$, for all nodes $v \in V_n$ there exists an edge $(i, j) \in E_n$ such that $D_{ij} < \zeta$ by definition of A_{ij}, hence G_n is complete and G is a disjoint union of complete graphs by construction.

Furthermore, suppose there are K partitions or disjoint complete graphs in the network induced by A, labelled from $1, ..., K$. The choice of label is arbitrary since any other permutation of labels will preserve adjacency and induce a graph G' which is isomorphic to G. □

The benefit of this method to cluster genetic sequences is that it is straightforward to implement and interpret. In practice one should choose a value ζ that is consistent with biological theory and prior beliefs. When considering genetic sequences sampled in some time frame, it may be suitable to choose ζ with respect to the expected evolutionary changes in a larger time frame. In Young et al. (2012) the authors estimate the evolution of a bacterial population of *S. aureus* to be a rate of 2.72 mutations per megabase per year, hence choosing $\zeta > 2.72\gamma$ would separate sequences with more than γ years of expected mutations.

Once we have calculated the adjacency matrix A from n_s sequences and observe $K \geq 1$ partitions according to the rule, we are able to assign the subtype groups as follows. Let η denote the vector of genetic groups such that $\eta_i = j$ if and only if sequence i belongs in sub genetic network j for $i = 1, ...n_s$ and $j = 1, ..., K$.

2.7.3.2 The k-means clustering and partitioning around medoids algorithms

The k-means clustering method seeks to partition objects into groups such that the objects within the group are sufficiently close, and the discordance between the data and a given partition $P(M, K)$ of M objects and K clusters is measured by an error $e\left(P(M, K)\right)$ (Hartigan, 1975). The main idea is to search for an optimal partitioning where each observation belongs to a cluster with the nearest mean by minimising the within-cluster variances (given by the error function e).

A caveat here is that the algorithm typically operates on a data matrix by minimizing squared Euclidean distances and as such it is not appropriate to apply the method for non-Euclidean distances. In the context of pairwise sequence alignment, the number of SNPs is referred to as the Hamming distance between the two sequences (Mohammadi-Kambs et al., 2017). Alternatively we look towards other techniques that partition objects irrespective to the distance metric, such as partitioning around medoids.

The idea of partitioning around medoids (PAM) was first described in Kaufmann and Rousseeuw (1987) as a more robust method than k-means clustering which seeks to find k 'representative' objects called medoids. After finding a set of k medoids, the k clusters are constructed by assigning each object of the data set to the nearest medoid (Kaufman and Rousseeuw, 2005). Since PAM performs clustering with respect to any distance metric, it allows a flexible definition for what it means for two elements to be 'close' (der Laan et al., 2003). A comprehensive comparison of k-means and PAM can be found in Mondal and Choudhury (2013) where the authors compare physical characteristics of different varieties of mango in India.

As with k-means clustering, the user must specify the number of clusters pre-analysis, which is usually situation dependent and purely a modelling choice. Choosing the ‘correct’ number of clusters is non trivial and will largely depend on subject specific knowledge and exploratory data analysis.

2.7.3.3 Spa typing

As an alternative to statistical based methodology for clustering using a pairwise distance matrix (or dissimilarity matrix in the literature), one can use biological techniques during the sequence analysis such as spa typing. The idea is to specifically look at the gene for *S. aureus* protein A (spa) due to the surface containing polymorphic regions (Frénay et al., 1994), which is where there are more than one variant genes that may be present.

Due to the high variability of gene expression in protein A it is then possible to classify different organisms by the number of repeating sequences and hence provide suitable discrimination for outbreak investigation (Shopsin et al., 1999).

Typically whole genome sequencing (WGS) provides much greater resolution than spa typing, however the data set discussed in Section 3.2 contains genetic spa types and hence can be considered and compared with other methods of genetic sequence clustering. The benefit here is that there is no need to make arbitrary choices as to what the clusters should be, either through choosing a cut off threshold or choosing the ideal number of clusters.

2.8 Construction of genetic networks

In this section we describe how to integrate genetic data with traditional epidemiological data (such as case incidence) in an outbreak scenario. Up to now, the probability distributions derived in Equations (2.4) and (2.6) are conditional on an underlying genetic network $\mathcal{G}$ which describes the evolution of the pathogen.

Recall from Section 2.6 that the genetic network describes the evolutionary relationship between observed sequences. The aim of this section is to construct a genetic network that is consistent with the epidemiological data, that is the transmission network.

Here we assume that individuals may be infected as a result of coinfection, that is individuals can be infected by multiple pathogen subtypes concurrently. Furthermore we do not make any attempt to model the introduction of these pathogen species and assume that every individual may possess any strain.

Let η denote the vector of genetic subtypes, such that $\eta_i = j$ if and only if genetic sequence i belongs to subtype j. Furthermore, we require the following assumptions.

1. Upon colonisation, the genetic information of the pathogen is transmitted from the source of colonisation to the recipient. The transmitted pathogen then evolves independently from the pathogen in the source of colonisation.

2. Each individual may have a genetic sequence for each subtype at the time of colonisation, which may be either observed or unobserved.

Assumption 1 provides us with a framework to perform any meaningful inference, since the genetic evolution models considered require the same pathogen to be observed at distinct points in time. In general, this is an oversimplification of the biological processes where we do not attempt to model any transmission bottleneck. Assumption 2 is a requirement to capture the shared stochastic evolution of the pathogen.

It should be noted that an individual's genetic sequence (and corresponding genetic distances) at the time of colonisation may not be directly observed and as such it is natural to impute these missing quantities in a Bayesian MCMC framework.

The main idea behind constructing genetic networks is to determine the genetic source for each genetic sequence. We use the term 'genetic source' rather than 'source' to avoid confusion with the source of infection. A genetic source for sequence i is the genetic sequence that i is assumed to have evolved from, also known as the ancestral sequence. We also define the term 'infectee' to be the recipient of an infection.

To determine the genetic source for a given sequence, at an individual level we look backwards in time to determine if there is a previously sampled genetic sequence. More precisely, we use the graph representation of the temporal transmission network described in Section 2.5 to trace the evolution of the pathogen backwards in time.

A genetic source sequence will either be a previously sampled sequence from the same individual, or a sequence at the time of infection from an infectee. If there are no genetic source sequences in the host, we then look at the source of infection and repeat the process.

The construction of the genetic network is unique and explicitly depends on the transmission tree, and therefore unobserved quantities such as the source and time of infection. If there are multiple sequences from the same individual at the same time, then we order the sequences by index. The lowest index will require calculating the genetic source as above, and all other indexes will assign the previous index to be the genetic source.

The genetic network $\mathcal{G}$ is constructed as follows. For each genetic sequence we wish to determine the source of the sequence, where the genetic source is defined as the ancestor sequence. In other words, a sequence j may have a source sequence i if and only if sequence i is assumed to have evolved to sequence j after some time. The evolutionary time between two sequences in the context of an outbreak is the difference in sampling times.

Suppose there are k genetic sequences assumed to belong to one of M genetic subtypes, we now briefly outline the method to determine the genetic source of a particular sequence, denoted by $i \in \{1, ..., k\}$.

For any sequence i we have the individual which the sequence has been sampled from denoted by ν_i at some time τ_i with genetic subtype η_i. Let $\mathcal{S}_i = \{(j, \tau_j) : \nu_j = \nu_i, \eta_j = \eta_i, i \neq j\}$ denote the set of sequences sampled from the same individual and genetic subtype, not including sequence i, and the corresponding time at which the sequence is sampled.

Then we order the set of sequences $\mathcal{S}_i$ by the corresponding sampling times τ and choose the sequence that has been sampled most recently. In the event of a tie, i.e. multiple sequences are sampled at the same time, then the sequences are ordered by index.

If there are no sequences in that particular individual with a time less than the sampling time of the current sequence, then we look at the source of colonisation s_{ν_i}. The idea is then to find the most recent sequence in the individual s_{ν_i} that belongs to the same genetic subtype. If there are no sequences to choose in this individual, then we look at their source of colonisation and so on.

This process is repeated until either an observed sequence is found or there is no more individuals to search through. If a genetic source is not found, then that particular sequence is considered to be an *imported sequence*, whereby the individual is considered to have imported that particular sequence from the community.

2.9 Modelling imported genetic sequences

In this section we wish to consider the possibility of multiple introductions of the pathogen, which are essential for epidemic models with a background transmission or importation (e.g. Deardon et al. (2010) and Worby et al. (2016)). The mutation models described thus far are only applicable when modelling the evolution of a single nucleotide or sequence. As a consequence, in order to model multiple introductions of the pathogen we introduce another probabilistic model which describes the likelihood of pairwise distances from multiple sequences that are unrelated.

Consider two distinct nucleotides observed at the same point in time t_C denoted by N_1,$N_2 \in \{A, T, G, C\}$. Typically from a phylogenetic viewpoint it is reasonable to assume that the two nucleotides have a common ancestor N_P at some time in the past t_P (Figure 2.7). Let $X = (N_1, N_2)$ denote the data which are the observed nucleotides and θ the model parameters for the particular mutation model under consideration. Then, conditional on the model parameters, the probability of the data is given by

$$\Pr(X|\theta) = \sum_k \pi_k p_{kN_1}(t_C - t_P) p_{kN_2}(t_C - t_P), \tag{2.7}$$

where $p_{ij}(t)$ is the probability of observing nucleotides i and j which are separated by t time units and π_k is the equilibrium frequency of nucleotide k (Felsenstein, 1981). The summation in Equation (2.7) is due to the fact that the ancestor nucleotide N_P is unobserved and as a result we sum over all possible choices. In the above formulation, the vector θ contains not only the parameters governing the rate of mutation but also the unknown time of evolutionary divergence t_P, which must also be estimated.

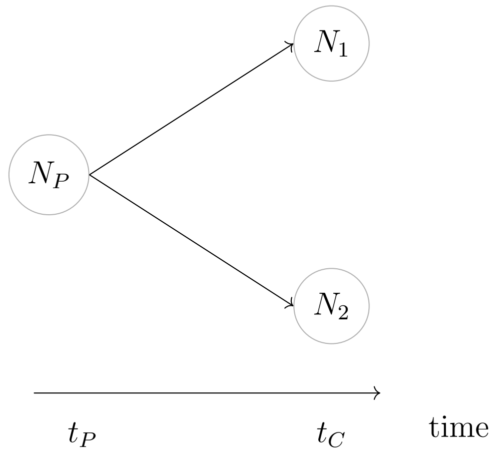


Figure 2.7: Phylogenetic approach for modelling distinct genetic sequences where nucleotides N_1,N_2 are observed at time t_C. The two nucleotides are assumed to have a common ancestor N_P at a previous point in time t_P. The nucleotides N_1,N_2 are then assumed to have evolved independently from a previous ancestral state N_P in $t_C - t_P$ time units. The ancestral nucleotide N_P and the time of evolutionary divergence t_P are both unobserved.

Assuming the JC69 mutation model, the above example is simple enough in the sense that there are only two unknown parameters, the mutation rate λ and the time of divergence t_P, however the problem quickly becomes cumbersome if there are more genetic sequences under consideration.

This type of problem is an active area of research in the Bayesian phylogenetics community and reconstructing phylogenetic networks is a non-trivial task and beyond the scope for this thesis. Instead we propose a simpler model to describe the introduction of multiple genetic sequences that are assumed to be unrelated.

Similar to the approach taken in Cassidy (2019), we wish to consider a generative process that explicitly models (i) the pairwise distances between directly connected sequences from the mutation model, (ii) the pairwise distances between imported sequences of the same subtype and (iii) the pairwise distances between imported sequences of a different subtype.

Suppose there are M distinct genetic subtypes defined by the vector η where each subtype contains m_i imported sequences for $i = 1, ..., M$. Let $\mathcal{I}^i = (\mathcal{I}^i_1, ..., \mathcal{I}^i_{m_i})$ denote the vector of imported sequence labels for genetic subtype i. For (i), Equation (2.4) is used to evaluate the probability of distances between directly connected nodes corresponding to the edges E in the genetic network $\mathcal{G}$ assuming the JC69 mutation model.

Similarly, for (ii) we assume that the pairwise distances for imported sequences of the same genetic subtype follow a Poisson distribution with mean μ_i for $i = 1, ..., K$ which represents the 'within-subtype importation distance' for genetic subtype i. For simplicity we assume that each genetic subtype has an identical parameter μ, i.e. $\mu_1, ..., \mu_K = \mu$.

Let $\mathcal{I}$ denote the complete set of imported sequences. In order to model the contribution of imported sequences of a different genetic subtype, we assume that the pairwise distances follow a Poisson distribution with mean ξ, then the probability to evaluate the pairwise distance matrix is given by

$$\begin{aligned}\Pr(D,\eta \mid \lambda,\mu,\xi) \propto & \prod_{(i,j)\in E(\mathcal{G})} \left(1-q\left(t_{ij}\right)\right)^{N-D_{ij}} \left(\frac{1}{3}q\left(t_{ij}\right)\right)^{D_{ij}} \\ & \times \prod_{i\in\mathcal{I}} \prod_{\substack{j\in\mathcal{I}\\ i\neq j}} \Bigg(\mathbb{1}_{\{\eta_j=\eta_k\}} \Pr\left[\mathrm{Po}(\mu) = D_{\mathcal{I}^i_j\mathcal{I}^i_k}\right] \\ & \qquad + \mathbb{1}_{\{\eta_j\neq\eta_k\}} \Pr\left[\mathrm{Po}(\xi) = D_{\mathcal{I}^i_j\mathcal{I}^i_k}\right] \Bigg) \end{aligned} \tag{2.8}$$

where λ is the mutation rate, $t_{ij} = t_j - t_i$ is the difference in sampling times of sequences i and j and $q\left(t_{ij}\right)$ is the probability of mutation in t_{ij} time units. The first line is the contribution of directly connected distances under the mutation model. The second line is the contribution of pairwise distances between imported sequences in the same genetic subtype where $\eta_j = \eta_k$. Finally, the last line is the contribution of pairwise distances between imported sequences in a different genetic subtype where $\eta_j \neq \eta_k$.

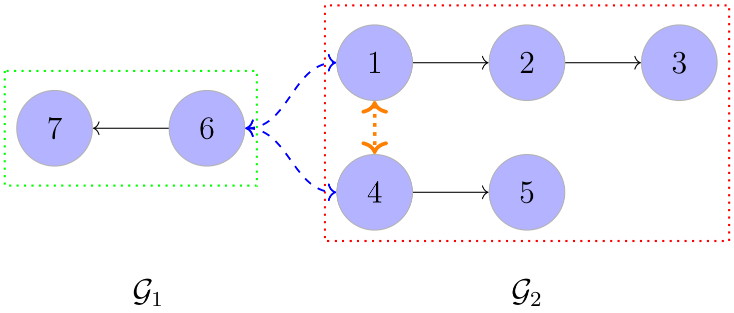


Figure 2.8: An example genetic network $\mathcal{G}$ consisting of 7 observed genetic sequences which is composed of $K = 2$ distinct genetic subtypes, where nodes 6 and 7 are in the sub genetic network $\mathcal{G}_1$ (dashed green line) with subtype 1 and the remaining nodes are in the sub genetic network $\mathcal{G}_2$ (dashed red line) with subtype 2. The solid black lines indicate directly connected nodes that are assumed to have evolved under the mutation model. The dotted orange line indicates pairs of sequences that are imported but within the same genetic subtype, and the dashed blue lines indicates pairs of sequences that are imported but in different subtypes.

We now make some remarks about the ‘between-subtype importation distance’ ξ. It is perhaps more suitable to have $K(K-1)/2$ independent parameters denoted by ξ_{jk} to model the average distance between imported sequences in distinct groups j and k. This is however not necessary for two reasons. Firstly, we are primarily interested in resolving transmission pathways by integrating genetic data, information of distances between subtypes offers little use as pairwise distances are likely to be large and therefore may be not indicative of transmission. Furthermore, we determine the genetic subtypes η pre-analysis which are fixed and as a consequence we do not attempt to infer ξ and therefore we may remove the last line in Equation (2.8). An example of the genetic model can be seen in Figure 2.8.

It is often the case that imported individuals do not have an observed sequence at the time of infection, hence these quantities must be imputed for the model described above. Inference in a Bayesian framework is able to handle missing data through data augmentation MCMC, details for imputing genetic distances are discussed in Section 3.5.3.

2.10 Discussion

In this chapter the primary goal was to develop a model for genomic data that is applicable in outbreak settings. Specifically, the aim was to work with pairwise genetic distances rather than the raw sequence data, under the assumption of a microevolution model. More precisely, we have derived the distribution of pairwise genetic distances implied under the assumption of a microevolution model.

We have considered two nucleotide substitution models, in particular the JC69 and K80 mutation models, and derived the joint distribution of pairwise genetic distances for sequences of length one.

The general strategy for deriving these distributions is to consider the set of possible nucleotides that give rise to the pairwise distances under consideration. Once we are able to assign nucleotide bases that are consistent with the genetic distances, then we are able to evaluate mutation probabilities conditional on an underlying genetic network that describes the evolution of the sequences.

We have then extended these ideas for more general sequences of arbitrary length N and derived an expression for this joint probability mass function. The exact distribution of the pairwise distances requires detailed information at the nucleotide level, specifically the number of unique bases at each site.

If the data are the matrix of pairwise genetic distances (i.e. we do not have the original raw sequence data), then we may only evaluate the probability up to a proportionality. From a practical viewpoint this does not pose an issue as the complete pairwise genetic distances are a sufficient statistic for the mutation rate parameters. Furthermore, in a Bayesian setting the terms that involve the data alone vanish in the posterior.

It is important to note that the derived distributions are conditional on an arbitrary genetic network. In Section 2.8 we discuss how to construct a genetic network that is consistent with the epidemiological data, i.e. the transmission network. The true transmission network is unobserved and a common approach among practitioners is to augment the parameter space with the unobserved data and sample these in a data augmentation MCMC framework.

When sampling from the space of transmission trees and proposing updates to the augmented data, there will now be a component in the likelihood that is the genetic data conditional on the transmission tree. In principle, proposed updates (e.g. Metropolis-Hastings steps) to the transmission tree are now informed by the genetic data.

The main strength of our method lies with the fact that we are working with pairwise distance matrices alone. This is particularly convenient as we do not need to store or impute the full genetic sequences, and the pairwise distance matrix acts as a summary statistic of the genetic data.

CHAPTER 3

Analysis of the Brighton data set

3.1 Introduction

In this chapter we focus on the transmission of nosocomial infections, specifically the spread of Methicillin-resistant *Staphylococcus aureus* (MRSA) within a hospital ward. The emergence of antibiotic resistant pathogens poses a serious public health risk to healthcare agencies worldwide. The burden of healthcare associated infections (HCAIs) is estimated to have cost the United Kingdom £2.1 billion in 2016/2017, arising from $653,000$ infections, and of that total $22,800$ deaths (Guest et al., 2020).

In the context of nosocomial infections, specifically MRSA, patients are referred to as 'colonised' rather than 'infected'. In the literature many authors use 'colonisation' and 'infection' interchangeably (Dani, 2014), however in fact colonisation is defined in Longe (2015) as "when a microorganism is found on or in a person without causing disease". Throughout we shall maintain the convention that colonisation is the presence of detectable levels of the pathogen, with no reference or indication to whether the patient has the disease.

Throughout we shall work in discrete time, a modelling assumption which is a natural choice when considering the type of data typically available. Hospital data and swab tests for the pathogen are often recorded by healthcare workers or administrators daily and the exact times are not always known.

We analyse a data set of an outbreak of *S. aureus* in the Royal Sussex County Hospital, an acute teaching hospital in Brighton during 2011-2012. In this data set patients, healthcare workers and the environment were systemically sampled for the presence of the pathogen. Isolates were cultured from positive swab results and their whole genome was sequenced. A comprehensive description of the study can be found in Price et al. (2017).

The aim of this chapter is to build upon the transmission model developed in Worby et al. (2016) and fit the genetic model developed in Chapter 2 to infer the mechanisms governing transmission for the Brighton data set. More precisely, we wish to extend the transmission model to incorporate healthcare worker data at an individual level.

This chapter is organised as follows. In Section 3.2 we provide an exploratory analysis of the Brighton data set. In Section 3.3 we provide a review of the literature surrounding the impact of healthcare workers on transmission. Then we introduce a model to incorporate healthcare worker transmission which builds upon the transmission model in Worby et al. (2016).

Next, in Section 3.4 we present the model likelihood and details for Bayesian inference, then we develop a Markov chain Monte Carlo algorithm to draw samples from the posterior distribution in Section 3.5.

In Section 3.6 we present a simulation study which details the simulation method, transmission network accuracy, parameter estimation, sensitivity analysis and MCMC diagnostics. In Section 3.7 we analyse the Brighton data set using the extended model developed throughout this chapter and perform model assessment. Finally, in Section 3.8 we provide a discussion of our analysis and highlight the limitations of our approach.

3.2 Brighton data set

In this section we present an overview of the Brighton data set, highlighting features of the data that motivated the need to incorporate multiple subtypes in the genetic model developed in Chapter 2. The data were collected as part of a study in the Brighton and Sussex University Hospital (BSUH), where the study was conducted between 31st October 2011 and 23rd December 2012, a total span for 14 months. During the study, data were collected from two locations, an adult intensive care unit (ICU) and high-dependency unit (HDU).

The adult ICU consisted of one 5-bedded area, one 4-bedded area, 3 double side rooms, and 1 single side room. The adult HDU was located two floors below the ICU and consisted of a 12-bedded unit with two 4-bedded areas, one 2-bedded areas, and 2 single side rooms (Price, 2014).

During the study period, patients were screened on admission to the ICU/HDU, weekly thereafter and on the day of discharge. Screens consisted of a single nasal swab, however in some cases swabs from the perineum, groin, sputum, urine and wounds may have been collected. All patients received antibacterial skin washes (Chlorhexidine) while on the ICU and patients identified as carrying MRSA received adjunctive antibacterial nasal ointment mupirocin (Price, 2014).

The study also consisted of screening data from healthcare workers on the wards. A total of 208 nurses were employed to work in the adult high-dependency settings at the BSUH. Up to 21 nursing staff worked on the ICU/HDU per shift depending on the number of patients, with shifts typically 12.5 hours long and allocation to either the ICU or HDU determined on a shift-by-shift basis.

Additional clinical staff were recruited in the final eight months of the study (16th April 2012 - 23rd December 2012), resulting in a total of 42 doctors and 9 physiotherapists who worked on the ICU/HDU during the study. Healthcare workers were screened weekly by nasal swabs, with a follow-up throat swab taken every six months.

Furthermore, environmental screening was undertaken on a monthly basis during the study period, with air sampling performed at 10 locations on ICU and HDU. Air sampling is a process which determines what airborne organisms are present in an environment. Throughout the study period 26 bed spaces were serially screened, where each screen consisted of 5 swabs representing areas of frequent staff contact (monitor button, computer keyboard, disposable curtains) and other less frequently touched areas such as the floor behind the bed and underside of the bed. Swabs were also taken of the communal blood gas machine located in the central ICU utility room.

All available screening swabs were included in the study and cultured for methicillin-resistant *S. aureus* (MRSA) and methicillin-susceptible *S. aureus* (MSSA). A timeline of swab results can be found in Figure 3.2.

The test sensitivity of patient screening for MRSA colonisation is partially dependent on the body sites sampled (Currie et al., 2008). For example, Senn et al. (2012) estimate the sensitivity for the nose, groin, throat as 48%, 63% and 61% respectively and as a consequence we wish to consider a model using only nasal swab data.

3.2.1 The data

The complete data set contains 1919 patient admissions which consists of 1760 unique individuals where 124 patients were admitted more than once. From the 1760 unique individuals, there are 1051 males (59.7%) and 709 females (40.3%) with an average age of 62.2 years (median 65.9, IQR [49.9,76.7]). The average length of stay during the ward is 5.2 days, with a median of 3 days (Figure 3.1).

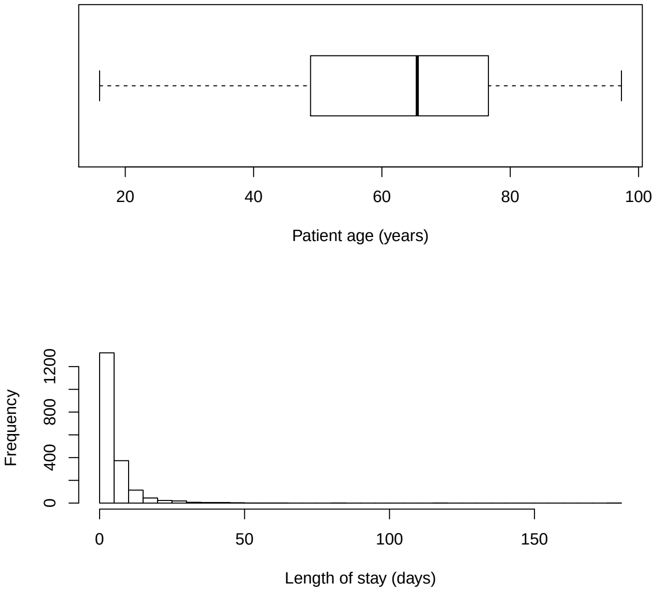


Figure 3.1: A summary of patients admitted to the ward in the Brighton data set. Top: Distribution of age among the admitted patients. Bottom: Length of stay for patients on the ward.

For the patient screening data there are 3851 recorded tests of which 759 returned a positive result indicating that the individual is colonised with *S. aureus* at the time of sampling. From the 759 positive test results, 552 are from the nose (72.7%), 91 from the perineum (12.0%), 30 from the sputum (3.96%) and the remaining 86 are from other sources such as urine, groin, wounds, etc. A timeline of screening results can be seen in Figure 3.2.

Motivated by the findings in Senn et al. (2012) where the authors found different test sensitivity across anatomical sites, we shall consider nasal carriage only and discard all test results from other anatomical sites. A timeline of the population of individuals present in the ward can be seen in Figure 3.3, with the colonised population overlaid where an individual is assumed to be colonised on day t if they have a positive test result on or before day t and we assume individuals remain colonised until discharge. From the timeline we can see that there is consistent carriage of the pathogen in

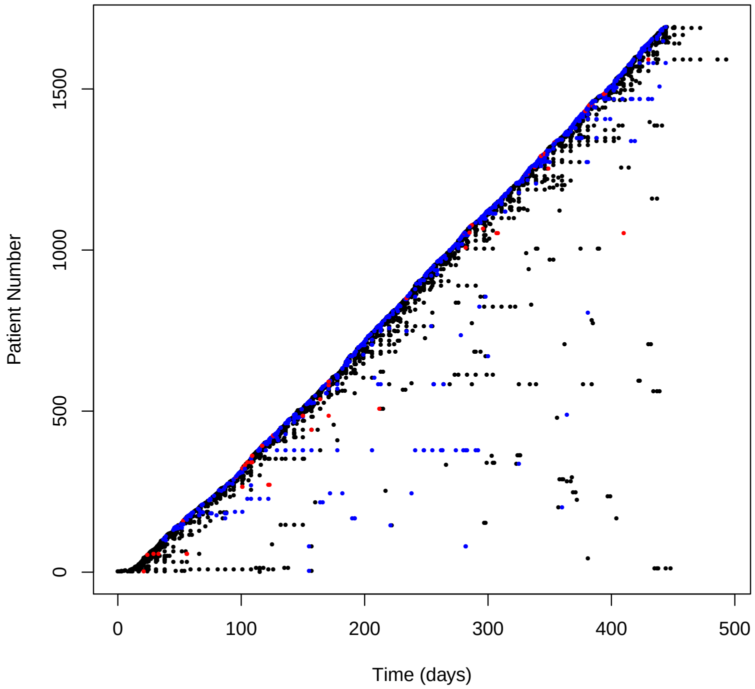


Figure 3.2: A timeline of screening results for patients in the ICU/HDU. Unique patient IDs are allocated based on ordered admission times (y-axis) with time in days on the x-axis. Points in black indicate negative results, blue indicate positive for MSSA and red is positive for MRSA.

the ward over the study period, highlighted by the fact that there is nearly always a colonised patient in the ward at any point in time.

Healthcare worker data Of the 198 hospital staff who consented and were involved in the study, 113 (57%) ever tested positive for colonisation of *S. aureus* during the study period. Out of the 113 healthcare workers to ever test positive, 24 (21.2%) only had one positive test with the remaining healthcare workers testing positive more than once.

From the distribution of the number of positive tests in Figure 3.4 it can be seen

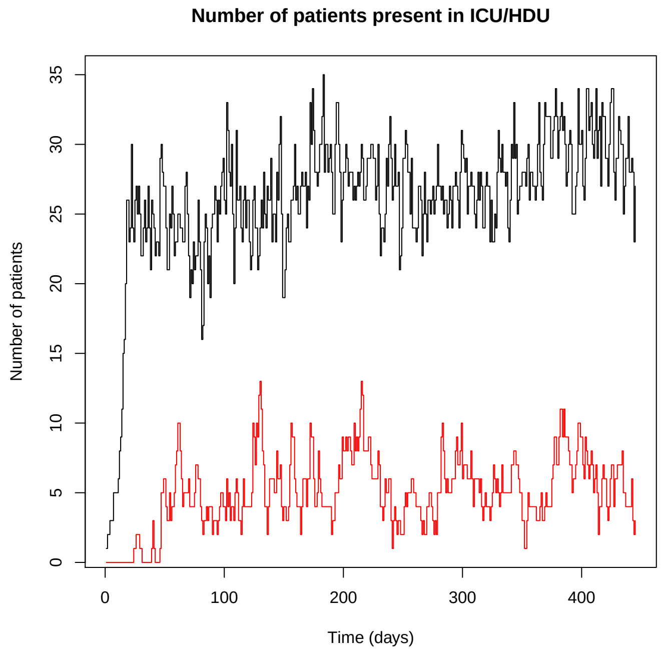


Figure 3.3: A timeline of the total number of patients in the ward (black) and the number of individuals who are assumed to be colonised (red). Individuals are assumed to be colonised on day t if they have a positive test result on or before day t.

that the majority of healthcare workers had more than one positive test result, with 30 healthcare workers having 10 or more positive results, suggesting fairly consist carriage of *S. aureus*.

By considering the time series data for positive tests (Figure 3.5), we assume that every individual with more than one positive result is colonised from the earliest result to the latest result. Furthermore, we assume that each healthcare worker remains colonised for α days following their last positive test result. This assumption only applies to healthcare workers since we do not have precise information of when the healthcare workers are present on the ward.

Genomic data In addition to the swab data, 1976 organisms were cultured and sequenced of which 867 are sampled from patients (43.9%), 929 from healthcare

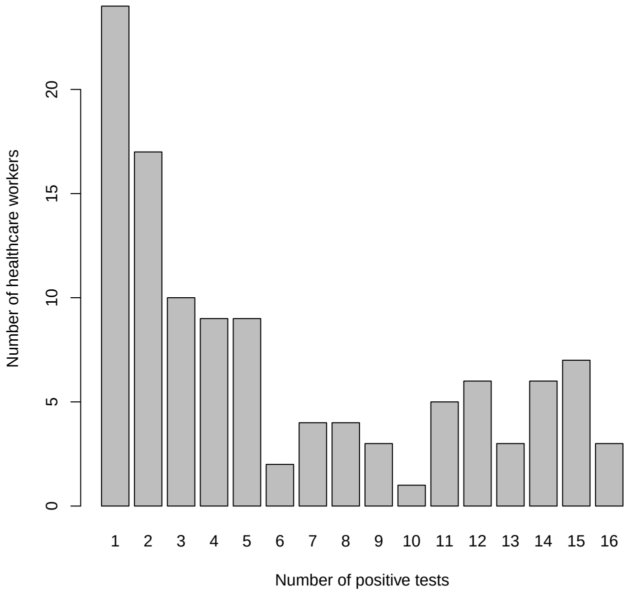


Figure 3.4: The number of positive tests on the x-axis with frequency of healthcare workers on the y-axis.

workers (47.0%) and the remaining 180 from the environment or unknown (9.1%). An overview of the genetic sequences collected in the study can be seen in Figure 3.6.

Since primary interest lies with integrating healthcare worker data and patient data, we focus on those sequences only and discard the rest. Out of the 1796 sequences under consideration that are sampled from either patients or healthcare workers, 1308 are sampled from the nose (72.8%) which will be the final sequences used for the rest of the analysis.

By excluding environmental and non-nasal swabs we are effectively discarding 668 of the total 1976 (33.8%) genetic sequences. While this constitutes a large proportion of the genetic data, the hope is that there is enough information left within the remaining sequences to answer the primary research question, which addresses if we are able to better determine transmission pathways with the addition of genetic data.

If we are interested in measuring the amount of within-host genetic diversity we can

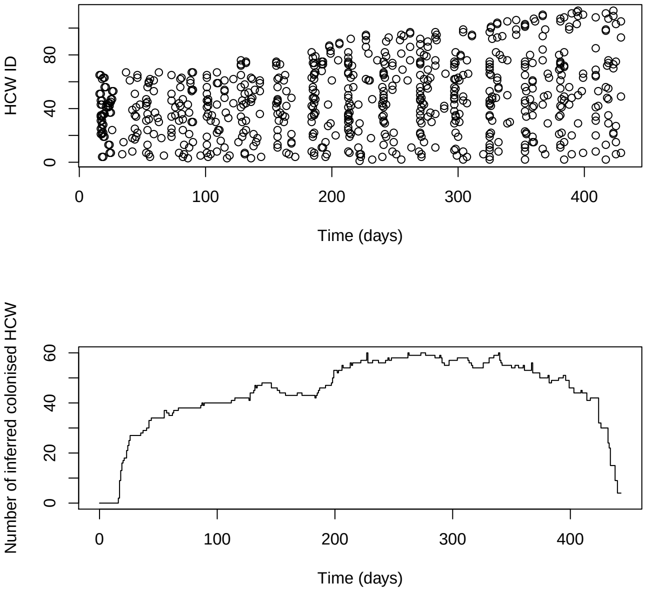


Figure 3.5: Top: Positive test results (black circle) for each healthcare worker. Bottom: Assumed number of colonised healthcare workers on the ward by assuming constant carriage in-between test results and for a fixed period of $\alpha = 14$ days thereafter.

look at every individual who has more than one genetic sequences sampled and look at the pairwise distances between those sequences only. We define the term *average within-host diversity* to be the mean of all pairwise distances from sequences sampled from the same individual, which is calculated as follows.

Suppose there are N genetic sequences sampled from one of k individuals where each individual has v_i genetic sequences for $i = 1, ..., k$. Let $A = \{i : v_i > 1\}$ be the set of individuals with more than one genetic sequence, then for each $j \in A$, let D_j denote the matrix of pairwise distances calculated for the sequences from individual j. There will be exactly $T_j = \frac{v_j(v_j-1)}{2}$ pairwise distances distances labelled $x_1, x_2, ..., x_{T_j}$ and hence the average pairwise distance for individual j is simply the arithmetic mean of these distances calculated by $\frac{1}{T_j}\sum_{i=1}^{T_j} x_i$. From this we can construct a distribution of the mean or average within-host diversity. Similarly, we can also calculate other summary statistics, such as the median.

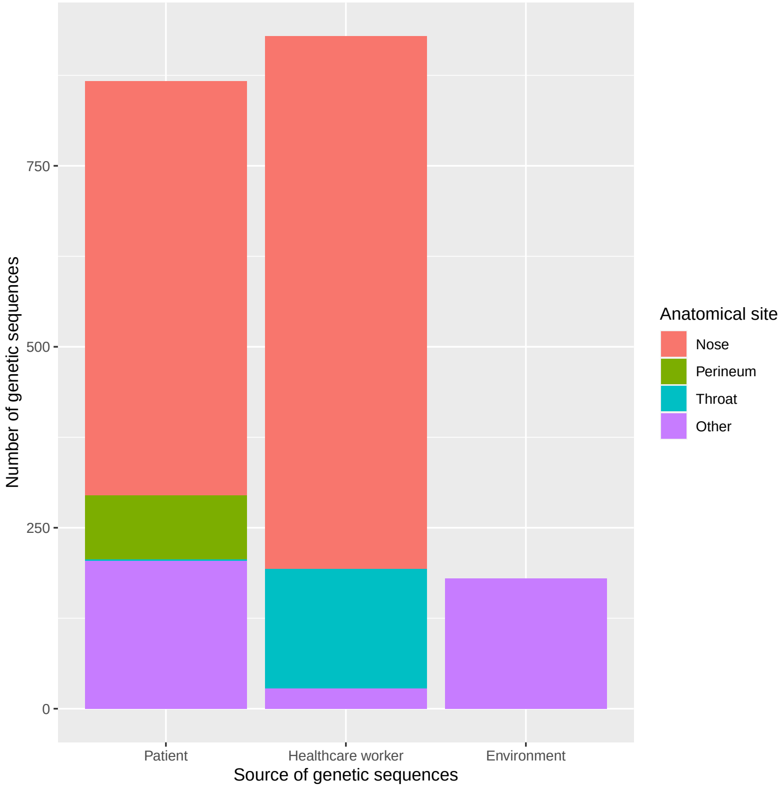


Figure 3.6: An overview of the number of sequences collected during the study from either (i) patients, (ii) healthcare workers or (iii) the environment. The anatomical site has been coloured in to show the relative frequency of each.

While this is a crude method to estimate the amount of within-host genetic diversity in the sample, it does offer some indication of the proportion of diverse genetic sequences. In Young et al. (2012) the authors estimate the number of mutations per megabase per year for *S. aureus* to be 2.72 ([1.64,4.42] 95% credible interval), hence one would expect to see a relatively small number of mutations, that is if sequences are genetically related.

From Figure 3.7 it can be seen that the majority of the average within-host distances fall within a reasonable range (less than four mutations), however it is clear there are a large portion of pairwise distances that are greater than $10,000$ SNPs and hence are almost certainly unrelated.

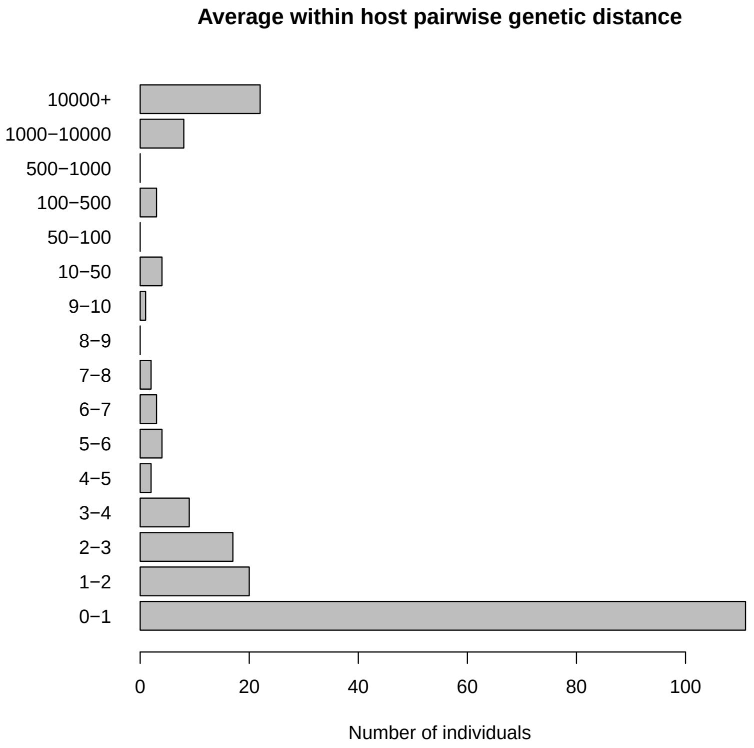


Figure 3.7: Mean genetic distance between pairs of nasal isolates sequenced from individuals with two or more sequences.

From the heat map of the pairwise distance matrix (Figure 3.8) it is somewhat difficult to deduce much information from the genetic data, with the exception of observing that there are extremely diverse genetic sequences with many distances greater than $10,000$ SNPs which corresponds to over 1800 years of evolution. It should be noted that these sequences, albeit extremely diverse, are still variations of the pathogen *S. aureus*.

Although not immediately obvious, upon closer inspection there are in fact small areas of the heat map with colours indicating < 5 SNPs, which ultimately may indicate epidemiological relatedness. We are not able to perform any meaningful analysis with the extremely diverse genetic distances, instead we focus on the similar genetic isolates. In a similar fashion to the work in Price et al. (2017) we define the term *genetic subtype* to refer to 'similar' sequences, details of how to calculate these

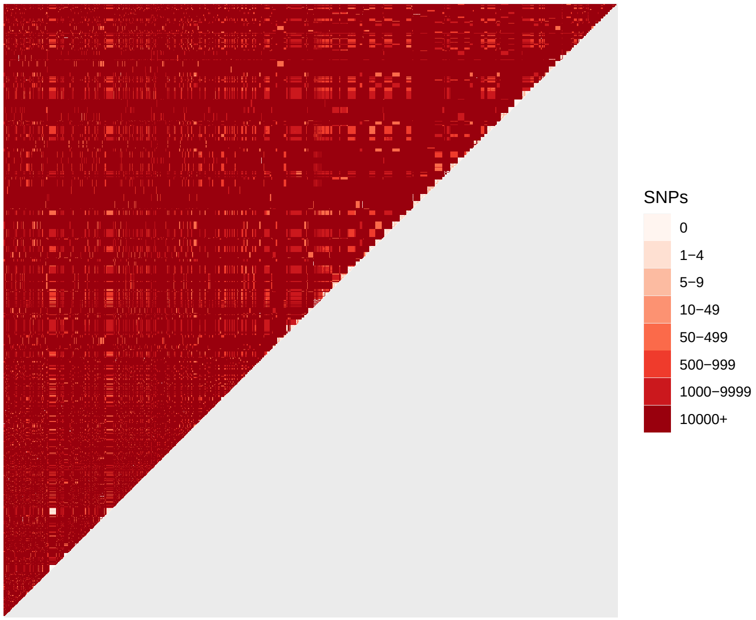


Figure 3.8: Heat map of the pairwise genetic distance matrix from the Brighton data set where samples are sorted by unique patient ID with only the upper triangular entries displayed due to symmetry. A colour (key indicated on the right) in position (i, j) is representative of the distance (or the number of SNPs) between sequences i and j.

genetic subtypes are discussed later in Section 2.7.3. Sequences within the same genetic subtype can be thought of as snapshots of the same organism, therefore the distances may be described by the mutation model.

We define the term *putative transmission pair* to be any pair of individuals that have observed distances with less than 5 SNPs and also spend time on the ward together at the same time. Intuitively if a pair of individuals have small genetic distances and are on the ward at the same time, it is considered likely that one may have colonised the other or both have been colonised by the same individual. We shall focus on patient-patient and healthcare worker-patient pairs since we are not directly modelling acquisition of healthcare workers.

A plot of the putative transmission pairs can be found in Figure 3.9 which highlights 14 potential clusters containing a total of 22 individuals that are assumed to be epidemiologically related.

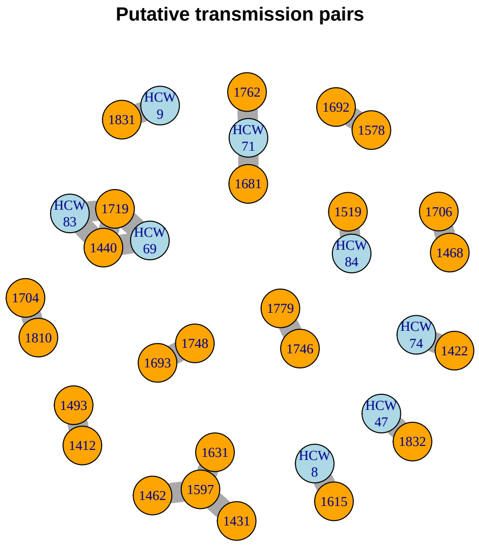


Figure 3.9: Putative transmission pairs for the Brighton data set which is defined as either two patients (orange) or a patient and a healthcare worker (blue) with at least one similar genetic sequence (a genetic distance less than 5), who are also on the ward together at the same time.

3.3 Incorporating healthcare worker data

In this section we develop the model described in Chapter 2 that readily incorporates minimal healthcare worker data. Throughout this section we shall consider healthcare workers as potential carriers of pathogens, with the objective to determine possible transmission pathways that may have been result of healthcare worker to patient contamination.

Consequently we are not going to attempt to model the acquisition of the pathogen from healthcare workers, even though in reality it may be the case that a healthcare worker becomes colonised from a patient or other members of staff.

First we will review the literature and previous attempts to model the impact of healthcare workers on the transmission of nosocomial infections and then present an updated transmission model which includes healthcare worker data at an individual level.

3.3.1 Background

Previous literature has indicated that transient carriage of *S. aureus* on the hands of hospital personnel appears to be the most important mechanism of serial patient-to-patient transmission (Thompson, 1982). The role of healthcare workers in transmission of nosocomial infections is not fully understood and needs to be investigated further, however prevalence of MRSA among healthcare workers has been previously estimated to be 4.6% ($[1.0, 8.2]$% 95% confidence interval) from pooled results of 127 investigations (Hawkins et al., 2011).

A more recent study discussed in Sassmannshausen et al. (2016) looked at estimating the prevalence among HCWs in nine hospitals in Germany. The authors estimated the prevalence of MRSA among HCWs to be 4.6% from 729 participants.

Another more recent study discussed in Sassmannshausen et al. (2016) estimated the overall prevalence of MRSA among HCWs in a non-outbreak situation to be 4.6%

A review and discussion of the literature can be found in Albrich and Harbarth (2008) where the authors identified 27 studies from a total of 106 that reported direct transmission from healthcare workers to patients through molecular and epidemiological evidence, and a further 52 studies that consider transmission likely.

Barnes et al. (2010) investigates transmission of MRSA using agent based modelling and simulation. In the study the authors simulate scenarios such as varying levels of transmissibility, number of colonised healthcare workers, isolation measures and hand-hygiene compliance, to name a few. The study found that frequent transmission occurs when the patient population is well mixed, which usually occurs when patients share the same healthcare worker.

In Goldstein et al. (2017) the authors present a simulation based network modelling approach to investigate the impact of hand hygiene on MRSA transmission within healthcare settings. A stochastic susceptible-colonised compartmental transmission model is used to describe the transmission of the pathogen indirectly via healthcare workers and observational data is used to construct a contact network. The authors conclude that the care network in the newborn intensive care unit described in the study is a major contributor to the transmission of MRSA, which is likely due to the care requirements in an intensive care setting which offer many opportunities for healthcare workers to be intermediaries or vectors for colonisation.

In Nübel et al. (2013) the authors integrate epidemiological and genomic data in order to investigate transmission pathways in a neonatal intensive care unit in Berlin. Genetic isolates were cultured and sequenced from 26 patients, 2 mothers of patients and 2 healthcare workers. Phylogenetic analysis identified two distinct clades which indicate close epidemiological linkage between outbreak isolates.

The authors conducted a matched case-control study to identify risk factors of MRSA transmission using controls such as positive swab results, weight at birth, gender, length of stay etc. Patients who test positive for the pathogen are compared to patients who have tested negative but have similar features (the ‘controls’). Admission times were not included when matching cases and controls to avoid over-matching for possible time-dependent factors affecting the entire ward, such as colonised healthcare workers. By constraining the phylogenetic tree to be consistent with the spatial and temporal epidemiological data, risk factor analysis using logistic regression was performed which found that contact with one specific healthcare worker increased the odds ratio of acquisition by 9.3 (p-value 0.03).

Worby et al. (2014) provide a framework for the usage of a pairwise genetic distance matrix to investigate disease transmission. The authors explicitly consider genetically diverse isolates by developing a model that includes a molecular clock, within-host pathogen population dynamics and a transmission bottleneck size. A Poisson mutation model is used to describe the number of mutations between lineage divergence and observation and a coalescent model is used to describe the number of mutations prior to lineage divergence. Coalescent models assume a constant population size which is violated by assuming a population bottleneck at the time of transmission, hence the authors consider a discrete-time population model as an approximation to the pathogen population size.

The authors apply the model to an outbreak at a special care baby unit in Cambridge where there were 15 newborn infants colonised with MRSA, each with a single genome sampled and sequenced. Furthermore, 20 genetic isolates were collected from a healthcare worker. The results indicate that the colonised healthcare worker played an important role in colonisation in the outbreak, with three patients having a posterior probability of colonisation from the healthcare worker to be $> 99\%$, and another two patients with $> 50\%$ probability. A more detailed description of the study can be found in Harris et al. (2013).

Popovich et al. (2020) present a detailed genomic analysis of patients, environments and healthcare workers in 4 adult intensive care units between September 2015 and February 2016. Swabs were taken from patients at various anatomical sites, environmental room surfaces, and healthcare workers before and after patient encounters. A pairwise distance matrix was constructed and a cutoff of 40 SNPs were used to determine putative transmission clusters. Isolates found to be part of genomic clusters that spanned different patient episodes were reviewed and further evaluated using spatial (location and room of the encounter) and temporal epidemiological data. The authors identified 2 instances of likely transmission between patients and healthcare workers as a result of genetic sequences differing by a maximum of 10 SNPs.

While the role of healthcare workers on transmission of nosocomial infections is not entirely understood, it is clear that further work must be done to better understand the impact on transmission in healthcare settings.

We now present a novel framework that incorporates healthcare worker data at an individual level. In line with the work developed before in Chapter 2, we wish to use the matrix of pairwise distances as summary statistics for the genetic data, which is made up from a combination of both patient and healthcare worker genetic samples.

3.3.2 Transmission model for nosocomial infections

In this section we first present the discrete time stochastic transmission model described in Worby et al. (2016) for patient data only before extending these ideas to incorporate healthcare workers at an individual level.

Suppose we observe n patients labelled $\{1, ..., n\}$ in a hospital ward over a period of L days where patient j has a time of admission t_j^a, a time of discharge, t_j^d, and a time of colonisation, t_j^c. If an individual j avoids colonisation throughout their entire stay on the ward, we set $t_j^c = \infty$.

As discussed in Section 3.1, time units are resolved at the level of days. At any time (or day) t, each patient is categorized as being either (1) susceptible, where they are able to contract the pathogen from other individuals or (2) colonised, where they have tested positive for the presence of the pathogen.

We assume that admission and discharge times are fixed and do not contribute ot the stochastic model - these times are directly provided in the data and are considered reliable. We further assume that once an individual is colonised, they will remain in that state until discharge. This seems reasonable since previous studies have estimated that the median duration of colonisation is 8.5 months (Scanvic et al., 2001), and the average length of stay for intensive care units has been estimated as 10.2 days (Toptas et al., 2018).

We assume that there is no background transmission, i.e. the only way to acquire the pathogen is through contact with an already colonised individual. It is important to note in our scenario, given the nature of intensive care units, contacts are invariably indirect and are usually through healthcare workers or hospital staff. We further assume that new introductions of the pathogen to the ward are imported cases, which are patients who are colonised on admission. We explicitly model this by assuming that each individual has a probability p of being colonised on admission, which is independent of every other patient. Let $\phi = (\phi_1, ..., \phi_n)$ be the vector of importations such that if a patient j is positive on admission, we set $\phi_j = 1$ and $t_j^c = t_j^a$, otherwise $\phi_j = 0$. We do not explicitly model patient readmissions, and consider any readmission to be a new patient.

We assume that patients on the ward mix homogeneously and that colonisation occurs upon contact with an already colonised individual. We note that patients do not 'mix' with other patients in the usual sense, but rather each colonised patient is equally likely to indirectly colonise each susceptible patient.

Suppose there are $C(t)$ colonised patients on day t, then the probability for a susceptible patient j to avoid colonisation on day t is given by $e^{-\beta C(t)}$, and it follows that

$$\Pr(\text{patient } j \text{ is colonised on day } t) = 1 - \exp(-\beta C(t)).$$

Mathematically this is motivated by a homogeneous Poisson processes where individuals mix at a constant rate β. It is important to note that the susceptible-colonised pairs of contact are formulated as a continuous-time model, which needs to be translated in a discrete-time setting. To construct a discrete-time model we assume that once an individual has been colonised, they are then able to colonise other individuals in the following day. This ensures that the total pressure from colonised individuals remains constant throughout each day. An illustration of the compartmental model can be seen in Figure 3.10.

If an individual becomes colonised at time t_j^c, it is assumed that they are able to

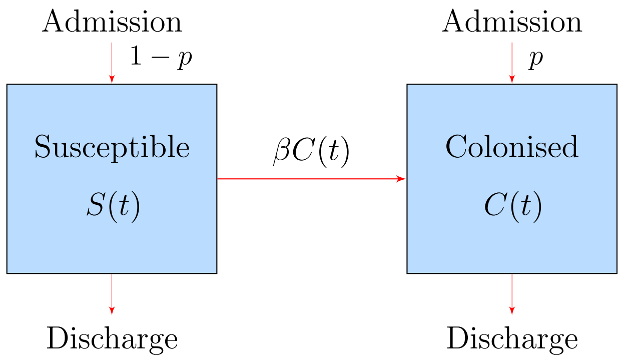


Figure 3.10: A stochastic compartmental model where individuals are either susceptible or colonised. Individuals may be admitted in a susceptible state with probability $1-p$, and admitted with a colonised state with probability p. Individuals transition from susceptible to colonised at a rate of $\beta C(t)$ and eventually patients are discharged.

colonise others from time $t_j^c + 1$ to t_j^d. Furthermore, each patient j may be tested regularly for presence of the pathogen during their stay. Let v_j be the number of tests for patient j, at times $t_j^t = \{t_{1,j}^t, ..., t_{v_j,j}^t\}$ with results $x_j = \{x_{1,j}, ..., x_{v_j,j}\}$ which are either positive or negative. Each pathogen test is assumed, independent of any other test, to have sensitivity z and specificity 1, which means a colonised individual has probability z to return a positive result, and an uncolonised individual will always return a negative result. The swab test results may also be referred to as 'screening data'.

It is also possible for an individual to have multiple tests on a given day, and is often common in real data scenarios. This is because swab testing carriage for MRSA is performed at multiple anatomical sites, for example the throat, groin and broken skin to name a few (Price et al., 2017). Patient tests are often performed on admission and regularly after that. It is possible that some hospital wards may adjust their test frequency depending on a patients clinical state, but in our analysis we do not assume that a patient who receives a test is in any way more likely to be colonised than a patient who did not receive a test on any given day.

A colonised patient j who receives positive screening results may also have m_j isolates of the pathogen sequenced. Genetic sequences, $g_j = \{g_{1,j}, ..., g_{m_j,j}\}$, for a patient j, are produced from isolates sampled at times $t^s_{g_j} = \{t^s_{g_{1,j}}, ..., t^s_{g_{m_j,j}}\}$.

Since we are interested in inferring the source of transmission, we introduce the vector $s = (s_1, \dots, s_n)$ to be the vector of sources such that $s_i = j$ implies that patient j is the source of colonisation for individual i, and we set $s_i = -1$ if patient i is colonised on admission and $s_i = \infty$ if the patient is never colonised.

3.3.3 Transmission model with healthcare worker data

In this section we extend the transmission model described in the previous section with the distinction that we now allow for the possibility of patient colonisation from healthcare workers. The aim here is to include healthcare worker data at an individual level, with the following assumptions:

1. Healthcare workers that ever become colonised are assumed to have imported the pathogen from the community, that is to say we assume that healthcare workers cannot become colonised by patients or other staff which is a consequence of not directly modelling healthcare worker acquisition.

2. Following a positive test result on day t, healthcare workers are able to colonise other individuals from day t onwards.

3. If a healthcare worker has more than one positive result, they are assumed to be colonised at every time in between positive test results and α days thereafter. If a healthcare worker has one test result, then they are able to colonise patients for α days thereafter.

These assumptions may not be entirely reasonable in practice, for example assumption 3 assumes that colonised healthcare workers are potentially able to colonise patients on the ward the entire time between positive swab results. This may be somewhat flawed in practice in the sense that some healthcare workers may not be present for some patient episodes (such as working on a different ward, annual leave, etc.), however they are considered by the model to exert a constant colonisation pressure at all times. Furthermore, with assumption 1 we are implicitly assuming that that healthcare works become colonised uniformly at random.

These assumptions are specific to the Brighton data set, that is to say we only have information on positive test results that have then be sequenced. There is no information in the data about negative test results, positive results that were not sequenced or shift patterns.

Another issue is that due to the nature of modelling nosocomial infections in an intensive care unit, forward transmission is implicitly assumed to occur indirectly through healthcare workers. We are now considering a model where patients are able to colonise other patients indirectly through healthcare workers or the environment, while healthcare workers are also able to colonise patients directly. Note that the respective parameters may be difficult to interpret due to the fact there is no information in the data to suggest whether colonisation is direct or otherwise.

Consider a general scenario similar to that in Section 3.3.2 whereby we have n patients labelled $\{1, ..., n\}$ monitored over a study period from $t = 0$ to $t = L$ days. Furthermore, suppose that there are n_H healthcare workers labelled $\{n+1, ..., n+n_H\}$ working on the ward over the study period.

We define β_H to be the rate of colonisation between healthcare worker and patient pairs and $H(t)$ to be the number of colonised healthcare workers in the ward at time t. Recall from Section 3.3.2 we define β to be the rate of contact between pairs of patients and $C(t)$ to be the number of colonised patients on the ward on day t. For a given susceptible patient, the total colonisation pressure that is exerted by colonised patients and healthcare workers is given by $A = \beta C(t) + \beta_H H(t)$, and therefore the probability of colonisation is given by

$$\text{Pr(susceptible patient j is colonised on day t)} = 1 - \exp\left(-A\right).$$

We are assuming that patients who acquire the pathogen in the ward on day t are

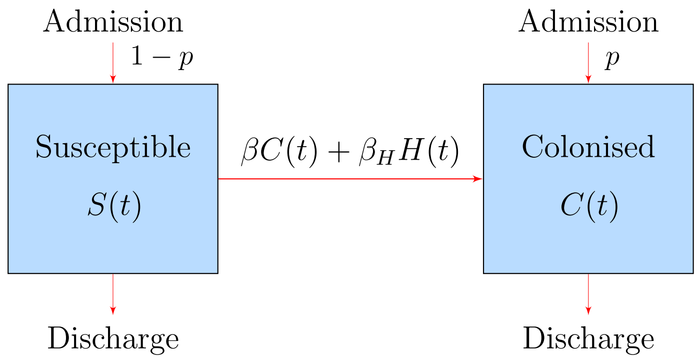


Figure 3.11: A stochastic compartmental model where individuals are either susceptible or colonised. Individuals are admitted in a susceptible state with probability $1-p$ or with a colonised state with probability p. Individuals on the ward recieve colonisation pressure from colonised patients and healthcare workers where individuals transition from susceptible to colonised at a rate of $\beta C(t) + \beta_H H(t)$ and eventually patients are discharged.

then able to colonise other patients on day $t+1$. Furthermore, we assume that on a given day, any susceptible patient that becomes colonised is independent of all other susceptible patients that become colonised.

In our model we wish to infer $C(t)$ and $H(t)$, the number of colonised patients and healthcare workers in the ward on day t respectively.

3.4 Inference

We have extended the transmission model in Worby et al. (2016) to incorporate healthcare worker data and also the genetic model in Chapter 2 by relaxing assumptions to allow for the possibility of multiple genetic subtypes. Now we outline the inference procedure in a Bayesian framework.

3.4.1 Data and notation

Suppose we have n patients and n_H healthcare workers ever admitted, or recorded to have worked on the ward in the case of healthcare workers, over a study period from $t = 0$ to $t = L$ days. Let $z^{obs} = (t^a, t^d)$ denote the set of admission and discharge times of both patients and healthcare workers where $t^a = (t^a_1, ..., t^a_n, t^a_{n+1}, ..., t^a_{n+n_H})$ and $t^d = (t^d_1, ..., t^d_n, t^d_{n+1}, ..., t^d_{n+n_H})$. The transmission dynamics are defined as $T = (t^c, s, \phi)$ which contains the colonisation times, t^c, vector of sources, s, and importation statuses, ϕ, for both patients and healthcare workers.

Recall from Section 3.3 that we assume healthcare workers import the pathogen from the community, i.e. they were colonised before admission and did not acquire the pathogen on the ward, therefore they are considered to be importations. Furthermore, from the Brighton data set we only have information on positive test results that were sequenced and therefore do not know the admission and discharge times, however for simplicity we set the admission time of healthcare workers to be $t = 0$ and the discharge time to be $t = T_i + \alpha$ where T_i is the latest positive test result for healthcare worker i and α is the colonisation period which is assumed to be fixed and known.

The observational data consists of the matrix, denoted by $x^s \in \{-1, 0, 1\}^{(n+n_H)\times(L+1)}$ of test results where the elements of x^s are defined as

$$(x^s)_{i,j} = \begin{cases} 1 & \text{individual } i \text{ tests positive on day } j-1 \\ 0 & \text{individual } i \text{ tests negative on day } j-1 \\ -1 & \text{individual } i \text{ has no test on day } j-1. \end{cases}$$

Note that above we have $j - 1$ because the study period begins at $t = 0$ and matrices are indexed from 1.

Given there are n_s genetic sequences, the genetic data is defined as $\psi = (D, \tau, \nu, \eta)$ where D is the matrix of pairwise distances, τ is the corresponding sampling times for the sequences and ν contains the ID of the individual which the sequence was sampled from, either a patient or healthcare worker. Define the vector η to contain the subtype number for each genetic sequence which is an arbitrary label in $\{1, ..., K\}$ where K is the number of distinct genetic subtypes. Furthermore, the genetic network $\mathcal{G}$ may now be composed of multiple sub genetic networks which explicitly is written as $\mathcal{G} = \cup_{i=1}^{K} \mathcal{G}_i$. The genetic network $\mathcal{G}$ is not a genuine parameter as the object is uniquely constructed depending on the unobserved transmission dynamics $\mathcal{G} = \mathcal{G}(T)$, however we consider it to be a separate parameter for clarity in notation. There also may be unobserved genetic data denoted by $\tilde{\psi} = (\tilde{D}, \tilde{\tau}, \tilde{\nu}, \tilde{\eta})$.

Let $F = (T, \tilde{\psi}, \mathcal{G})$ denote the vector of unobserved quantities that must be imputed and $\rho = (z, p, \beta, \beta_H, \lambda, \mu)$ the vector of model parameters, where we ultimately aim to infer these quantities. The key data and notation used in the model is summarised in Table 3.1.

3.4.2 Model likelihood

In this section we present the joint likelihood of the transmission dynamics, screening results and genetic data conditional the model parameters. The joint likelihood $\pi(z^{obs}, \psi, x^s|\rho)$ in Equation (3.1) is intractable as it involves summing over all possible unobserved values of F,

$$\pi(z^{obs}, \psi, x^s|\rho) = \sum_F \pi(z^{obs}, \psi, F, x^s|\rho). \tag{3.1}$$

Consequently, we consider augmenting the parameters space in a Bayesian framework with the unobserved quantities F which in turn leads to an augmented tractable likelihood. Note that we can decompose this likelihood into independent components, such that

$$\pi(z^{obs}, \psi, F, x^s|\rho) = \pi(\psi, \tilde{\psi}|z^{obs}, T, \mathcal{G}, \rho)\pi(x^s|z^{obs}, T, \rho)\pi(T|z^{obs}, \rho), \tag{3.2}$$

where the first term in Equation (3.2) is the likelihood of the genetic data, the second term is the likelihood of observational data arising from the pathogen screening tests and lastly the third term in is the likelihood of the transmission dynamics.

	Notation used for Chapter 3
n	Total number of patients admitted to the ward over the study
L	Length of study in days
$z^{obs} = (t^a, t^d)$	Observed admission and discharge times that do not contribute to the stochastic model
t^a	Vector of admission times where t_i^a is the admission time for the i^{th} individual
t^d	Vector of discharge times where t_i^d is the discharge time for the i^{th} individual
$T = (t^c, s, \phi)$	Vector of quantities that describe the transmission dynamics
t^c	Vector of colonisation times where t_i^c is the colonisation time for the i^{th} individual
s	Vector of sources
ϕ	Vector of importation statuses
x^s	The swab times and results encoded in a matrix
$\psi = (D, \tau, \nu, \eta)$	Observed genetic data consisting of the pairwise distance matrix D, sequence sample times τ, IDs ν and subtype numbers η.
$\tilde{\psi} = (\tilde{D}, \tilde{t}^s, \tilde{\nu}, \tilde{\eta})$	Unobserved genetic data consisting of the pairwise distance matrix $\tilde{D}$, sequence sample times $\tilde{\tau}$, IDs $\tilde{\nu}$ and subtype numbers $\tilde{\eta}$.
$\mathcal{G}$	The underlying genetic network that describes the structure of the isolates
$F = (T, \tilde{\psi}, \mathcal{G})$	Vector of unobserved quantities that consists of the transmission dynamics T, the unobserved genetic data $\tilde{\psi}$ and the underlying genetic network $\mathcal{G}$.
$\rho = (z, p, \beta, \beta_H, \lambda, \mu)$	Vector of model parameters
z	Test sensitivity parameter
p	Importation probability parameter
β	Transmission rate parameter
β_H	Healthcare worker transmission rate parameter
λ	Mutation rate parameter
μ	Average imported sequence distance parameter

Table 3.1: Summary of the data and notation used in the extended model which models the contribution of healthcare worker data and genetic diversity.

Epidemiological component The term $\pi(T|z^{obs}, \rho)$ in Equation (3.2) is the likelihood of the data given the admission and discharge times and the model parameters. Recall that the total number of patients who are ever admitted to the ward over the course of the study is denoted by n and the health care workers in the study is denoted by n_H. The total number of individuals in the study are labelled $i = 1, ..., n, n+1, ..., n+n_H$ where the first n individuals correspond to patients and individuals $n+1$ to $n+n_H$ correspond to healthcare workers, the likelihood is written as

$$\begin{aligned}\pi(T|z^{obs}, \rho) &= p^{\sum_{i=1}^{n}\phi_i}(1-p)^{n-\sum_{i=1}^{n}\phi_i} \\ &\times \prod_{i=1}^{n}\left[\exp\left(-\mathbb{1}_{\{t_i^c \neq t_i^a\}}\sum_{t=t_i^a}^{\min(t_i^c-1, t_i^d)} \beta C(t) + \beta_H H(t)\right)\right] \\ &\times \prod_{\substack{j: t_j^c \neq \infty \\ \phi_j = 0}}\left[\left(\frac{1-e^{-\left(\beta C(t_j^c)+\beta_H H(t_j^c)\right)}}{\beta C(t_j^c)+\beta_H H(t_j^c)}\right)\left(\mathbb{1}_{\{s_j \leq n\}}\beta + \mathbb{1}_{\{s_j > n\}}\beta_H\right)\right], \end{aligned} \tag{3.3}$$

where the source of colonisation s_j necessarily must be able to colonise patient j. The first term is the contribution of the importation status of individuals. The first product is the probability of individuals avoiding colonisation events. The second product is the probability that a given individual is colonised which depends on whether the source of colonisation is a patient or healthcare worker. Specifically this joint probability can be split into the following:

$$\begin{aligned} &\Pr(\text{Colonisation on day } t \text{ and colonised by type } i) \\ =\,&\Pr(\text{Colonised by type } i \mid \text{colonised on day } t)\Pr(\text{colonised on day } t) \end{aligned}$$

where the type i can either be a patient or healthcare worker. The probability of colonisation on day t is given by

$$\Pr(\text{colonised on day } t) = 1 - e^{-(\beta C(t)+\beta_H H(t))}.$$

If we first consider the probability of colonisation from any other patient given that the patient has been colonised on day t, we assume that all colonised individuals are equally likely to be the source of transmission and we obtain the following

$$\Pr(\text{Colonised by patient} \mid \text{colonised on day } t) = \frac{\beta C(t)}{\beta C(t)+\beta_H H(t)} \times \frac{1}{C(t)}$$

where the first term is the probability of colonisation from all colonised patients and the second term is the probability of the source of transmission, since each of the $C(t)$ colonised individuals are considered to be equally likely. In a similar fashion, the probability to become colonised by a specific healthcare worker is given by

$$\Pr(\text{Colonised by patient} \mid \text{colonised on day } t) = \frac{\beta_H H(t)}{\beta C(t) + \beta_H H(t)} \times \frac{1}{H(t)}.$$

Recall that s denotes the vector of sources, therefore if there are n patients then it follows that $s_j \leq n$ for any patient j who is colonised by another patient, otherwise $s_j > n$ which indicates colonisation by a health care worker. After simplifying terms and inclusion of the indicator functions which determine whether the source of colonisation is a patient or healthcare worker, we arrive at the second product in Equation (3.3).

Screening component The contribution of the observational screening data $\pi(x^s|z^{obs}, T, \rho)$ is given by

$$\pi(x^s|z^{obs}, T, \rho) = z^{\mathrm{TP}(x^s)}(1-z)^{\mathrm{FN}(x^s,T)}, \tag{3.4}$$

where $\mathrm{TP}(x^s)$ and $\mathrm{FN}(x^s, T)$ are the number of true positive and false negative test results. Explicitly these quantities are

$$\mathrm{TP}(x^s) = \sum_{i=1}^{n} \sum_{j=1}^{L+1} \mathbb{1}_{\{x^s_{ij}=1\}}$$

$$\mathrm{FN}(x^s, T) = \sum_{i=1}^{n} \sum_{j=1}^{L+1} \mathbb{1}_{\{x^s_{ij}=0, t^c_i \leq j+1\}}$$

Since we assume that tests have perfect specificity, the number of true positives depends on the observed data alone and is simply the number of all positive test results. On the other hand, the number of false negatives depends on the unobserved colonisation times and is not known from the observed data.

We note that in the Brighton data set healthcare workers do not contribute to the number of false negatives and true positives due to lack of negative test data.

Genetic component In Section 2.9 we describe the genetic model in an outbreak setting. Suppose there are K distinct genetic subtypes defined by the vector η where each subtype contains m_i imported sequences for $i = 1, ..., K$ and let $\mathcal{I}^i = (\mathcal{I}^i_1, ..., \mathcal{I}^i_{m_i})$ denote the vector of imported sequence labels for genetic subtype i.

The contribution to the likelihood from the complete genetic data $(\psi, \tilde{\psi})$ given the admission and discharge times, z^{obs}, unobserved transmission dynamics T, genetic network $\mathcal{G}$ and model parameters ρ is given by

$$\pi(\psi, \tilde{\psi}|z^{obs}, T, \mathcal{G}, \rho) \propto \prod_{(i,j)\in E(\mathcal{G})} \left(1 - q\left(t_{ij}\right)\right)^{N-D_{ij}} \left(\frac{1}{3}q\left(t_{ij}\right)\right)^{D_{ij}} \\ \times \prod_{i=1}^{K}\prod_{j=2}^{m_i}\prod_{k=1}^{j} \frac{\mu^{D_{\mathcal{I}^i_j\mathcal{I}^i_k}} e^{-\mu}}{D_{\mathcal{I}^i_j\mathcal{I}^i_k}!} \tag{3.5}$$

where $E(\mathcal{G})$ is the set of edges in the genetic network $\mathcal{G}$, $q(t_{ij})$ is the probability of observing a mutation between nodes i and j in t_{ij} time units. The first term is the contribution of directly connected distances assuming a mutation model and the second term is the contribution of all distances between imported sequences within the same genetic subtype.

Furthermore, the proportionality in Equation (3.5) is required due to the fact we only have knowledge of the full matrix of pairwise distances and do not know the precise number of mutations at each site.

3.4.3 Bayesian inference

Our primary interest is to make inferences about the model parameters ρ and unobserved data F, and to do so we wish to explore the joint posterior density

$\pi(\rho, F|z^{obs}, \psi, x^s)$. From Bayes' Theorem we obtain

$$\begin{aligned}
\pi(\rho, F|z^{obs}, \psi, x^s) &\propto \pi(\psi, \tilde{\psi}|z^{obs}, T, \mathcal{G}, \rho)\pi(x^s|z^{obs}, T, \rho)\pi(T|z^{obs}, \rho)\pi(\rho) \\
&= z^{\mathrm{TP}(x^s)}(1-z)^{\mathrm{FN}(x^s,T)} \\
&\times \prod_{(i,j)\in E(\mathcal{G})} \left(1-q\left(t_{ij}\right)\right)^{N-D_{ij}} \left(\frac{1}{3}q\left(t_{ij}\right)\right)^{D_{ij}} \\
&\times \prod_{i=1}^{K}\prod_{j=2}^{m_i}\prod_{k=1}^{j} \frac{\mu^{D_{\mathcal{I}^i_j\mathcal{I}^i_k}} e^{-\mu}}{D_{\mathcal{I}^i_j\mathcal{I}^i_k}!} \\
&\times \prod_{i=1}^{n}\left[\exp\left(-\mathbb{1}_{\{t^c_i \neq t^a_i\}} \sum_{t=t^a_i}^{\min(t^c_i-1, t^d_i)} \beta C(t) + \beta_H H(t)\right)\right] \\
&\times \prod_{\substack{j:t^c_j\neq\infty \\ \phi_j=0}} \left[\left(\frac{1-e^{-\left(\beta C(t^c_j)+\beta_H H(t^c_j)\right)}}{\beta C(t^c_j)+\beta_H H(t^c_j)}\right)\left(\mathbb{1}_{\{s_j\leq n\}}\beta + \mathbb{1}_{\{s_j>n\}}\beta_H\right)\right] \\
&\times p^{\sum_{i=1}^n \phi_i}(1-p)^{n-\sum_{i=1}^n \phi_i} \times \pi(\rho), \qquad (3.6)
\end{aligned}$$

where the first three terms denote the likelihood of the observational, genetic and epidemiological data given by Equations (3.3)-(3.5) and $\pi(\rho)$ is the joint prior density of the parameters. If the outbreak is fully observed then it is straightforward to sample from the posterior distribution using Markov chain Monte Carlo algorithms. If the outbreak is not fully observed (which is invariably the case for real data sets), then we must update the augmented data in an efficient manner to explore the space. We assign

$$\begin{aligned}
z &\sim \mathrm{Beta}(\alpha_z, \beta_z) \\
p &\sim \mathrm{Beta}(\alpha_p, \beta_p) \\
\beta &\sim \mathrm{Exp}(\nu_\beta) \\
\beta_H &\sim \mathrm{Exp}(\nu_{\beta_H}) \\
\lambda &\sim \mathrm{Exp}(\nu_\lambda) \\
\mu &\sim \mathrm{Gamma}(\alpha_\mu, \beta_\mu)
\end{aligned}$$

a priori and assume that for the priors the parameters are independent, i.e. $\pi(\rho) = \pi(z, p, \beta, \beta_H, \lambda, \mu) = \pi(z)\pi(p)\pi(\beta)\pi(\beta_H)\pi(\lambda)\pi(\mu)$. Details of the MCMC algorithm can be found in Section 3.5.

3.5 Markov chain Monte Carlo procedure

In order to fit the model outlined in Section 3.4.3, we employ a data-augmented MCMC routine to sample the parameters ρ and the unobserved data F from the target density in Equation (3.6).

At each iteration our MCMC routine updates each parameter sequentially, and then proposes updates to the augmented data y times. Details of the routine can be found in Algorithm 3 and parameter updates can be found in Section 3.5.1. The mechanisms for imputing genetic distances and updating the augmented data described later in Section 3.5.2.

Algorithm 3 Structure of the MCMC algorithm

```
Initialise the chain with initial values $z^{(1)}, p^{(1)}, \beta^{(1)}, \beta_H^{(1)}, \lambda^{(1)}, \mu^{(1)}$, maximum number of iterations $M$, proposal variances $\sigma^2_\beta, \sigma^2_{\beta_H}, \sigma^2_\lambda$ and prior parameters $a_z, b_z, a_p, b_p, \nu_\beta, \nu_{\beta_H}, \nu_\lambda, a_\mu, b_\mu$.
    Repeat the following steps
Sample $z^{(i)}$ from the full conditional distribution.
Sample $p^{(i)}$ from the full conditional distribution.
Update $\beta^{(i)}$ using a random walk Metropolis-Hastings step with proposal variance $\sigma^2_\beta$.
Update $\beta_H^{(i)}$ using a random walk Metropolis-Hastings step with proposal variance $\sigma^2_{\beta_H}$.
Update $\lambda^{(i)}$ using a random walk Metropolis-Hastings step with proposal variance $\sigma^2_\lambda$.
Sample $\mu^{(i)}$ from the full conditional distribution.
Update the augmented data $y$ times.
```

3.5.1 Parameter updates

In this section we explicitly show how each of the parameters are updated in our MCMC algorithm. Let $X = (z^{obs}, \psi, x^s)$ denote the observed data, ρ the model parameters, F the unobserved data and ρ_{-x} to be the vector of parameters not including x. We may sample p, z and μ from the following full conditional distributions,

$$\begin{aligned}
z \mid \rho_{-z}, F, X &\sim \text{Beta}\left(\alpha_z + \text{TP}(x^s), \beta_z + \text{FN}(x^s, T)\right) \\
p \mid \rho_{-p}, F, X &\sim \text{Beta}\left(\alpha_p + \sum_{i=1}^{n} \phi_i, \beta_p + n - \sum_{i=1}^{n} \phi_i\right) \\
\mu \mid \rho_{-\mu}, F, X &\sim \text{Gamma}\left(\alpha_\mu + \sum_{i=1}^{K} \sum_{j=2}^{m_i} \sum_{k=1}^{j} D_{\mathcal{I}_j^i \mathcal{I}_k^i}, \beta_\mu + R\right),
\end{aligned}$$

where the triple sum is the total of all pairwise distances and R is the number of pairwise distances between imported sequences within the same subtype. For the remaining parameters which are the patient transmission rate β, healthcare worker transmission rate β_H and mutation rate λ, we may update these parameters using a Metropolis-Hastings random-walk with a normal proposal distribution.

Let β', β'_H and λ' be the proposal values for the parameters and ρ' be the vector of proposal parameters where $\rho' = (p, z, \beta', \beta_H, \lambda, \mu)$, $\rho' = (p, z, \beta, \beta'_H, \lambda, \mu)$ or $\rho' = (p, z, \beta, \beta_H, \lambda', \mu)$, depending on which parameter we are trying to update. Explicitly the proposal values are given by

$$\begin{aligned}
\beta' &\sim N(\beta, \sigma_\beta^2) \\
\beta'_H &\sim N(\beta, \sigma_{\beta_H}^2) \\
\lambda' &\sim N(\lambda, \sigma_\lambda^2),
\end{aligned}$$

which are accepted with probability

$$\min\left(1, \frac{\pi(\rho', F | z^{obs}, \psi, x^s)}{\pi(\rho, F | z^{obs}, \psi, x^s)}\right).$$

3.5.2 Imputing genetic distances

In this section we present the proposal mechanisms for imputing genetic distances. Recall from Section 2.8 that we are assuming that each individual has a genetic sequence at the time of colonisation, which may be unobserved. If the time of colonisation for an individual occurs prior to the first observed genetic sequence then a new set of distances may need to be imputed.

The general setup is as follows. Suppose we wish to impute genetic distances for some genetic sequence denoted by X for patient i at time $t = t_X$. Begin by constructing the genetic network $\mathcal{G}$ which includes the imputed node. Define the *parent node* as the node which corresponds to the genetic source. Suppose the parent node of X is node j, then we write this by $P_a(X) = j$. If there is no genetic source, then the imputed node X is imported, and we set the parent node $P_a(X) = -1$.

Define a *child node* for node X to be a genetic sequence with X as the genetic source, then the set of child nodes for X is defined as $C_h(X) = \{j : P_a(j) = X\}$. Let $|C_h(X)|$ denote the number of child nodes such that X is the parent. If $|C_h(X)| \leq 1$, i.e. there are no children or only one child node, then the imputation is not necessary.

We begin by defining two distinct scenarios for imputing genetic nodes, which are *interior nodes* and *exterior nodes*. We define an imputed node X to be an interior node if and only if $P_a(X) \neq -1$, otherwise it is an exterior node. An example setup with an interior and exterior node is given in Figure 3.12.

Before looking at the specific details of the imputation, recall a key fact for Poisson processes. Let $N(t)$ be the number of mutations in t time units. Suppose there are $N(t) = n$ mutations observed in the interval $[0, t]$, then the n points are distributed as independent and identically distributed uniform random variables on $[0, t]$. It follows that for $s < t$,

$$(N(s)|N(t) = n) \sim \text{Bin}(n, s/t). \tag{3.7}$$

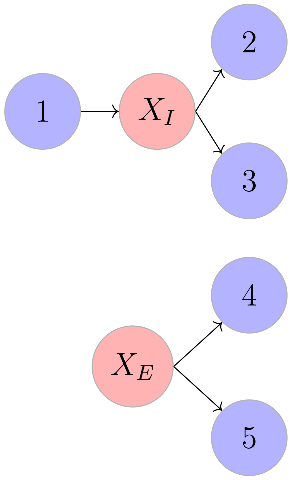


Figure 3.12: An example setup with an interior node, denoted by X_I, and an exterior node denoted by X_E. Since the node X_I has a parent node given by $P_a(X_I) = 1$ and children nodes $C_h(X_I) = \{2, 3\}$, then by definition X_I is an interior node. Similarly, node X_E has no parent and thus $P_A(X_E) = -1$ with children nodes $C_h(X_E) = \{4, 5\}$, hence X_E is an exterior node by definition.

Imputing interior nodes Suppose we wish to impute a set of genetic distances for a genetic sequence labelled X at time $t = t_X$. The term interior node comes from the fact there exists a parent node $P_a(X)$ and at least two child nodes $C_h(X)$. Let $k = |C_h(X)|$ be the number of child nodes and begin by labelling the parent node P which is observed at time $t = t_P$ and the child nodes C_i for which is observed at times $t = t_{C_i}$ for $i = 1, ..., k$.

Let $\alpha \in C_h(X)$ denote the sequence that exhibits the smallest amount of genetic evolution between the parent node P and the child nodes $C_h(X)$, i.e. $D_{P\alpha} \leq D_{PC_i}$ for all $i = 1, ..., k$. We are interested in the number of mutations between nodes P and X (denoted by D_{PX}) in the smaller interval $t = t_P - t_X$, given that we have observed $D_{P\alpha}$ mutations in the larger interval $t = t_\alpha - t_X$.

In line with the notation in Equation (3.7), we have $n = D_{P\alpha}$, $s = t_P - t_X$ and $t = t_\alpha - t_X$. Therefore we can propose the genetic distance between the parent node

P and the imputed node X by

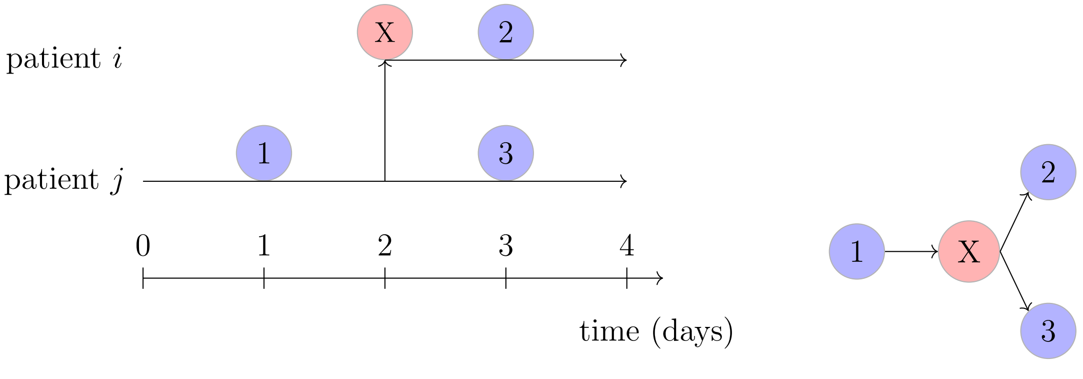
$$(D_{PX}|D_{P\alpha} = n) \sim \text{Bin}\left(n, \frac{t_P - t_X}{t_\alpha - t_X}\right)$$

By assuming that any nucleotide cannot mutate more than once, the distances between the imputed node X and the child nodes C_i for $i = 1, ..., k$ are calculated by

$$D_{XC_i} = D_{PC_i} - D_{PX}.$$

Although this assumption is not entirely consistent with continuous time Markov chains, we find this assumption to be reasonable for large N and small t. Verifying the validity of this assumption is discussed more in detail in Section 3.6.7.

The genetic distance between the imputed node X and all other nodes is calculated by simply summing the total distance along the branches to the node. An example outbreak that requires imputing an interior node is presented in Figure 3.13.



(a) Transmission tree **(b)** Genetic network

Figure 3.13: An example outbreak with genetic data (a) and the corresponding genetic network (b) for the scenario of imputing interior nodes. Patient j colonises patient i at time $t = 2$, with observed genetic sequences denoted in blue, and the unobserved sequence X in red. The imputed node is denoted by X, with the parent node $P_a(X) = 1$ and child nodes $C_h(X) = \{2, 3\}$.

Imputing exterior nodes Suppose we wish to impute a set of genetic distances for a sequence labelled X at time $t = t_X$ such that node X has no parent, i.e.

$P_a(X) = -1$ and therefore X is an exterior node. Let $C_h(X)$ denote the set of child nodes and $k = |C_h(X)|$ denote the number of child nodes, where $k \geq 2$. Label the child nodes C_i which is observed at time $t = t_{C_i}$ for $i = 1, ..., k$.

Let $\alpha, \beta \in C_h(X)$ denote the pair of sequences from the child nodes that exhibits the smallest amount of genetic evolution, where $\alpha \neq \beta$. In a similar fashion to before, we wish to simulate the distance $D_{X\alpha}$ mutations in the interval $t = t_\alpha - t_X$, given that there are $D_{\alpha\beta}$ in the larger interval $t = (t_\alpha - t_X) + (t_\beta - t_X)$. By considering the setup in Equation (3.7), we obtain

$$(D_{X\alpha} | D_{\alpha\beta} = n) \sim \text{Bin}\left(n, \frac{t_\alpha - t_X}{t_\alpha + t_\beta - 2t_X}\right).$$

Assuming that nucleotides cannot mutate more than once in the time interval, the distance between the imputed node X and child node C_i is given by

$$D_{XC_i} = D_{\alpha C_i} - D_{X\alpha}, \quad i = 1, ..., k \quad C_i \neq \alpha.$$

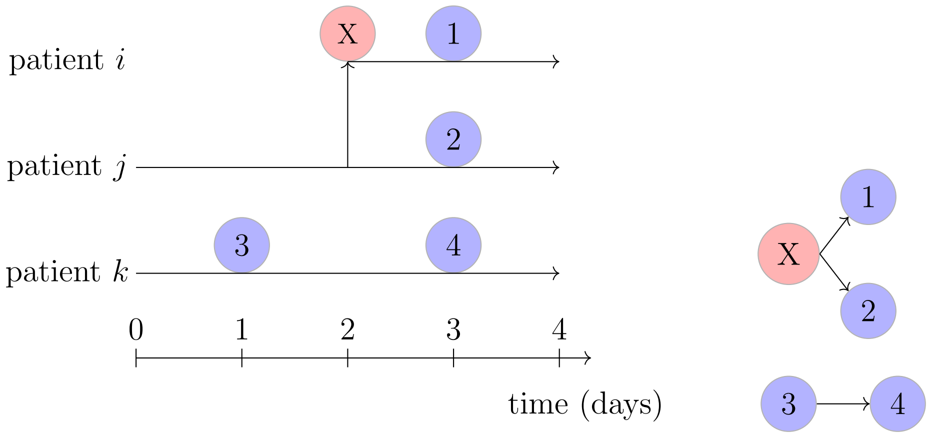
Now we must consider the distance between the imputed node X, and other nodes in the genetic network $\mathcal{G}$ that are roots of their genetic tree. Let $S_X = \{i : P_a(i) = -1, i \neq X\}$ denote the set of imported sequences that excluding the imputed node X and $k_X = |S_X|$ denote the size of this set. Given that $k_X > 0$, we calculate these distances by

$$D_{iX} = D_{i\alpha} - D_{X\alpha}, \quad i \in S_X.$$

The genetic distance between the imputed node X and all other nodes is calculated by simply summing the total distance along the branches to the node. An example outbreak that requires imputing an exterior node is presented in Figure 3.14.

3.5.2.1 Proposal ratios when imputing distances

In this section we briefly outline the contribution to the proposal ratio when updating genetic data. After proposing a change to the transmission tree it may be necessary to update the augmented genetic data.



(a) Transmission tree

(b) Genetic network

Figure 3.14: An example outbreak with genetic data (a) and the corresponding genetic network (b) for the scenario of imputing exterior nodes. Patient j colonises patient i at time $t = 2$, patient k is colonised on admission. Observed genetic sequences denoted in blue, and the unobserved sequence X in red. The imputed node is denoted by X, with the parent node $P_a(X) = -1$ and child nodes $C_h(X) = \{1, 2\}$.

Let $\mathcal{N}, \mathcal{N}^*$ denote the set of nodes that require imputation given the current and proposal configuration respectively. Then we simulate distances for each $i \in \mathcal{N}^*$ using the mechanisms in Section 3.5.2. The contribution to the proposal ratio is given by

$$\mathcal{Y}_{gen} = \frac{\prod_{i\in\mathcal{N}} \left(\mathbb{1}_{\{P(i)\neq -1\}} f_I(i) + \mathbb{1}_{\{P(i)=-1\}} f_E(i)\right)}{\prod_{i\in\mathcal{N}^*} \left(\mathbb{1}_{\{P(i)\neq -1\}} f_I(i) + \mathbb{1}_{\{P(i)=-1\}} f_E(i)\right)},$$

where $f_I(i)$ and $f_E(i)$ are the probability mass functions associated with generating interior and exterior distances respectively. Explicitly these are given by

$$f_I(i) = \Pr\left[\text{Bin}\left(D_{P\alpha}, \frac{t_P - t_i}{t_\alpha - t_i}\right) = D_{iP}\right],$$

where for each imputed node i, P denotes the parent node and α denotes the child that exhibits the minimum amount of genetic evolution. Similarly, for exterior distances we obtain

$$f_E(i) = \Pr\left[\text{Bin}\left(D_{\alpha\beta}, \frac{t_\alpha - t_i}{t_\alpha + t_\beta - 2t_i}\right) = D_{i\alpha}\right],$$

where for each imputed node i, α and β denote the pair of nodes that exhibit the minimum amount of genetic evolution.

In each step we are proposing to update all of the unobserved genetic distances simultaneously, i.e. a block update.

3.5.3 Augmented data updates

In this section we present the motivation for updating the augmented data and the specific mechanisms for this, which is based on the work in Cassidy (2019) and Worby et al. (2016). The general idea is as follows: any individual who receives a positive test result must have a colonisation time since we are assuming that the test specificity is 100% and there are no false positives. The colonisation times are partially observed in the sense that the exact time of colonisation is not known, but the latest time an individual could have been colonised is the time of the first positive test result. Therefore we propose to update the colonisation times by moving them.

In principle any patient may have a finite colonisation time and will either be colonised on admission or colonised by another patient. For those patients with no positive test results, addition and deletion of colonisation times can occur in addition to moving the times.

After sampling a new vector of model parameters we proceed to update the augmented data F, which consists of the set of colonisation times, importation statuses, sources of colonisation and genetic data. In order to update the augmented data, we randomly choose one of the possible moves with equal probability. We may also update the importation status of an individual and the source of colonisation.

Once we have proposed a colonisation time, source of colonisation and importation status, we may also update the genetic data. Genetic sequences (and corresponding distances) need only be imputed in cases which are necessary to capture the shared stochastic evolution of the pathogen. After identifying the set of sequences and times that require imputation, distances can be generated using the scheme discussed in Section 3.5.2.

Let F^* denote the complete candidate data set such that $F^* = \{t^{c*},\ \phi^*, s^*,\ \tilde{D}^*,\ \tilde{\tau}^*,\ \tilde{\nu}^*,\ \mathcal{G}^*\}$ and define the proposal ratio $q_{F,F^*} = \Pr(F^* \to F)/\Pr(F \to F^*)$, which is the probability of making the reverse move divided by the probability of making the forward move.

First we propose to update the transmission tree by proposing a move with corresponding candidate values (t^{c*}, ϕ^*, s^*). Depending on the move proposed we then identify sequences that should be imputed (or removed) and update the candidate genetic network $\mathcal{G}^*$. Finally, we then update genetic distances and propose values for $(\tilde{D}^*, \tilde{\tau}^*, \tilde{\nu}^*)$. Hence we accept the proposal data set $F*$ with probability

$$\min\left(1, \frac{\pi(\rho, F^*|z^{obs}, \psi, x^s)}{\pi(\rho, F|z^{obs}, \psi, x^s)} q_{F,F^*}\right).$$

For the rest of this section we shall outline the specific moves and the corresponding proposal ratios.

3.5.3.1 Moving a colonisation time

Select one of the colonised patients uniformly who are considered to have ever contracted the pathogen during their stay on the ward, called patient j. With probability w the patient is proposed to be an importation and we set $\phi_j^* = 1$ and $t_j^{c*} = t_j^a$, otherwise with probability $1 - w$ the patient is proposed to have been colonised by another patient. Let l_j be the last possible day that patient j could have been colonised, which is either the time of first positive swab or the time of first colonisation with patient j as the source. Then uniformly sample the colonisation time t_j^{c*} from the day of admission to the last day, $\{t_j^a, ..., l_j\}$ and set $\phi_j^* = 0$. Select a source of colonisation uniformly from one of the $C(t_j^{c*})$ individuals that are able to colonise on day t_j^{c*}, if $C(t_j^{c*}) = 0$ then we reject this move.

Finally, given the proposal transmission tree we update the genetic data introduced in Section 3.5.2.1 with a contribution of $\mathcal{Y}_{gen}$ to the proposal ratio.

The mechanisms to propose updates for colonised individuals depend on the current configuration of the chain and the type of proposal. The proposed moves and corresponding proposal ratios q_{F,F^*} for patient j can be summarised in Table 3.2.

	Acquisition	Importation
Acquisition	$\frac{C(t_j^{c*})}{C(t_j^c)}\mathcal{Y}_{gen}$	$\frac{1-w}{w(l_j-t_j^a+1)C(t_j^c)}\mathcal{Y}_{gen}$
Importation	$\frac{w(l_j-t_j^a+1)C(t_j^{c*})}{1-w}\mathcal{Y}_{gen}$	$\mathcal{Y}_{gen}$

Table 3.2: Proposal ratios for the possible moves when updating an individual's time of colonisation, importation status and source. The leftmost column is for the current state of the patient, and the top row is the proposal state. Importations are the case where patients are positive on admission, and acquisitions the case of patients being colonised by other patients.

3.5.3.2 Adding a colonisation time

In this move we randomly select a currently uncolonised patient, j, and propose to add a colonisation for them. If there are no uncolonised patients to select, then no move is made. The number of uncolonised patients is the number of patients to have never tested positive, n_s, minus the number of patients added by the algorithm, n_a. For a given patient j, propose them to be an importation with probability w, in which case we set $t_j^{c*} = t_j^a$ and $\phi_j^* = 1$. Otherwise patient j is proposed to have been colonised by another patient with probability $1 - w$, and the colonisation time is uniformly sampled from the day of admission to the day of discharge, $\{t_j^a, ..., t_j^d\}$. The source of the colonisation is then sampled randomly from the $C(t_j^{c*})$ currently colonised patients. If there are no colonised patients at the proposed time of colonisation, then no move is made.

The proposal ratio for adding an importation is given by

$$q_{F,F^*} = \frac{n_s - n_a}{w(1 + n_z)},$$

and the proposal ratio for adding an acquisition is

$$q_{F,F^*} = \frac{C(t_j^{c*})(n_s - n_a)(t_j^d - t_j^a + 1)}{(1 - w)(1 + n_z)}.$$

3.5.3.3 Removing a colonisation time

In this move we uniformly select one of the n_a patients who have been added by the algorithm but are not considered to be a source of colonisation. If there are no available individuals to remove, then no move is made. If we are removing an individual who is assumed to be imported, then the proposal ratio is

$$q_{F,F^*} = \frac{n_z w}{n_s - n_a + 1}.$$

If we are removing an individual who is assumed to be an acquisition, then the proposal ratio is

$$q_{F,F^*} = \frac{n_z(1 - w)}{(t_j^d - t_j^a + 1)(n_s - n_a + 1)(C(t_j^c) - 1)}.$$

We note that in each of our proposal moves, the quantity w is simply a tuning parameter to ensure sufficient mixing and exploration of the state space.

3.5.3.4 Moving a colonisation time of a healthcare worker

Similar to moving the colonisation time of patients, this move is motivated by the fact that colonisation times are not known exactly and there is uncertainty due to the test sensitivity and testing frequency. We found that in simulated scenarios it may be the case that a healthcare worker colonises a patient on day t and then tests positive on day $t + \delta$ for some $\delta > 0$. If the healthcare worker colonisation times were fixed then it could be possible that the patient would never be colonised by the 'true' source, hence we propose small updates to the colonisation times that help resolve these issues and explore the state space of the chain.

Let n_H be the number of healthcare workers with a finite colonisation time. Uniformly at random sample one of the n_H workers, denoted by j. Let t_j^c denote the current colonisation time for the healthcare worker, we then update this time by adding 1 with probability $\frac{1}{2}$ or subtract 1 with probability $\frac{1}{2}$, and denote this proposal colonisation time as t_j^{c*}. If the proposal time is outside of the study period we reject the move, in other words reject if $t_j^{c*} < 0$ or $t_j^{c*} > L$.

Finally, given the proposal transmission tree we update the genetic data (see Section 3.5.2.1 for details) and hence the proposal ratio is $q_{F,F^*} = \mathcal{Y}_{gen}$, where $\mathcal{Y}_{gen}$ is contribution of the genetic imputation to the proposal ratio.

3.5.3.5 Update a transmission cluster

The idea behind this move is fix the root of a transmission tree (the imported individual) and propose to update every other individual in the tree. The rationale behind this is that in some scenarios the chain would get stuck in a local mode and some configurations were very unlikely to ever be explored from updating colonisation times and sources one at a time. Therefore to address this we are proposing a block update to a variable number of colonisation times depending on the size of the transmission tree.

Let n_{imp} the set of individuals who are considered to have imported the pathogen from the community which will also all colonised healthcare workers, and sample individual j uniformly at random from this set. Let ω_j denote the set of ‘complete offspring’ for imported individual j, which is defined as all offspring originating from the root of the tree. If there are no offspring originating from individual j we reject the move. Let f_i denote the time of the first positive test result or first observed sequence for individual i. Then for each offspring $i \in \omega_j$, sample a colonisation time uniformly at random from the set $\{t_i^a, ..., f_i\}$.

After sampling a colonisation time for each offspring we have a proposal vector of colonisation times t^{c*}, now we propose to update the source of colonisation for each individual. For each offspring $i \in \omega_j$, sample a source of colonisation uniformly at random from the $C(t_i^{c*}) + H(t_i^{c*})$ possible colonised individuals at the proposal time of colonisation for patient i. If there are no possible sources to sample for any of the offspring we reject the move.

Finally, given the proposal transmission tree we update the genetic data (see Section 3.5.2.1 for details) and hence the proposal ratio is

$$q_{F,F^*} = \prod_{i \in \omega_j} \frac{C(t_i^{c*}) + H(t_i^{c*})}{C(t_i^{c}) + H(t_i^{c})} \mathcal{Y}_{gen},$$

where $\mathcal{Y}_{gen}$ is contribution of the genetic imputation to the proposal ratio.

3.5.3.6 Swap a patient with the source

Similar to the work in Cassidy (2019), we outline a move to swap a patient with their source of colonisation. The idea behind this is to ‘reshuffle’ parts of the transmission chain in order to avoid getting stuck in local modes. In some situations it may be impossible to move the colonisation time of specific targets due to that target being a source of colonisation, i.e. they are assumed to colonise someone else at a later time.

Suppose the pair i, j are in fact epidemiologically linked, i.e. one colonised the other and they may have similar genetic sequences (and hence small distances), in which case there should be a high probability in favour of one to colonise the other. Further suppose that i colonises j however the current state of the chain is that j colonised i. It is generally unlikely to move one or the other using a single proposal, since any move that attempts to update a source will ‘break’ a link in the genetic network and the proposal will invariably have a lower likelihood.

Select one of the patients, denoted by j, on the ward who are assumed to have acquired the pathogen from another individual uniformly at random. Define i to be the source of colonisation for j such that $s_j = i$. If i is a healthcare worker we immediately reject since healthcare workers cannot acquire the pathogen on the ward. Next if the colonisation time for i is before the admission time for j, i.e. $t_i^c < t_j^c$, we reject the move. Furthermore, if there is a positive test result or an observed sequence for i prior to the time of colonisation for j we reject since an individual should not have a colonisation time later than a positive test result. Finally, if i colonises any other individual prior to the time of colonisation for j, we reject since it is not possible to colonise someone before the time of colonisation. Otherwise swap colonisation times and set $t_j^{c*} = t_i^c$ and $t_i^{c*} = t_j^c$ and also set the source of colonisation for i to be j, and the source of colonisation for j to be the previous source of colonisation for i, i.e. $s_i^* = j$ and $s_j^* = s_i$.

Finally, given the proposal transmission tree we update the genetic data (see Section 3.5.2.1 for details) and hence the proposal ratio is $q_{F,F^*} = \mathcal{Y}_{gen}$, where $\mathcal{Y}_{gen}$ is contribution of the genetic imputation to the proposal ratio.

3.6 Simulation study

In this section we present a comprehensive simulation study and examine the performance of our methods through parameter estimation, sensitivity analysis and network accuracy.

The main goal is to simulate outbreaks with various parameter choices and determine for which the parameters can be inferred using the algorithm detailed in Algorithm 3. We also would like to explore the valuation of genetic data, in other words we wish to examine the improvement when inferring the transmission network with and without genetic data. Finally, we would also like to assess the performance of the sampler through various MCMC diagnostic tools.

3.6.1 Simulation method

Simulating outbreaks of nosocomial infections in a hospital setting is achieved through three independent steps. First we simulate an outbreak from the transmission model, then we screen individuals on admission and then at regular intervals from the observational model, and finally we generate a pairwise genetic distance matrix that corresponds to the positive swab results such that every positive swab will correspond to a set of genetic distances.

Simulating outbreak data In order to simulate an outbreak of a nosocomial infection on a hospital ward, we specify the number of patients in the study, n, the number of healthcare workers to ever become colonised, n_H, the length of the study, L, and the average length of stay for patients admitted to the ward, μ_{LOS}.

For each of the n patients, their admission time is sampled uniformly at a random time from $t = 0$ to $t = L$, and their length of stay is drawn from a Poisson distribution with mean μ_{LOS}. Furthermore, each patient has some probability p to be colonised on admission. All other patients admitted to the ward are susceptible from the day of admission. For each of the n_H healthcare workers a colonisation time is sampled uniformly from the days $\{0, ..., L\}$.

Simulating observational data For a given individual i, the patient is tested on admission at time $t = t_i^a$ and every κ days thereafter until the day of discharge, t_i^d. Healthcare workers are tested at time $t = 0$ and every κ_H days thereafter. For a given patient or healthcare worker i that is tested on day t, a positive result is returned with probability z if the individual is colonised such that $t \geq t_i^c$, otherwise a negative result is returned.

Simulating genetic data We will assume that every positive swab test has a corresponding genetic sequence sampled. This is not necessarily true in practice as individuals may have a positive result but no genetic isolate sequenced, however here we shall assume complete coverage of genetic testing.

In order to generate the genetic data we must specify the mutation rate λ, importation distance μ and average distance between imported sequences in distinct subtypes, denoted by δ. Furthermore, we require the genome size of the pathogen, N, and the number of distinct genetic subtypes K. We wish to generate genetic data that is consistent with the true simulated transmission network.

From the positive results we have the nodes in the genetic network denoted by V. For each node $v \in V$ in the genetic network, uniformly at random sample the subtype number η_v from the set $\{1, ..., K\}$. First we construct the sub genetic networks $\mathcal{G}_i$ outlined in Section 2.8, with the added requirement that all colonised individuals have a sequence at the time of colonisation. Given the true underlying genetic network $\mathcal{G} = \cup_{i=1}^{\mathcal{G}_{max}} \mathcal{G}_i$ and the corresponding sequence times τ, we simulate the number of mutations along each edge $(i, j) \in E(\mathcal{G})$ from the genetic evolution model outlined in Section 2.4. Let $q(t)$ be the probability of mutation in t time units, then the distance between nodes i and j is simply

$$D_{ij} \sim \text{Bin}(N, q(t_j - t_i)), \tag{3.8}$$

where N is the length of the genetic sequence and $t_j \geq t_i$. To generate distances from nodes that are in the same genetic group but separate trees, i.e. imported sequences, we draw pairwise distances between the imported sequences from

$$D_{ij} \sim \text{Po}(\mu), \tag{3.9}$$

where μ can be interpreted as the average distance between imported sequences within the same genetic group.

If we wish to generate distances between sequences from different genetic groups, we draw the distances of the root nodes from

$$D_{ij} \sim \text{Po}(\xi), \tag{3.10}$$

where ξ is required to generate a complete matrix of pairwise distances, but we note that this parameter is not inferred.

To generate the pairwise distances between nodes that are not directly connected, we assume that the same nucleotide cannot mutate twice (see Section 3.6.7), therefore the distance is a sum of the distances along the path that connects the nodes. This assumption may not hold for diverse genetic sequences, however should be reasonable for within subtype distances which are of primary interest. An example simulated complete genetic network can be found in Figure 3.15.

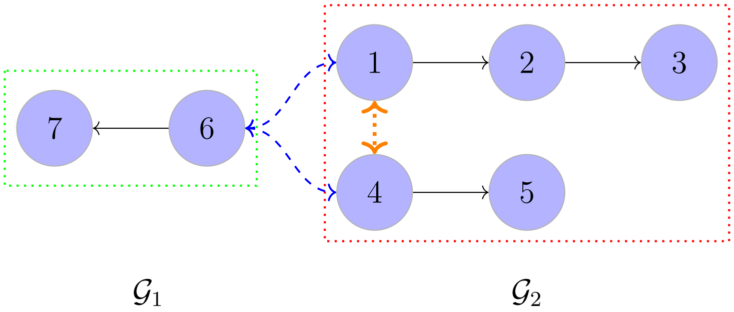


Figure 3.15: An example genetic network $\mathcal{G}$ consisting of 7 observed genetic sequences which is composed of $K = 2$ distinct genetic subtypes, where nodes 6 and 7 are in the sub genetic network $\mathcal{G}_1$ (dashed green line) with subtype 1 and the remaining nodes are in the sub genetic network $\mathcal{G}_2$ (dashed red line) with subtype 2. Solid black lines between directly connected nodes i and j have their distances D_{ij} simulated from the mutation model in Equation 3.8. The orange dotted line connects nodes that are in the same group, but are importations of the sequence and therefore we simulate distances D_{ij} from Equation 3.9. Dashed lines in blue are root nodes that are in distinct sub genetic networks and therefore we simulate the distances from Equation 3.10. All other pairwise distances are calculated by summing all connected distances in the network.

Note that instead of simulating genetic distances it is possible to instead generate entire sequences, calculate the number of mutations along each branch in the genetic network and then uniformly sample and replace nucleotides in the sequences. While this method would more accurately describe the biological process, generating large sequences is computationally intensive. The above method generates a close enough approximation, assuming that a nucleotide at any single site cannot mutate twice.

3.6.2 Network accuracy

In this section we describe methods to determine the accuracy of the transmission network inferred by the algorithm. At each iteration of the MCMC algorithm we store the current state of the transmission network which includes the source of colonisation for each patient. Recall that the vector of sources is denoted by s where $s_i = j$ if and only if the source of colonisation for individual i is individual j. From the output of the algorithm, for each patient i we are able to construct the marginal posterior distribution of the source of colonisation, given by $s_i \mid X$ where $X = (z^{obs}, \psi, x^s)$ denotes the observed data, by simply calculating the proportion of iterations any individual is considered the be the source of colonisation for i.

For a given individual, the 'most probable source' is defined as the person who is considered to be the most frequent source, i.e. the posterior mode. The proportion of iterations that individual i is the source for individual j is the posterior probability of $s_i = j$, and for any given data set we can construct a weighted transmission network such that the edges are assigned to be the posterior probability of that transmission link.

Let $\hat{S}_i$ denote the 'most probable source' for individual i who is considered to be the most common source of colonisation, i.e. the posterior mode. The proportion of iterations that individual i is the source for individual j is the posterior probability of $s_i = j$ and for any data set we can construct a weighted transmission network such that the edges are assigned to be the posterior probability of that transmission link. Mathematically this is defined as

$$\hat{S}_i = \arg\max_y \; \Pr(s_i = y \mid z^{obs}, \psi, x^s).$$

Furthermore, if an individual i is considered to have imported the pathogen we set $s_i = -1$ in addition to $\phi_i = 1$ such that importations can be determined from the vector of sources.

Constructing a posterior probability transmission network is informative to gauge transmission routes, however this is purely a visual technique and therefore we wish to explore summary statistics to categorise the accuracy of the inferred transmission network.

Definition 3.6.1 (Source accuracy)**.** The term *source accuracy*, denoted by $\mathcal{A}_S$, is the proportion of sources such that the true source is the most probable source, or the posterior mode. Let v_p denote the set of patients who ever receive a positive test result on the ward. For any $i \in v_p$, let $\hat{S}_i$ denote the most probable source for individual i and S_i denote the true simulated source. The source accuracy is calculated as

$$\mathcal{A}_S = \frac{1}{|v_p|} \sum_{i \in v_p} \mathbb{1}_{\{\hat{S}_i = S_i\}},$$

where $|v_p|$ is the number of elements in v_p, or simply the number of patients to ever receive a positive test result. This is a straight forward estimate for the proportion of patients that are correctly inferred and is a rough indicator of the performance of the algorithm.

Definition 3.6.2 (Cluster accuracy)**.** The term *cluster accuracy*, denoted by $\mathcal{A}_C$, is the proportion of clusters in the genetic network that are correctly inferred. We define the term cluster to be the set of individuals in the same transmission tree. Recall that the transmission network (or transmission forest) may be composed of multiple disjoint transmission trees, where the root of each transmission tree corresponds to an individual who has imported the pathogen from the community.

Let $\hat{S} = (\hat{S}_1, ..., \hat{S}_n)$ denote the vector of posterior modes for the sources of colonisation and $\hat{v}_{imp} = \{i : \hat{S}_j = -1, j = 1, ..., n\}$ denote the set of individuals who are inferred to have imported the pathogen to the ward, where $\hat{d} = |\hat{v}_{imp}|$ is the number of individuals in this set and n is the number of patients in the study. Let $\hat{C} = \{\hat{C}_1, ..., \hat{C}_{\hat{d}}\}$ denote the set of inferred clusters where $\hat{C}_i$ contains the individuals in cluster i for $i = 1, ..., \hat{d}$.

Let v_{imp} denote the set of known imported individuals from simulated data, which is of size $d = |v_{imp}|$. Let $C = \{C_1, ..., C_d\}$ denote the set of true simulated clusters where C_i contains the individuals in cluster i.

An inferred cluster $\hat{C}_i$ is considered correctly inferred if there exists the same cluster in the set of true clusters C. We define two clusters $\hat{C}_i$ and C_j to be equal if and only if they have the same elements. Mathematically this is equivalent to writing $\hat{C}_i = C_j$ if and only if $\hat{C}_i \subset C_j$ and $C_j \subset \hat{C}_i$. Define the function $\Phi(\hat{C}_i \mid C)$ to equal 1 if cluster i is correctly inferred and 0 otherwise, then the cluster accuracy is calculated as

$$\mathcal{A}_S = \frac{1}{\hat{d}} \sum_{i=1}^{\hat{d}} \Phi(\hat{C}_i \mid C).$$

The notion of inferred correct clusters is motivated by the idea that it may not be possible to correctly infer the root of each cluster simply because there is not enough information in the data. In some simulated scenarios an individual may import the pathogen and later colonise other individuals, however if this imported individual tests positive at either similar or a later time to the other colonised individuals then it may be too difficult to correctly resolve the true root of the tree.

In order to account for this we require each individual in an inferred cluster to be the same as the true simulated data. The idea behind this is that genetic distances for related individuals should be small and hence informative of transmission clusters. Therefore the algorithm may infer related individuals within the cluster but may not be able to successfully determine the structure within the cluster. Intuitively this metric is informative as there lies interest in determining clusters of transmission which is useful from an infection control standpoint.

3.6.3 Parameter estimation

In order to demonstrate the performance of the algorithm, we begin by first simulating data sets with suitable parameter values motivated by the work in Cassidy et al. (2020) and attempt to infer the true values.

In our simulations we simulate outbreaks that consists of $n = 100$ patients over a period of $L = 150$ days, with an average length of stay set to $\mu_{LOS} = 7$. Furthermore, we suppose that there are $n_H = 5$ healthcare workers who ever become colonised during the study period. We impose that patients and healthcare workers are tested on admission and every $\kappa = 7$ days thereafter.

For each positive test result a genetic sequence of length $N = 2800000$ is returned and the genetic group of the sequence is sampled uniformly from $K = 3$ subtypes.

The parameter values for the simulated outbreaks were set as the following:

- $z = 0.7$
- $p = 0.05$
- $\beta = 0.03$
- $\beta_H = 0.003$
- $\lambda = 9.04 \times 10^{-9}$
- $\mu = 20.$

We simulated 1000 data sets and inferred the model parameters using the MCMC routine outlined in Algorithm 3. In the algorithm we set the maximum number of iterations to be $M = 25000$, the importation tuning parameter to be $w = 0.3$ and at the end of each iteration we update the augmented data $y = 25$ times. For the test sensitivity parameter z and importation probability p we assigned vague Beta$(1, 1)$ priors. For the patient transmission rate parameter β, healthcare worker transmission rate β_H, mutation rate parameter λ and importation distance parameter μ we assigned vague Exp(0.001) priors.

For each of the 1000 simulations we excluded the first 5000 iterations as a burn in and then thinned the remaining 20000 posterior samples by a factor of 4, resulting in 5000 samples from the posterior distribution. Furthermore, we output the colonisation times and vector of sources at each iteration which is needed to construct a weighted posterior transmission network, details of which can be found in Section 3.6.2.

As a summary statistic for the posterior distributions, we calculate the median for each parameter and plot the distribution of posterior median estimates in Figure 3.16. For each of the parameters we can see that the true simulated value has been recovered well and lies within the 95% credible intervals.

In an ideal world one would monitor every single parameter, including all colonisation times, but this approach is infeasible due to the large number of parameters. Therefore to demonstrate efficient exploration of the Markov chain we assess the relative error of the sum of colonisation times. Consider a specific simulated data set and let T denote the sum of colonisation times and $\hat{T}$ denote the posterior median estimate of T calculated from the MCMC output. The relative error is then defined as

$$\epsilon = \frac{T - \hat{T}}{T},$$

and the distribution of errors from simulated data can be found in Figure 3.17.

Lastly for each data set we also fit the transmission and observational model without genetic data. The idea behind this is to compare the accuracy of transmission network that is inferred with and without the genetic data. For any network accuracy metric, the difference between the full model and the simpler model provides an indication of the valuation of genetic data. The distribution for the two network accuracy metrics, specifically source and cluster improvement, can be found in Figure 3.18. Out of the 1000 simulated data sets, 842 (84.2%) demonstrated an improvement in source accuracy and 959 (95.9%) showed an improvement in cluster accuracy.

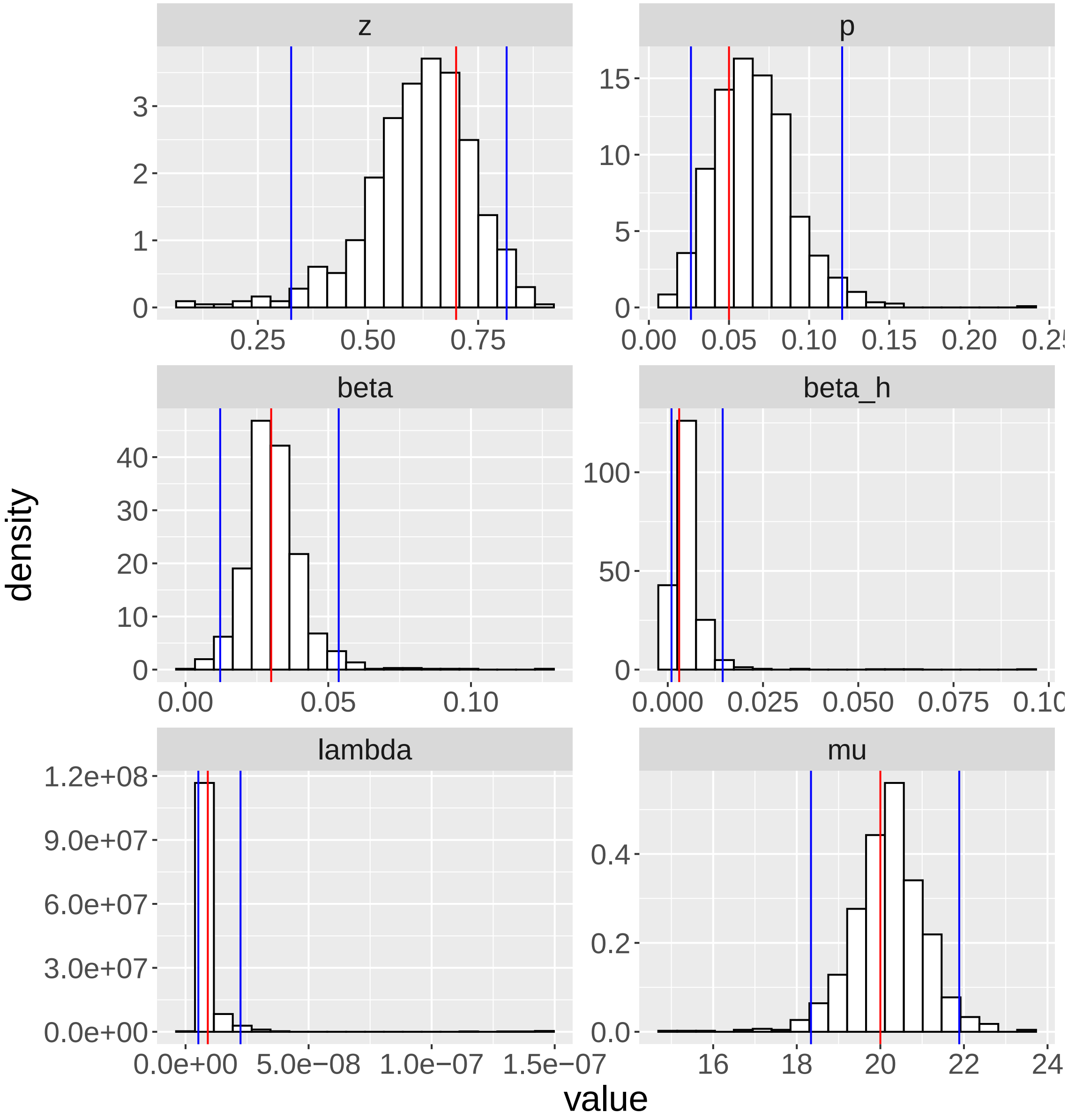


Figure 3.16: The distributions of the posterior median estimates for z (top left), p (top right), β (middle left), β_H (middle right), λ (bottom left) and μ (bottom right) for the simulated data sets. The blue lines are the 95% (equal-tailed) credible intervals and the red line is the true simulated value.

3.6.4 MCMC diagnostics

In this section we demonstrate some tools used to check that the algorithm is sampling from a stationary distribution and there is sufficient mixing. For each of the simulated

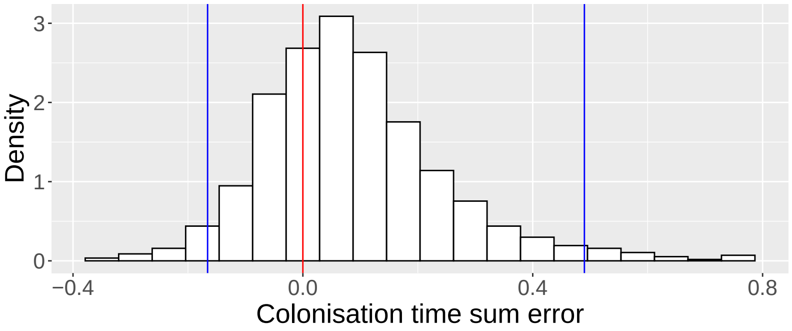


Figure 3.17: The distribution of the difference between the true sum of colonisation times and the posterior median of the sum of the colonisation times. The blue lines are the 95% (equal-tailed) credible intervals the red line is the true value.

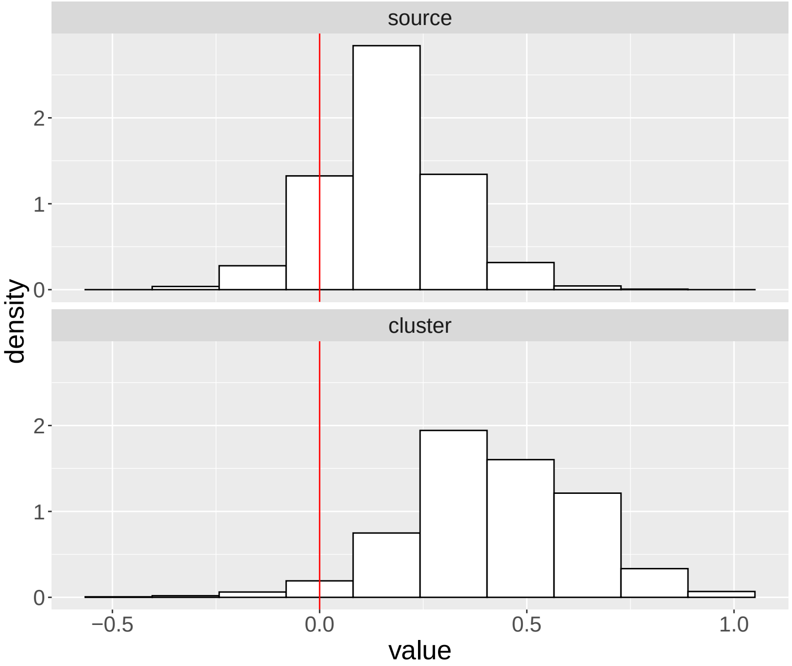


Figure 3.18: Transmission network accuracy improvement on simulated data for two metrics, source accuracy (above) and cluster accuracy (below). Here ‘improvement’ is defined as the difference between the accuracy with and without genetic data.

data sets we check the parameter trace plots as a method to check convergence. The parameter trace plots for a specific data set can be found in Figure 3.19. For nearly all parameters the truth is well recovered by the algorithm, apart from β_H which is slightly overestimated however the true value still lies within the 95% credible interval. Furthermore, we summarise the mean and median effective sample size (ESS) of all parameters and the acceptance rates of the Metropolis-Hastings updates in Table 3.3.

	Acceptance rate		ESS	
	Mean	Median	Mean	Median
z	–	–	251.38	180.46
p	–	–	2572.23	2367.96
β	0.44	0.42	414.60	354.55
β_H	0.39	0.45	441.20	369.87
λ	0.30	0.29	1156.93	1186.81
μ	–	–	952.26	853.49

Table 3.3: Mean and median acceptance rate and effective sample size (ESS) for the 1000 simulated data sets.

Another commonly used technique to evaluate the convergence of MCMC schemes is to start multiple chains in parallel from overdispersed starting points, with the idea that in theory the chains should converge to the same stationary distribution. Figure A.19 shows the trace plots of four parallel chains of the parameters updated by Metropolis-Hastings steps, where visually each eventually converges to the same stationary distribution.

Finally, we may evaluate the other features of the MCMC when ran in parallel, such as the trace plot of the log-likelihood and also the sum of colonisation times (Figure A.20).

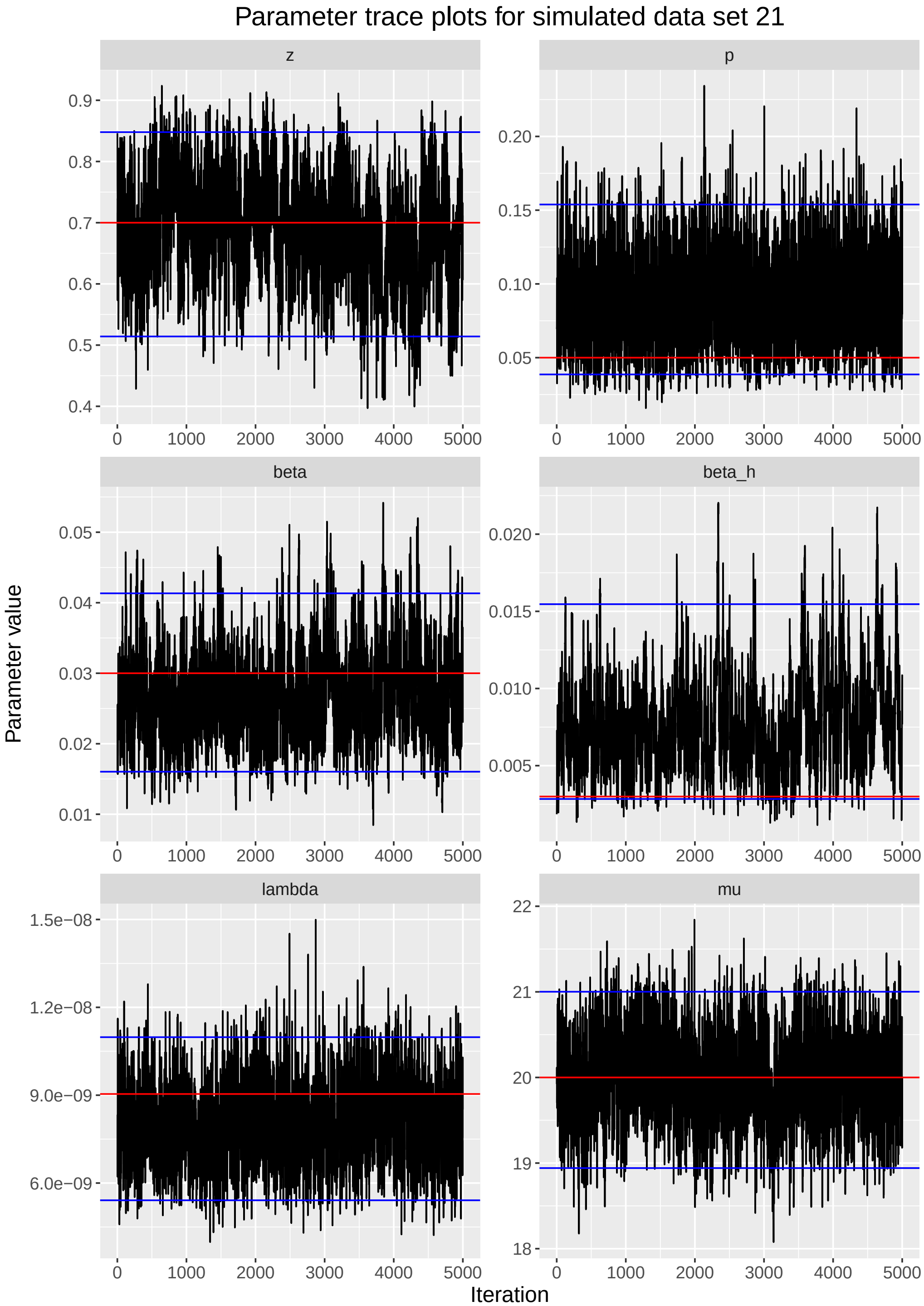


Figure 3.19: Parameter trace plots of the posterior samples for simulated data set 21. The line in red indicates the true simulated value and the blue lines are the 95% (equal-tailed) credible intervals.

3.6.5 Sensitivity analysis

In this section we provide a sensitivity analysis to evaluate the output of the model under various parameter choices. We wish to vary the parameters which are representative of plausible scenarios and attempt to recover the true simulated values.

For each parameter we choose five values, two of which are lower and two are higher than parameter values used in Section 3.6.3. Specifically these parameter values are given by

- $z = (0.25, 0.5, 0.7, 0.75, 1)$
- $p = (0.01, 0.025, 0.05, 0.15, 0.3)$
- $\beta = (0.01, 0.02, 0.03, 0.04, 0.05)$
- $\beta_H = (0.001, 0.002, 0.003, 0.004, 0.005)$
- $\lambda = (1.80, 4.52, 9.04, 18.08, 45.20) \times 10^{-9}$
- $\mu = (5, 10, 20, 30, 35)$.

For each parameter that is varied, all other parameters are set to the 'baseline' values which are parameter choices motivated by the results in Cassidy et al. (2020). For each of the 30 parameter choices we simulate 100 data sets and infer the parameters using the MCMC algorithm (Algorithm 3) for $M = 25000$ iterations and update the augmented data $y = 10$ times. The first 5000 samples are excluded as a burn-in and the remaining samples are thinned by a factor of 4, leaving a total of 5000 posterior samples. We then plot the medians for each parameter that is varied.

Finally, we also fit the transmission and observational model without genetic data and compare the change in source and cluster accuracy.

Results when varying the test sensitivity parameter z For each of the parameter choices of z the true simulated value is slightly underestimated but generally well recovered for $z = 0.25, 0.5, 0.7, 0.75$, with the true value in the interquartile range of the posterior median estimates (Figure A.1). For $z = 1$ the parameter is underestimated which is perhaps not surprising since we are sampling from a Beta distribution with an approximate median $\frac{TP-1/3}{TP+FN-2/3}$ where TP and FN are the number true positives and false negatives respectively. Recall that the number of true positives are fixed, but the number of false negatives are calculated using the current state of the unobserved data augmented by the algorithm and will constantly be updated. Therefore due to the uncertainty around detection it is unlikely our estimates will recover the parameter value of 1.

There does not appear to be any significant patterns with the source accuracy improvement when varying z (Figure A.2), however each scenario shows an improvement for determining the source of colonisation when including genetic data.

On the other hand there appears to be a clear positive association between the simulated parameter values and cluster accuracy improvement (Figure A.3), as z increases so does the cluster accuracy. This is explained by the fact that more positive tests result in an increase of observed genetic sequences (and hence genetic distances) and less uncertainty in the true time of colonisation. Consequently the algorithm is able to successfully determine clusters of transmission, which intuitively is achieved by grouping similar individuals that have small pairwise genetic distances.

Results when varying the importation probability parameter p For the importation probability p the true simulated parameter is well recovered for $p = 0.05, 0.15, 0.3$ but is slightly underestimated for $p = 0.01, 0.025$ (Figure A.4). This is likely due to the fact that for low values of p, there will less imported cases and hence less overall transmission.

For both the source and cluster accuracy (Figures A.5-A.6) there appears to be a negative association between the simulated parameter values and accuracy and in general as p increases, accuracy for both metrics decreases. This is not surprising as increasing the importation probability p will result in more imported cases, and hence more transmission overall and consequently it may be difficult to infer the correct source for each individual. Since imported individuals are the root of transmission clusters, an increase in the importation probability will increase the number of transmission clusters and hence it may be harder to correctly infer each of them.

Results when varying the patient transmission rate parameter β The algorithm tends to successfully recover the true simulated values of β for each scenario (Figure A.7). The source accuracy tends to decrease as the transmission rate increases (Figure A.8) which makes sense as there will be more transmission overall and hence it may be difficult to correctly infer sources of colonisation.

The cluster accuracy does not appear to significantly change when varying the parameter β (Figure A.9). This demonstrates that if the number of transmission clusters remain roughly the same, the algorithm is able to successfully infer individuals within each cluster, even if the transmission rate is high and therefore a higher average number of individuals within each cluster.

Results when varying the healthcare worker transmission rate parameter β_H The algorithm tends to slightly overestimate the true simulated values for each parameter choice (Figure A.10), with the true value on the edge of the interquartile range of the posterior median estimates. This could potentially be a consequence of low levels of healthcare worker transmission and therefore identifiability issues since there may not be enough information in the data to estimate β_H well.

The source accuracy tends to increase when β_H increases (Figure A.11), in contrast to the patient transmission rate parameter β where increasing the parameter value resulted in a decrease of source accuracy. This is potentially because although there is more healthcare worker transmission, it is generally easier to choose a healthcare worker as the source of colonisation since they are assumed to exert a constant colonisation pressure and are nearly always able to colonise patients.

The cluster accuracy remains fairly consistent, with $\beta_H = 0.005$ indicating the best overall improvement (Figure A.12). Similar to the cluster accuracy for β, the average number of clusters between scenarios remains the same and the algorithm is able to successfully infer individuals within each cluster.

Results when varying the mutation rate parameter λ The true simulated values are generally well recovered by the algorithm (Figure A.13). There are however many outliers (indicated by the black dots), which are a consequence of the algorithm incorrectly inferring sources of colonisation and hence individuals that are unrelated (with higher than expected pairwise distances) are considered to be related, which increases the estimated mutation rate. This is essentially a mixing issue and in an ideal world these errors can be mitigated with more iterations of the MCMC algorithm and more updates each iteration for the augmented data.

There does not appear to be any significant trend for the source accuracy for the various values of λ (Figure A.14). On the other hand there appears to be a positive association with cluster accuracy and the mutation rate λ (Figure A.15). Intuitively this can be explained by the fact that closely related individuals will have small genetic distances and therefore should be easier to infer, however we would also expect the source accuracy to increase which is not the case here.

Results when varying the average importation distance parameter μ In general the importation distance parameter μ is generally well recovered, except for when $\mu = 5$ in which case the parameter is slightly overestimated (Figure A.16). There does not appear to be any obvious patterns for the source accuracy (Figure A.17), however there is a positive association between cluster accuracy and the true simulated values of μ (Figure A.18). This is perhaps not surprising as individuals in distinct transmission clusters will have larger pairwise distances as μ increases and consequently it is easier to identify similar individuals within the same transmission cluster.

3.6.6 Model misspecification

In this section we wish to investigate the robustness of the model through simulation-based techniques. In particular we wish to examine model performance under misspecification, that is where the data generating process for simulation is different to that assumed in the inference procedure. The real world biological mechanisms that govern processes such as mutation are significantly more complex than that assumed in most mathematical models. Therefore, it would be of interest to examine the performance of our inference algorithm when the genetic data are generated from models that are (i) similar and (ii) significantly different to the JC69 model.

The probability distributions derived in Chapter 2 assume that each nucleotide site evolves independently of all other sites and the parameters governing the mutation process are equal. In general this is a large oversimplification of the mutation process as it well known that nucleotides mutate at varying rates, for example there are interactions between neighbouring sites (Benzer, 1961). We now introduce the notion of variable mutation rates across sites for the mutation models described in Chapter 2 (see (Yang, 2014, Chapter 1) for more details).

One method to incorporate rate heterogeneity between sites is to assume that the parameters governing the mutation process, θ, are now random variables with probability density or mass function denoted by $g(\theta \mid \xi)$ where ξ are the associated parameters.

More precisely, we simulate genetic distances from the probability mass function $f(\mathbf{d} \mid \lambda, \mathcal{G})$ in Theorem 2.7.3 where the mutation rate is a random variable such that $\lambda \sim g$. In this scenario we are directly violating the assumption that distance matrices are identically distributed in Equation 2.4. Assume that the mutation rate λ follows a Gamma distribution with shape α and rate γ and density function

$$g(\lambda \mid \alpha, \gamma) = \frac{\lambda^{\alpha-1} e^{-\gamma\lambda} \gamma^{\alpha}}{\Gamma(\alpha)}, \quad \text{for } \lambda > 0 \quad \alpha, \gamma > 0.$$

We wish to consider the scenarios where g generates mutation rates that are similar to the JC model and those that are different. For simplicity we fix the shape parameter to $\alpha = 1$ in which case we obtain $\mathbb{E}[\lambda] = 1/\gamma$ and $\mathrm{Var}(\lambda) = 1/\gamma^2$. In the first instance we may choose $\gamma = 1/9.04 \times 10^{-9}$ so that the mean mutation rate is the equal to the mutation rate value chosen for the simulation study described in Section 3.6.3 and denote this model as M_1.

Next, we wish to choose γ such that the mutation rates (and hence genetic distances) are significantly different to those previously considered. One choice is to choose $\gamma = 1/9.04 \times 10^{-9} \times 1/1000$ such that the mean rate is 1000 times higher than previously assumed and denote this model as M_2.

We repeat the procedure described in Section 3.6.1 however now we simulate the genetic data from model M_1 or M_2. We repeat this process 100 times for each model and then summarise the MCMC output by the posterior median estimates (Figure 3.20). For each of the simulations we fix the genome size to $N = 8000$ and the parameters to the following,

- $z = 0.7$
- $p = 0.05$
- $\beta = 0.04$
- $\beta_H = 0.003$
- $\lambda = 9.04 \times 10^{-9}$
- $\mu = 5$.

In order to simulate the genetic data from the variable rate substitution model, we replace the distance calculation between directly connected nodes in Equation (3.8) with $D_{ij} = \sum_{k=1}^{N} d_{ij}^{[k]}$, where

$$d_{ij}^{[k]} \sim \mathrm{Bernoulli}(q(t_j - t_i \mid \lambda_k)),$$

and λ_k denotes the mutation rate for the kth nucleotide and $\lambda_k \sim \text{Gamma}(1, \gamma)$. In other words, we explicitly model the mutation (or lack thereof) of each nucleotide conditional on a random mutation rate which has been drawn from a Gamma distribution with rate γ, dependent on models M_1 or M_2.

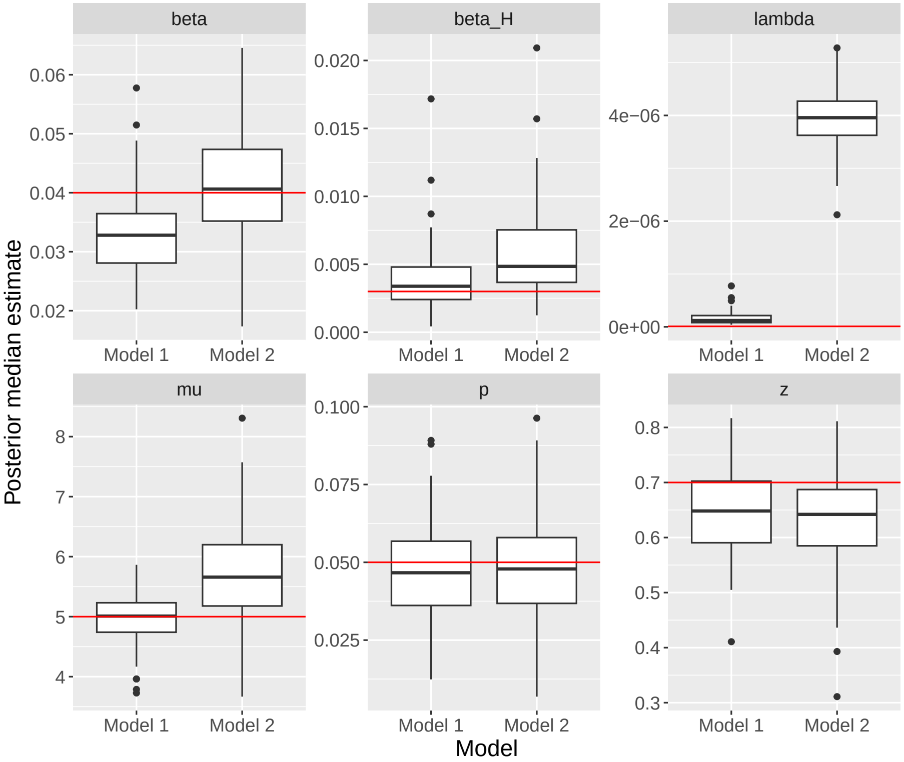


Figure 3.20: Posterior median estimates for 100 simulated data sets under model M_1 and M_2 for each of the model parameters ρ. The red line indicates the true simulated parameter values.

In general we find that the parameters governing the transmission process are reasonably well recovered, indicating that the model performs well under misspecification of the genetic process. The posterior median estimates for the mutation rate parameter λ is similar to the truth for model M_1, and significantly higher for M_2 which is expected due to the higher mean mutation rate.

3.6.7 Checking the mutation assumption

In the genetic distance imputation and simulation methods we explicitly assume that a nucleotide cannot mutate more than once at a single site, and consequently we wish to determine if this assumption is reasonable.

Consider a genetic sequence of N nucleotides evolving under the JC69 mutation model under a known genetic network $\mathcal{G} = (V, E)$. We assume that the network is fully connected and there are $k+1$ vertices and k edges. Each vertex $v \in V(\mathcal{G})$ has a corresponding time at which the sequence is sampled denoted by τ_v.

Under the JC69 mutation model, the number of mutations can be described by a Poisson process. Hence the number of mutations for a single nucleotide in some time interval $[0, t]$ can be described by a Poisson random variable with mean λt where λ is the mutation rate. Let X denote the number of mutations for a single nucleotide such that

$$X \sim \text{Po}(\lambda t).$$

Let Y be the indicator function of the event that at least one nucleotide mutates more than once. Since we have N independent Poisson processes, $Y = 0$ if and only if each of the N independent Poisson processes that govern the mutations have either 0 or 1 mutations in $[0, t]$. In other words,

$$\begin{aligned}\Pr(Y = 0) &= (\Pr(X = 0) + \Pr(X = 1))^N \\ &= e^{-N\lambda t}(1 + \lambda t)^N,\end{aligned}$$

hence it follows that

$$\Pr(Y = 1) = 1 - e^{-N\lambda t}(1 + \lambda t)^N.$$

The evolution of MRSA in a hospital environment was studied in Harris et al. (2010) and the mutation rate of the pathogen was estimated to be 9.04×10^{-9} per site per day, 95% confidence interval $(6.85 \times 10^{-9}, 1.10 \times 10^{-8})$. Furthermore, previous studies have calculated the genome length of MRSA to be approximately $N = 2800000$ (Ali et al., 2019).

If we wish consider the mutation rate λ and size of the genetic sequences N to be fixed, then we wish to find the largest value t such that the probability of any nucleotide mutates more than once is greater than some predefined threshold, α say.

Assuming a threshold of $\alpha = 0.001$ and sequence length of $N = 2800000$, we can plot the probability as a function of time for various mutation rates λ (Figure 3.21). The time at which $\Pr(Y = 1) > \alpha$ is calculated to be $t = 2957$.

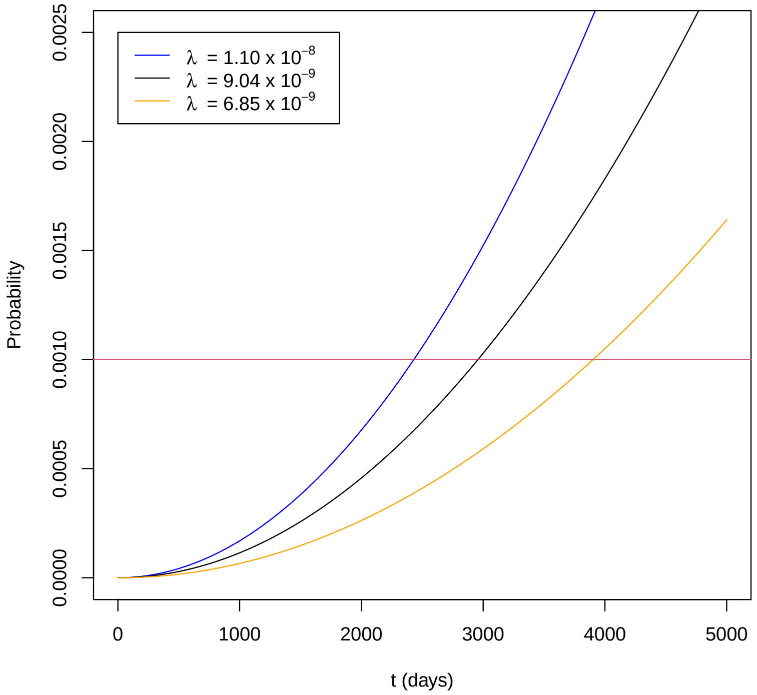


Figure 3.21: The probability that any nucleotide mutates more than once against time, assuming sequences of length $N = 2800000$ and that each nucleotide mutates according to some constant rate λ. The three increasing lines correspond to low (orange), middle (black) and high (blue) mutation rates. The red line indicates the assumed threshold of $\alpha = 0.001$.

Note that t has been rounded down so that it is consistent with the discrete time model presented in this chapter. In the context of our problem, t is the sum of all branch lengths in the genetic network $\mathcal{G}$, which explicitly is $\sum_{(i,j)\in E} |\tau_j - \tau_i|$. For most scenarios this value is unlikely to be very large, however for larger data sets under a longer study period this assumption may not hold.

3.7 Analysis of the Brighton data set

In this section we analyse the Brighton data set first introduced in Section 3.2. In order to investigate if there are any significant differences between MRSA and MSSA we provide an analysis into the separate data sources and also the combined results.

The genetic subtype groups η were calculated pre-analysis using the group by threshold method (Section 2.7.3.1) with threshold of $\zeta = 40$ which is consistent with the analysis in Price et al. (2017).

We ran the algorithm for 250,000 iterations, set the importation tuning parameter to be $w = 0.3$ and updated the augmented data 50 times at each iteration. The first 50,000 iterations are excluded as a burn in and the output is then thinned by a factor of 20, resulting in 10,000 remaining posterior samples.

For the test sensitivity parameter z and importation probability p we assigned uninformative $\text{Beta}(1, 1)$ priors. For the patient transmission rate parameter β, healthcare worker transmission rate β_H and mutation rate parameter λ we assigned uninformative Exp(0.001) priors. For importation distance parameter μ we assigned a informative $\text{Gamma}(0.001, 0.001)$ prior.

The MCMC algorithm was coded in the C++ language and was run on a server with an Intel (R) Xeon (R), E5-2450 2.1GHz CPU. Parts of the algorithm, specifically the imputation of genetic distances and calculation of the genetic network, were implemented by parallel processing using OpenMP (Dagum and Menon, 1998).

We repeated this five times for the three separate analyses (MRSA, MSSA and combined) to give 15 outputs. The separate analyses were done to compare inferred parameter estimates between the data sources, and the replications were done in order to determine if the chains do in fact converge to the same stationary distributions.

3.7.1 Results

From the box plots of the posterior samples (Figure 3.22) it is immediately obvious that the estimates when considering only MRSA are significantly different from the MRSA or combined data sources. Furthermore, there is much higher uncertainty with the parameter estimates for MRSA compared with the other two data sources which is likely due to the fact that MRSA data constitutes a small percentage of the combined data.

Specifically there are 34 distinct patients who ever test positive for MRSA out of 388 for the combined data (8.76%) and 169 genetic sequences out of 1308 for the combined data set (12.9%). Furthermore, it is not surprising that the estimates for MSSA and combined are similar since the majority of the combined data are MSSA organisms.

From the posterior box plots for each data source (Figure 3.22), the test sensitivity z and importation probability p are significantly lower in the model that contains only MRSA samples compared to MSSA and combined data sources. Furthermore, both the patient and healthcare worker transmission rates, denoted by β and β_H respectively, are higher in the model with only MRSA compared to the other two. In the cases above the significantly different estimates and larger uncertainty is a likely consequence of a lack of observed data, with only 34 patients out of 1919 (1.77%) to become colonised, there simply is not enough information to draw any meaningful inference.

The mutation rate λ is similar across all three analyses which is expected, however the average importation distance μ when considering MRSA only is estimated to be approximately three times higher than the other data sources and reasons for why are discussed later in Section 3.8

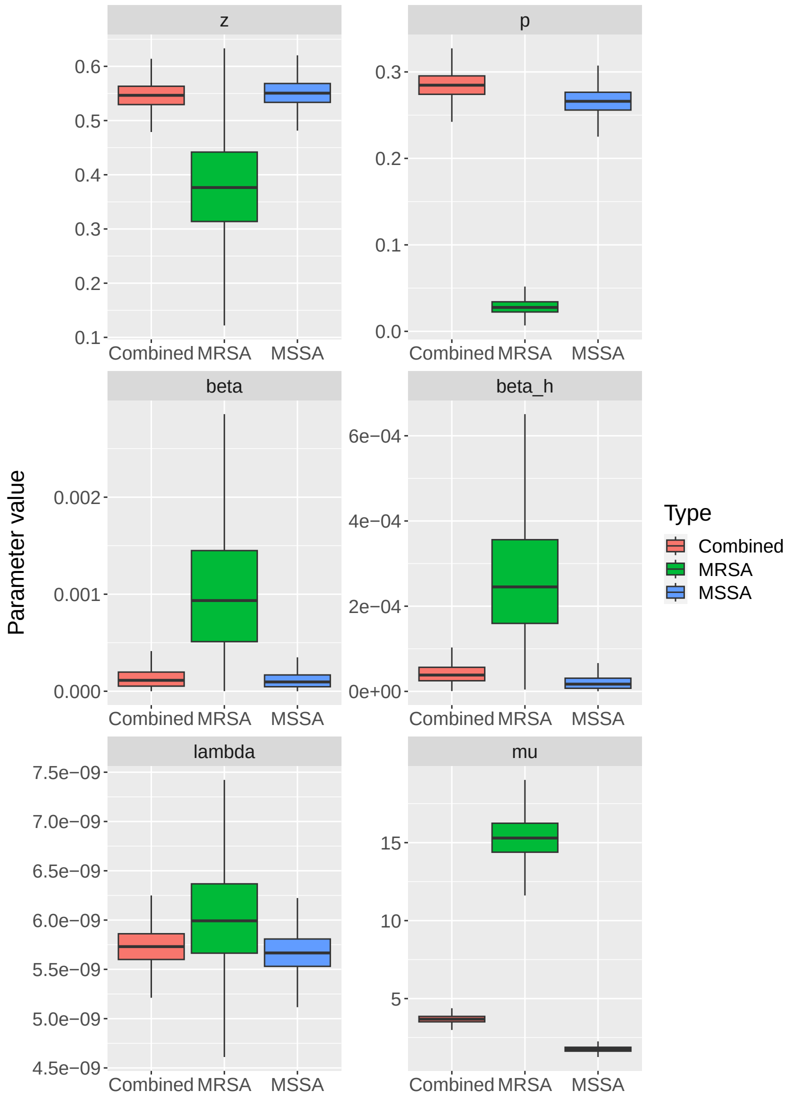


Figure 3.22: Posterior distributions of each parameter when fitting the full model to the Brighton data set where the genetic data is composed of either MSSA (blue), MRSA (green) or combined (red). The parameters are explicitly the test sensitivity z (top left), importation probability p (top right), patient transmission rate β (middle left), β_H (middle right), λ (bottom left) and μ (bottom right).

Since the results of the combined data set are similar to the data set containing only MSSA organisms, we shall now turn out attention to the combined data set. The

Parameter	MRSA		MSSA		Combined	
z	0.381	(0.202,0.569)	0.552	(0.501,0.601)	0.546	(0.496,0.594)
p	0.029	(0.015,0.053)	0.266	(0.238,0.297)	0.287	(0.257,0.32)
$\beta \times 10^{-4}$	10.9	(1.17, 27.9)	1.22	(0.043,3.79)	1.54	(0.058,4.66)
$\beta_H \times 10^{-5}$	27.2	(5.95,67.7)	2.17	(0.09,6.6)	2.94	(0.166,8.29)
$\lambda \times 10^{-9}$	6.02	(5.05,7.08)	5.65	(5.26,6.06)	5.96	(5.32,6.07)
μ	15.4	(12.7,18.2)	1.76	(1.4,2.14)	3.63	(3.15,4.14)

Table 3.4: Posterior mean estimates for the model parameters along with 95% (equal-tailed) credible intervals.

posterior distributions for each parameter can be found in Figure 3.23 and the table of posterior mean estimates can be found in Table 3.4.

3.7.2 Inferred transmission network

From the output of the MCMC algorithm we are able to construct a weighted posterior transmission network (Section 3.6.2). Recall that a transmission network contains all colonised individuals in the outbreak, however this may be difficult to interpret if we plot the complete transmission tree since there are 501 distinct individuals to ever test positive for the pathogen. For clarity we shall only display patient acquisitions in the inferred transmission network (Figure 3.24), from which there are three patient acquisitions.

The link between healthcare HCW 9 and patient 1831 is explained by the having distances of 0 between a sequence observed on day 368 for the healthcare worker and three sequences for the patient on days $367, 367$ and 371 respectively.

The link between HCW 71 and patient 1672 is explained by having a distance of 1 between a sequence sampled on day 419 for the healthcare worker and two sequences for the patient on days 343 and 346 respectively.

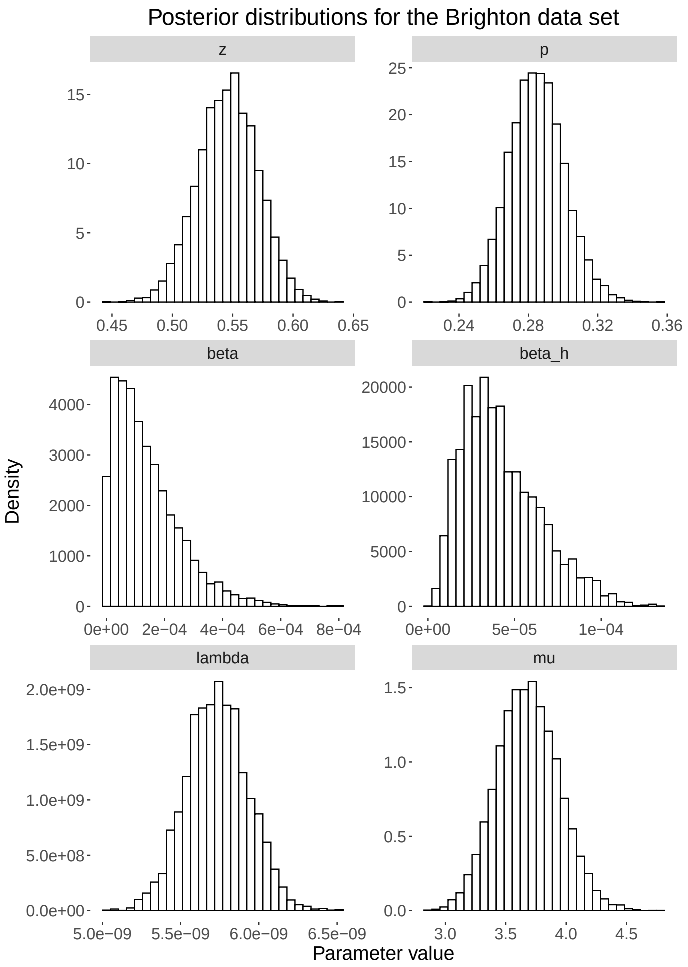


Figure 3.23: Posterior distributions of each parameter when fitting the full model to the Brighton data set where the genetic data contains both MRSA and MSSA organisms. The parameters are explicitly the test sensitivity z (top left), importation probability p (top right), patient transmission rate β (middle left), healthcare worker transmission rate β_H (middle right), mutation rate λ (bottom left) and average importation distance μ (bottom right).

The link between patients 1693 and 1748 can be explained by a distance of 0 between a sequence observed on day 338 for patient 1693 and on day 339 for patient 1748.

The algorithm has found transmission pairs which are considered likely by the model, supported by small pairwise distances indicating similar genetic sequences across the individuals. On the other hand, from the exploratory analysis there are 22 putative

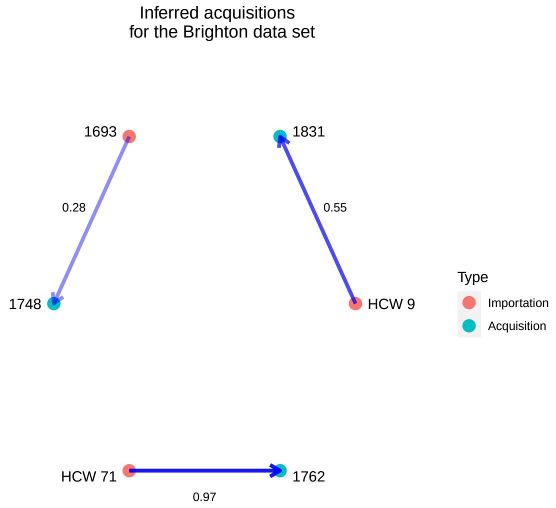


Figure 3.24: Inferred transmission network for the Brighton data set. Each patient is either an importation (red) or acquisition (blue). Individual 1963 is assumed to be the source of colonisation for individual 1748 with posterior probability 0.28. Healthcare worker 9 is assumed to be the source of colonisation for individual 1831 with posterior probability 0.55. Healthcare worker 71 is assumed to be the source of colonisation for individual 1762 with posterior probability 0.97.

transmission pairs that one may hope to recover by the algorithm, possible reasons for why are discussed in Section 3.8.

3.7.3 MCMC diagnostics

In order to ensure we are sampling from the target density in Equation (3.6), we shall utilise diagnostic tools to check the quality of the samples generated by the algorithm. We specifically are interested in the convergence and mixing of the Markov chains, which may be determined by visual inspection of trace plots. We may also start several chains to run in parallel and check whether the chains converge to the same mode.

We ran the Brighton analysis for each of the data sources (MRSA, MSSA and combined) for a total of five times each. From the posterior distributions in Figures (B.1)-(B.6) and trace plots in Figure B.7 we can see that the posterior distributions are similar within each data source and hence the chains have converged to the same stationary distribution by visual inspection. Furthermore we calculate the multivariate potential scale reduction factor for each of the data sources and find these values to be 1.00, 1.01 and 1.10 for the MRSA, MSSA and combined chains respectively (Brooks and Gelman, 1998). Finally a summary of the effective sample size (ESS) for each parameter is displayed in Table 3.5 and acceptance rates are displayed in Table 3.6.

	z	p	β	β_H	λ	μ
MSSA 1	1756.98	1531.70	360.51	176.62	703.74	6656.66
MSSA 2	1815.36	1529.17	428.12	194.26	916.43	7518.18
MSSA 3	1837.30	1546.44	442.25	178.15	1191.23	7122.54
MSSA 4	1818.20	1514.33	342.73	130.28	647.77	6920.98
MSSA 5	1578.93	1509.07	314.28	119.83	635.17	7049.84
MRSA 1	1497.15	1368.50	1050.94	684.84	3746.48	9484.13
MRSA 2	1223.16	1186.96	1000.52	912.17	3706.07	9788.43
MRSA 3	1177.08	882.05	699.22	477.12	2836.83	9655.08
MRSA 4	1308.01	1161.57	1135.44	1023.03	3377.99	9670.17
MRSA 5	1251.23	1183.93	1154.51	762.18	2766.21	9403.57
Combined 1	1302.40	969.27	228.33	163.03	844.01	1409.69
Combined 2	1537.11	1225.95	256.10	250.45	653.49	1727.16
Combined 3	1304.21	1029.36	286.68	158.17	712.07	1668.81
Combined 4	1215.46	1010.30	235.00	161.74	543.23	1303.04
Combined 5	1649.17	1266.44	188.42	190.21	556.04	2154.60

Table 3.5: Effective sample size for each parameter for each of the MCMC outputs.

	β	β_H	λ
MSSA 1	0.61	0.15	0.61
MSSA 2	0.62	0.16	0.62
MSSA 3	0.58	0.16	0.61
MSSA 4	0.61	0.15	0.60
MSSA 5	0.61	0.15	0.61
MRSA 1	0.32	0.76	0.91
MRSA 2	0.31	0.76	0.90
MRSA 3	0.31	0.76	0.90
MRSA 4	0.30	0.75	0.90
MRSA 5	0.31	0.76	0.91
Combined 1	0.62	0.17	0.58
Combined 2	0.62	0.22	0.59
Combined 3	0.66	0.17	0.59
Combined 4	0.67	0.16	0.59
Combined 5	0.63	0.18	0.60

Table 3.6: Acceptance rates for the parameters updated by Metropolis-Hastings steps for each of the MCMC outputs.

We may also check the trace plot of the log-likelihood (Figure B.8) without a burn in and observe that the chains start in regions of low likelihood and increase until convergence. Finally, due to the large amount of data augmentation and therefore parameters inferred in the model, it is necessary to check that the chains are well mixed and are able to efficiently explore the state space. Summary trace plots for the augmented data can be found in Figure B.9 which include the sum of colonisation times and the number of patients who have been added as colonised by the algorithm.

3.7.4 Informative priors

We also wish to consider the scenario where we use more informative priors for the importation distance parameter μ. The rationale behind this is that we would expect this parameter to be reasonably high.

Recall that the importation distance parameter μ describes the average pairwise distance between genetic sequences of the same subtype that have been imported by the community. In general we would expect this value to be higher than values that could be explained by direct mutation. Recall that in Young et al. (2012) the authors estimate the evolution of a bacterial population of *S. aureus* to be a rate of 2.72 mutations per megabase per year, therefore we would expect the average distance to be higher than this.

We assume *a priori* that $\mu \sim \text{Gamma}(10, 1)$ such that $\mathbb{E}[\mu] = 10$ and the variance is large enough to cover a wide range of values. A comparison of the inferred posterior distributions with and without informative priors is in Figure 3.25.

Finally, the inferred transmission network with a stronger prior for μ is shown in Figure 3.26. In general we see more patient acquisitions, reasons for this are discussed in Section 3.8. Reassuringly, we recover the transmission links previously found in Figure 3.24 and also recover 12 of the 21 putative transmission links one would hope to see from the exploratory analysis (Section 3.2) which is summarised in Table 3.7.

3.7.5 Model assessment

In this section we shall assess the model fit to the Brighton data, which in a Bayesian framework can be naturally done via *posterior predictive checking*. The purpose of posterior predictive checking is that replicated data generated under the model should look similar to the observed data (Gelman et al., 2013). In other words, the observed data should look plausible under the posterior predictive distribution, which

Link	Posterior probability
HCW 9 → 1831	0.98658
HCW 74 → 1422	0.89184
HCW 71 → 1671	0.72868
HCW 71 → 1422	0.85040
HCW 8 → 1615	0.95104
1692 → 1578	0.95926
1704 → 1810	0.91624
1597 → 1631	0.76562
1746 → 1779	1
HCW 84 → 1519	0.59984
1693 → 1748	0.66602
1706 → 1468	0.17322
1468 → 1706	0.59522

Table 3.7: Inferred transmission links from the putative transmission pairs and the posterior probability of each link.

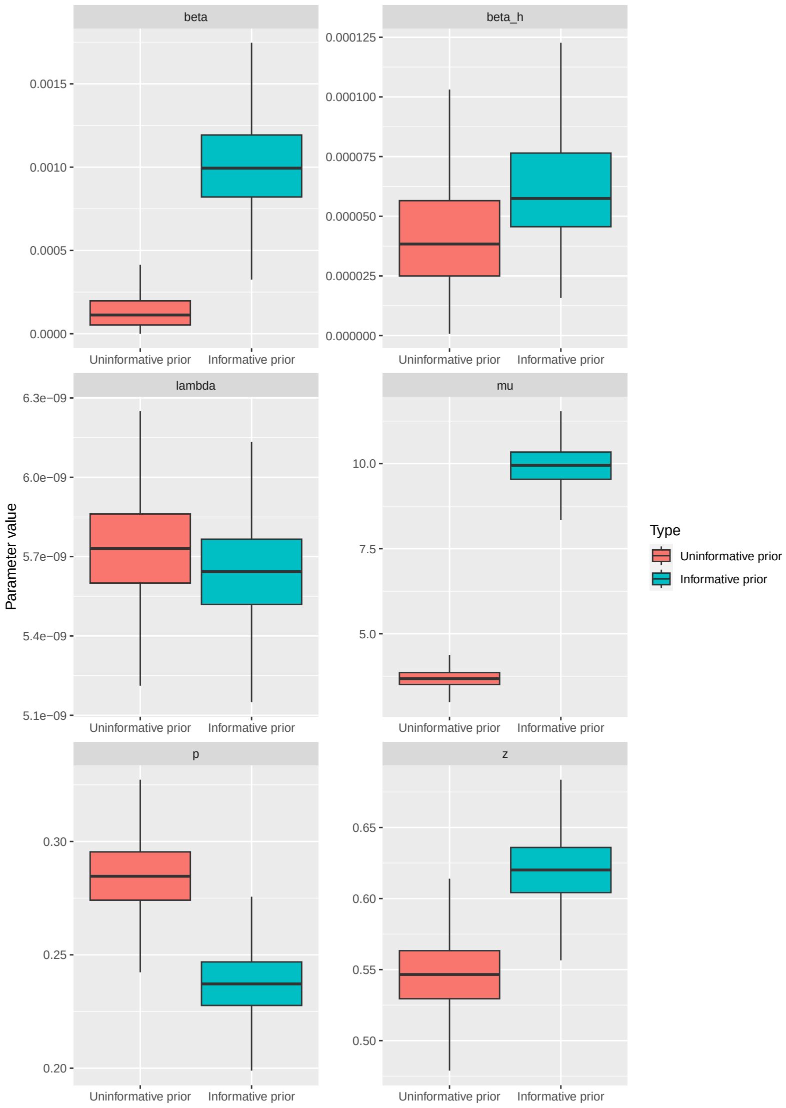


Figure 3.25: Comparison of posterior estimates with (blue) and without (red) informative priors. The parameters are explicitly the test sensitivity z (top left), importation probability p (top right), patient transmission rate β (middle left), healthcare worker transmission rate β_H (middle right), mutation rate λ (bottom left) and average importation distance μ (bottom right).

is defined as follows.

Let y be the observed data and θ be the vector of model parameters and define y^{rep} to be data that has been generated (or replicated) by the model. Then the posterior

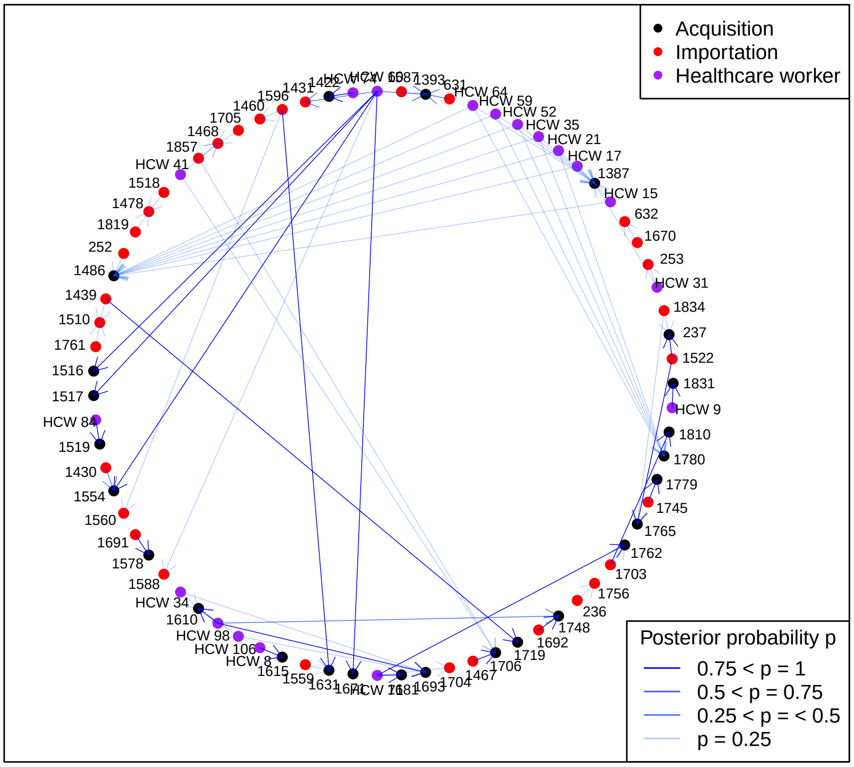


Figure 3.26: Inferred transmission network for the Brighton data set. Each patient is either an importation (red) or acquisition (blue) and healthcare workers are denoted by purple.

predictive distribution (PPD) is defined by

$$\pi(y^{\text{rep}} \mid y) = \int \pi(y^{\text{rep}} \mid \theta)\pi(\theta \mid y)d\theta.$$

In order to measure the discrepancy between the data and model we define the notion of *test quantities* which are aspects of the data we wish to check. A test quantity, $T(y, \theta)$ is a scalar statistic of some feature of the data which explicitly depends on the data and model parameters.

The general idea of posterior predictive checking is as follows. From the MCMC output we have M samples from the posterior distribution of the model parameters, denoted by $\theta_1, ..., \theta_M$. We then simulate N data sets from the model, denoted by $y_1^{\text{rep}}, ..., y_N^{\text{rep}}$, from the distribution $y^{\text{rep}} \mid \theta$ where the simulated model parameters are random samples from the posterior distribution $\theta \mid y$. We then compare the observed test statistic $T(y, \theta)$ to the distribution of the test statistic under replicated data $T(y^{\text{rep}}, \theta)$. The Bayesian p-value p^B, is then defined as the probability that the replicated data is more extreme than the observed data, measured by the test quantity:

$$p^B = \Pr(T(y^{\text{rep}}, \theta) \geq T(y, \theta)|y).$$

Extreme p^B values that are close to 0 or 1 are evidence for poor model fit, whereas if the model is true or close to true, the posterior predictive p-value will almost certainly be very close to 0.5 (Gelman, 2013).

3.7.5.1 Epidemic model assessment

Similar to the work in Cassidy (2019) and Worby et al. (2016) we use three summary statistics that capture key features of the transmission process, specifically they are (i) the total number of positive tests over the study period, (ii) the number of patients to ever test positive on the ward and (iii) the number of colonised patients in the ward over time.

From the posterior predictive distributions (Figure 3.27) it is clear that the true observed number of positive tests and patients to ever test positive lies within the 95% credible intervals. Furthermore, looking at the number of patients colonised in the ward over time, the observed value in general lies within the 95% credible intervals. Explicitly we may calculate the proportion of days where the true observed $C(t)$ lies within the credible interval, which in this case is 96%. All of these posterior predictive distributions indicate that after simulating data from the model, using parameters inferred by the algorithm, we obtain data sets that are similar to the observed data and hence there is no evidence of lack of fit.

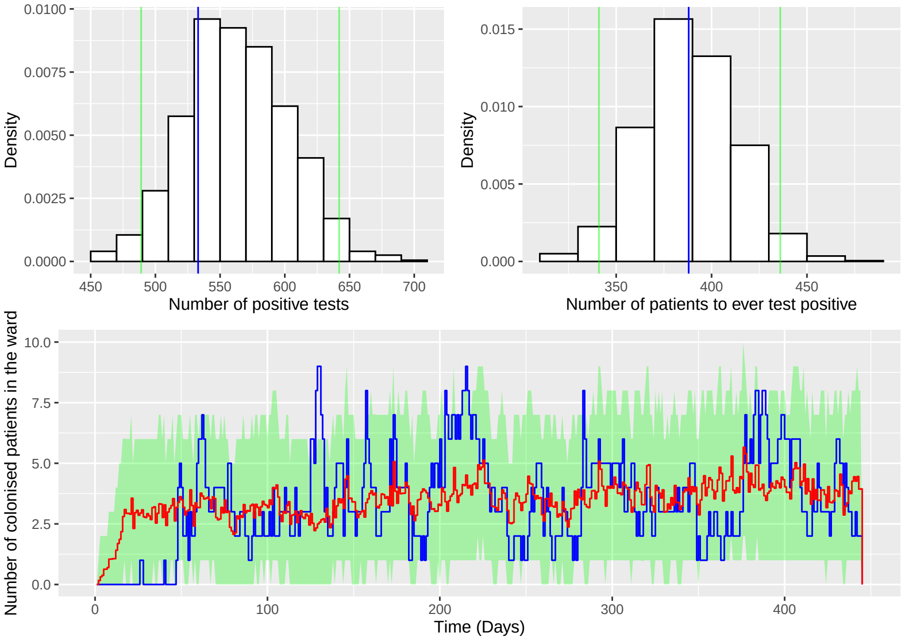


Figure 3.27: Posterior predictive distributions for the number of positive tests (top left), number of patients to ever test positive (top right) and the number of colonised patients in the ward (bottom). The true observed values are in red. The green region indicates the 95% (equal-tailed) credible interval and red line is the mean.

3.7.5.2 Genetic model assessment

In this section we present the framework to perform model assessment for the genetic data using the approach described in Cassidy et al. (2020). The differences here are that we use a different generative model for the distances with explicit time dependence and the possibility of multiple genetic subtypes.

Recall that the genetic model consists of a model for (i) pairwise distances between directly connected sequences from the mutation model, (ii) pairwise distances between imported sequences in the same genetic subtype and (iii) pairwise distances between imported sequences in distinct genetic subtypes (see Section 2.9 for details). Since we are interested in pairwise distances within the same genetic subtype, we do not attempt to model the distances between distinct subtypes, hence the model assessment will only include pairwise distances within the same genetic group.

In order to simulate a pairwise distance matrix we require knowledge of the underlying genetic network $\mathcal{G}$ which is constructed from the current state of the transmission network. For convenience we output the genetic network at each iteration in the algorithm to avoid extra computation.

Given a genetic network $\mathcal{G}$, we may simulate pairwise distances of directly connected sequences from the mutation model given the mutation rate λ and sampling time between sequences. Next we can simulate pairwise distances of sequences in the same genetic subtype using the Poisson model given the average importation distance μ, and in distinct genetic subtypes also using the Poisson model with parameter ξ. Finally, we calculate pairwise distances between two sequences that are not connected by summing the number of mutations along the path in the genetic network under the assumption that any nucleotide cannot mutate more than once.

Suppose we wish to produce N simulated data sets of genetic distances $D_1^{\text{rep}}, ..., D_N^{\text{rep}}$ from M available posterior samples. The general procedure is for $i = 1, ..., N$, select a posterior sample (ρ, F) uniformly at random and then simulate a distance matrix D_i^{rep} using the genetic data $(\lambda, \mu, \mathcal{G})$ and a specified between subtype importation distance ξ.

In order to compare the replicated data sets $D_1^{\text{rep}}, ..., D_N^{\text{rep}}$ to the true observed matrix D we look at the marginal distributions of distances. In a framework similar to above, suppose there k within-subtype pairwise distances indexed by some i, j. Then by considering each of the marginal distances, there will be k (univariate) marginal PPDs denoted by D_{ij}^{rep} where now it is straightforward to compare to the observed distance D_{ij}.

We then may use the p-value definition above and for any marginal PPD indexed by i, j, we may compute the probability that the replicated data is more extreme than the observed data, which explicitly is written as $p_{ij}^B = \Pr\left(D_{ij}^{\text{rep}} \geq D_{ij}\right)$. Finally, we may define a P-matrix to contain the p-values of each marginal distribution.

One potential caveat with the P-matrix is that depending on the number of pairwise distances under consideration k, the matrix may be difficult to interpret. There may also be incorrect conclusions drawn for discrete random variables on a non-negative support. For example, consider a discrete random variable X with an observed value $x = 0$ and suppose there are five samples generated from the PPD denoted by $X^{\text{rep}} = (0, 0, 0, 0, 1)$. The posterior predictive distribution clearly is similar to the observed data, however the p-value is $\Pr(X^{\text{rep}} \geq 0) = 1$ which is contradictory as an extreme p-value indicates poor model fit. Consequently notions of p-values may not be useful for discrete random variables that are particularly concentrated around low values, which certainly may be the case for genetic distances.

Alternatively we may opt for a simpler graphical representation of the posterior predictive summaries and instead use binary matrices that satisfies any criterion. For example we may define a Q-matrix, where elements are equal to 1 if and only if the observed genetic distance D_{ij} lies within the 95% credible interval of the marginal PPD D_{ij}^{rep}. More generally we may define a Q_α-matrix for any confidence level α where any element $[Q_\alpha]_{ij}$ is equal to 1 if and only if D_{ij} lies within the $100(1-\alpha)\%$ credible interval.

For the Brighton data there are 210 distinct genetic subtypes that have more than one observed sequence, with a total of 3954 within group pairwise distances. Furthermore, there are a total of 1308 genetic sequences and hence $854,778$ total pairwise distances. Consequently presenting all 3954 posterior predictive checks at the same time may be difficult to visualise and interpret due to the fact the Q-matrix is sparse, instead we may look at individual matrices for a particular genetic subtype.

From Figure 3.28 we see the posterior predictive summary for a particular genetic subtype which contains sequences from a specific patient. There are a total of 13 pairwise distances where the observed distance lies within the 95% credible interval out of 21 (61.9%) which demonstrates reasonably good model fit.

By considering all 3954 within group pairwise distances, there are 1630 marginal

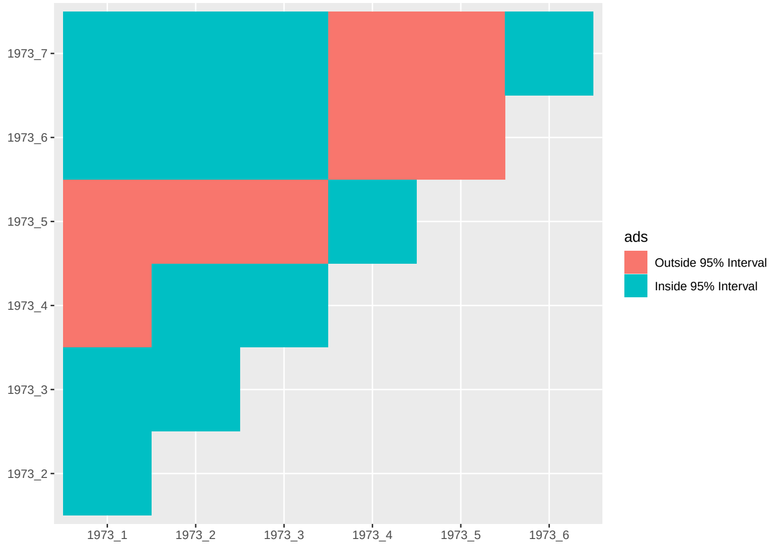


Figure 3.28: Posterior predictive summary for all sequences in genetic subtype 65. All genetic sequences within this subtype were sampled from the same patient (ID 1973). The blue boxes indicate that the true observed distance lies within the 95% credible interval of the simulated distances, and red indicates the true distances lies outside of the interval.

posterior predictive distributions (41.2%) such that the observed distance lies within the 95% credible interval. This indicates that these distances can be well explained by the model, conversely there are a significant proportion of distances that are not captured by the model which highlights the need to develop alternative models to accurately describe genetic data in an outbreak.

3.8 Discussion

In this chapter we have presented a novel framework for analysing outbreak data which may contain within-host genetic diversity and healthcare worker data. These extensions naturally build upon the work described in Worby et al. (2016) and Chapter 2 by modelling healthcare workers on an individual level and partitioning genetic sequences that are assumed to have evolved from independent stochastic processes.

The two extensions are motivated by the Brighton data set, where a significant portion of the data are sequences from healthcare worker swab tests. Furthermore, the data set exhibits a significant amount of within-host genetic diversity, where some individuals have multiple sequences from different strains of the organism.

Genetic sequences are partitioned into groups such that the pairwise distances within a group are similar and hence represent snapshots of the same organism at various points in time. The underlying mutation model describes the evolution of a single nucleotide through time, therefore separating sequences into groups is necessary where each group corresponds to a different nucleotide. The motivation here is that modelling multiple strains evolving independently more accurately describes the underlying biological process. Typically individuals test positive for colonisation if the pathogen is detected on a particular anatomical site, which usually presents in the form of a colony which will likely contain variations of the pathogen. Since multiple 'versions' or more suitably strains of the pathogen exist concurrently, hence we model each of these as independent stochastic processes.

The results in this chapter are natural generalisations of those derived in Chapter 2 where the contributions of the genetic data for each subtype are independent of one another. Furthermore, we discuss how to construct a genetic network that is consistent with the transmission tree and genetic subtypes.

Including healthcare workers at an individual level provides us with a method to better determine transmission routes through the population. This is particularly useful in a Bayesian framework where we frequently update the augmented data which consists of the unobserved transmission dynamics, particularly the source of colonisation. In our MCMC algorithm we continuously update the source of colonisation for each individual which now may be either other patients or healthcare workers.

When updating sources of colonisation in the algorithm, proposals that involve individuals with similar genetic sequences are more likely. In principle, updates that 'join' two individuals together (i.e. one of the individuals are proposed to be the source of colonisation for the other) are informed by the genetic data. If the two individuals have similar genetic sequences (indicated by low pairwise distances), then the genetic model should consider these two to be epidemiologically related and hence more likely to accept the proposed move. Inclusion of healthcare worker data now allows those individuals to be proposed as a source of colonisation for patients which should determine transmission pathways by proposing to randomly 'join' individuals together and eventually finding the most likely transmission network.

We then examined the performance of the model through a simulation study and sensitivity analysis. When simulating data sets with parameter choices informed by real world choices, we were able to successfully infer the true simulated parameter values and also demonstrated significant improvement in inferring the correct transmission network compared to without the genetic data. Furthermore, we demonstrated that the algorithm is able to infer times and sources of colonisation.

We also tested the robustness of the parameter estimation by varying the true simulated parameters across a range of values and showed that the algorithm was able to recover the true value most of the time. The model seemed to consistently underestimate the healthcare worker transmission rate β_H, reasons why are discussed below. For all parameter choices in simulated scenarios the algorithm demonstrated significant improvement of accuracy when reconstructing the transmission network with genetic data, compared to when inferring the transmission network without the genetic data. Consequently this gives support that inclusion of genetic data is worthwhile to better determine transmission pathways through the population.

Next we fit the model to the Brighton data set and where we inferred the parameter values, times and sources of colonisation. From the posterior samples from the MCMC we were able to construct posterior distribution for each parameter and also a weighted transmission network. We found that there were three instances of transmission, two of which were healthcare worker to patient (posterior probability of 55% and 97%) and one of patient to patient (posterior probability 28%).

The reason for inferring little transmission is likely due to identifiability issues with the Poisson model to describe the genetic distance between imported sequences. Put another way, there are multiple ways to explain the same data. Suppose we have three individuals that are separated by low genetic distances (less than 5 SNPs). One explanation is that all three imported the pathogen from the community, in which case the average distance parameter μ is low. On the other hand, another explanation is that all three are directly linked in one way or another, hence those genetic distances now inform the mutation rate and the average distance parameter μ will be higher.

We see this in the posterior median estimates for MSSA and the combined data sources, where the median of μ is estimated to be 1.76 and 3.63 respectively. There are a few ways to ameliorate this, such as imposing informative priors or fixing the parameter. Since we have a genuine belief that this parameter is higher, we feel it is appropriate in this case to use stronger priors to reflect this and we find that the model captures more transmission rather than assuming each individual imported the pathogen from the community.

We also investigated the impact of organism type, specifically by applying the analysis separately to the three genetic data sources, specifically data that contained MRSA, MSSA and combined. The MSSA and combined data sources produced similar results which is expected since the MRSA data contributes a small portion of the combined data. On the other hand, the results when considering only MRSA were significantly different to the other two, which is likely due to a lack of data, indicated by the amount of uncertainty in the parameter estimates.

Furthermore, we replicated the analysis five times for each data source to verify that independent runs of the same data under a different starting seed converge to the same mode. We also visually inspected various trace plots, such as the log likelihood, sum of colonisation times and number of patients added by the algorithm, to verify that the chains have converged to the same stationary distribution.

Finally, we performed model assessment using Bayesian posterior predictive checks to determine the extent to which the observed data departs from replicated data simulated from the model. We found that posterior predictive distributions for the epidemiological data are consistent with the observed data, specifically the observed number of positive tests, number of patients to ever test positive and the number of colonised patients in the ward are reasonable under the posterior predictive distributions.

On the other hand, for the genetic data only 41.2% of the marginal posterior predictive distributions contained the true distance within the 95% credible intervals, indicating that over half of the within-group distances could not be well explained by the model.

Ultimately the genetic model should certainly be improved, however when modelling complex biological systems there will invariably be discrepancies between an assumed model and the true underlying process. The strength in our model lies with the ability to summarise the genetic data using the matrix of pairwise distances alone, where the evolutionary relationship of these distances are (partly) inferred by the algorithm. This is particularly relevant for the Brighton data set where there are 1308 genetic sequences. If we were to store the whole genome sequences then we are working with 1308 N-dimensional vectors, which translates to approximately 3.66 billion nucleotide bases to store. Alternatively the matrix of pairwise distances requires $854,778$ elements stored which is over 4000 times smaller than if we were to store the sequences.

3.8.1 Limitations

By treating healthcare workers at an individual level and without knowledge of actual time spent on the ward (shift pattern data), we are implicitly assuming that HCWs apply constant colonisation pressure at all points in time between positive test results.

In reality it may be the case that for some patient episodes a particular healthcare worker may never have come into contact with a specific individual simply due to not working in the ward on those specific dates. While the healthcare worker transmission model may have some simplifying assumptions, we believe that they are necessary without more detailed data and ultimately provides a starting point for analysis.

Furthermore, in the analysis of the Brighton data set we have assumed that there exists one large ward in the hospital where patients are admitted and eventually discharged, however in fact that data was collected from two separate wards. This may violate certain assumptions in the transmission model, particularly with regards to homogeneous mixing of patients (indirectly via healthcare workers). In principle the correct approach would be to run two separate analyses, one per ward, however this data were not available and as a consequence it is natural to pool them together.

When analysing the Brighton data set we determined the genetic subtype pre-analysis using the group by threshold method with a cut-off of 40 single nucleotide polymorphisms (SNPs). This choice was to be consistent with the analysis accompanying the data set, however the decision is still somewhat arbitrary.

As noted about we found that the healthcare worker transmission rate was typically underestimated in nearly all of the results of the simulation study. Patient to patient transmission in intensive care units typically occurs indirectly via healthcare workers, therefore including a healthcare worker to patient transmission rate is potentially difficult to interpret as there may be some identifiability issues. Another reason is that healthcare worker transmission is typically low and there may not be enough information in the data to well estimate the healthcare worker transmission rate parameter β_H.

In the exploratory analysis of the Brighton data set we identified 21 putative transmission pairs informed by low distances sampled at similar times which the algorithm should recover. The analysis of the Brighton data set inferred only 3 of these 21 pairs and therefore there may be some acquisitions of the pathogen that were not correctly inferred by the algorithm and we believe that this is a consequence of the Poisson model for imported sequences.

The modelling of imported sequences is intended to provide a generative model to give contribution to distances that are in distinct transmission chains where the standard mutation model may not be used. The idea behind this is that distances within the same genetic subtype by definition are lower than an arbitrary threshold ζ, but there may still be some natural clustering within the subtypes. Intuitively larger distances (typically greater than 5 SNPs) within the same genetic subtype are considered too large to occur by mutation alone, hence this supports the idea that these two observed groups are related since they are in the same subtype, but occur as a result of multiple introductions of the pathogen.

These ideas borrow from the phylogenetic approach where we assume that these sequences evolved from a most recent common ancestor and instead of inferring the phylogenetic network we assume a Poisson model for the pairwise distances between these sequences.

CHAPTER 4

Inference for foot-and-mouth disease (FMD)

4.1 Introduction

Foot-and-mouth disease (FMD) is a contagious viral disease that primarily affects cattle and swine, however the virus can be transmitted by other cloven hoofed animals such as sheep and goats (Dunbar et al., 2009). Livestock infected with the disease typically exhibit fever, shivering, drooling of saliva, lameness and vesicular lesions on the tongue, nose, coronary bands and teats due to the replication of FMD (Meyer and Knudsen, 2001).

The virus is considered one of the most highly contagious diseases due to the rapid replicated of the pathogen within animals and the ease of transmission between other susceptible livestock in close proximity (Grubman and Baxt, 2004). The virus may also be spread through aerosolised droplets that come into contact with the respiratory tract of recipient livestock (Alexandersen et al., 2003).

In addition to direct animal-to-animal contact, FMD can also be spread through indirect contact, such as contaminated fomites. Fomites are objects or materials that can becoming contaminated with infectious microorganisms, such as vehicles, clothing or equipment that have come into contact with infected animals or their secretions (Woodbury, 1995).

In the context of foot-and-mouth disease, studies have estimated a £3.1 billion GBP loss to the agriculture and food chain and £355 million to agricultural produces in the 2001 UK outbreak alone (Thompson et al., 2002).

Modelling outbreak data provides an opportunity to understand transmission dynamics, evaluate control measures (Hayama et al., 2013), predict observed transmission (Jewell et al., 2008), model vaccination strategies (Keeling et al., 2002) and infer transmission networks (Cottam et al., 2008; Morelli et al., 2012; Lau et al., 2015).

This chapter is organised as follows. In Section 4.2 we review previous modelling studies of FMD. In Section 4.3 we introduce the motivating data set for this chapter and present an exploratory analysis. Next in Section 4.4 we outline the stochastic models we have developed to describe the data.

Later in Section 4.5 we describe how to perform inference in a Bayesian framework, in particular the Markov chain Monte Carlo (MCMC) algorithm used to draw samples from the posterior distribution and steps to sample the unobserved data and in Section 4.6 we present a simulation study to demonstrate the performance of the algorithm.

Furthermore, we present an analysis of an outbreak of FMD in Darlington in the north east of the United Kingdom in Section 4.7 for various competing models informed by the exploratory data and then assess the goodness of fit of each of these models in Section 4.8. Finally, we present a discussion into the models and methods developed in this chapter in Section 4.9 including the limitations of our approach and potential future research directions.

4.2 Literature review

In this section we review previous attempts at modelling FMD, in particular the 2001 outbreak in the United Kingdom which began in February 2001 and lasted for over six months and saw the destruction of four million animals (Davies, 2002). The work in Keeling (2005) provides a detailed description of three popular models developed at the time by different groups, that is the Cambridge-Edinburgh model (Keeling et al., 2001), Imperial model (Ferguson et al., 2001) and InterSpread (Morris et al., 2001), however in this review we only focus on the former.

In Keeling et al. (2001) the authors provide an analysis of the dynamics of the 2001 foot-and-mouth epidemic in the UK. The model described (later coined as the 'Cambridge-Edinburgh' model (Keeling, 2005)) is a discrete time stochastic Susceptible-Exposed-Infected-Removed (SEIR) model where the rate of transmission depends on the distance between infected-susceptible farm pairs, the number and type of infected livestock and their relative susceptibility.

In their formulation, at time t each susceptible farm j receives infectious pressure from each infective farm i, independent of all other farms, denoted by β_{ij}. Exposure to the pathogen is then modelled by contacts with infective farms and contact with other farms are determined by the time points of independent Poisson processes. Hence, the probability of exposure in the time interval $(t, t+1]$ is given by

$$\text{Pr}(\text{farm } i \text{ is exposed on day } t) = 1 - \exp\left(-\sum_{i \in I_t} \beta_{ij}\right),$$

where I_t is the set of infected farms at the start of day t. The latent and infectious periods are considered to be fixed and are 5 and 9 days respectively.

The authors employ a two-step approach whereby they first estimate the model parameters using maximum likelihood estimation to generate an initial fit, and then refine their estimates on a regional level using Monte Carlo simulation. The general procedure was to split the data and model parameters at a county level and then tune parameter estimates which minimise the differences between the observed data and the simulated epidemic curves using a least squares approach. The main findings were the importance of rapid implementation of control strategies and the importance of quantifying the spatial infection kernel that contributed to local and non-local spread. Using the same model, a detailed analysis of vaccination strategies can be found in Keeling et al. (2002).

Cottam et al. (2006) examines the molecular epidemiology of FMD by sequencing the genomes of 21 samples collected from infected premises during the 2001 outbreak in the United Kingdom. The authors reconstructed a phylogenetic tree using maximum likelihood methods under the assumption of the Hasegawa–Kishono–Yano (HKY) mutation model, which is similar to the Kimura mutation model with the assumption of equal base frequencies relaxed. The authors provide an estimate for the mutation rate of the virus, which was found to be 2.26×10^{-5} (95% CI $[1.75 \times 10^{-5}, 2.80 \times 10^{-5}]$) and an estimated date for the start of the outbreak, 7th February 2001, which is consistent with clinical evidence. The authors also assess the feasibility of contract tracing when informed by genetic data. They found that phylogenetic analysis agreed with traditional tracing methods and also provided new insights, such as linking farms that were previously considered unrelated.

In Savill et al. (2006) the authors look at the spatial dynamics that can be attributed to the spread of FMD in the 2001 UK outbreak. Specifically they use contract tracing data from the Department for Environment, Food and Rural Affairs (DEFRA) alongside UK road network data to compare the risk of infection between susceptible and infective farm pairs using Euclidean distance versus the shortest and average distance using roads. They conclude that a transmission kernel based on Euclidean distance is sufficient to estimate transmission risk compared with the shortest path via road travel.

In their discussion the authors suggest possible reasons for the results, which are (i) there are more transmission routes than just the main road, such as farms sharing a common boundary or private paths not included in the road data and (ii) infections may occur by social networks.

Tildesley et al. (2008) provides a detailed analysis of the Cambridge-Edinburgh model (Keeling, 2005) and develops techniques to assess the accuracy of the model, in particular its predictive ability to correctly identify cases and culls. By simulating data at a farm level the authors are able to compare the status of any particular farm at the end of the outbreak with the true status, where a farm may either be (i) reported, (ii) culled or (iii) susceptible. In the usual fashion, a straightforward measure of accuracy is then given by

$$\text{accuracy} = \frac{N_{RR} + N_{CC} + N_{SS}}{\text{total number of farms}},$$

where N_{RR}, N_{CC} and N_{SS} are the number of farms that were correctly predicted to be reported, culled and susceptible respectively. The authors calculated the accuracy to be 92.46% (95% of simulations lie within 91.66-93.16%), suggesting good predictive power in general but only a small part of the overall picture.

In particular, it is difficult to summarise the accuracy of a models ability to capture the spatio-temporal dynamics down to a single statistic, hence the authors consider the accuracy of predicting reported and culled cases separately.

4.3 The Darlington data set

The data set we apply the models and methods developed in this chapter to is the 'Darlington cluster' (Cottam et al., 2008). The data set includes information on the location, date and type of premises affected, as well as the number of animals slaughtered and the dates on which they were slaughtered. The data set is called the 'Darlington cluster' because it refers to a specific geographic area in the north-east of England where a large number of farms were affected by the outbreak. The original data was provided by DEFRA and a description can be found in (Cottam et al., 2008).

More precisely, the observed data are the location (longitude and latitude of infected premises), date of culling, date of examination, date of first lesion appearance and the pairwise transitions and transversions for all isolates that were sequenced, the corresponding farm and time of sampling.

In this work we do not account for the species and number of each infected animal on the premises, however it is possible to account for farm level covariates (Firestone et al., 2020). An overview of the epidemiological and genetic data can be found in Figures 4.1 and 4.2 respectively.

We also note that there are other farms in the nearby area that were never infected and are not included in this data set. The inclusion of susceptible farms will invariably have an impact on the inference and consequently we later consider a separate analysis with susceptible farms.

Immediately from the overview of the genetic data we observe sequences with a high degree of similarity, indicated by the white cells in Figure 4.2. Intuitively, genetic sequences that are similar suggest that the farms which they were sampled from are potentially linked in terms of transmission.

From the pairwise transitions, we notice similarity between farms A and N, with a distance of 6 transitions. Furthermore, the next farm that is similar to A is K with a distance of 17 transitions, and the next most similar farm to N is also K with a distance of 21. Hence it is clear that the genetic sequences sampled from A and N are distinct from all other observed farms, which gives evidence that these two may be related and in a separate transmission cluster. It can also be seen that these two premises are also geographically linked, which further suggests that transmission between the two is plausible.

Next we observe multiple small genetic distances for farm K where the corresponding farms with these similar sequences are F, B, M, G and P with $(5, 6, 8, 8, 9)$ transitions respectively, which may be an indicator for a small cluster of farms which are potentially related.

Moreover there is a strong link between farms M and D with a transition distance of

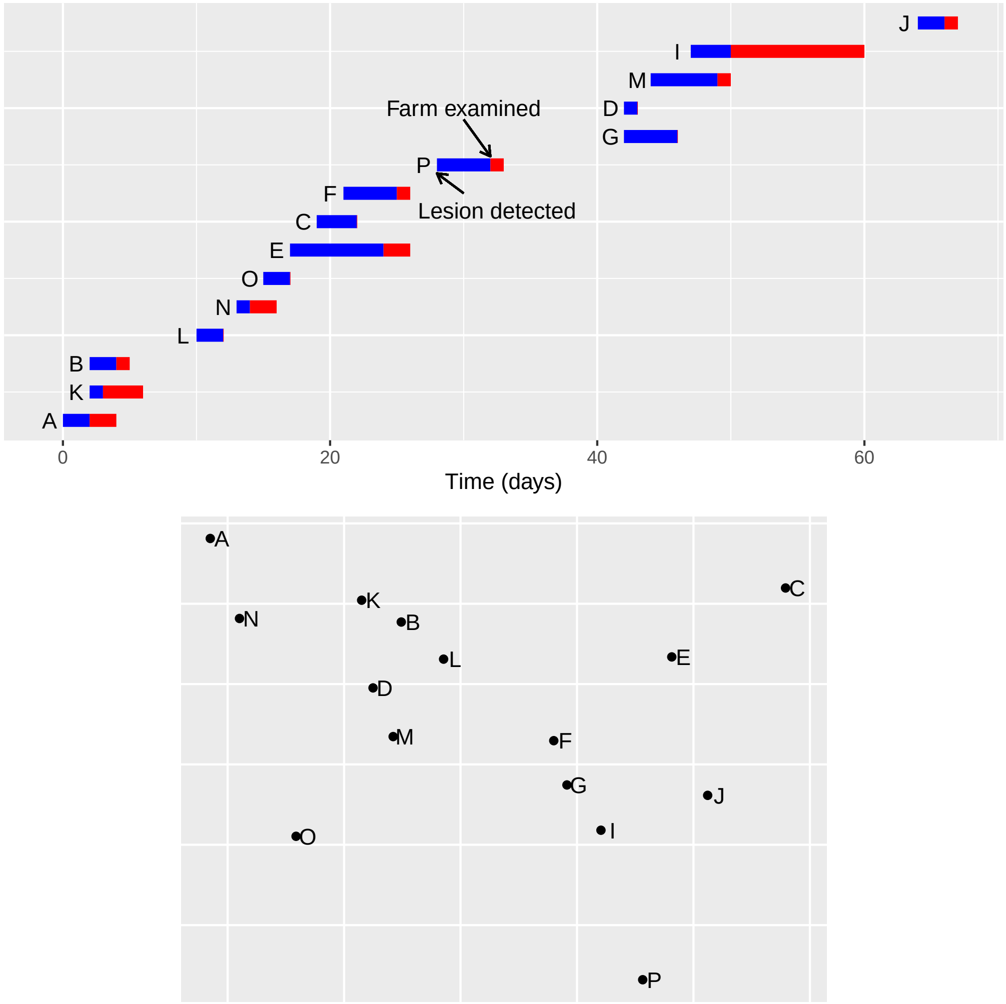
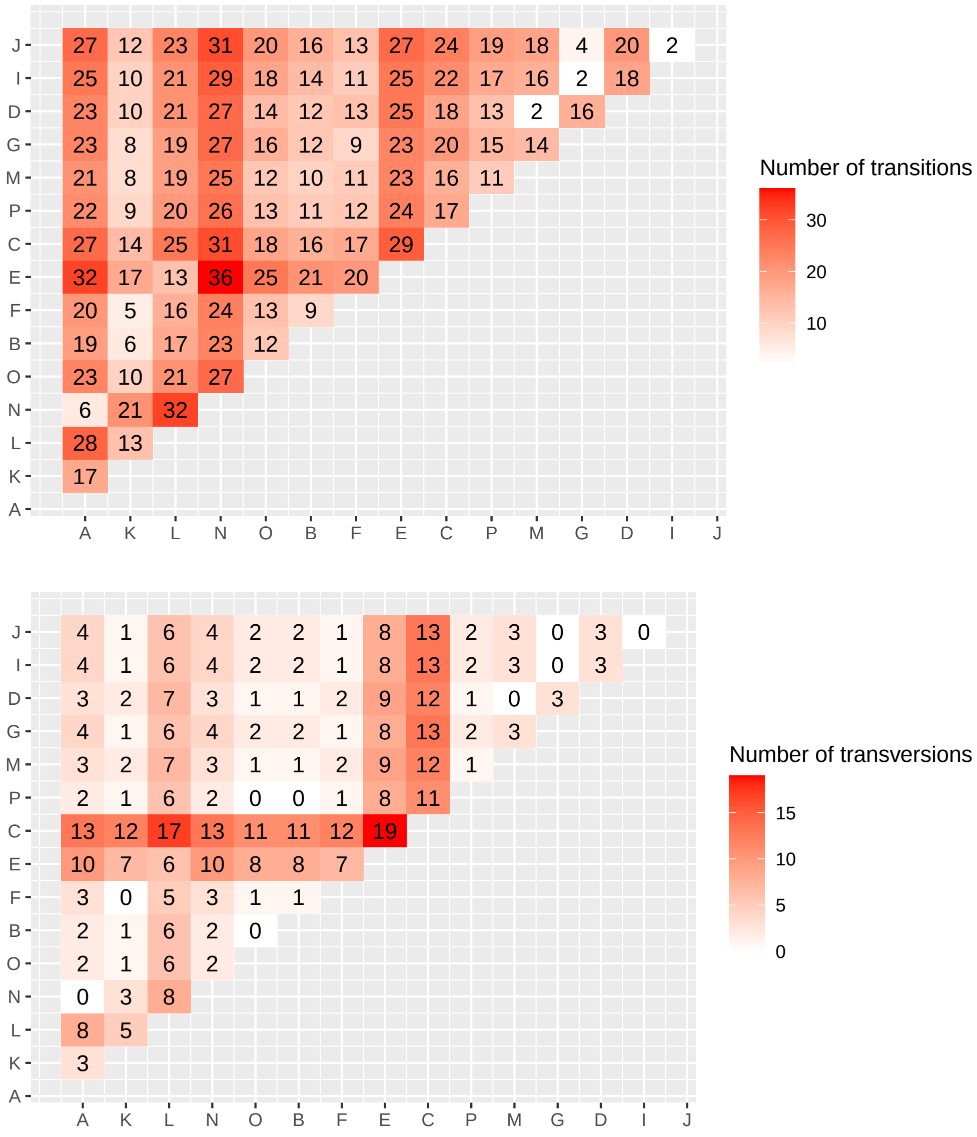


Figure 4.1: An overview of the epidemiological data for the Darlington cluster. Above: Temporal trends with the infected premises where the segments in blue indicate the time where lesions were detected on the livestock but before examination. The segments in red indicate the time after examination to the time of culling. Below: The location of each farm in space.

2 which suggests strong support for transmission. Finally, we see another cluster of transmission between farms G, I and J with pairwise transitions $(T_{GI}, T_{GJ}, T_{IJ}) = (2, 4, 2)$, again indicating suggesting that these farms may be related in transmission, such as one infecting the others or all three infected by an external farm. It can also be seen that these three farms are also geographically linked, in the sense that these



	A	K	L	N	O	B	F	E	C	P	M	G	D	I	J
J	27	12	23	31	20	16	13	27	24	19	18	4	20	2	
I	25	10	21	29	18	14	11	25	22	17	16	2	18		
D	23	10	21	27	14	12	13	25	18	13	2	16			
G	23	8	19	27	16	12	9	23	20	15	14				
M	21	8	19	25	12	10	11	23	16	11					
P	22	9	20	26	13	11	12	24	17						
C	27	14	25	31	18	16	17	29							
E	32	17	13	36	25	21	20								
F	20	5	16	24	13	9									
B	19	6	17	23	12										
O	23	10	21	27											
N	6	21	32												
L	28	13													
K	17														
A															

	A	K	L	N	O	B	F	E	C	P	M	G	D	I	J
J	4	1	6	4	2	2	1	8	13	2	3	0	3	0	
I	4	1	6	4	2	2	1	8	13	2	3	0	3		
D	3	2	7	3	1	1	2	9	12	1	0	3			
G	4	1	6	4	2	2	1	8	13	2	3				
M	3	2	7	3	1	1	2	9	12	1					
P	2	1	6	2	0	0	1	8	11						
C	13	12	17	13	11	11	12	19							
E	10	7	6	10	8	8	7								
F	3	0	5	3	1	1									
B	2	1	6	2	0										
O	2	1	6	2											
N	0	3	8												
L	8	5													
K	3														
A															

Figure 4.2: An overview of the genetic data for the Darlington cluster, summarised as a heat map of the number of pairwise transitions (above) and transversions (below). A colour (key indicated on the right) in position (i,j) is representative of the number of transitions or transversions (text in individual cells) between sequences i and j.

farms are all close to one another.

From the pairwise transversions in Figure 4.2, the data are less informative due to the transversion rate being significantly lower than the transition rate and hence we see many sequences that are identical in terms of transversions, or differ only by a few. The heat map for the number of pairwise transitions is in general darker than the heat map for the number of pairwise transversions, which is to be expected as the transition transversion ratio has previously been estimated to be 7.61 (Cottam et al., 2006).

One feature we may discern from the transversion matrix is that farm C appears to be significantly different from all other farms, perhaps indicating that C did not play a role in the Darlington cluster. Similarly, farm E is also perhaps unrelated to other farms, however to a much lesser degree.

We now make some remarks as to possible explanations for the observed data in Figure 4.1. Firstly, it should be noted that the true exposure time will be before the time at which the lesions are detected and we will attempt to infer these in our analysis. One explanation is that the true time of exposure occurs shortly before the developing symptoms, and hence the period between infection and symptoms is small. In this scenario, if infectious periods are short then there may be multiple transmission clusters or unobserved transmission events that are not present in the data, i.e. unreported premises.

On the other hand, if the infectious periods are long, then it is possible that the outbreak is a result of a single initial infective. The aim of this chapter is to incorporate multiple sources of information, namely the recorded time of culling, examination and symptoms with geographical location and genetic data to reduce uncertainty in the inference of (i) key epidemiological parameters and (ii) the transmission network.

Since we have extra information of the events on the farm prior to culling, such as the date of lesion onset and the date of examination, we are motivated to extend the general SEIR modelling framework to incorporate this extra information.

4.4 Model description

In this section we outline the model and assumptions used to analyse data from an outbreak of food-and-mouth disease virus (FMD). We propose a stochastic model where we are modelling farms at an individual level in the context of FMD and throughout the rest of this chapter we shall use the terminology ‘farm’ and ‘premises’ interchangeably.

In general this is an oversimplification of the transmission process since we are explicitly assuming that within-farm infections occur rapidly (Thornley and France, 2009). Furthermore, a farm is considered to have been exposed to the pathogen if any of the livestock show clinical symptoms of FMD.

We use a compartmental stochastic epidemic model to describe the transmission process where each farm may belong to exactly one of six compartments: susceptible (S), exposed (E), presymptomatic (I), postsymptomatic (L), examined (X) and removed (R).

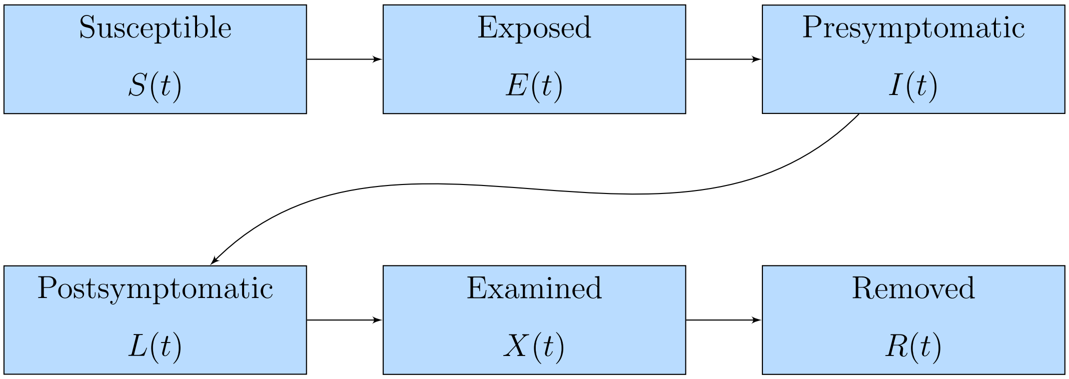


Figure 4.3: A stochastic compartmental model where each farm may belong to one of six states, which are susceptible, exposed, presymptomatic, postsymptomatic, examined and removed. Initially farms are susceptible to the disease and transition to the exposed state upon contact with an infected farm.

Any particular farm that is susceptible, denoted by i, receives infectious pressure denoted by $P_i(t)$ at any time t, which depends on parameters governing the transmission process and also the data, for example geographical distance. We assume that farms become infectious and hence exert infectious pressure on susceptible farms after entering the presymptomatic stage I.

Following standard assumptions in the literature (Andersson and Britton, 2012, Chapter 2), we assume that exposure to the pathogen for a susceptible farm j is a result of contact with an infectious farm i. Contacts between pairs of farms can be modelled as a time homogeneous Poisson process with intensity β_{ij}, with the assumption that all Poisson processes are mutually independent. The infectious pressure a susceptible farm i receives is therefore given by

$$P_i(t) = \sum_{j \in \mathcal{Y}(t)} \beta_{ij},$$

where $\mathcal{Y}(t) = \{j : I_j < t < R_j\}$ is the set of farms who are infected at time t.

In light of the previous modelling attempts discussed in Section 4.2 we are motivated to adopt a heterogeneous transmission rate that depends on the Euclidean distance between farms. More precisely, we wish to define a transmission kernel where the rate of contact between farms i and j is $\beta_{ij} = \beta K(d)$, that is the transmission rate depends on the distance between the two farms through the kernel K. Throughout the rest of this chapter we shall assume the *exponential kernel* (Lau et al., 2014), which is explicitly given by $K(d_{ij}) = \exp(-\kappa d_{ij})$. The choice of transmission kernel is largely arbitrary, however one may circumvent this decision and instead opt to estimate the infection rate non-parametrically (Seymour et al., 2022).

The intuition behind the kernel is that the rate of transmission is high when individuals are close to one another and this decays as distance increases, where κ is the parameter that specifies the rate of decay. Example transmission kernels with various values of κ are shown in Figure 4.4.

As opposed to Lau et al. (2014) we do not consider background exposures but instead consider the possibility of multiple initially infected farms. In the context of Foot-and-Mouth disease, clearly farms cannot contact others in the usual sense, rather

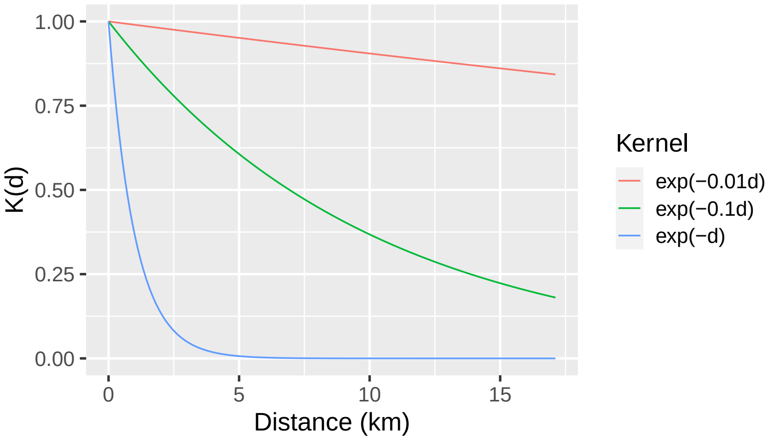


Figure 4.4: Spatial transmission kernels of the form $K(d) = \exp(-\kappa d)$ where d is the Euclidean distance and κ is the decay parameter. The line in red correspond to $\kappa = 0.01$, green corresponds to $\kappa = 0.1$ and blue corresponds to $\kappa = 1$.

contacts are indirect and occur through various means, such as contact with fomites (contaminated objects), airborne transmission, contaminated feed and water (Paton et al., 2018).

Susceptible farms that are exposed to the pathogen and hence transition from $S \to E$ remain in state E for a fixed amount of time which defined as the latent period. During this period the farm has contracted the disease but are not yet able to further transmit the pathogen.

After the latent period the farm will transition from $E \to I$, whereby the farm now exerts infectious pressure on susceptible farms. The amount of time spent in state I is defined as the presymptomatic period, after which the farm will transition to the lesion state $(I \to L)$ upon developing symptoms. This duration may also be referred to as the ‘subclinical‘ period in the literature (Mardones et al., 2010).

We define the time spent in state L to be the post symptomatic period where there are observable symptoms on the livestock, but there has not been a formal examination yet. Intuitively this can be thought of as the response time, which may be high if an outbreak is severe and there are not enough veterinary doctors to attend all infected premises. Naturally it is of interest for the government and health agencies to keep this as low as possible. After examination the farm transitions compartments ($L \to X$).

In the United Kingdom, after developing symptoms a farm is considered an infected premises and is subject to various restrictions. In particular, livestock will be humanely culled in accordance with The Animal Health Act 1981 and the European Communities Act 1972 and an epidemiological investigation is carried out in an attempt to control the spread of the disease. Hence after being culled a farm transitions from state X to R, where it is considered to be removed and no longer exerts infectious pressure on the population. The time spent in class X can intuitively be thought of as the time until culling after detection of the pathogen. Usually farms are culled immediately, however in some instances this may take a day or two. Furthermore, it is of interest from an infection control perspective to estimate this time period.

We assume that the latent period is constant and assumed to be known for all farms and the presymptomatic, post-symptomatic and removal periods are modelled by Gamma distributions. Furthermore, these periods are assumed to be independent of one another and the periods for different farms are assumed to be independent. Note that in general these random periods can be modelled by any arbitrary but specified non-negative distribution.

Furthermore, we assume that genomic sampling has been done where isolates have been cultured and sequenced from infected livestock, usually at the time of examination. Similar to Section 2.8 we assume that there exists a single dominant lineage in the population at any particular time, and upon exposure the pathogen is 'copied' from the source of infection to the recipient. That is to say, suppose farm i infects farm j at time t, then pathogen in farms i and j will be identical at time t, i.e. they will have the same whole genome sequence.

Finally, we model the genetic data at the level of nucleotide substitutions and assume that nucleotides evolve according to the Kimura model (Section 2.4.2). Recall that substitutions between the two pyrimidines ($T \leftrightarrow C$) or the two purines ($A \leftrightarrow G$) are called transitions and substitutions between a pyrimidine and a purine ($T, C \leftrightarrow A, G$) are called transversions. Finally, we assume that each nucleotide evolves independently of each other.

4.4.1 Data and notation

Following from the transmission and genetic models described above, we shall now outline some key notation that shall be used throughout the chapter. Consider a closed population of N farms, labelled $\{1, ..., N\}$, and let n_E denote the number of premises to ever become exposed to the pathogen. Due to the compartmental nature of the model, farms may only belong to one of the six classes at any particular moment in time, hence we necessarily must have $S(t) + E(t) + I(t) + L(t) + X(t) + R(t) = N$.

For $j = 1, ..., N$ we denote E_j, I_j, L_j, X_j and R_j to be the precise time of exposure, infectiousness, lesion onset, examination and removal where necessarily we must have $E_j < I_j < L_j < X_j < R_j \leq T$ where $T = \max_j R_j$. If a particular farm j is never exposed to the pathogen we set $A = \infty$ for all $A \in \{E_j, I_j, L_j, X_j, R_j\}$. Recall that we assume there is a fixed latent period which we denote by L and hence $E_j = I_j - L$ for all $j = 1, ..., n_E$.

Let $\mathcal{K}$ denote the set of farms that are assumed to be the root of the transmission clusters, where the number of transmission clusters, n_c, is the size of this set. Furthermore, define s_j to be the source of infection such that $s_j = k$ if and only if farm k infects farm j. If a particular farm j is considered to be the root of the transmission cluster we set $s_j = -1$ for all $j \in \mathcal{K}$, and $s_j = \infty$ if farm j is never exposed to the pathogen.

Let $\mathbf{E}, \mathbf{I}, \mathbf{L}, \mathbf{X}, \mathbf{R}$ denote the complete set of times for the exposures, infectiousness, lesion onset, examination and removal respectively and $\mathbf{s} = (s_1, ..., s_N)$ denote the set of source of infection for each farm. Note that the true time of exposure, infection and the source of infection are typically unobserved, hence we denote $\mathbf{Y} = (\mathbf{E}, \mathbf{I}, \mathbf{s})$ to be the unobserved quantities. We assume that the lesion, examination and removal times are observed and hence we write $\mathbf{Z} = (\mathbf{L}, \mathbf{X}, \mathbf{R})$ to be the observed data. Note that there does exist some uncertainty around the estimated date of lesion onset where the data are originally from the Defra data warehouse (Cottam et al., 2008), however for simplicity we shall treat them as known.

Furthermore, suppose we have observed n_s genetic sequences, we define $T \in \mathbb{N}_0^{n_s \times n_s}$ and $V \in \mathbb{N}_0^{n_s \times n_s}$ to be the matrices of pairwise transitions and transversions respectively (for example Figure 4.2). For any genetic sequence label $j = 1, ..., n_s$ we denote ν_j to be the farm where the sequence was collected, η_j to be the genetic subtype number for the sequence and τ_j to denote the corresponding time of sampling. Let ν, η and τ denote the complete set of sequence locations, subtype numbers and sampling times, then we may write the complete genetic data as $\psi = (T, V, \nu, \eta, \tau)$. There may also be unobserved genetic sequences, hence the unobserved genetic data can be written as $\tilde{\psi} = (\tilde{T}, \tilde{V}, \tilde{\nu}, \tilde{\eta}, \tilde{\tau})$. Finally, let $\mathcal{G}$ denote the underlying genetic network that describes the evolution of the pathogen. Recall from Section 3.4.1 that the genetic network $\mathcal{G}$ is not a genuine parameter and is constructed from the data.

Let $F = (\mathbf{Y}, \tilde{\psi}, \mathcal{G})$ denote the unobserved data which consists of the times of exposure, infection, source of infection, genetic data and the underlying genetic network.

We define $\rho = (\alpha_1, \gamma_1, \alpha_2, \gamma_2, \alpha_3, \gamma_3, \beta, \kappa, \lambda_T, \lambda_V)$ to be the vector of model parameters which contain the shape and rate parameters for the distributions of the random times spent in the compartments, the parameters governing the transmission process and finally parameters for the mutation model.

4.4.2 Model likelihood

In this section we present the joint likelihood of the transmission process and genetic data given the model parameters. Let $\mathcal{K}$ denote the farms that are the roots of each transmission clusters and $E_{\mathcal{K}}$ denote their respective exposure times. The density of the observed data $f(\mathbf{Z}, \psi \mid \rho, E_{\mathcal{K}}, \mathcal{K})$ in Equation (4.1) is intractable as it involves integrating over the unobserved data F,

$$f(\mathbf{Z}, \psi \mid \rho, E_{\mathcal{K}}, \mathcal{K}) = \int_F f(\mathbf{Z}, \psi, F \mid \rho, E_{\mathcal{K}}, \mathcal{K}) dF. \tag{4.1}$$

Consequently we consider augmenting the parameter space with the unobserved quantities F which leads to a tractable likelihood. Note that we may decompose this likelihood into two independent components,

$$f(\mathbf{Z}, \psi, F \mid \rho, E_{\mathcal{K}}, \mathcal{K}) = f_G(\psi, F \mid \rho_G, \mathbf{Z}, E_{\mathcal{K}}, \mathcal{K}) f_T(\mathbf{Z}, F \mid \rho_T, E_{\mathcal{K}}, \mathcal{K}), \tag{4.2}$$

where $f_G(\psi, F \mid \rho_G, \mathbf{Z}, E_{\mathcal{K}}, \mathcal{K})$ is the density of the genetic data, conditional on the outbreak data and model parameters, $f_T(\mathbf{Z}, F|\rho_T, E_{\mathcal{K}}, \mathcal{K})$ is the density of the transmission data conditional on the model parameters and the complete parameter vector is $\rho = (\rho_G, \rho_T)$.

For the rest of the chapter we shall drop the conditioning on the transmission cluster roots $\mathcal{K}$ for notational convenience.

4.4.2.1 Epidemiological component

The term $f_T(\mathbf{Z} \mid F, \rho, E_{\mathcal{K}})$ in Equation (4.2) is the density of the transmission process, conditional on the model parameters. We may decompose this into independent components,

$$f_T(\mathbf{Z}, F \mid \rho, E_{\mathcal{K}}) = L_T \times L_S \times L_X \times L_R, \tag{4.3}$$

where L_T, L_S, L_X and L_R are the contributions from the (i) transmission process, (ii) presymptomatic periods, (iii) postsymptomatic periods and (iv) removal periods respectively. We define the random period of time spent in classes I, L, X and R for each farm to be the sojourn times.

Transmission contribution We do not directly model introductions of the pathogen into the population and instead assume that the number of transmission clusters to be constant and known, hence the likelihood of the transmission process consists of the probability of susceptible premises avoiding infection and the probability of farms becoming infected.

As discussed in Section 4.4, any susceptible farm denoted by i receives infectious pressure from already infective farms, j say, at a rate $\beta_{ij} = \beta \exp(-\kappa d_{ij})$.

The contribution of the transmission process can be written as

$$L_T = \prod_{\substack{j=1 \\ j \notin \mathcal{K}}}^{n_E} \left(\beta e^{-\kappa d_{s_j j}} \right) \times \exp\left(-\beta S\right) \tag{4.4}$$

$$= \beta^{n_E - n_c} \times \exp\left(-(\kappa \bar{d} + \beta S)\right), \tag{4.5}$$

where $\bar{d}$ is the distance sum between infected susceptible pairs and S is the total infectious pressure during the epidemic and these quantities are explicitly given by

$$\bar{d} = \sum_{\substack{j=1 \\ j \notin \mathcal{K}}}^{n_E} d_{s_j,j}, \tag{4.6}$$

$$S = \sum_{j=1}^{n_I} \sum_{k=1}^{N} e^{-\kappa d_{ij}} \left((R_j \wedge E_k) - (I_j \wedge E_k) \right). \tag{4.7}$$

The product in Equation (4.4) is the probability of exposure from a particular source and the exponential term is the probability to avoid exposure from infectious farms.

Sojourn time contribution The contribution of the sojourn times are similar hence we present a general framework here. Suppose we wish to determine the contribution of random time spent in a particular class, X, before transitioning to the next compartment, Y. Recall that we assume the random time spent in a class can be modelled by some arbitrary but non-negative distribution, Q, with probability density function $f_Q(Y_j - X_j \mid \omega)$ for all $j = 1, ..., n_E$. Since all sojourn times are independent of one another, the contribution to the likelihood is a product of the independent densities, given by

$$L_k = \prod_{j=1}^{n_E} f_Q(Y_j - X_j \mid \omega), \tag{4.8}$$

where $k \in \{E, S, X, R\}$ and ω are the relevant model parameters and Y_j, X_j are the random variables. The specific choice of Q is a modelling decision and may be dependent on the particular pathogen under consideration. We shall outline some common choices in the literature (Vergu et al., 2010; Streftaris and Gibson, 2012).

If we assign $Q \sim \text{Gamma}(\alpha, \gamma)$, then Equation (4.8) becomes

$$L_k = \gamma^{\alpha n_E} \exp\left\{-\gamma \sum_{j=1}^{n_E}(Y_j - X_j)\right\} \prod_{j=1}^{n_E} \frac{(Y_j - X_j)^{\alpha-1}}{\Gamma(\alpha)}. \tag{4.9}$$

Alternatively we may set $\alpha = 1$, then we have $Q \sim \text{Exp}(\alpha)$, then we have

$$L_k = \gamma^{n_E} \exp\left\{-\gamma \sum_{j=1}^{n_E}(Y_j - X_j)\right\}.$$

Finally, we may consider $Q \sim \text{Weibull}(\alpha, \gamma)$ with density function $f(x \mid \alpha, \gamma) = (\alpha/\gamma)\,(x/\gamma)^{k-1}\, e^{-(x/\gamma)^{\alpha}}$. Then we arrive at

$$L_k = \left(\frac{\alpha}{\gamma^{\alpha}}\right)^{n_E} \exp\left\{-\frac{1}{\gamma^{\alpha}} \sum_{j=1}^{n_E}(Y_j - X_j)^{\alpha}\right\} \prod_{j=1}^{n_E} (Y_j - X_j)^{\alpha-1}$$

Throughout the rest of this chapter we shall assign

$$\begin{aligned} L_i - I_i &\sim \text{Gamma}(\alpha_1, \gamma_1), \\ X_i - L_i &\sim \text{Gamma}(\alpha_2, \gamma_2), \\ R_i - X_i &\sim \text{Gamma}(\alpha_3, \gamma_3), \end{aligned}$$

to the sojourn times and the vector of parameters for the transmission process is therefore $\rho_T = (\beta, \kappa, \alpha_1, \gamma_1, \alpha_2, \gamma_2, \alpha_3, \gamma_3)$.

4.4.2.2 Genetic model

In Section 2.7.2 we describe the joint probability mass function for genetic distances under the assumption of the Kimura mutation model. In the current formulation of the transmission model we do not attempt to model the incursion of the virus, and instead we allow for the possibility of multiple introductions of the pathogen.

Since we assume that the number of transmission clusters are fixed and known, we do not need to consider the contribution of genetic sequences in unrelated transmission chains.

Let $\psi = (T, V, \nu, \eta, \tau)$ denote the genetic data which are the matrix of pairwise transitions, transversions, farm where the sequence was sampled, genetic subtype and sampling times respectively.

From Section 2.7.2, the density of the genetic data, conditional on the genetic network $\mathcal{G}$ and model parameters ρ_G is proportional to

$$f_G(\psi, F \mid \rho_G, \mathbf{Z}, E_{\mathcal{K}}) \propto \prod_{(i,j)\in E} q(\tau_{ij})^{N-T_{ij}-V_{ij}} p_T(\tau_{ij})^{T_{ij}} \left(p_V(\tau_{ij})/2\right)^{V_{ij}} \tag{4.10}$$

where $q(\tau_{ij})$, $p_T(\tau_{ij})$ and $p_V(\tau_{ij})$ are the probabilities for a nucleotide to not mutate, transition and transversion in τ_{ij} time units and $q(\tau_{ij}) + p_T(\tau_{ij}) + p_V(\tau_{ij}) = 1$.

The proportionality in Equation (4.10) is a consequence of knowing only the matrices of pairwise transitions and transversions, rather than the sequence data directly.

4.5 Bayesian inference

Since we augment the parameter space to include all unobserved quantities, it is then natural to perform inference in a Bayesian framework whereby we treat the augmented parameters as random variables, conditional on the data. Using Bayes' theorem, we obtain

$$\begin{aligned}
\pi(\rho, F, E_{\mathcal{K}} \mid \mathbf{Z}, \psi) &\propto f_G(\psi, F \mid \rho_G, \mathbf{Z}, E_{\mathcal{K}}) f_T(\mathbf{Z}, F \mid \rho_T, E_{\mathcal{K}}) \pi(\rho, E_{\mathcal{K}}) \\
&= \prod_{(i,j)\in E} q(\tau_{ij})^{N-T_{ij}-V_{ij}} p_T(\tau_{ij})^{T_{ij}} \left(p_V(\tau_{ij})/2\right)^{V_{ij}} \\
&\times \beta^{n_E - n_c} \times \exp\left(-(\kappa\bar{d} + \beta S)\right) \\
&\times \gamma_1^{\alpha_1 n_E} \exp\left\{-\gamma_1 \sum_{j=1}^{n_E}(L_j - I_j)\right\} \prod_{j=1}^{n_E} \frac{(L_j - I_j)^{\alpha_1 - 1}}{\Gamma(\alpha_1)} \\
&\times \gamma_2^{\alpha_2 n_E} \exp\left\{-\gamma_2 \sum_{j=1}^{n_E}(X_j - L_j)\right\} \prod_{j=1}^{n_E} \frac{(X_j - L_j)^{\alpha_2 - 1}}{\Gamma(\alpha_2)} \\
&\times \gamma_3^{\alpha_3 n_E} \exp\left\{-\gamma_3 \sum_{j=1}^{n_E}(R_j - L_j)\right\} \prod_{j=1}^{n_E} \frac{(R_j - L_j)^{\alpha_3 - 1}}{\Gamma(\alpha_3)} \\
&\times \pi(\rho, E_{\mathcal{K}}), \qquad (4.11)
\end{aligned}$$

where $f_G(\psi, F \mid \rho_G, \mathbf{Z}, E_{\mathcal{K}})$ and $f_T(\mathbf{Z}, F \mid \rho_T, E_{\mathcal{K}})$ are the likelihoods of the genetic and transmission data respectively and $\pi(\rho, E_{\mathcal{K}})$ is the joint distribution of the prior parameters and exposure times for each transmission cluster.

We assume that the model parameters and exposure times that initiate the transmission clusters are independent *a priori*, and that the model parameters have the following distributions:

$$\begin{aligned}
\beta &\sim \text{Gamma}(a_\beta, b_\beta) \\
\kappa &\sim \text{Uniform}(0, 100) \\
\alpha_i &\sim \text{Uniform}(0, 100) \qquad i = 1, 2, 3 \\
\gamma_i &\sim \text{Gamma}(a_i, b_i) \qquad i = 1, 2, 3 \\
\lambda_T &\sim \text{Uniform}(0, 1) \\
\lambda_V &\sim \text{Uniform}(0, 1).
\end{aligned}$$

For the prior distribution on the exposure times that initiate the transmission clusters, $\pi(E_{\mathcal{K}})$, in a similar fashion to O'Neill and Becker (2001) we assign improper uniform priors to each of the exposure times that initiate the transmission clusters.

4.5.1 Markov chain Monte Carlo procedure

In this section we describe how we draw approximate samples from the target density of interest, $\pi(\rho, F, E_{\mathcal{K}} \mid \mathbf{Z}, \psi)$, i.e. the posterior distribution. We provide an overview of the MCMC procedure in Algorithm 4 and then explicitly discuss how to update the parameters (Section 4.5.2) and augmented data (Section 4.5.3).

Algorithm 4 Structure of the MCMC algorithm

```
Initialise the chain with initial values $\beta^{(1)}, \kappa^{(1)}, \alpha_1^{(1)}, \gamma_1^{(1)}, \alpha_2^{(1)}, \gamma_2^{(1)}, \alpha_3^{(1)}, \gamma_3^{(1)}, \lambda_T^{(1)}, \lambda_V^{(1)}$,
maximum number of iterations $M$, proposal variances $\sigma^2_\kappa, \sigma^2_{\alpha_1}, \sigma^2_{\alpha_2}, \sigma^2_{\alpha_3}, \sigma^2_{\lambda_T}, \sigma^2_{\lambda_V}$
and prior parameters $a_\beta, b_\beta, a_1, b_1, a_2, b_2, a_3, b_3$.
    Repeat the following steps for $i$ = 2,...,$M$
Sample $\beta^{(i)}$ from $\pi(\beta \mid \rho_{-\beta}, F, \mathbf{Z}, \psi)$.
Sample $\gamma_1^{(i)}$ from $\pi(\gamma_1 \mid \rho_{-\gamma_1}, F, \mathbf{Z}, \psi)$.
Sample $\gamma_2^{(i)}$ from $\pi(\gamma_2 \mid \rho_{-\gamma_2}, F, \mathbf{Z}, \psi)$.
Sample $\gamma_3^{(i)}$ from $\pi(\gamma_3 \mid \rho_{-\gamma_3}, F, \mathbf{Z}, \psi)$.
Update $\kappa^{(i)}$ using a random walk Metropolis-Hastings step.
Update $\alpha_1^{(i)}$ using a random walk Metropolis-Hastings step.
Update $\alpha_2^{(i)}$ using a random walk Metropolis-Hastings step.
Update $\alpha_3^{(i)}$ using a random walk Metropolis-Hastings step.
Update $\lambda_T^{(i)}$ using a random walk Metropolis-Hastings step.
Update $\lambda_V^{(i)}$ using a random walk Metropolis-Hastings step.
Update the augmented data $y$ times.
```

4.5.2 Parameter updates

In this section we explicitly show how each of the parameters are updated in our MCMC algorithm. Let $(\mathbf{Z}, \psi)$ denote the observed data, ρ the model parameters, F the unobserved data and ρ_{-x} to be the vector of parameters not including x.

Steps 2-5: From Equation (4.11) it is straightforward to see that we may sample β and γ_i for $i = 1, 2, 3$ from the following full conditional distributions,

$$\begin{aligned}\beta \mid \rho_{-\beta}, F, \mathbf{Z}, \psi \sim{} & \text{Gamma}\left(n_E - n_c + a_\beta, S + b_\beta\right) \\ \gamma_i \mid \rho_{-\beta}, F, \mathbf{Z}, \psi \sim{} & \text{Gamma}\left(\alpha_i n_E + a_{\gamma_i}, \sum_{j=1}^{n_E}(X_i - Y_i) + b_{\gamma_i}\right), \qquad i = 1, 2, 3\end{aligned}$$

where S is the double sum in Equation (4.7) and the quantity $\sum_{j=1}^{n_E}(X_i - Y_i)$ refers to the sojourn period associated with the rate parameter γ_i where X_i and Y_i are the appropriate variables.

Steps 6-11: For the remaining parameters, which are the shape parameters for the sojourn time distributions α_1, α_2 and α_3, the transmission kernel decay κ and transition and transversion rates λ_T and λ_V, we may update these parameters using a Metropolis-Hastings random-walk with a normal proposal distribution.

We may update the all remaining parameters using a Metropolis-Hastings random-walk with a normal proposal distribution. Let $\theta = \{\alpha_1, \alpha_2, \alpha_3, \kappa, \lambda_T, \lambda_V\}$ denote this parameter set, then for each $x \in \theta$ we propose

$$x' \sim N(x, \sigma_x^2),$$

where x' denotes the proposed parameter value which is drawn from a normal distribution with the mean set to the current parameter value x and σ_x^2 is the variance. Finally, the proposed value x' is accepted with probability

$$\min\left(1, \frac{\pi(\rho', F, E_{\mathcal{K}} \mid \mathbf{Z}, \psi)}{\pi(\rho, F, E_{\mathcal{K}} \mid \mathbf{Z}, \psi)}\right),$$

where ρ' is the parameter set with the proposal parameter value x'.

4.5.3 Augmented data updates

In this section we describe how to sample the unobserved quantities $F = (\mathbf{Y}, \tilde{\psi}, \mathcal{G})$ in Algorithm 4, step 12 which are the unknown infection times, the source of infection and unobserved genetic distances. Since the transmission tree is defined by the event times and source of infection, updating these quantities is equivalent to sampling from the space of transmission networks.

The general procedure is as follows. Let F^* denote the proposed data set such that $F^* = (\mathbf{Y}^*, \tilde{\psi}^*, \mathcal{G}^*)$ and define the proposal ratio $q_{F,F^*} = \Pr(F^* \to F)/\Pr(F \to F^*)$, which is the probability of making the reverse move divided by the probability of making the forward move.

Any proposal to the transmission tree may require the genetic data to also be updated where we use an identical process to that described in Section 3.5.2, with the difference that here we sample the number of transitions and transversions independently from the proposal distributions f_I and f_E for interior and exterior nodes respectively. Once a candidate data set has been proposed, we accept this move with probability

$$\min\left(1, \frac{\pi(\rho, F^*|z^{obs}, \psi, x^s)}{\pi(\rho, F|z^{obs}, \psi, x^s)} q_{F,F^*}\right),$$

otherwise we reject and move to the next step.

For the following sections we denote $v_E = \{i : E_i < \infty\}$ to be the set of individuals to ever become exposed to the pathogen and hence have a finite exposure time and $C(E_j) = \{i : I_i < E_j < R_i\}$ to be the set of individuals who are currently infective at time E_j. Finally, let $O(j) = \{i : s_i = j\}$ denote the set of offspring for farm j where we define the term offspring to be the set of farms who are considered to have been infected by j.

Sampling the unobserved data is not a straightforward task and standard data augmentation moves are not efficient to explore the space of transmission networks. Instead we develop carefully designed proposals to update the augmented data for more efficient sampling which we describe below.

4.5.3.1 Update an infection time

First we sample one of these exposed farms uniformly at random from v_I and denote this individual as j. Next we propose a new infection time $I_j^* = L_j - X$ where $X \sim \text{Gamma}(\alpha_1, \gamma_1)$ and (α_1, γ_1) are the current values of the shape and rate parameters for the presymptomatic period distribution and we set $E_j^* = I_j^* - L$, where L is the known constant latent period.

Note that if $E_j^* < I_{s_j}$, that is the proposed exposure time is before the infection time for the source, we reject immediately. Similarly if $E_j^* > R_{s_j}$ where the proposed exposure occurs after the source has been removed we also reject immediately. Next we check that the proposed infection time is before the exposure time of any offspring, i.e. $I_j^* < E_k$ for all $k \in O(j)$. Finally, we check that that the proposed infection time I_j^* is before any sampled sequence time associated with farm j. In all of these scenarios we arrive at an outbreak that is impossible and hence has probability zero.

After proposing E_j^* we may update the genetic data if necessary by imputing any unobserved sequences (and hence corresponding distances) at the time of exposure, and we denote the proposal ratio for the genetic data to be $\mathcal{Y}_{gen}$. The proposal ratio for this move is explicitly given by

$$\begin{aligned} q_{F,F^*} &= \frac{\Pr\left[\text{Gamma}(\alpha_1, \gamma_1) = I_j - E_j\right]}{\Pr\left[\text{Gamma}(\alpha_1, \gamma_1) = I_j - E_j^*\right]} \mathcal{Y}_{gen} \\ &= \left(\frac{I_j - E_j}{I_j - E_j^*}\right)^{\alpha_1 - 1} e^{-\gamma_1 (E_j^* - E_j)} \mathcal{Y}_{gen}. \end{aligned}$$

4.5.3.2 Update an infection time and source

This move is similar to above, except now we also propose to update the source of infection jointly with the infection time. First we sample one of these exposed farms uniformly at random from v_I and denote this individual as j. Next we propose a new infection time $I_j^* = L_j - X$ where $X \sim \text{Gamma}(\alpha_1, \gamma_1)$ and (α_1, γ_1) are the current values of the shape and rate parameters for the presymptomatic period distribution and we set $E_j^* = I_j^* - 2$.

Next we check that the proposed infection time is before the exposure time of any offspring, i.e. $I_j^* < E_k$ for all $k \in O(j)$. Finally, we check that that the proposed exposure time E_j^* is before any sampled sequence time associated with farm j. In all of these scenarios we arrive at an outbreak that is impossible and hence has probability zero.

Then we sample a source of infection s_j^* from the set of farms exerting infectious pressure $C(E_j^*)$ with probability

$$\Pr(s_j^* = i | E_j^*) = \frac{\beta e^{-\kappa d_{i,j}}}{\sum_{k \in C(E_j^*)} \beta e^{-\kappa d_{k,j}}},$$

however if this set is empty we reject the move.

After proposing E_j^* we may update the genetic data if necessary by imputing any unobserved sequences (and hence corresponding distances) at the time of exposure, and we denote the proposal ratio for the genetic data to be $\mathcal{Y}_{gen}$. The proposal ratio for this move is explicitly given by

$$\begin{aligned} q_{F,F^*} &= \frac{\Pr\left[\text{Gamma}(\alpha_1, \gamma_1) = I_j - E_j\right] \Pr(s_j = i | E_j)}{\Pr\left[\text{Gamma}(\alpha_1, \gamma_1) = I_j - E_j^*\right] \Pr(s_j^* = i | E_j^*)} \mathcal{Y}_{gen} \\ &= \left(\frac{I_j - E_j}{I_j - E_j^*} \right)^{\alpha_1 - 1} e^{-\gamma_1 (E_j^* - E_j)} \frac{\sum_{k \in C(E_j^*)} e^{-\kappa d_{kj}}}{\sum_{k \in C(E_j)} e^{-\kappa d_{kj}}} e^{-\kappa (d_{s_j, j} - d_{s_j^*, j})} \mathcal{Y}_{gen}. \end{aligned}$$

4.6 Simulation study

In this section we examine the performance of the algorithm on simulated data. We begin by simulating data sets with parameter values motivated by existing analyses (e.g. Mardones et al. (2010); Charleston et al. (2011); Lau et al. (2015) to name a few) and examine the properties of the algorithm, in particular the ability to recover the true simulated parameter values and the accuracy in determining the source of infection.

We also with to determine how well the algorithm can identify clusters of transmission, therefore we consider scenarios of $n_c = 1, 2, 3$ clusters and compare the parameter estimates and source accuracy for each.

4.6.1 Simulation method

In order to simulate from the model with a population of size n we begin by drawing a pair of (x, y) coordinates uniformly at random from the unit square for each farm to describe their geographical location in space. These coordinates are considered to be part of the observed data and hence we do not model how these are generated. We also specify the number of transmission clusters n_c and sample the initially infective farms uniformly at random from all susceptible premises.

We use an extension of the Gillespie algorithm for stochastic simulation (Bratsun et al., 2005) to generate outbreaks from non-Markovian epidemic models. The general idea is to simulate the time until exposure for each susceptible farm where the time until exposure is a random draw from an exponential distribution with rate to be the total infectious pressure received at that time and then choose the minimum of these as the time until next exposure. Then we compare the time until next exposure with the time of the next transition for all other compartments and choose the minimum of these as the next event.

In the event of an exposure, we must also draw the sojourn times for each compartment from the relevant Gamma distribution and then update the state of the system and move time forward. Otherwise, in the event of any other transition, we simply update the state of the system and move time forward.

For a population of size n with y removed farms by time t, there are exactly $n - y$ total transitions that may occur. Suppose the system is currently at time t, then the $K = n - y$ event times are drawn from the relevant distributions, denoted by τ_i for $i = 1, ..., K$. Then let $\tau = \min_i \tau_i$ denote the smallest of these event times and P

denote the farm where the next transition will occur. We set the new time $t = t + \tau$ and update the state of farm P.

Susceptible farms receive infectious pressure at time t from each of the $j \in C(t)$ where $C(t) = \{j : I_j < t < R_j\}$ is the set of currently infectious farms. Note that $C(t)$ may be empty, in which case the farm cannot be exposed to the pathogen and hence the outbreak has ceased.

If the transition for farm P is $S \to E$, i.e. an exposure has occurred, we update the state of the farm and also generate the future infection, symptom, examination and removal time. Next we sample a source of infection where the probability of choosing an infectious farm $i \in C(t)$ is proportional to $\beta e^{-\kappa d_{ij}}$. Otherwise we simply update the state of the farm for any other type of transition.

Next we generate genetic data consistent with the transmission tree and genetic model. More precisely we wish to simulate matrices of pairwise transitions, T, and transversions, V. We assume that genetic isolates are collected and sequenced at the examination times $\mathbf{X}$, which constitute the observed genetic data. However, to generate these distances we must construct the genetic network $\mathcal{G}$ from the true transmission network, which also requires simulating unobserved sequences at the times of exposure $\mathbf{E}$.

After constructing the 'true' genetic network, we simulate genetic distances from the mutation model between pairs of connected sequences in the network, i.e. for each $(i, j) \in E$ we draw

$$(T_{ij}, V_{ij}, N - T_{ij} - V_{ij}) \sim \text{Multinomial}(N; p_T(\tau_{ij}), p_V(\tau_{ij}), 1 - p_T(\tau_{ij}) - p_V(\tau_{ij})),$$

where τ_{ij} is the absolute difference in sampling times of the sequences, $p_T(\tau_{ij})$ and $p_V(\tau_{ij})$ are the probabilities of transition and transversion in τ_{ij} time units respectively and N is the size of the genome.

This may be achieved by simulating from successive Binomial distributions (Linderman et al., 2015) by first drawing the number of transitions $T_{ij} \sim \text{Bin}(N, p_T(\tau_{ij}))$ and then, conditional on the number of transitions, we draw the number of transversions $V_{ij} \mid T_{ij} \sim \text{Bin}\left(N - T_{ij}, \frac{p_V(\tau_{ij})}{1-p_T(\tau_{ij})}\right)$.

Similar to Section 3.6.1, we assume that for any nodes $(i, j) \notin E$ that are not directly connected we simply sum along the branches. We do not explicitly model the distance between unrelated sequences that are imported, however to generate complete matrices of pairwise transitions and transversions we draw these distances from a Poisson distribution with mean μ.

Finally, after generating the complete set of pairwise transitions and transversions, we discard the distances associated with the unobserved sequences at the time of infection and return only the observed data denoted by $\psi = (T, V, \nu, \eta, \tau)$.

4.6.2 Parameter estimation

We first simulate data sets with parameter values consistent with those in the literature and assess the performance of the algorithm. For each outbreak we set the population size to be $n = 30$ where the transmission parameters were chosen such that a 'typical' outbreak had a final size of roughly 20 farms. We start the outbreak at time $t = 0$ and run until there are either no infectious farms or no susceptible farms left to infect.

For person-to-person transmission we set the rate of contact between pairs i and j to be $\beta_{ij} = 0.01 \exp(-0.2 d_{ij})$, where d_{ij} denotes the Euclidean distance.

For the latent period we chose a fixed length of 2 days and for the pre-symptomatic, post-symptomatic and removal periods we assign Gamma(2,1), Gamma(1,1), Gamma(1,2) distributions respectively.

For the genetic data the length of the genome was $N = 8000$, the transition and transversion rates were set to be $\lambda_T = 4 \times 10^{-6}$ and $\lambda_V = 2 \times 10^{-6}$ respectively. The average transition and transversion distance between imported sequences were set to be $\mu_T = 20$ and $\mu_V = 20$.

For all future inference we fix the distance kernel parameter κ, the shape parameters $\alpha_1, \alpha_2, \alpha_3$ (similar to Jewell et al. (2008)) to the truth because we found that trying to infer all of these parameters, in addition to infection times, led to undesirable properties of the Markov chain, such as poor mixing and convergence.

For each of the parameters $\beta, \gamma_1, \gamma_2, \gamma_3$, we assign independent uninformative exponential prior distributions with mean 0.001 and for λ_T and λ_V we assign Uniform(0,1) priors. We simulated 300 data sets using the procedure and parameters outlined above and then attempted to infer the parameter values and exposure times using Algorithm 4. We repeat this procedure for $n_c = 1, 2, 3$ number of initially infected farms, resulting in a total of 900 simulated data sets and 900 outputs from the MCMC algorithm.

In each iteration we randomly select 50% of the infected premises and propose to update the infection time without changing the source, and then randomly select another 50% and attempt to update both the infection time and the source. We found that a combination of these moves aided the mixing properties of the algorithm such that the sampler was able to better explore the space of transmission networks.

Finally, we also run a similar algorithm assuming that only the epidemiological data are available. The aim here is to compare the posterior distribution of sources with and without genetic data as rough indicator of the extra information gain from using whole genome sequence data.

4.6.2.1 Results

For each simulated data set we ran the algorithm for 5000 iterations as a burn in and then store every 20th iteration until we have 1000 samples from the posterior. As a summary we calculate the posterior median of each parameter and summarise the distribution of these estimates in Figure (4.5). We first look at the results for the single transmission cluster scenario and then we compare these to the results for the two and three cluster case.

Single transmission cluster In general the parameters are well recovered by the algorithm seen clearly by the distribution of posterior medians estimates that are centered on the true simulated values in red. There are some posterior median estimates for the mutation rate parameters λ_T and λ_V which are significantly higher than the true simulated values. The reason is that in some of those outbreaks, farms

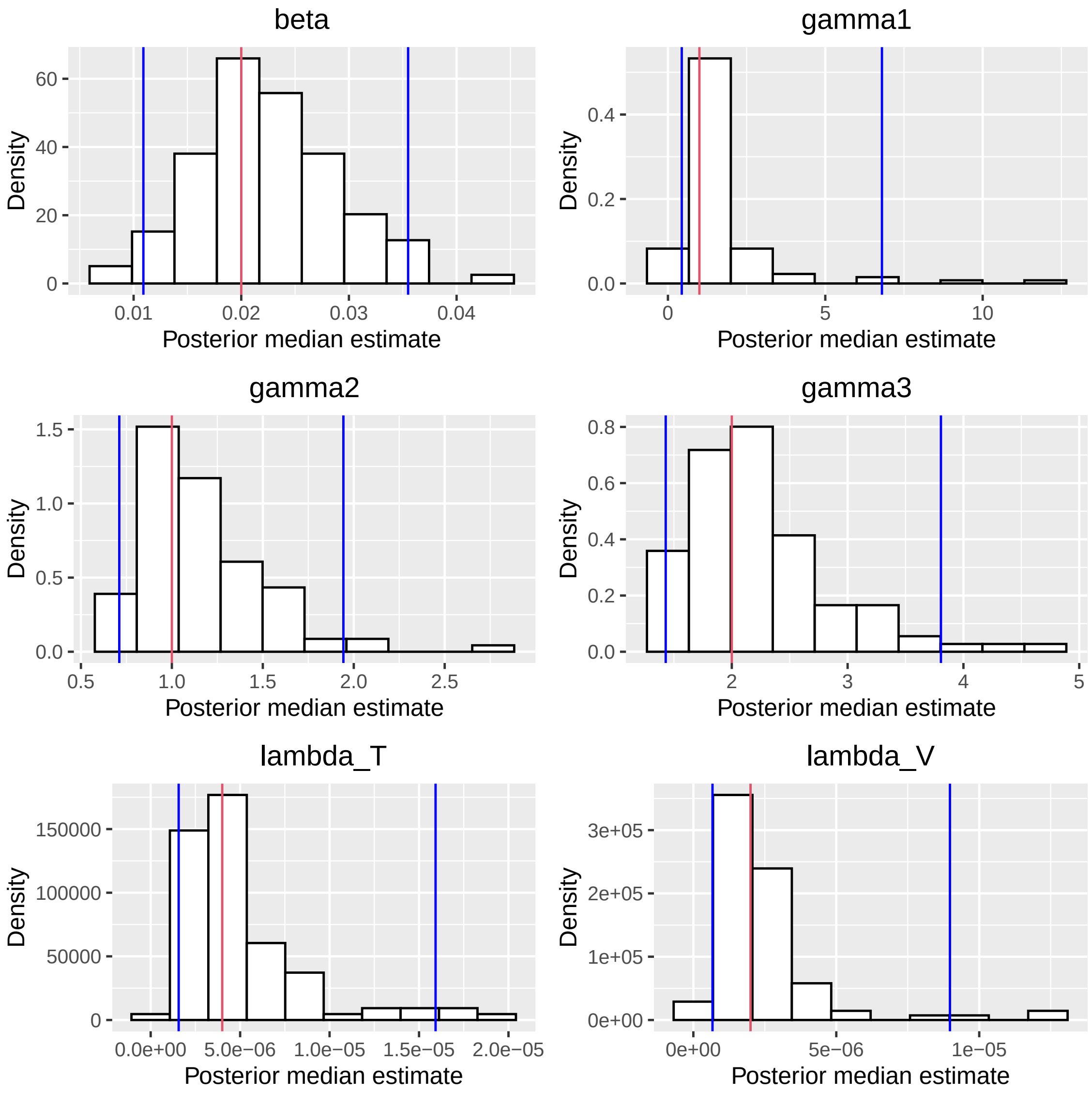


Figure 4.5: Posterior median estimates of the model parameters for 300 simulated data sets for the single cluster scenario. The red line indicates the true simulated value and the blue lines are the 95% credible intervals.

which are unrelated and hence have high genetic distances are considered to be related by the algorithm and the high genetic distances therefore drive up the estimate for the mutation rates.

In addition to the inference for model parameters, we are also interested in the determining the accuracy of the inferred infection times and the source of infection. In principle one would check each of these parameter individually for convergence and mixing, however in practice we instead consider a summary statistic of the infection times. For the accuracy of the inference of infection times, we consider the relative error between the inferred sum of infection times and the true sum of infection times. Let $\bar{I} = \sum I_j$ denote the true simulated sum of the infection times and $\hat{I}$ to denote the inferred sum of infection times which are calculated from the current state of the Markov chain in the algorithm. Finally, let $\hat{I}_M$ denote the median of the posterior samples, then the infection error is defined as

$$\epsilon = \frac{\bar{I} - \hat{I}_M}{\bar{I}}.$$

We use the term *source accuracy* (defined in Section 3.6.2) to be the proportion of correctly inferred sources, where an inferred source is the marginal posterior mode for a particular individual.

Finally, we define the term 'source improvement' to be the difference between the source accuracy inferred using the full model and the source accuracy inferred using only epidemiological data. This can be considered as a crude metric to estimate the information gain when including genetic data in this model.

In Figure 4.6 we can see that the error of the infection times sum is centered on zero which demonstrates that in general the algorithm is able to reasonably well infer the times at which the farms are infectious. For the source improvement we see that distribution has low density for values less than zero, specifically 32 of the 300 simulations (10.7%), suggesting that in general the algorithm does not perform worse than when using epidemiological data alone. Furthermore, from the 300 simulations we have 208 (69.3%) which show an improvement in accuracy after the inclusion

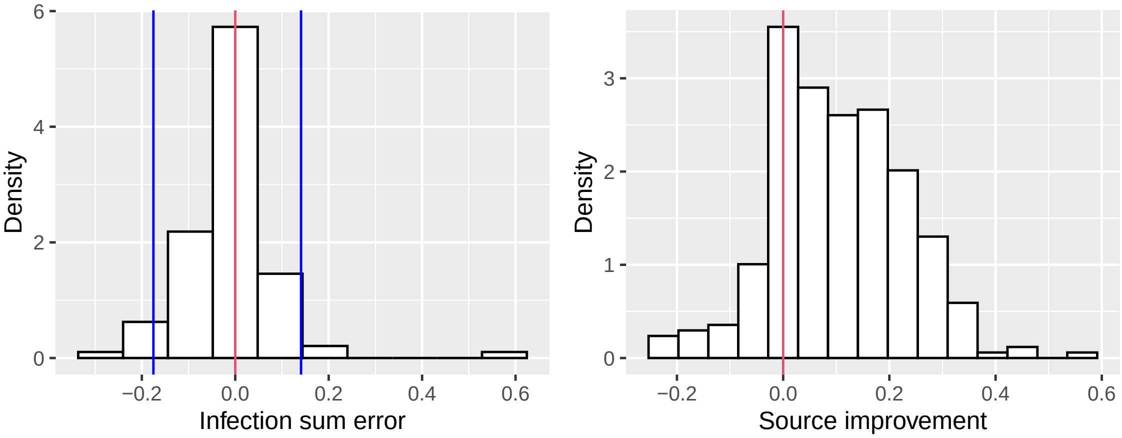


Figure 4.6: A summary of the augmented data for the transmission model, specifically the unobserved infection times summarised by the sum (left) and the source accuracy as a summary of the inferred transmission network (right).

of genetic data, indicating that the genetic data provides some insight into the transmission pathways of the pathogen through the population.

Multiple transmission clusters We have demonstrated that the algorithm tends to perform well in the event of a single initially infectious farm and hence a single transmission cluster, however it is also of interest to assess the performance of the algorithm in the event of multiple transmission clusters, i.e. $n_c > 1$. From the box plots in Figure 4.7 we see that the algorithm is able to successfully recover the true simulated parameter values well and there does not appear to be any real discernable differences between the inferences.

Furthermore, we see from Figure 4.8 that the error of the inferred sum of the infection times is also centered on zero, suggesting that the algorithm is in general able to successfully recover the true simulated times.

The source improvement increases as the number of clusters increase, which is unsurprising given the nature of the model. We assume that the average distance

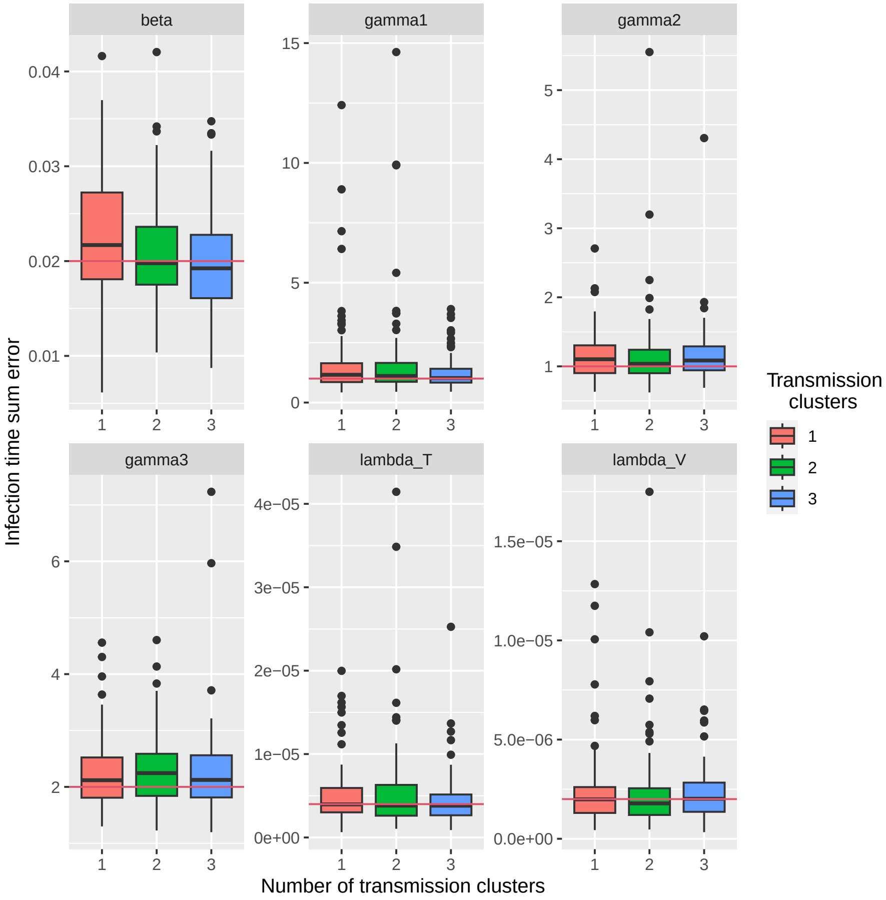


Figure 4.7: Posterior median estimates for multiple transmission clusters where we have one cluster(red), two clusters (green) and three clusters (blue). For each scenario we simulated 300 data sets and plot the posterior medians of each.

between sequences in distinct transmission clusters is $\mu_T = \mu_V = 20$, hence when proposing a new source of infection, some moves will be much less likely due the fact that some proposed configurations will have a high number of mutations between sequences in the genetic network. Each of the outputs from the algorithm were visually inspected for mixing and convergence. Example trace plots for a particular simulated data set can be seen in Appendix C.1 and the inferred posterior transmission network for this particular data set in Appendix C.2.

In conclusion, the MCMC algorithm (Algorithm 4) developed in this chapter is able to successfully recover the true simulated parameter values when the data are generated

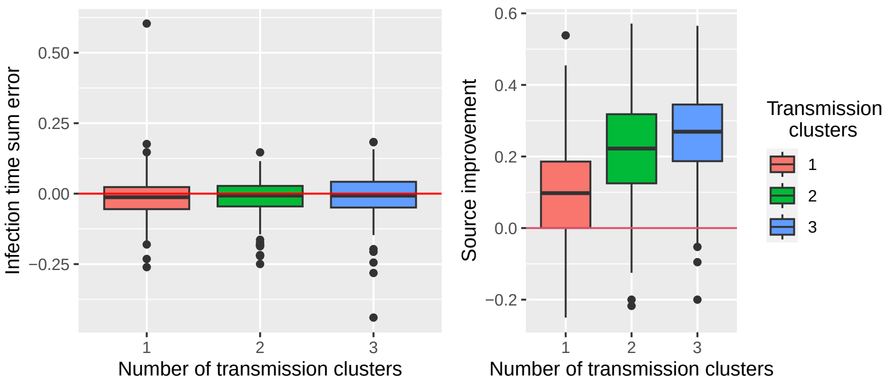


Figure 4.8: Summary of the augmented transmission data for simulated data for multiple transmission clusters where we have one cluster(red), two clusters (green) and three clusters (blue). Specifically we have the unobserved infection times summarised by the sum (left) and the source accuracy as a summary of the inferred transmission network (right).

from the SEILXR transmission model. Furthermore, we have demonstrated that, assuming a fixed and known latent period, we are able to reasonably well infer the exact times and source of infection. Furthermore, we have demonstrated that genetic data reduces uncertainty when inferring the source of infection on simulated data.

4.7 Analysis of the Darlington data set

In this section we wish to apply Algorithm 4 on the real data and compare the results to previous attempts at modelling FMD in the UK, in particular the work in Cottam et al. (2008); Morelli et al. (2012); Lau et al. (2015).

In our model we assume a fixed number of transmission clusters, therefore the specific choice of these will invariably have a significant impact on the analysis. Therefore we wish to consider competing models where each model reflects our belief about the transmission clusters in the data. We shall write these competing models as $M_i(\mathcal{A})$ where i denotes the number of independent transmission clusters and $\mathcal{A}$ is the set of

Model Number	Model	Description
1	$M_1(A)$	One cluster with A as the root
2	$M_1(B)$	One cluster with B as the root
3	$M_1(K)$	One cluster with K as the root
4	$M_2(A, K)$	Two clusters with A and K as the roots
5	$M_2(A, B)$	Two clusters with A and B as the roots
6	$M_3(A, B, K)$	Three clusters with A, B and K as the roots

Table 4.1: Competing models for the Darlington data set where $M_i(\mathcal{A})$ denotes $n_c = i$ transmission clusters with the roots of these clusters to be $\mathcal{A}$. Specifically we consider up to the three cluster scenario which is supported by exploratory analysis and previous modelling attempts.

farms that are assumed to have been the first case of the transmission cluster, where the size of this set is equal to i.

From the exploratory analysis (Section 4.3) there is strong genetic support that premises A and N are unrelated to all other transmission, hence we wish to consider various models that include A as the root of a transmission tree. Furthermore, due to the similarity in symptom times between farms A, B and K, it is plausible that all of these may have been independently infected from an outside site, hence we wish to consider scenarios where K or B are the roots of the clusters. An overview of all competing models can be found in Table 4.1. Similar to Section 4.6.2 we assign vague uninformative Gamma(1,0.0001) priors for parameters $\beta, \gamma_1, \gamma_2, \gamma_3$ and Uniform$(0, 1)$ priors for λ_T and λ_V. We fixed the distance kernel parameter to be $\kappa = 0.02$ and the latent period to be 5 days.

We ran the algorithm for 250,000 iterations with the first 50,000 discarded as a burn in and the remaining thinned by a factor of 200, resulting in 1000 samples from the posterior distribution. In each iteration we update the infection time without changing the source for all infected premises, and then propose to update the infection time and the source for all premises again. Finally, we propose to update 10 genetic distances at each iteration.

4.7.1 Results

We can see that the posterior distributions for each parameter are broadly similar across the three models (Figure 4.9). One striking feature of our inference is the posterior samples for the rate parameter γ_1 in the pre-symptomatic period ranges from 0.1 to 0.3, hence the expected values range between 6.6 and 20 days since the shape parameter is fixed to $\alpha_1 = 2$. The large differences indicate that the inference is sensitive to the specific model under consideration since the inferred parameter values offer different explanations for the data.

Finally, we compare the results for the mutation rate parameters in Figure 4.10 with a previous analyses of the virus. In Cottam et al. (2008) the authors estimate the overall mutation rate to be 2.26×10^{-5} (95% CI, 1.75×10^{-5} to 2.80×10^{-5}) and the transition transversion ratio to be 7.61, both of which are higher than what has been inferred in our analysis.

On the other hand, in Yoon et al. (2011) the authors estimate the overall mutation rate to be 4×10^{-6} (95% HPD, 2.9×10^{-6} to 5.1×10^{-6}) and the transition transversion ratio estimates range from 2.30 to 5.34. These estimates are lower than those obtained in our analysis, however since our inferred mutation rates are somewhere between previous estimates, we have some confidence that our results are plausible. To put it another way, if the inferred rates were significantly different to estimates in the literature, then this may suggest that our model offers a poor fit to the data.

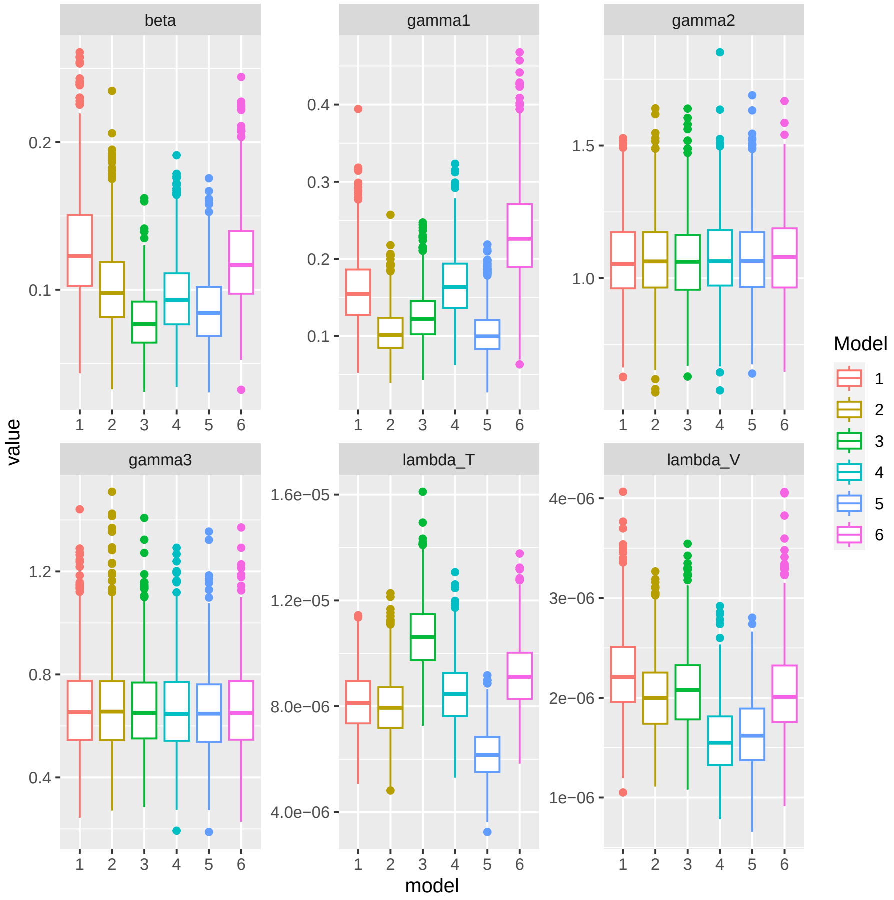


Figure 4.9: Posterior distributions for the inferred model parameters for the Darlington data set. The distributions are broadly similar and details of the competing models can be found in Table 4.1.

4.7.2 Inferred transmission networks

In this section we analyse the posterior transmission networks inferred by the algorithm. The networks in Figures (4.11)-(4.13) are weighted such that a directed edge from i to j with weight w indicates that farm i was considered the source of infection for farm j for w% of the iterations, i.e. the marginal posterior probability $\Pr(s_j = i|\cdot) = w$.

For $M_1(A)$, the model with a single cluster and A as the root of transmission, we notice that some expected links are present ($A \to N$ and $I \to J$), however the algorithm has failed to identify potential transmission sources, such as the G with I

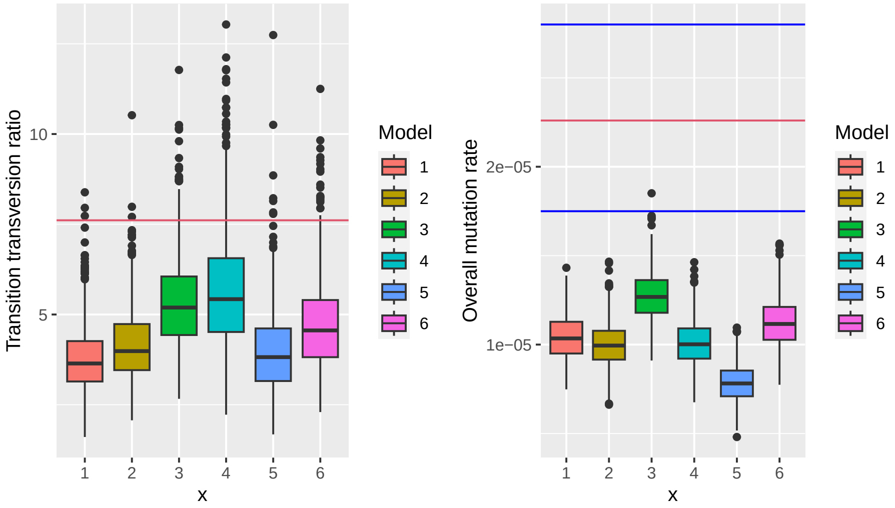


Figure 4.10: Validation of the model when considering the inferred mutation rates. Left: posterior distributions for the transition transversion ratio λ_T/λ_V with the estimated ratio from Cottam et al. (2008) overlaid in red. Right: posterior distributions for the overall mutation rate $\lambda_T + \lambda_V$ with the estimated rate from Yoon et al. (2011) overlaid in red and the HPDR overlaid in blue.

or J and any links between K and F, B, M, G, P.

For $M_1(B)$ with B as the root of the transmission cluster, we have $A \rightarrow N$ and $I \rightarrow J$ identified again. However this time there are extra links which are informed by genetic data, such as $B \rightarrow K$ and $B \rightarrow F$.

For the final single cluster model $M_1(K)$ with K as the root of the transmission tree, we see that the algorithm has managed to capture some strong links such as $A \rightarrow N$ and $G \rightarrow I \rightarrow J$. There are also other links that can be explained by the genetic data, such as $K \rightarrow B$ and $K \rightarrow F$.

For each of the three competing models, from a qualitative perspective we find that the analysis with K as the root of the transmission cluster provides the most likely transmission network in the sense that we have found several links that were identified as potential transmission in the exploratory analysis (Section 4.3).

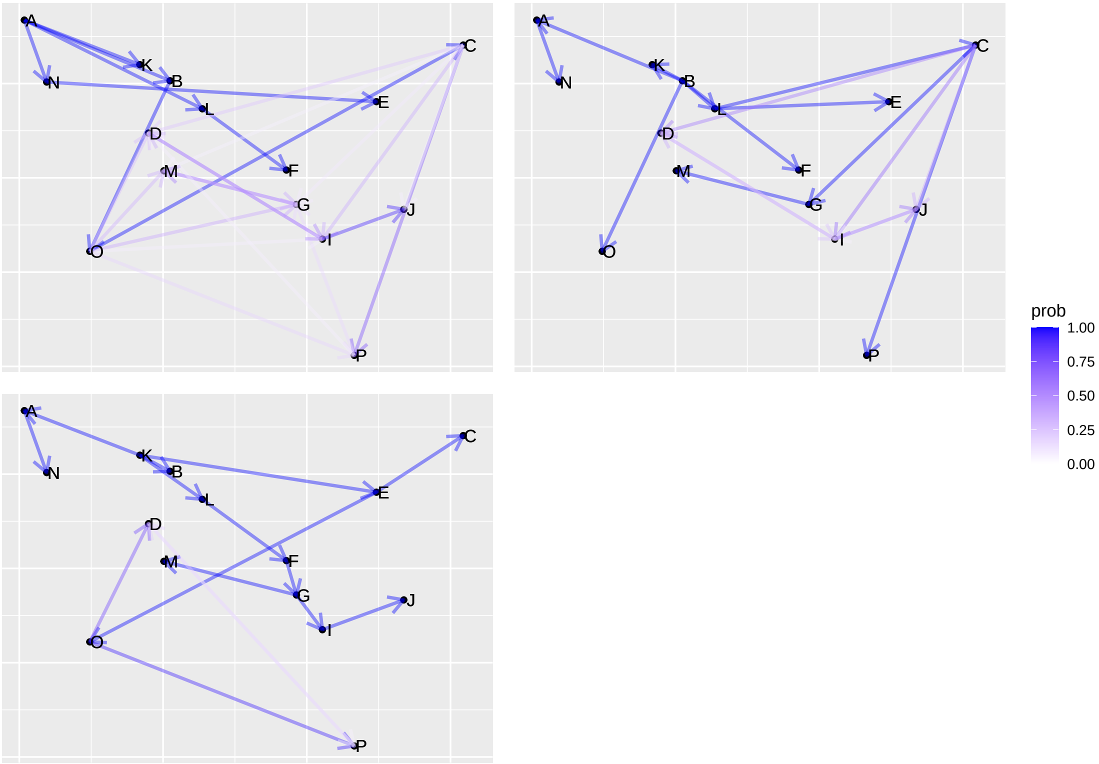


Figure 4.11: Inferred transmission networks for the single cluster scenario with (top left) $M_1(A)$, (top right) $M_1(B)$ and (bottom left) $M_1(K)$. The opacity of the edges are weighted by the posterior probability and stronger links are darker and blue.

For the two cluster scenario (Figure 4.12) we can immediately see in both scenarios than the (A, N) link has been identified as a separate cluster of transmission. We also find that model 4 ($M_2(A, K)$) has identified the (G, I, J) cluster suspected from the exploratory analysis.

For the three cluster scenario (Figure 4.13) we once more identify the (A, N) and (G, I, J) clusters. There are also some similarities to the inferred networks in Lau et al. (2015); Morris et al. (2001), such as the (K, L), (A, N) and (G, I, J) links, however the overall structure has some differences.

Our analysis for $M_3(A, B, K)$ indicates that the three clusters of transmission are (A, N), (K, L, O) and the third cluster consisting of the remaining farms. However previous studies have inferred the transmission network with clusters are (A, N), (B) and the third cluster again consisting of the remaining farms (Cottam et al., 2008; Lau et al., 2015).

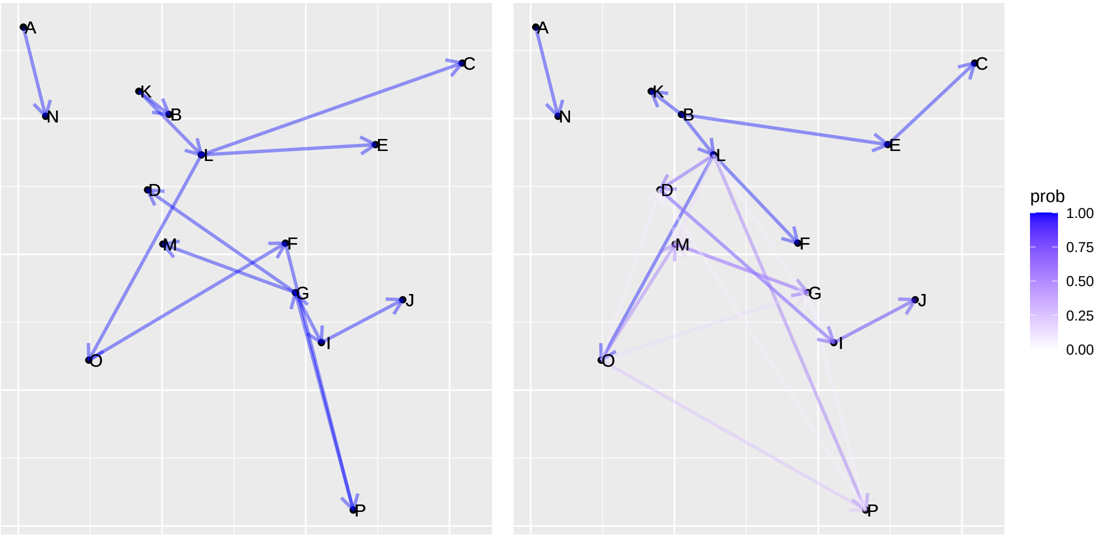


Figure 4.12: Inferred transmission networks for the two cluster scenario with (left) $M_2(A, K)$ and (right) $M_2(A, B)$. The opacity of the edges are weighted by the posterior probability and stronger links are darker and blue.

4.7.3 Including unreported susceptibles

The Darlington data set only includes information of premises that were confirmed infected and consequently we do not have data on farms that remained susceptible in the region during the outbreak.

In a similar fashion to Lau et al. (2015), we propose a similar methodology whereby we simulate 300 susceptible farms with a geographical location drawn uniformly at random from the study area. We repeat the analysis for all six models under consideration and use the same location of each farm for the models.

From the posterior distributions of the inference with susceptibles included (Fig-

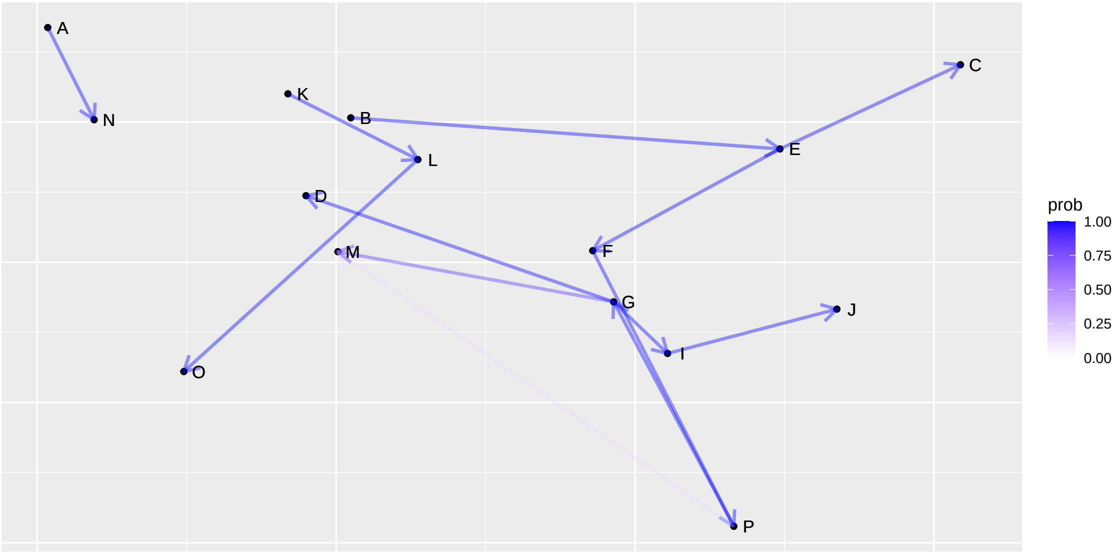


Figure 4.13: Inferred transmission networks for the three cluster scenario where the model is $M_3(A, B, K)$. The opacity of the edges are weighted by the posterior probability and stronger links are darker and blue.

ure 4.14), immediately we see that the transmission rate β is significantly lower, however this is unsurprising considering previously we had every farm in the population infected (15 out of 15), however in this case we have less than 5% of farms infected (15 out of 315). Consequently this drives down the estimate of the transmission rate.

In principle the other parameters should be largely unaffected by the inclusion of extra susceptibles since susceptible premises only contribute to the transmission part of the likelihood. Any other differences can be attributed to differences in the estimation of the unobserved data, in particular the infection times and transmission network, where the Markov chain ends up in a slightly different local mode.

4.8 Model assessment

In this section we assess the quality of our model fit through posterior predictive checking (see Section 3.7.5 and Gelman et al. (2013)). In particular, we are interested

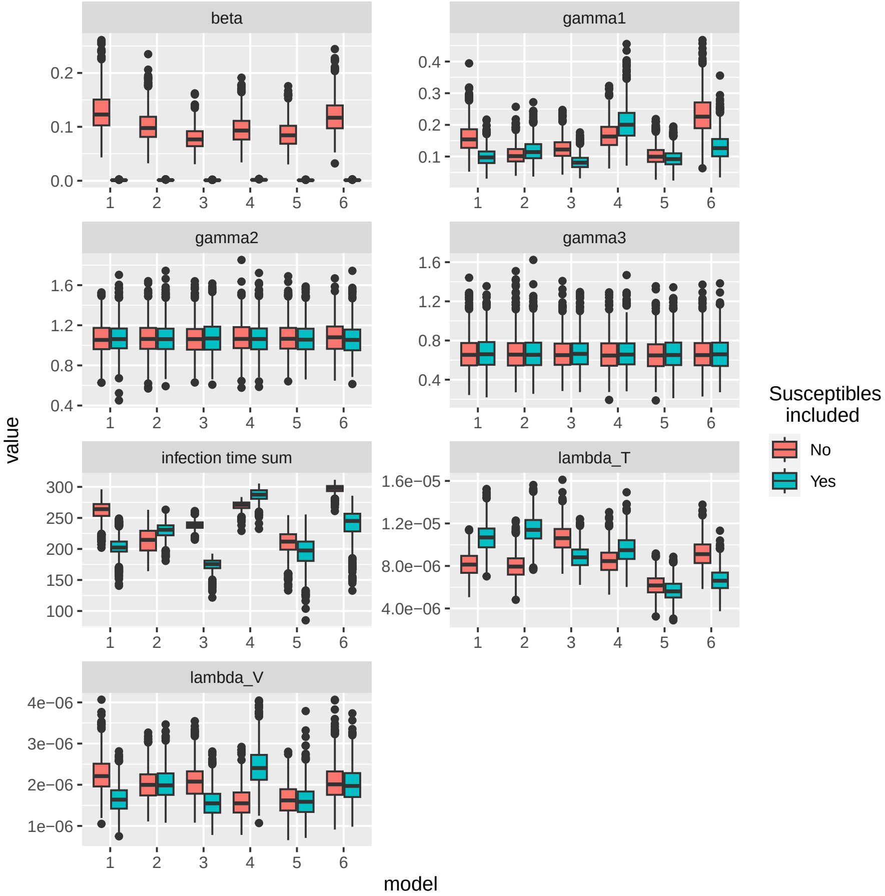


Figure 4.14: The posterior distributions of the model parameters for each of the competing models with and without unreported susceptible premises. The red plots indicate without susceptibles and the blue plots are with susceptible farms included.

in assessing whether each of our proposed models accurately describes the observed data, rather than identifying the 'best' of the competing models.

We shall now briefly describe posterior predictive checking in the context of epidemic modelling. Let $\mathbf{Z}$ denote the observed data, $\pi(\mathbf{Z} \mid \rho)$ the data generating process of the model and $\pi(\rho \mid \mathbf{Z})$ the posterior density of the model parameters ρ. Then we may draw replicated data $\mathbf{Z}^{\text{rep}}$ from the posterior predictive distribution, denoted by $\pi(\mathbf{Z}^{\text{rep}} \mid \mathbf{Z})$, which is explicitly given by

$$\pi(\mathbf{Z}^{\text{rep}} \mid \mathbf{Z}) = \int \pi(\mathbf{Z}^{\text{rep}} \mid \rho)\pi(\rho \mid \mathbf{Z})d\rho. \tag{4.12}$$

The posterior distribution is analytically intractable hence we describe how to compute Equation (4.12) numerically. Suppose we have K samples from the posterior distribution denoted $\rho_{[k]}$ for $k = 1, ..., K$, then for each sample from the posterior $\pi(\rho \mid \mathbf{Z})$ we simulate replicate outbreaks from the model $\pi(\mathbf{Z} \mid \rho)$ and denote these samples $\mathbf{Z}^{\text{rep}}_{[k]}$.

Comparison between observed data $\mathbf{Z}$ and replicates $\mathbf{Z}^{\text{rep}}$ can be complicated for high-dimensional data and it is therefore of interest to focus on real valued statistics or *test quantities*. Define the test quantity $T(\mathbf{Z}, \rho)$ to be a scalar statistics that captures features of the data, then we wish to compare the observed test statistic $T(\mathbf{Z}, \rho)$ to the distribution of the test statistic under replicated data $T(\mathbf{Z}^{\text{rep}}, \rho)$.

Finally, we define the Bayesian p-value similar to Section 3.7.5 as the probability that the replicated data is more extreme than the observed data, measured by the test quantity

$$p^B = \Pr(T(\mathbf{Z}^{\text{rep}}, \rho) \geq T(\mathbf{Z}, \rho) \mid \mathbf{Z}).$$

Recall that extreme p^B values close to 0 or 1 are evidence for poor model fit whereas values close to 0.5 indicate goodness of fit.

Our model can be decomposed into two parts, a model describing (i) the transmission of the pathogen and (ii) the evolution of the genetic sequences. We will assess the quality of these two separately and determine which of the competing models provided an adequate fit for the data.

4.8.1 Epidemic model assessment

In order to assess the goodness of fit for the transmission model we use the distance method described in Aristotelous et al. (2022) which looks at disease progression curves which explicitly assess the models ability to capture temporal features of the outbreak. However, rather than looking at removal curves, we look at *symptom curves* as a summary of the temporal progression of the outbreak. In principle one may also choose to look at *examination curves*, or indeed curves of any observed data, however we believe that these will be similar and offer little information gain.

The method is described as follows. Define $z_t(\mathbf{L}) = \sum_{i=1}^{n_E} \mathbb{1}_{\{L_i \leq t\}}$ as the *symptom curve*, which is simply the number of premises that have detected any symptomatic livestock up to time t. The test statistic is defined as $T(z_t) = d(z_t, \mathbb{E}z_t^{\text{rep}})$ on the space of symptom curves $\mathcal{L}$, where $\mathbb{E}z_t^{\text{rep}}$ is the mean of the replicated symptom curves and d is the Euclidean distance which is given by

$$d(z_t, z_t^*) = \left(\int_{\min(L_1, L_1^*)}^{\max(L_{n_E}, L_{n_E}^*)} (z_t - z_t^*)^2 dt \right)^{\frac{1}{2}},$$

where $z_t, z_t^* \in \mathcal{L}$ with corresponding time-ordered symptom vectors (lesion times) $\mathbf{L}, \mathbf{L}^* \in \mathbb{R}^{n_E}$.

The posterior predictive mean symptom curve is defined as $\mathbb{E}z_t^{\text{rep}} = z_t(\mathbb{E}\mathbf{L}^{\text{rep}}) = \sum_{i=1}^{n_E} \mathbb{1}_{\{\mathbb{E}(L_i^{\text{rep}}) \leq t\}}$ where

$$\mathbb{E}(L_i^{\text{rep}}) = \int L_i^{\text{rep}} \pi(L_i^{\text{rep}} \mid \mathbf{L}^{\text{obs}}) dL_i^{\text{rep}}, \quad i = 1, ..., n_E. \tag{4.13}$$

We may approximate the integral in Equation 4.13 by its Monte Carlo estimate $\frac{1}{K} \sum_{k=1}^{K} (L_i^{\text{rep}})_{[k]}$ for $k = 1, ..., K$.

Aristotelous et al. (2022) also introduce the idea of a ‘folded posterior predictive value‘ (fppp-value) which is defined as

$$\text{fppp-value} = \Pr(|\mathbf{Z}^{\text{rep}} - \mathbb{E}(\mathbf{Z}^{\text{rep}})| < |\mathbf{Z}^{\text{obs}} - \mathbb{E}(\mathbf{Z}^{\text{obs}})|),$$

which may be again approximated by the Monte Carlo estimate $\frac{1}{K} \sum_{k=1}^{K} \mathbb{1}_{\{T_{[k]}^{\text{rep}} < T^{rep}\}}$, i.e. the proportion of test statistics that have been calculated from outbreaks simulated from the posterior predictive distribution that are less than the observed test statistic, which in this case is the Euclidean distance between symptom curves. Folded ppp-values near 0 indicate goodness of fit and values near 1 indicate poor model fit. A more comprehensive overview of model assessment techniques and extensive simulation studies can be found in Aristotelous (2020).

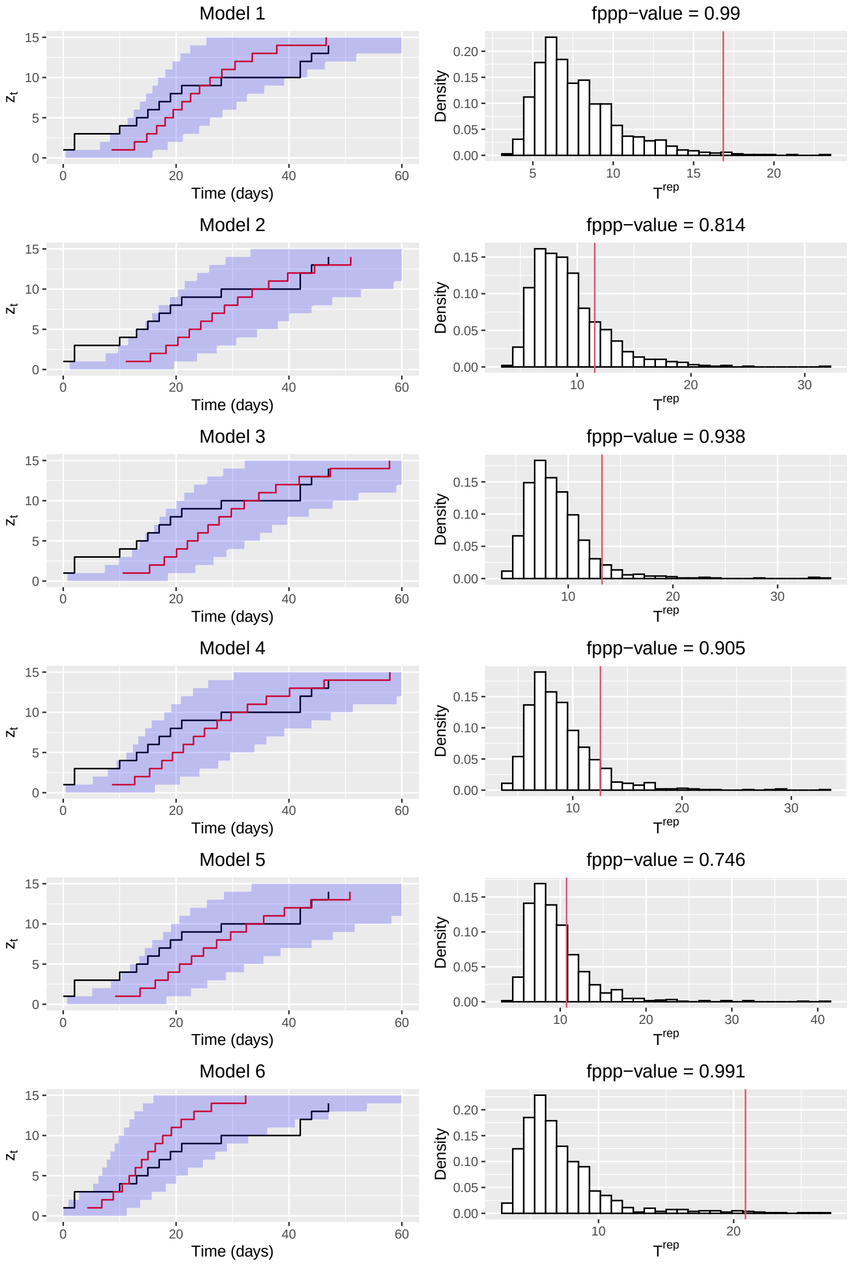


Figure 4.15: Epidemic model assessment for the Darlington data set for each of the competing models. Left column: observed symptom curve in black z_t^{obs}, the mean of the replicated symptom curves $\mathbb{E}z_t^{\text{rep}}$ and the 95% quantiles of the simulated replicate symptom curves are overlaid in blue. Right column: Histograms of 1000 replications from the posterior predictive distribution T^{rep} with the observed distance T^{obs} overlaid in red.

We apply this methodology for each of the competing models described in Section 4.7 and produce the simulated symptom curves and calculate the folded ppp-values.

The replicated symptom curves and distribution of the distances can be found in Figure 4.15. For the models with a single cluster we find that model 1 and model 3 have extreme ppp-values of 0.993 and 0.957 respectively, indicating poor model fit. All other models are slightly better, with fppp-values of 0.818, 0.886, 0.788 and 0.788 for models 2, 4, 5 and 6 respectively, indicating that all of these manage to capture some of the temporal aspects in the data.

Finally, we observe from visual inspection of the symptom curves that the model is able to capture the general progression of the epidemic, with the exception of the start which is always underestimated.

4.8.2 Genetic model assessment

In order to assess the goodness of fit of the genetic model we use the method described in Cassidy (2019). We shall briefly recall the procedure which is discussed in Section 3.7.5, with the modification that we are looking at the posterior predictive distributions for the matrix of pairwise transitions and transversions separately.

Suppose there are n_s genetic sequences, it follows that there are $N = \frac{1}{2}n_s(n_s - 1)$ pairwise transition and transversion distances denoted by T_{ij}^{obs} and V_{ij}^{obs} respectively for $j = 2, ..., N$, $i = 1, ..., j-1$. We wish to consider the marginal genetic distances separately for each i, j and calculate the posterior predictive p-value which is defined as

$$p_{ij}^{B[X]} = \Pr(X_{ij}^{\text{rep}} \geq X_{ij}^{\text{obs}}), \tag{4.14}$$

where $X \in \{T, V\}$ and X_{ij}^{rep} is the random posterior predictive distance. We may approximate these probabilities using the Monte Carlo estimate $\frac{1}{K}\sum_{k=1}^{K} \mathbb{1}_{\{X_{ij}^{\text{rep}} \geq X_{ij}^{\text{obs}}\}}$ where each X_{ij}^{rep} are samples from the posterior predictive distribution.

Finally, we define the term 'accepted proportion' (AP) to be the proportion of

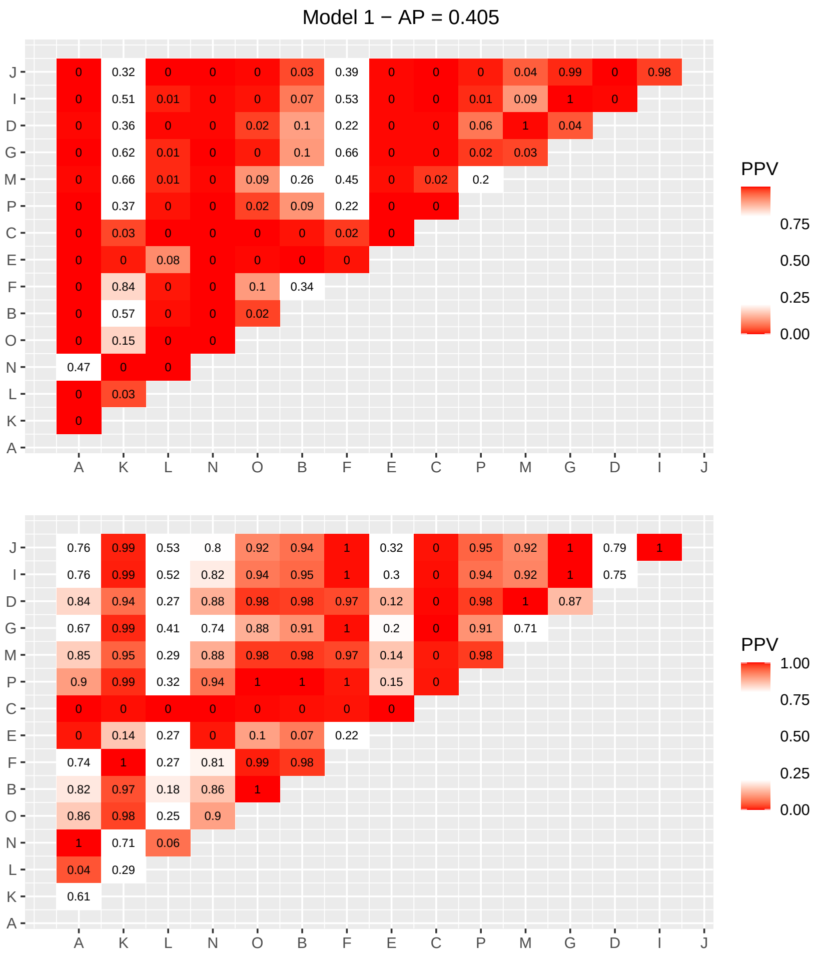


Figure 4.16: Genetic model assessment for model 1 for the Darlington data set. Matrix summary of pairwise marginal transition distances (above) and transversion distances (below) simulated from the posterior predictive distribution where the values in each cell are the Bayesian posterior predictive p-value (Equation (4.14)).

posterior predictive p-values which lie within the $100(1-\alpha)\%$ credible interval of the appropriate marginal distribution, where α is a value set to define what p-values are 'acceptable'.

More precisely, suppose there are $2N$ marginal pairwise posterior predictive distributions in total, then the accepted proportion is given by

$$AP = \frac{1}{N}\sum_{j=2}^{N}\sum_{i=1}^{j-1} \mathbb{1}_{\{p_{ij}^{B[T]}\in C(\alpha)\}} + \mathbb{1}_{\{p_{ij}^{B[V]}\in C(\alpha)\}},$$

Model Number	Model	AP
1	$M_1(A)$	0.405
2	$M_1(B)$	0.419
3	$M_1(K)$	0.481
4	$M_2(A, K)$	0.429
5	$M_2(A, B)$	0.395
6	$M_3(A, B, K)$	0.429

Table 4.2: Competing models for the Darlington data set where $M_i(\mathcal{A})$ denotes $n_c = i$ transmission clusters with the roots of these clusters to be $\mathcal{A}$. Specifically we consider up to the three cluster scenario which is supported by exploratory analysis and previous modelling attempts.

where $p_{ij}^{B[T]}$ and $p_{ij}^{B[V]}$ are the posterior predictive p-values for the transition and transversion distance between sequences i and j, and $C(\alpha) = \{p : \alpha/2 < p < 1-\alpha/2\}$ is the set of acceptable p-values. For the sake of clarity we present the posterior predictive matrices for model 1 in Figure 4.16 and the posterior predictive matrices for models 2-6 in Appendices (D.1)-(D.5).

From the summary of AP values in Table 4.2 we find that each of the competing models are broadly similar when considering the number of observed pairwise genetic distances that are captured by the model.

Model 3 which has a single transmission cluster K seems to provide a better fit than all other models. One possible reason for this is that the inferred transmission network, and hence the genetic network is more accurate.

It is important to note that in order to simulate genetic data from the model we must simulate distances between unrelated genetic sequences which are described by Poisson distributions with mean μ. The specific choice of μ significantly impacts the pairwise genetic distances between sequences in distinct transmission chains, however in this model assessment we set $\mu = 0$ in order to generate pairwise distances in distinct transmission clusters.

4.9 Discussion

In this chapter we have analysed a data set of small cluster, named the Darlington cluster, of foot-and-mouth disease in the United Kingdom in 2001. We have extended the general SEIR modelling framework to incorporate extra covariates, such as the timing of symptoms and examination which can naturally be accommodated in a stochastic compartmental model.

In order to incorporate the sequence data we have applied the genetic model developed in Chapter 2 where the data are pairwise transitions and transversions. Then we present novel Markov chain Monte Carlo techniques which allow for the imputation of unobserved genetic distances at the time of transmission.

We present a detailed simulation study that determines (i) how well the algorithm is able to recover the true simulated parameter values and (ii) how much information the genetic data is able to provide in the context of reconstructing the transmission network. We found that the algorithm was able to infer the parameter values well, with the distribution of posterior median estimates centered on the true value.

Furthermore, the algorithm was able to infer the infection times with good accuracy, also demonstrated by the distribution of the infection sum errors also centered on zero. More often than not we found that inclusion of the genetic data and the genetic model resulted in an increase of the proportion of correctly inferred sources, which is a nice summary statistic for the accuracy of the transmission networks.

We repeated the above simulations for the two and three cluster scenario, in order to assess how well the model performs under the assumption of multiple initial invectives. We found that in general the parameter estimates were consistent with the single cluster scenario, and in some cases there was less uncertainty of the posterior median estimates. We also found that the the error of the infection time sum was consistent with the single cluster analysis, and source accuracy increased with the number of transmission clusters.

In our model we do not allow for the possibility of background exposure, that is we specify the number of initially infective farms pre analysis. Consequently we consider various competing models that are informed by the exploratory analysis (Section 4.3) and previous modelling attempts (Lau et al., 2015; Morelli et al., 2012; Cottam et al., 2008) and analyse the results separately.

We found some links in our inferred transmission network to be the same in previous analyses, however there were also several important distinctions. Naturally the inferred transmission networks in our analysis and other previous analyses will not be identical due to differences in the model and inference, however it would be of interest to determine the strengths and weaknesses of these algorithms more closely.

Inference for transmission networks, and in our model genetic networks, is in general a non-trivial task due to the complex high dimensional nature of the problem. A key challenge when designing any MCMC algorithm is a balance between efficiency and complexity, one must design moves that are able to effectively explore the parameter space but also get accepted a reasonable amount of times.

Often when designing proposal distributions to sample the unobserved data (infection times, sources, genetic distances, etc.) we often choose to update a single or just a few data points. The rationale here is that the proposed move must be small enough such that it will be accepted enough times, however it also needs to be large to fully explore the state space. Designing more complex proposal distributions and sampling the unobserved data more efficiently is an avenue for future research, such as proposing block updates to the infection times and sources.

A key strength of our model remains the ability to capture important features of the genetic information using pairwise transitions and transversions alone, rather than working with raw sequence data. This strength is not completely utilised in the Darlington data set due to this small cluster having only 15 genetic sequence sampled, compared to say Chapter 3 where there are over 1800 genetic sequences. Nevertheless the algorithm is able to scale reasonably well and with the increased frequency of data collection and genetic sampling, scalable inference for large outbreaks offers an interesting avenue for future research.

Finally, we perform model assessment to determine the goodness of fit of the competing models and found that two of the six models were a poor fit for the transmission model and that the genetic models were broadly similar in terms of model fit. Interestingly, the model with the best fit for the genetic data performed comparatively worse for the epidemiological model fit.

In conclusion, from the model assessment we find that models 2, 3, 4 and 5 are all plausible explanations for the data and model selection should involve external factors such as expert opinion when explaining the outbreak. Exploratory analysis of the genetic data suggests that the A, N cluster is distinct from all other premises, therefore on balance we conclude that models 4 and 5 both provide a plausible explanation for the observed data. A more principled approach would be to perform formal model selection using methods such as Bayes factors (Alharthi et al., 2019), latent residuals (Lau et al., 2014) and Deviance Information Criterion (Deeth et al., 2015), to name a few.

4.9.1 Limitations

The biggest limitation with our implementation is the requirement of a fixed number of transmission clusters that are specified pre-analysis. A more flexible approach would be to include a constant background infectious pressure as a way to explain multiple introductions of the pathogen. With a constant background pressure, one can carefully design sophisticated proposals to the augmented data to fully explore the space of transmission trees. Preliminary results (not presented) indicate that this is very challenging and will be the topic of future research.

For example, we may propose to reduce or increase the number of transmission clusters (similar to the hospital data in Chapter 3 or Lau et al. (2015)) where such moves are informed by the sequence data. In principle these approaches provide a systematic way to identify the clusters of transmission within an outbreak. Furthermore, another clear benefit is that we do not need to make arbitrary decisions pre-analysis about the number of clusters, the idea being that the algorithm is able to infer these. We also would circumvent the need to compare various competing models and can analyse a single model directly.

Next another limitation is the assumption of a constant latent period rather than random. We are once again making an arbitrary decision about our belief of the length of the latent period, a more flexible approach would be to have this interval follow some distribution and then we could infer the parameters the characterise the distribution.

It is straightforward to simulate from and draw inference for a model with a random latent period, however this is only the case when the exposure times are fixed and known. We found that it would incredibly difficult to infer both the exposure and infection time for each individual, in addition to the rate parameters governing these distributions. Intuitively this is difficult since we are trying to infer two data points (E_j and I_j) for a farm j from a single data point L_j, the symptom time. Furthermore, these parameters and times are naturally correlated.

Next it would have also been interesting to also infer the shape parameters of the sojourn time distributions (namely $\alpha_1, \alpha_2, \alpha_3$), however we found that it was also difficult to estimate both the shape and rate of data that are assumed to be Gamma distributed, perhaps due to the fact that the total number of infected farms is $n_E = 15$ and hence there is not enough information in the data.

CHAPTER 5

Conclusion

In this thesis we have developed novel models and methods to integrate genetic data with traditional epidemiological data in a structured Bayesian framework. In particular, we have extended the work of Worby et al. (2016) and Cassidy et al. (2020) where the genetic data are represented by the matrix of pairwise distances.

In Chapter 2, our approach began by deriving the joint distribution of pairwise genetic distances under the assumption of a mutation model for sequences of length one. We then extended these ideas for sequences of length N and derive an expression for the joint distribution of these distances.

These distributions derived are all conditional on an underlying genetic network, in order to evaluate the probabilities of mutation we need some idea of how the observed sequences are related. We then provide a general methodology to construct genetic networks given the transmission network. Next we consider the contribution of genetic distances that are unrelated which is useful for modelling the multiple introductions of pathogen.

To show the generality of the proposed modelling framework developed we demonstrate the performance with two distinct scenarios; a discrete time model for nosocomial infections in Chapter 3 and also a continuous time model for the spread of foot-and-mouth disease virus in Chapter 4.

In Chapter 3 we extend both the genetic and transmission model motivated by features in the Brighton data set. Specifically, we provide a framework to incorporate healthcare worker data at an individual level and also extended the genetic model to allow for the possibility of multiple distinct genetic subtypes.

We demonstrate the performance of the MCMC algorithm on simulated data where we are able to successfully recover the simulated parameter values and also demonstrate that the addition of genetic data nearly always results in a more accurate inferred transmission network.

Finally, we present a detailed analysis of the Brighton data set and attempt to infer the model parameters and transmission dynamics for the MRSA and MSSA data separately and combined. We also present a comparison when using informative priors for the importation distance parameter and the inferred transmission networks in each case.

In Chapter 4 we apply the genetic model to a continuous time transmission model for the spread of foot-and-mouth disease virus. We extend the traditional SEIR framework to incorporate extra data sources (estimated date of lesion onset and examination) and attempt to infer the transmission network.

We show the algorithm has good performance on simulated data, that is we are able to reasonably well recover the true simulated parameter values and transmission networks. We then provide an analysis of the Darlington cluster data set and compare our inference to existing approaches.

The strength of our model lies with the fact we are working with pairwise genetic distances rather than raw sequence data. This is especially evident with the Brighton data set in Chapter 3 genetic data consist of 1976 genetic sequences, which invariably has significant computational costs when working with raw sequence data. The fact that our model uses the pairwise genetic distances as a summary of the data allows for fast and scalable inference.

5.1 Limitations and future work

The biggest avenue for future work and a limitation of the current model is the contribution of imported sequences that are considered unrelated. For example, in Lau et al. (2015) the authors assume that imported sequences are a result of a 'master' sequence that is unobserved and to be imputed. In our work we assume that pairwise genetic distances from unrelated sequences can be modelled by a Poisson distribution with parameter μ.

Both approaches are reasonable and pragmatic however ultimately are arbitrary, therefore it may be of interest to extend these approaches to include the inherent phylogenetic information contained in the sequences. A natural extension would be to derive the distribution of pairwise genetic distances assuming an underlying phylogenetic tree, however the phylogenetic tree is unobserved and inference for tree requires expensive MCMC algorithms (Nascimento et al., 2017). A more detailed introduction to Bayesian phylogenetics can be found in (Yang, 2014, Chapter 8). An approach similar to Klinkenberg et al. (2017) may provide a good starting point for future analysis where the authors perform simultaneous inference of both the transmission and phylogenetic tree.

Another limitation with the work in Chapter 4 is the lack of a model to explain background transmission. A consequence of this is that we specify the number of transmission clusters pre-analysis, however it would be of interest to infer the number of transmission clusters which in principle are informed by the genetic data.

We also would like to derive the distribution of pairwise distances under the most general nucleotide substitution model and evaluate the information gain by using more biologically accurate models (if any).

In Chapter 3 we could look at inferring the genetic subtypes η within the algorithm. At present we determine these groups pre-analysis and fix them for the duration, however in principle one should be able to propose to update these subtypes within the algorithm. This certainly would be useful in the sense that the arbitrary choice of partitioning the genetic groups would have less impact on the inference, assuming that the inferred groups converge to the truth from any starting point.

We are motivated to extend the transmission model described in Chapter 4 to naturally accommodate a constant background pressure to explain multiple introductions of the pathogen. This is straight forward to implement for transmission data, however care must be taken when modelling the genetic data.

Next, in all of the MCMC algorithms featured in the thesis we utilise simple proposal distributions for each of the Metropolis-Hastings updates. Certainly these algorithms can be improved to utilise more efficient sampling techniques, e.g. adaptive methods (Spencer, 2021) or Hamiltonian Monte Carlo techniques (Andrade and Duggan, 2020).

Finally, another interesting research direction one could take is the development of model assessment techniques for genetic data. The advantage of having a generative model for the genetic data is that we may assess model fit through Bayesian posterior predictive checking, a technique that has received very little attention apart from the work in Cassidy et al. (2020).

CHAPTER 6

Appendix

APPENDIX A

Simulation study results

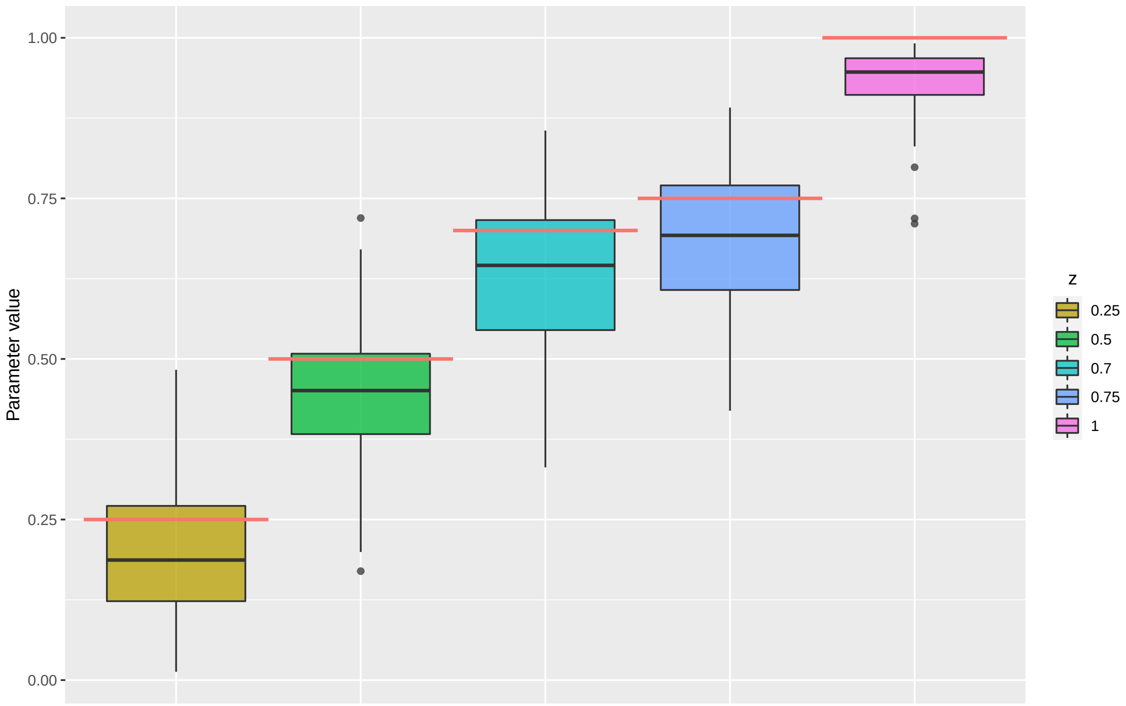


Figure A.1: Posterior median estimates of the parameter z from fitting the full model to 100 simulated data sets for each parameter choice of $z = 0.25, 0.5, 0.7, 0.75, 1$ with the true simulated value is overlaid in red.

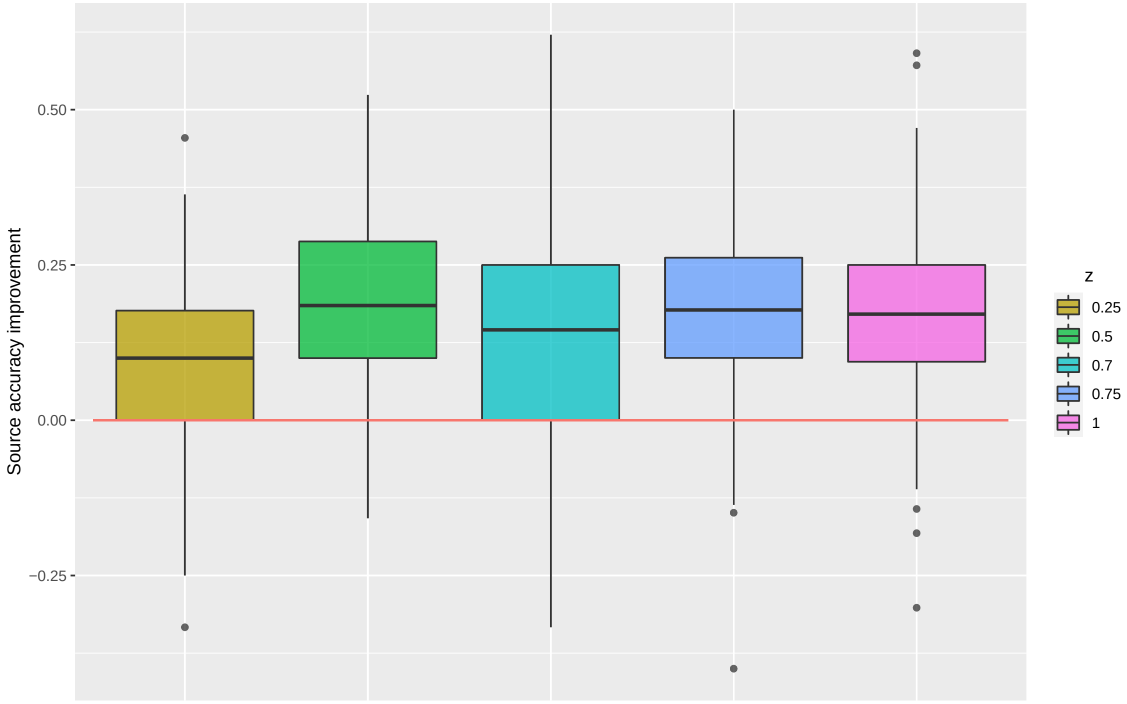


Figure A.2: Source accuracy improvement from fitting the model with and without genetic data to 100 simulated data sets for each parameter choice of $z = 0.25, 0.5, 0.7, 0.75, 1$. The line $y = 0$ has been overlaid in red which would indicate no improvement from the genetic data.

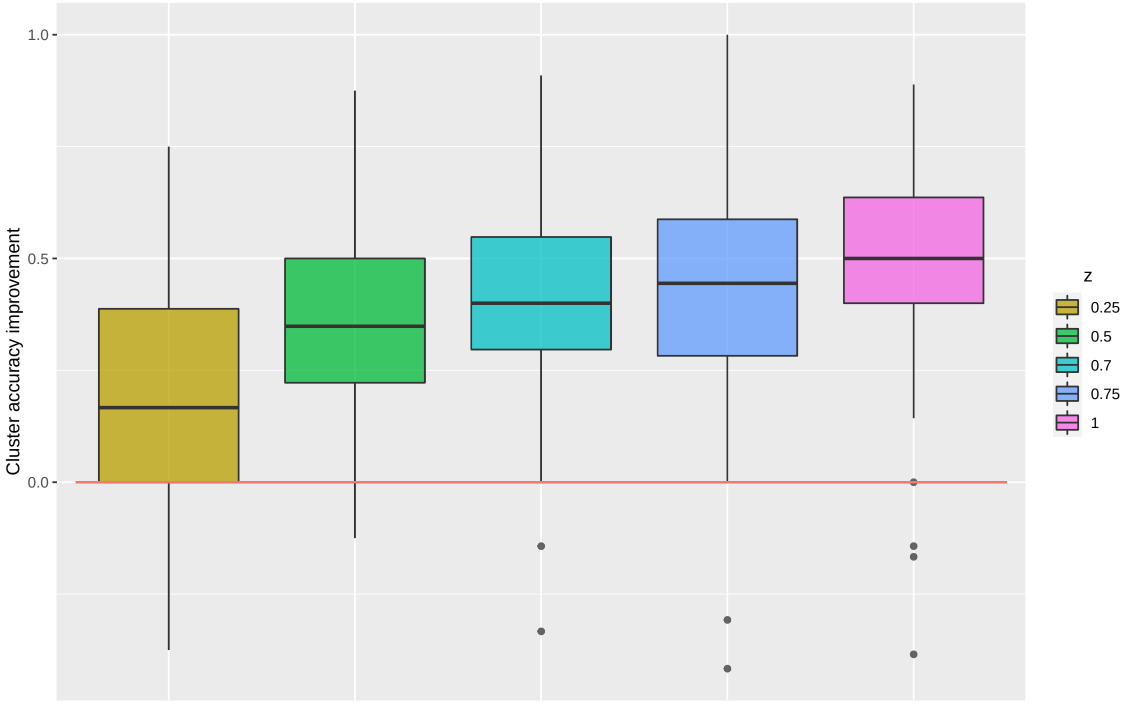


Figure A.3: Cluster accuracy improvement from fitting the model with and without genetic data to 100 simulated data sets for each parameter choice of $z = 0.25, 0.5, 0.7, 0.75, 1$. The line $y = 0$ has been overlaid in red which would indicate no improvement from the genetic data.

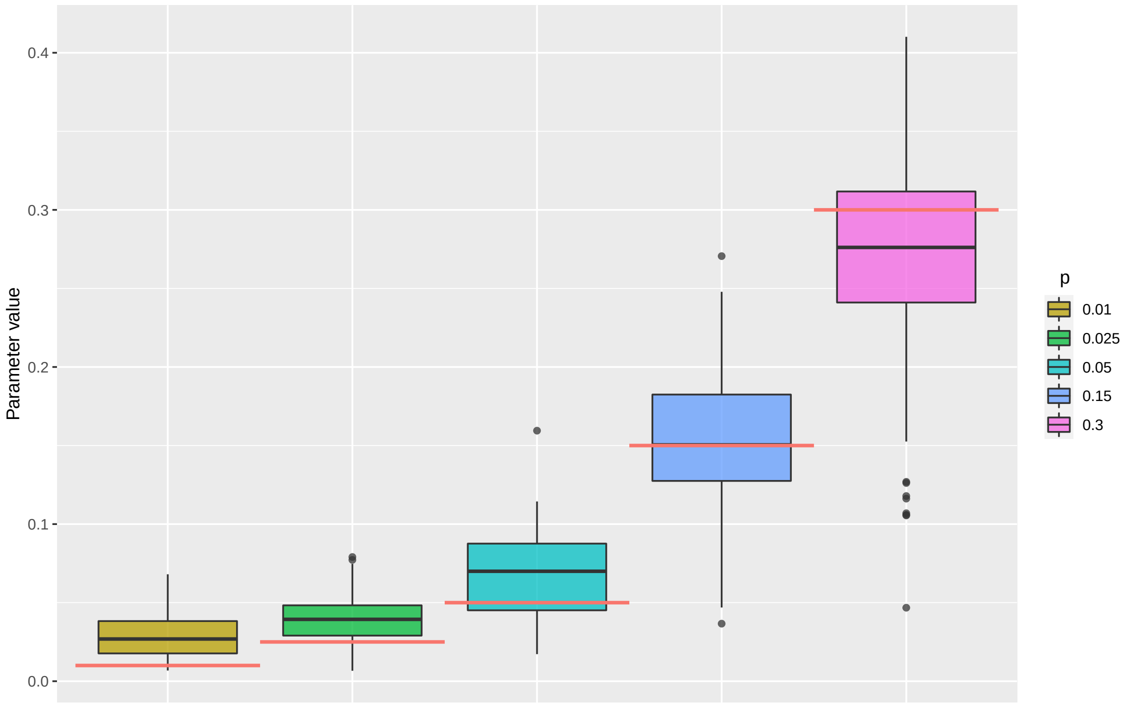


Figure A.4: Posterior median estimates of the parameter p from fitting the full model to 100 simulated data sets for each parameter choice of $p = 0.01, 0.025, 0.05, 0.15, 0.3$ with the true simulated value is overlaid in red.

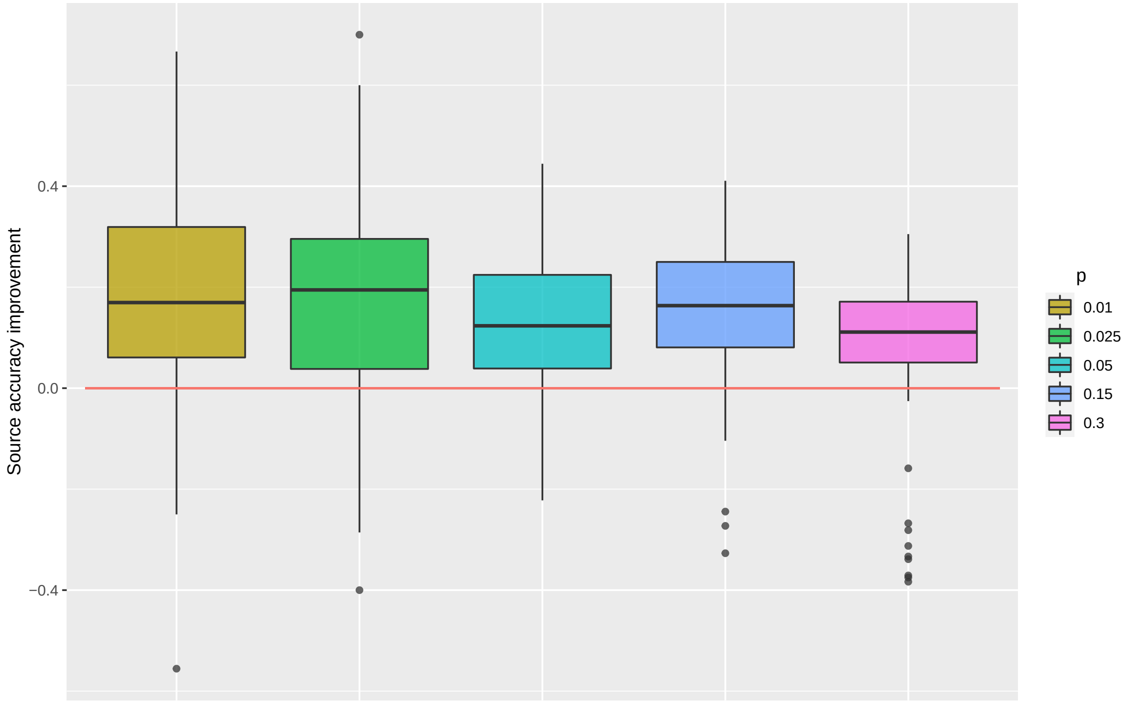


Figure A.5: Source accuracy improvement from fitting the model with and without genetic data to 100 simulated data sets for each parameter choice of $p = 0.01, 0.025, 0.05, 0.15, 0.3$. The line $y = 0$ has been overlaid in red which would indicate no improvement from the genetic data.

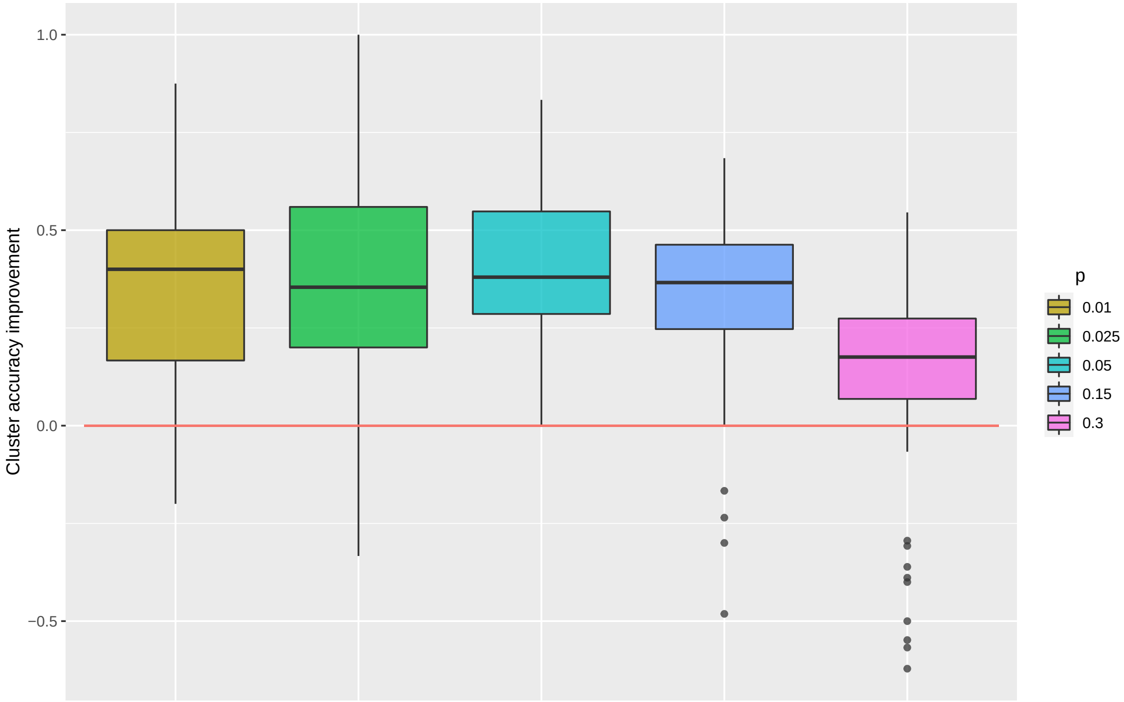


Figure A.6: Cluster accuracy improvement from fitting the model with and without genetic data to 100 simulated data sets for each parameter choice of $p = 0.01, 0.025, 0.05, 0.15, 0.3$. The line $y = 0$ has been overlaid in red which would indicate no improvement from the genetic data.

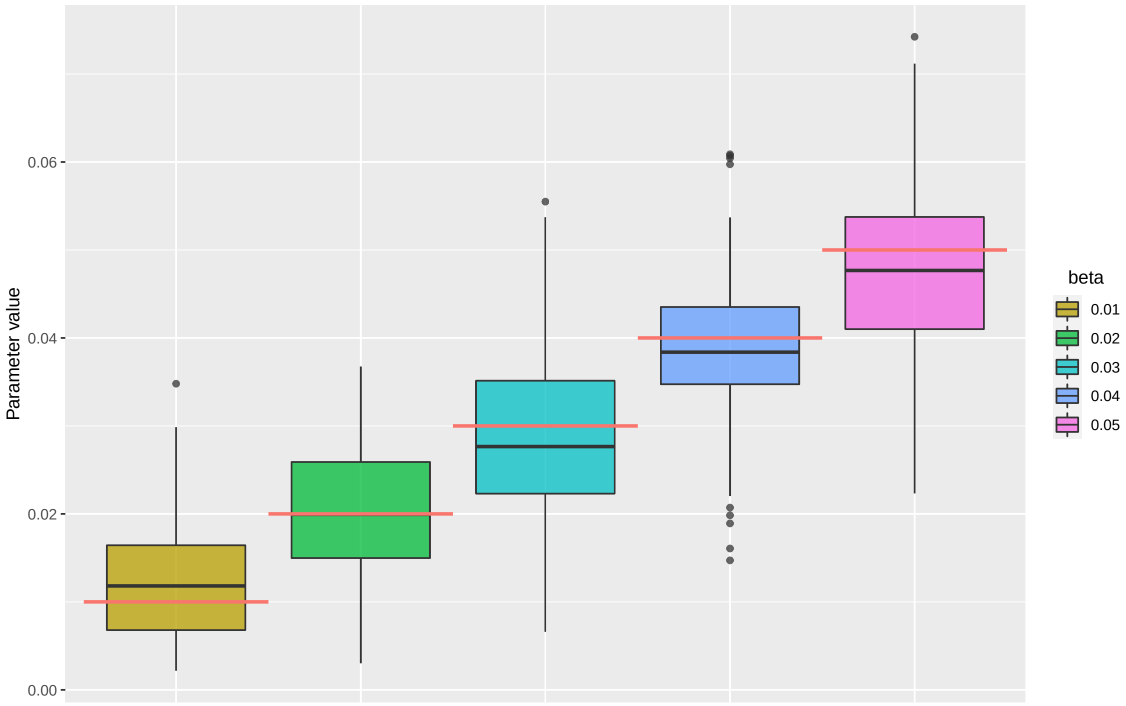


Figure A.7: Posterior median estimates of the parameter β from fitting the full model to 100 simulated data sets for each parameter choice of $\beta = 0.01, 0.02, 0.03, 0.04, 0.05$ with the true simulated value is overlaid in red.

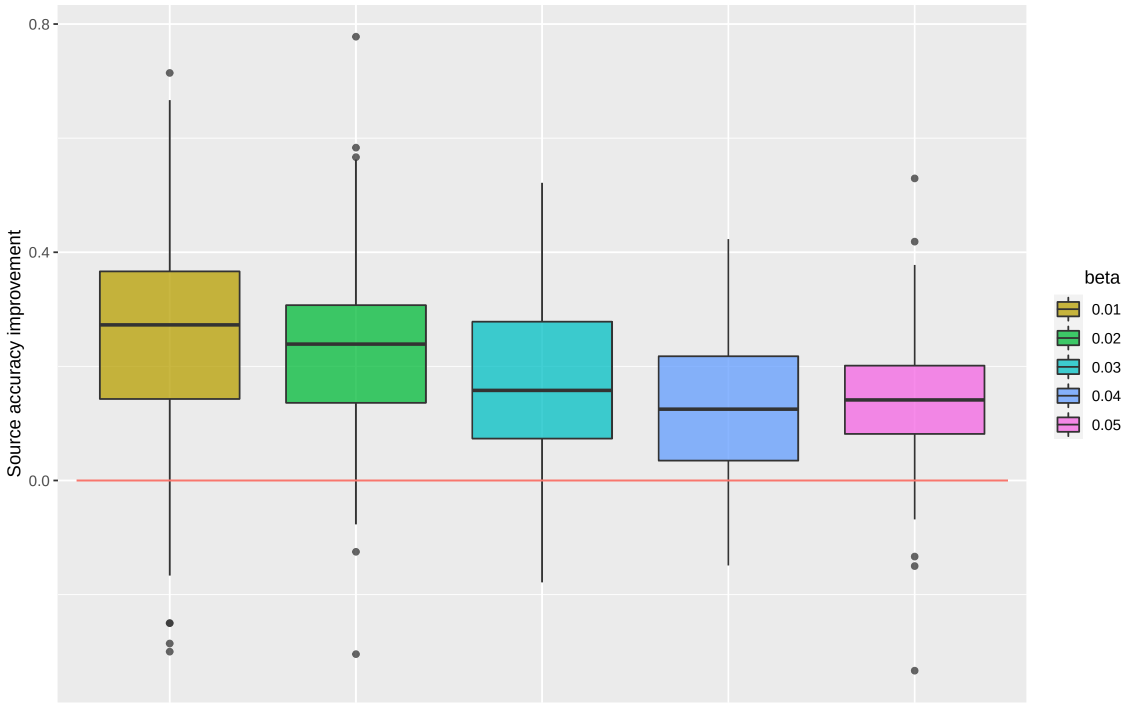


Figure A.8: Source accuracy improvement from fitting the model with and without genetic data to 100 simulated data sets for each parameter choice of $\beta = 0.01, 0.02, 0.03, 0.04, 0.05$. The line $y = 0$ has been overlaid in red which would indicate no improvement from the genetic data.

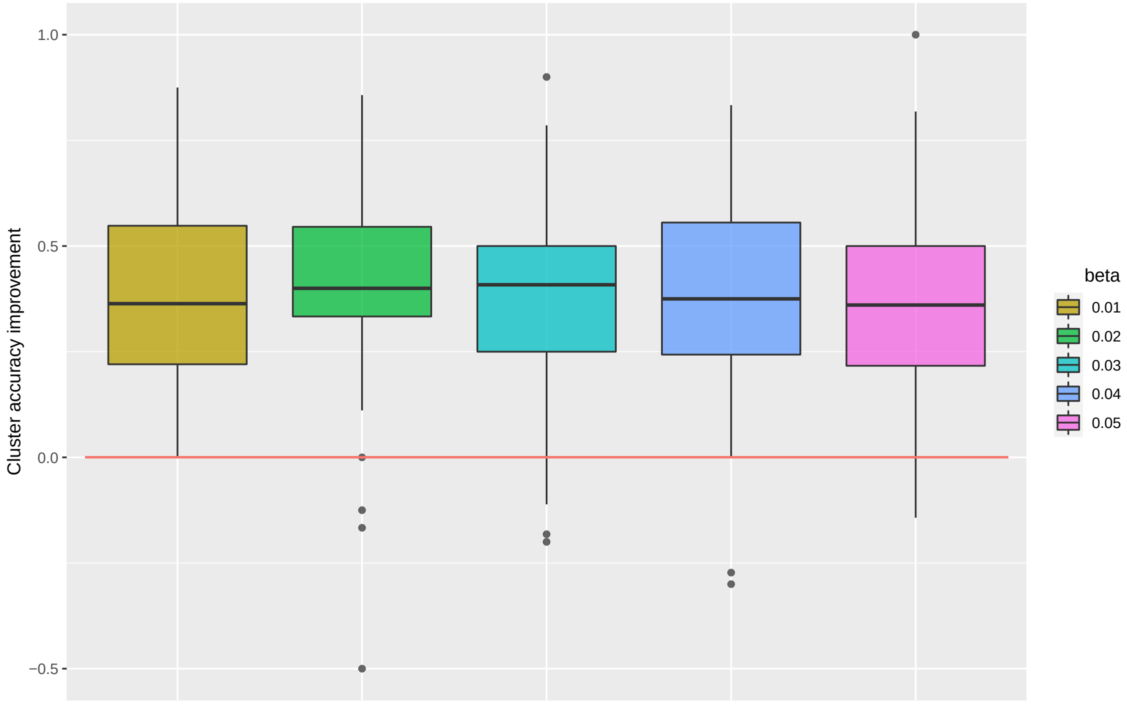


Figure A.9: Cluster accuracy improvement from fitting the model with and without genetic data to 100 simulated data sets for each parameter choice of $\beta = 0.01, 0.02, 0.03, 0.04, 0.05$. The line $y = 0$ has been overlaid in red which would indicate no improvement from the genetic data.

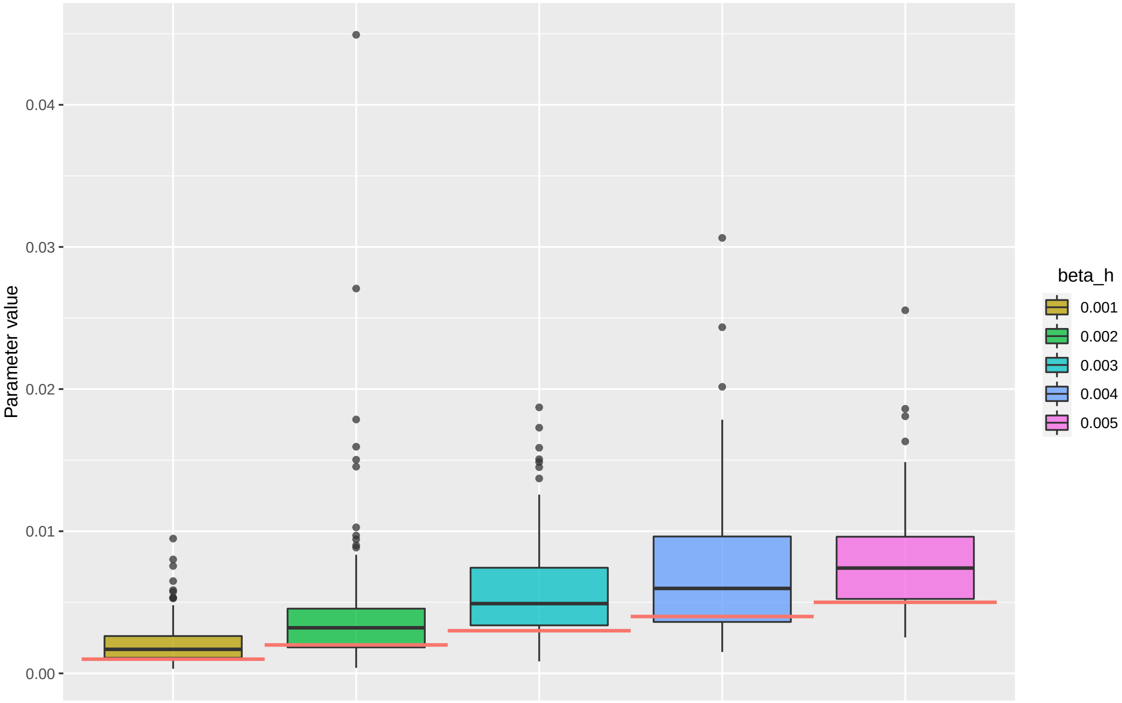


Figure A.10: Posterior median estimates of the parameter β_H from fitting the full model to 100 simulated data sets for each parameter choice of $\beta_h = 0.001, 0.002, 0.003, 0.004, 0.005$ with the true simulated value is overlaid in red.

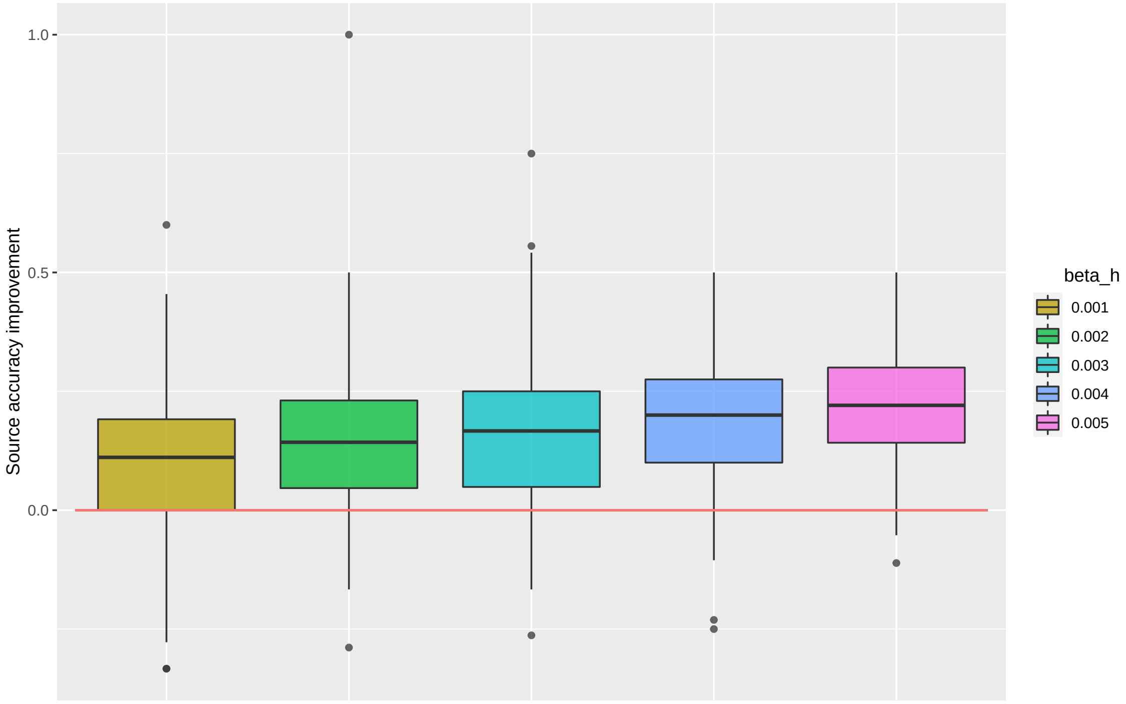


Figure A.11: Source accuracy improvement from fitting the model with and without genetic data to 100 simulated data sets for each parameter choice of $\beta_H = 0.001, 0.002, 0.003, 0.004, 0.005$. The line $y = 0$ has been overlaid in red which would indicate no improvement from the genetic data.

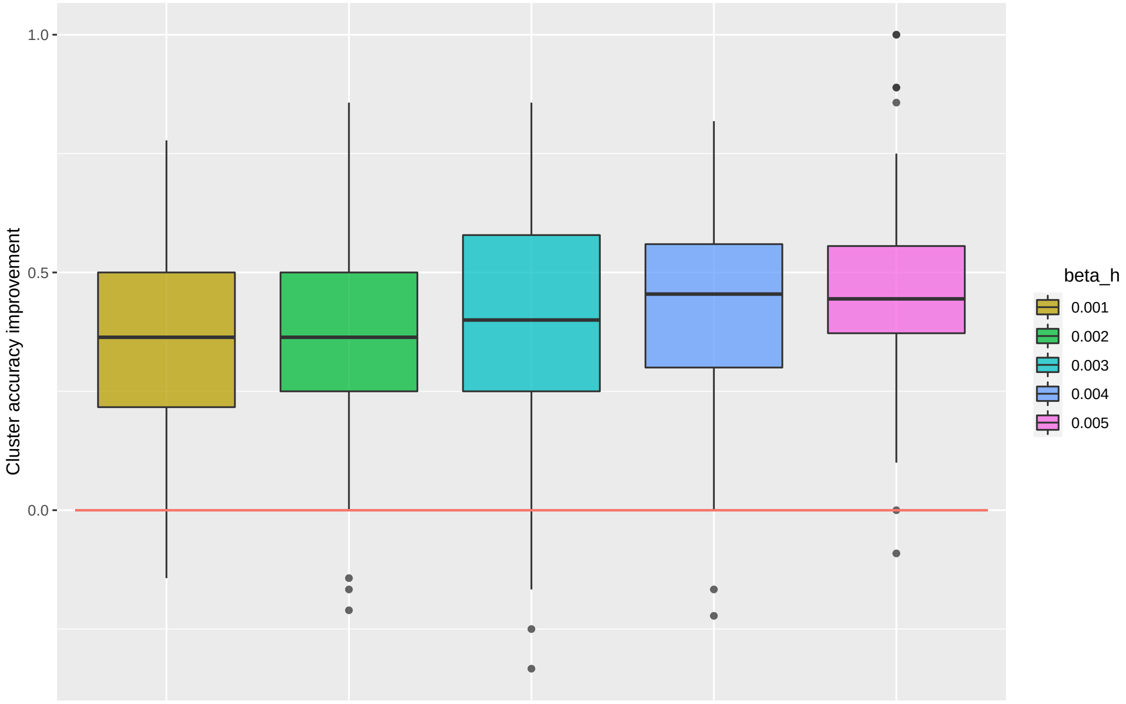


Figure A.12: Cluster accuracy improvement from fitting the model with and without genetic data to 100 simulated data sets for each parameter choice of $\beta_H = 0.001, 0.002, 0.003, 0.004, 0.005$. The line $y = 0$ has been overlaid in red which would indicate no improvement from the genetic data.

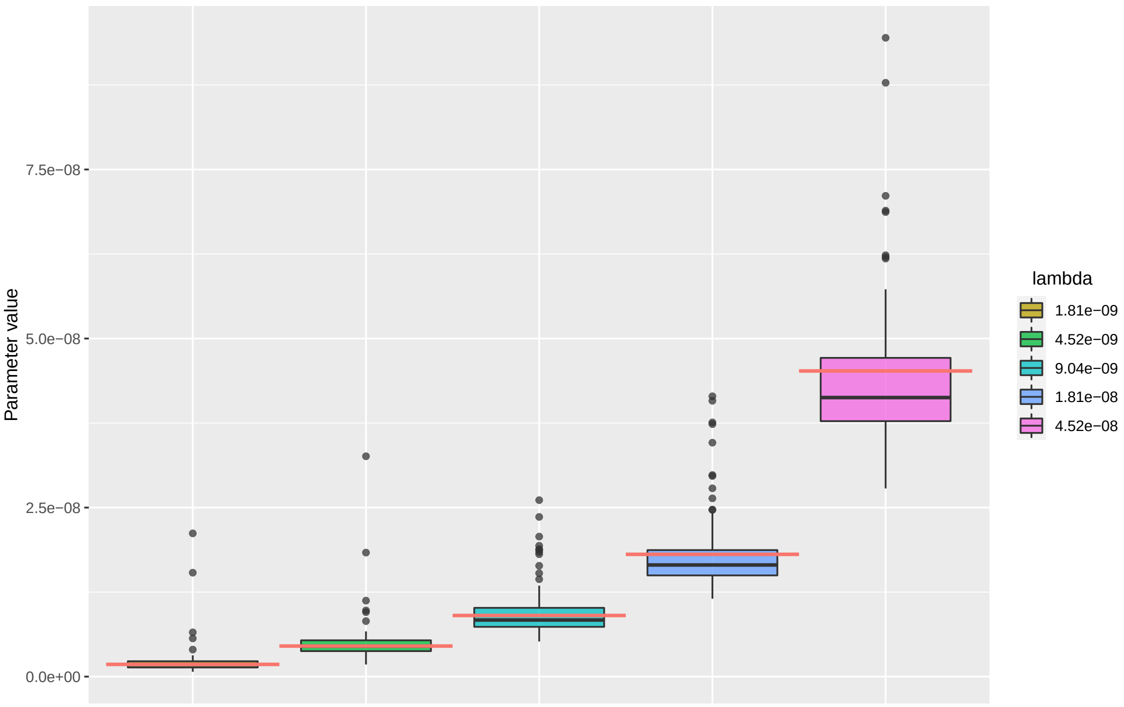


Figure A.13: Posterior median estimates of the parameter λ from fitting the full model to 100 simulated data sets for each parameter choice of $\lambda = (1.80, 4.52, 9.04, 18.08, 45.20) \times 10^{-9}$ with the true simulated value is overlaid in red.

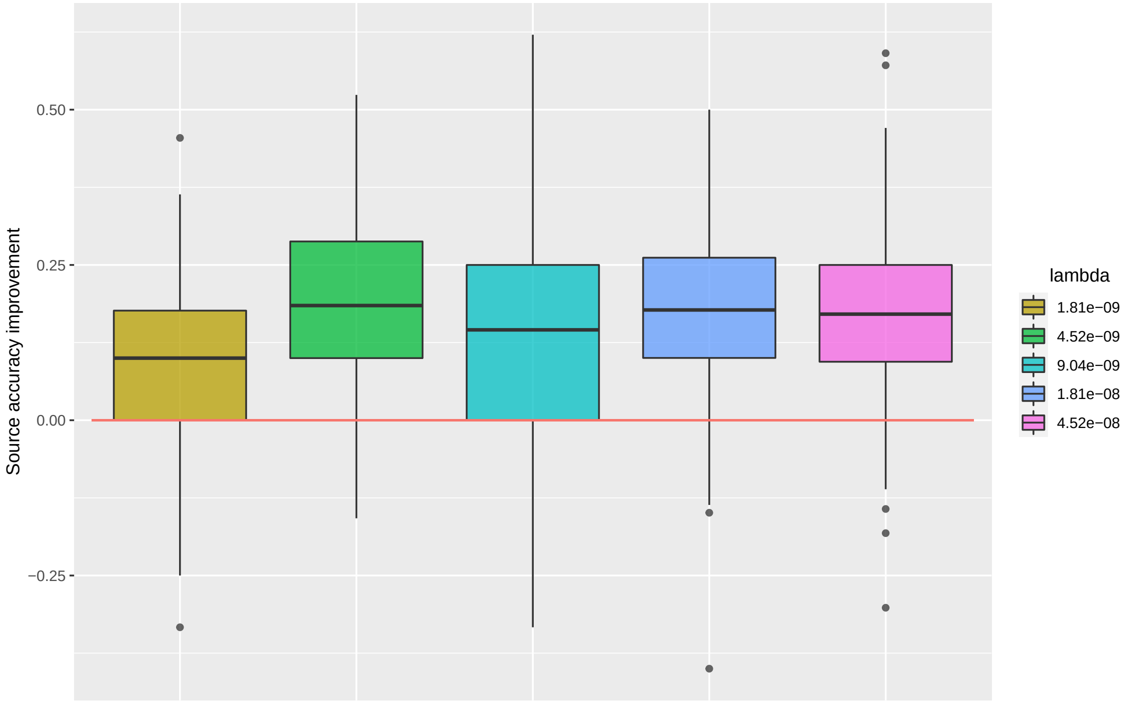


Figure A.14: Source accuracy improvement from fitting the model with and without genetic data to 100 simulated data sets for each parameter choice of $\lambda = (1.80, 4.52, 9.04, 18.08, 45.20) \times 10^{-9}$. The line $y = 0$ has been overlaid in red which would indicate no improvement from the genetic data.

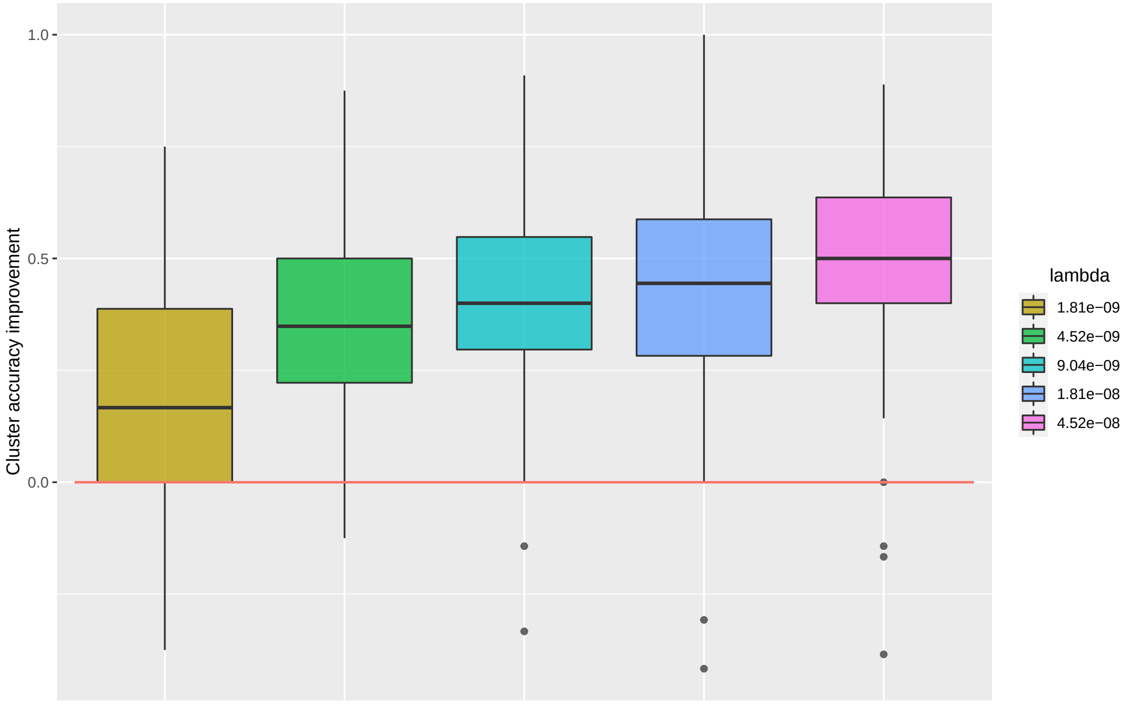


Figure A.15: Cluster accuracy improvement from fitting the model with and without genetic data to 100 simulated data sets for each parameter choice of $\lambda = (1.80, 4.52, 9.04, 18.08, 45.20) \times 10^{-9}$. The line $y = 0$ has been overlaid in red which would indicate no improvement from the genetic data.

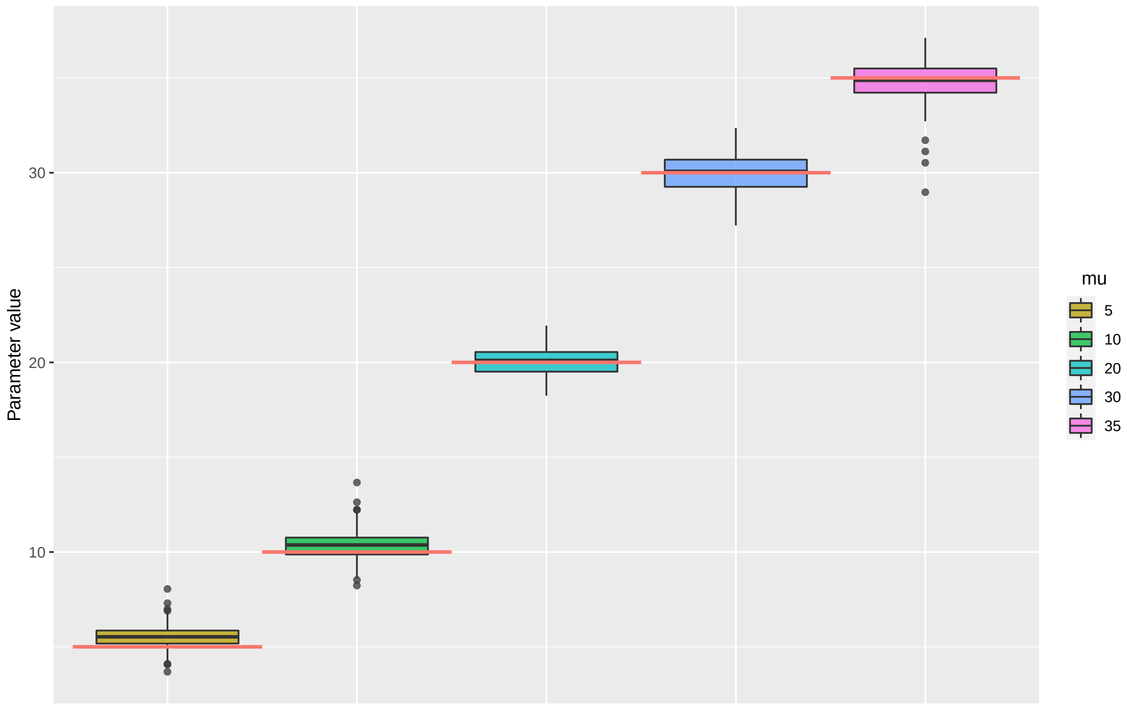


Figure A.16: Posterior median estimates of the parameter μ from fitting the full model to 100 simulated data sets for each parameter choice of $\mu = 5, 10, 20, 30, 35$ with the true simulated value is overlaid in red.

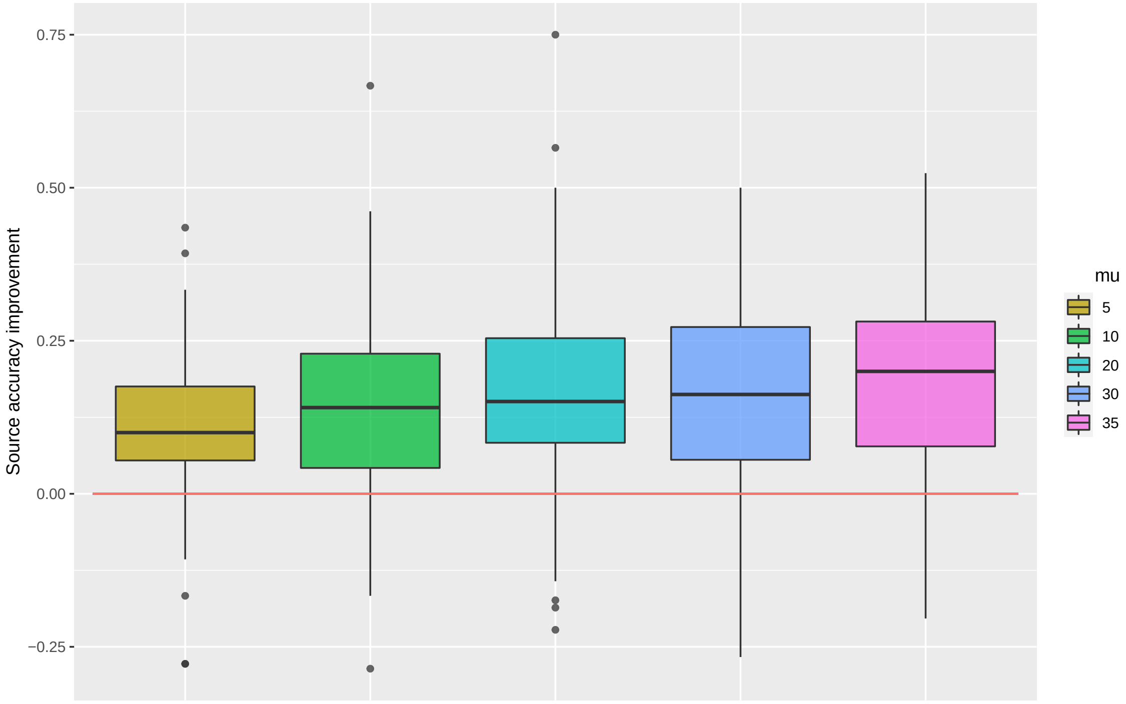


Figure A.17: Source accuracy improvement from fitting the model with and without genetic data to 100 simulated data sets for each parameter choice of $\mu = 5, 10, 20, 30, 35$. The line $y = 0$ has been overlaid in red which would indicate no improvement from the genetic data.

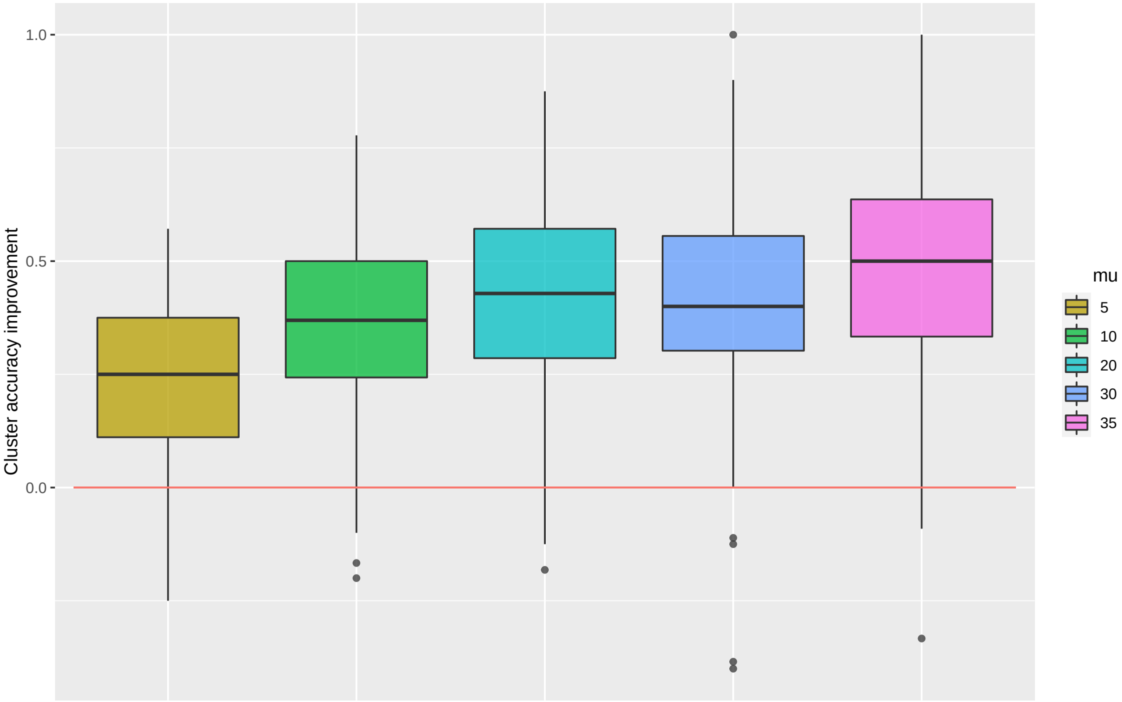


Figure A.18: Cluster accuracy improvement from fitting the model with and without genetic data to 100 simulated data sets for each parameter choice of $\mu = 5, 10, 20, 30, 35$. The line $y = 0$ has been overlaid in red which would indicate no improvement from the genetic data.

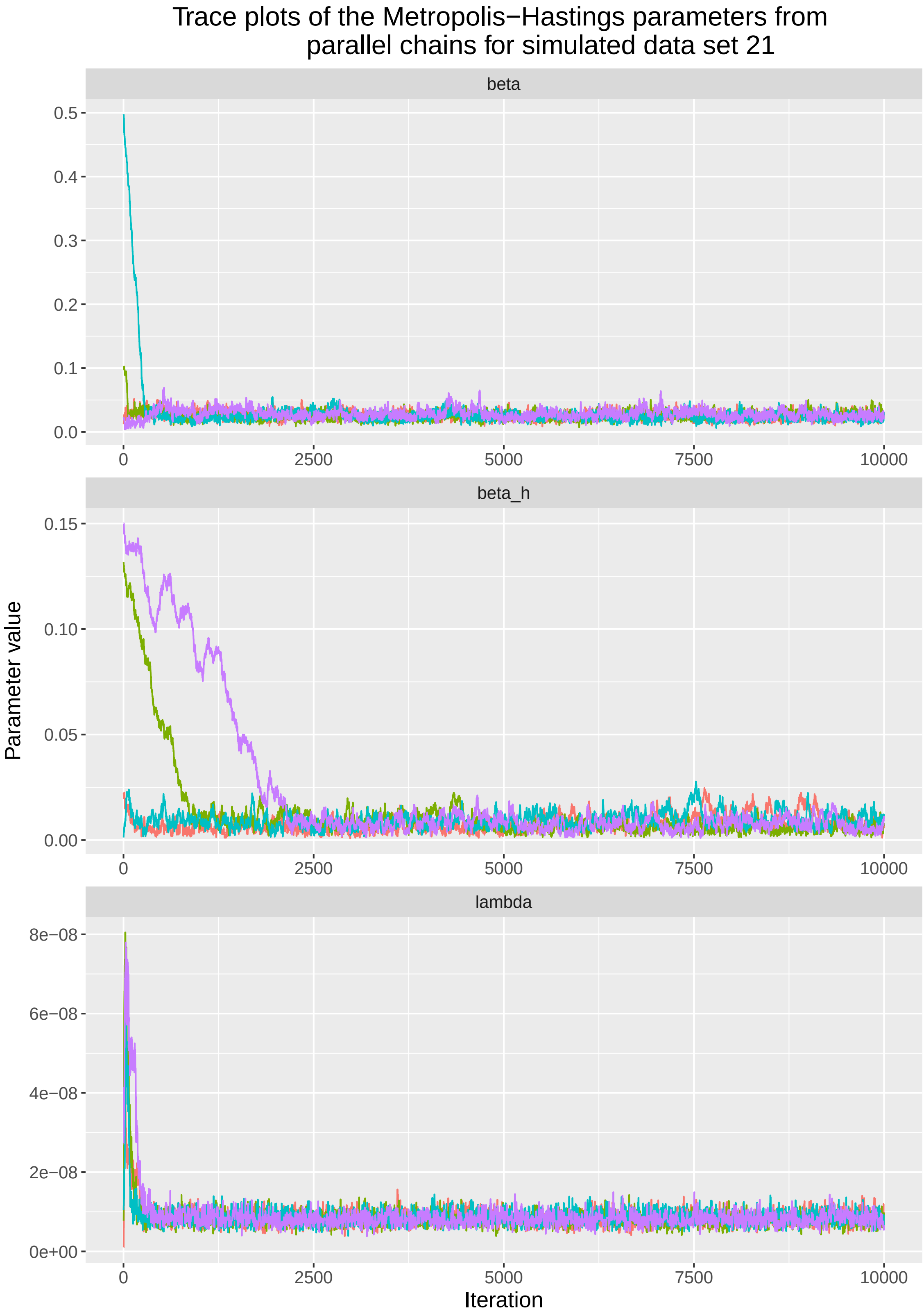


Figure A.19: Trace plots of four chains run in parallel on simulated data set 21 with the burn in period included. The specific parameters here are those updated by a Metropolis-Hastings step, which are the patient transmission rate β, healthcare worker transmission rate β_H and mutation rate λ.

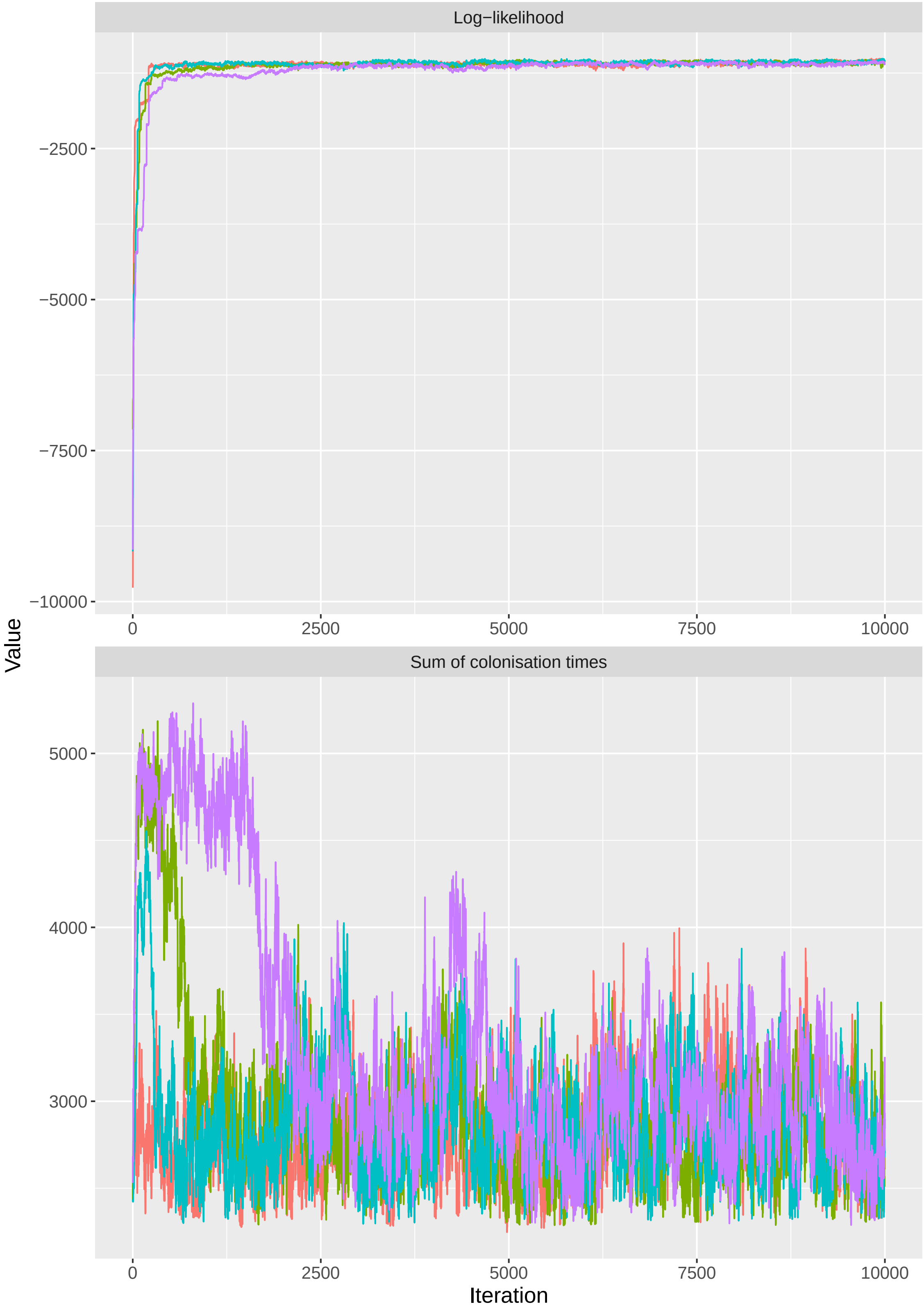


Figure A.20: Log-likelihood and colonisation time sum trace plots for simulated data set 21.

APPENDIX B

Results from the Brighton data set

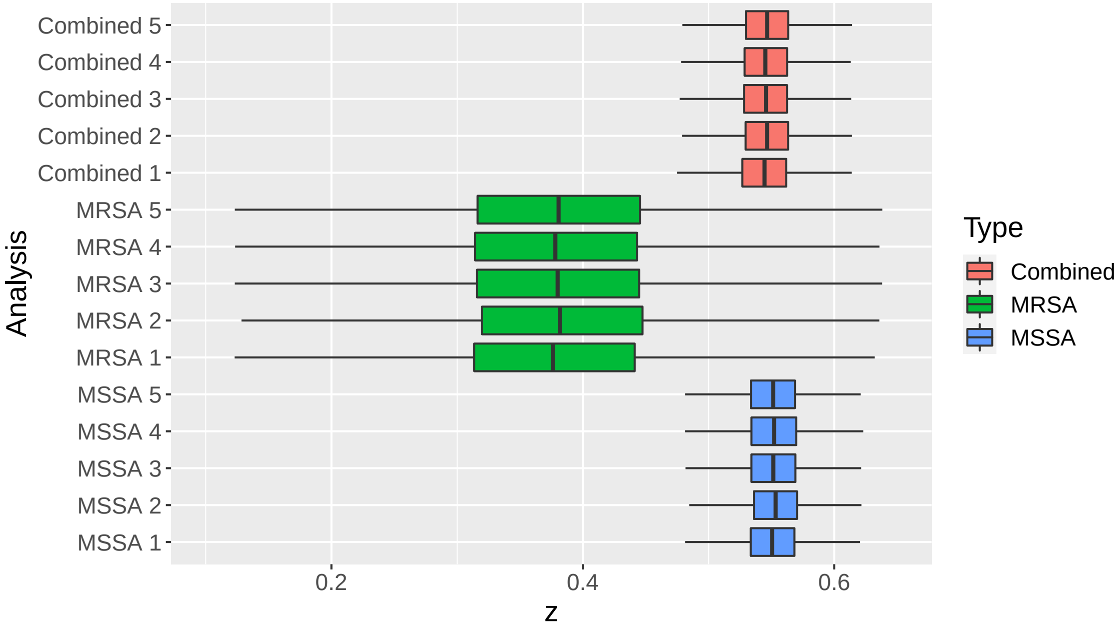


Figure B.1: Posterior box plots for the test sensitivity parameter z for the MRSA (red), MSSA (blue) and combined (green) data sources. For each data source there are 5 separate runs of each analysis.

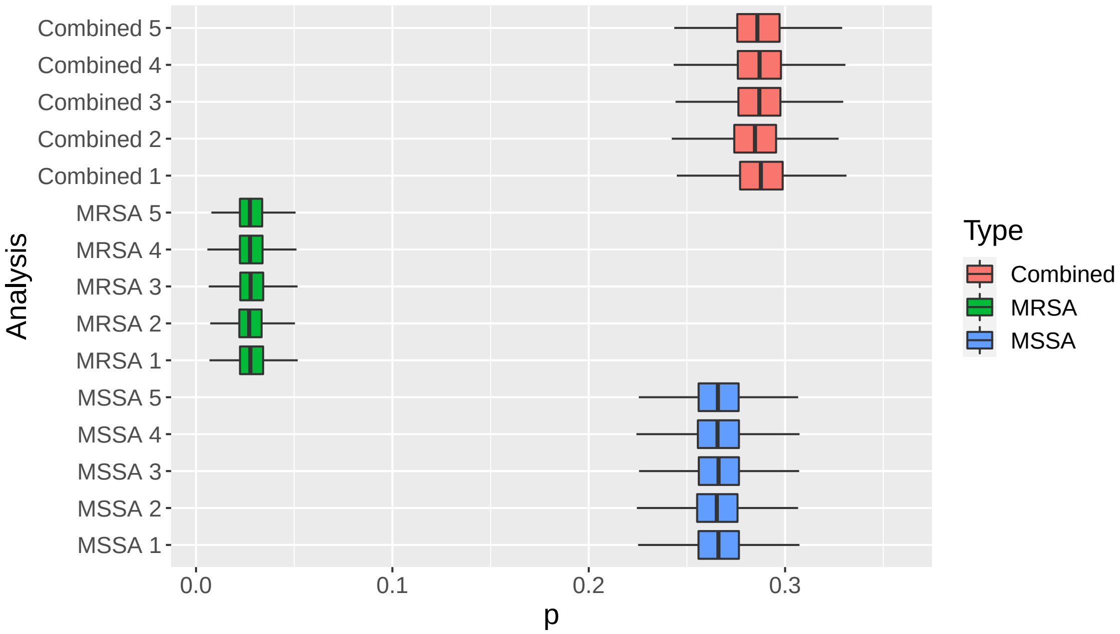


Figure B.2: Posterior box plots for the importation probability parameter p for the MRSA (red), MSSA (blue) and combined (green) data sources. For each data source there are 5 separate runs of each analysis.

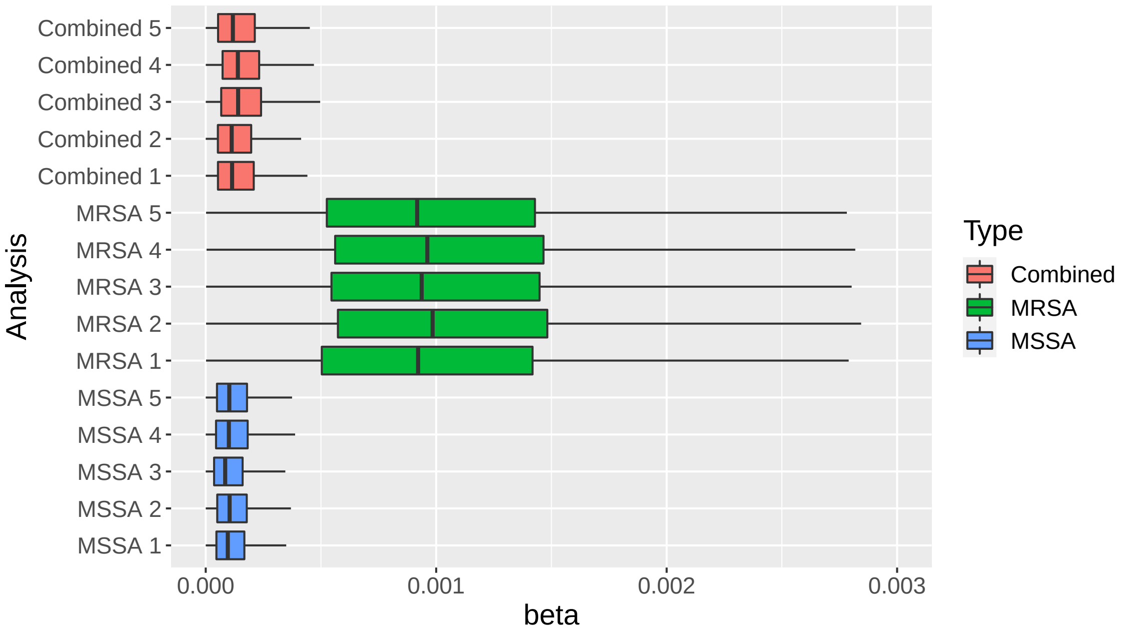


Figure B.3: Posterior box plots for the patient transmission parameter β for the MRSA (red), MSSA (blue) and combined (green) data sources. For each data source there are 5 separate runs of each analysis.

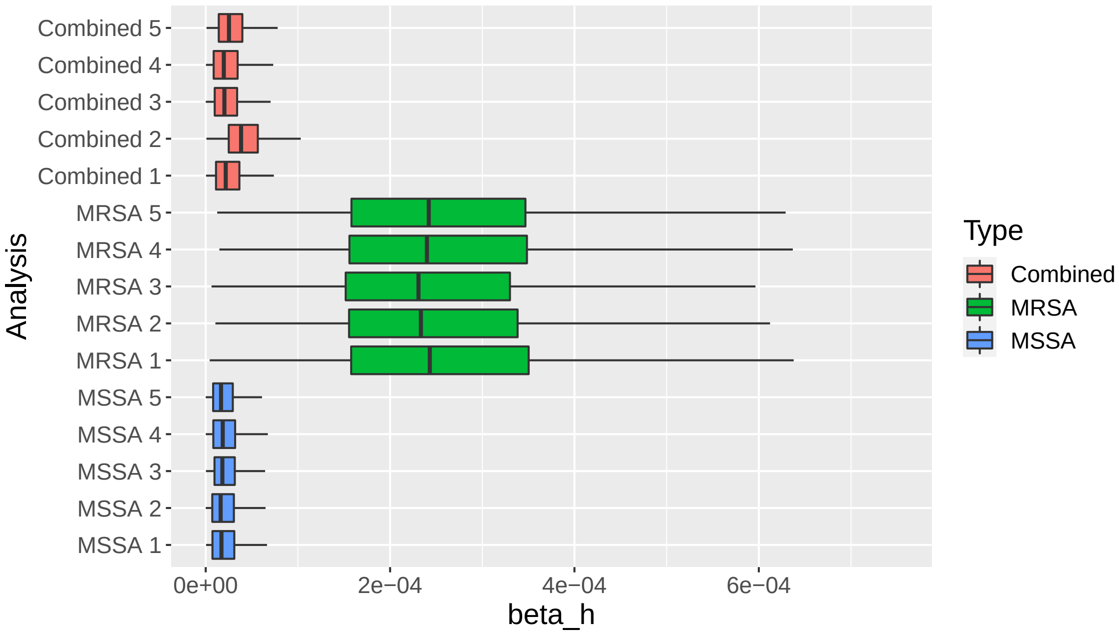


Figure B.4: Posterior box plots for the healthcare worker transmission parameter β_H for the MRSA (red), MSSA (blue) and combined (green) data sources. For each data source there are 5 separate runs of each analysis.

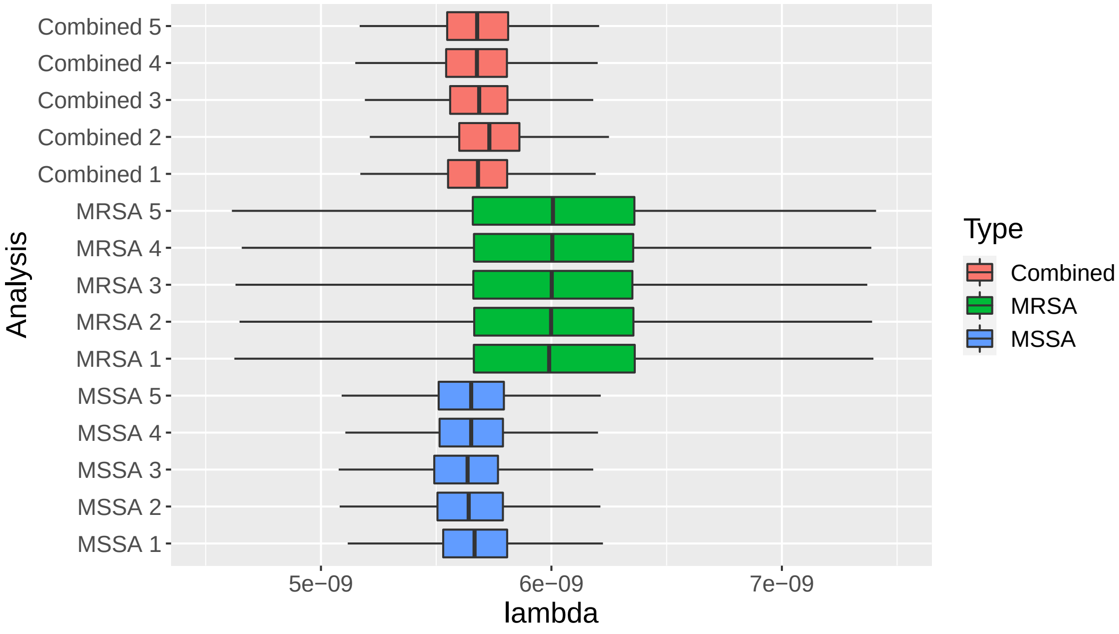


Figure B.5: Posterior box plots for the mutation rate parameter λ for the MRSA (red), MSSA (blue) and combined (green) data sources. For each data source there are 5 separate runs of each analysis.

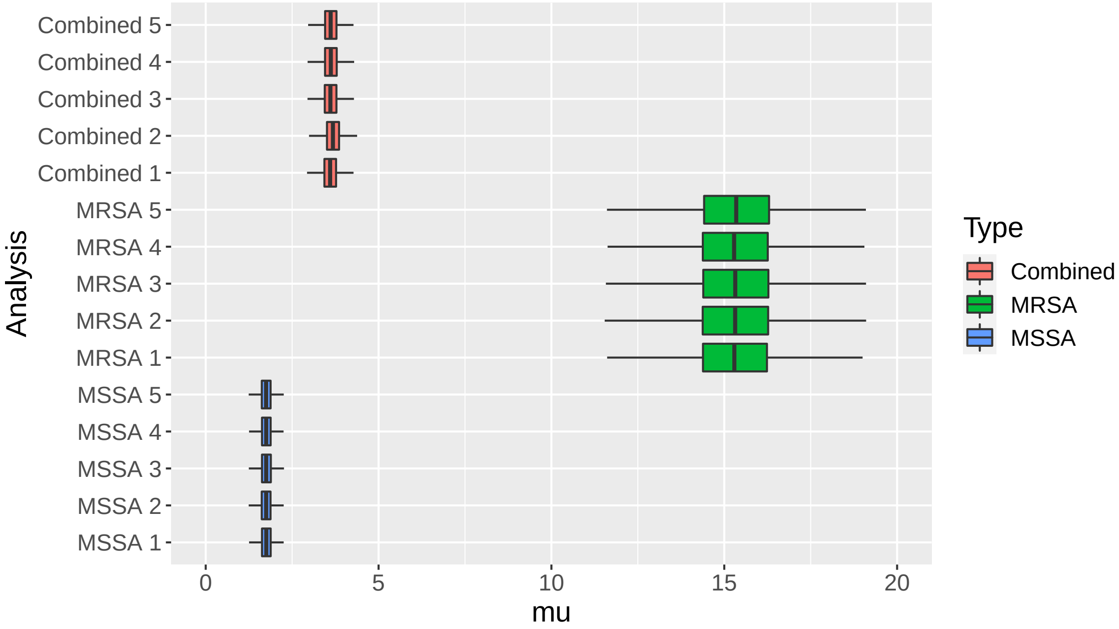


Figure B.6: Posterior box plots for the importation distance parameter μ for the MRSA (red), MSSA (blue) and combined (green) data sources. For each data source there are 5 separate runs of each analysis.

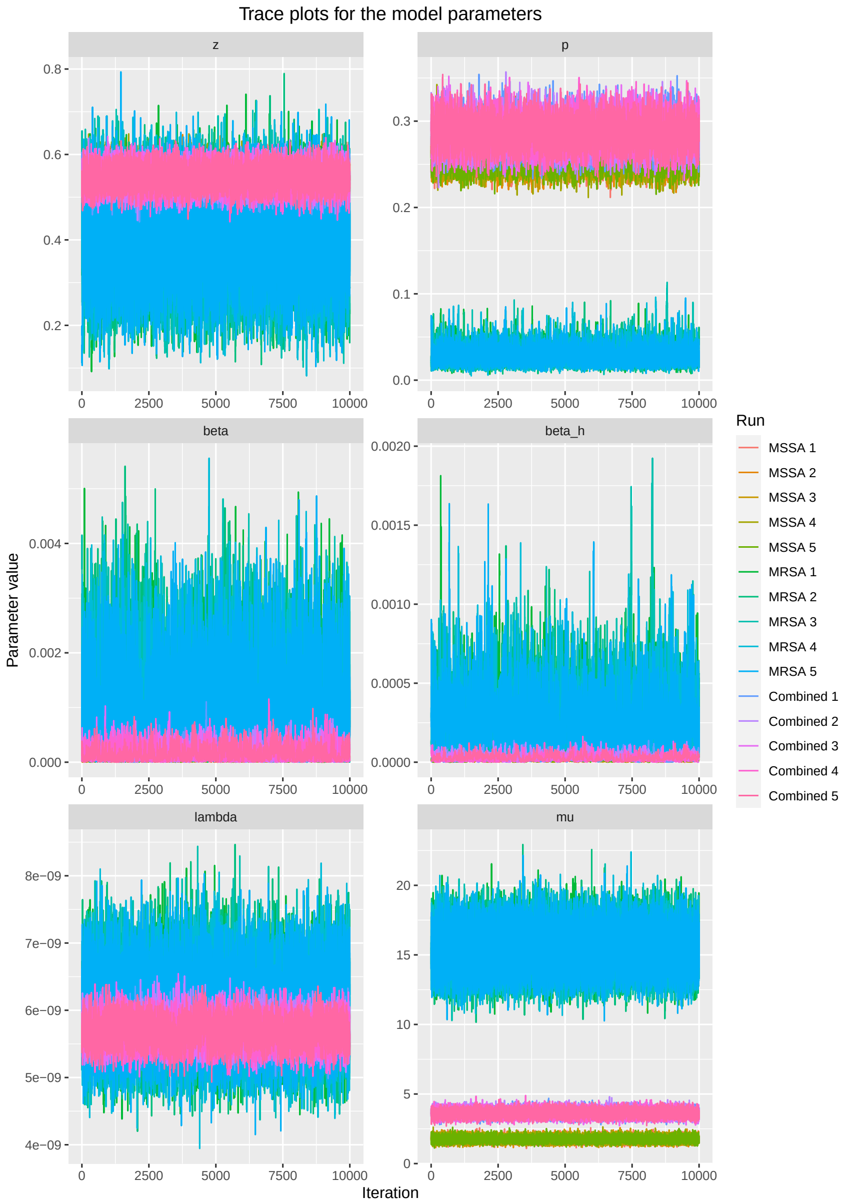


Figure B.7: Posterior trace plots for the model parameters when fitting the full model to the Brighton data set where the genetic data contains either MRSA, MSSA or both. The parameters are explicitly the test sensitivity z (top left), importation probability p (top right), patient transmission rate β (middle left), healthcare worker transmission rate β_H (middle right), mutation rate λ (bottom left) and average importation distance μ (bottom right).

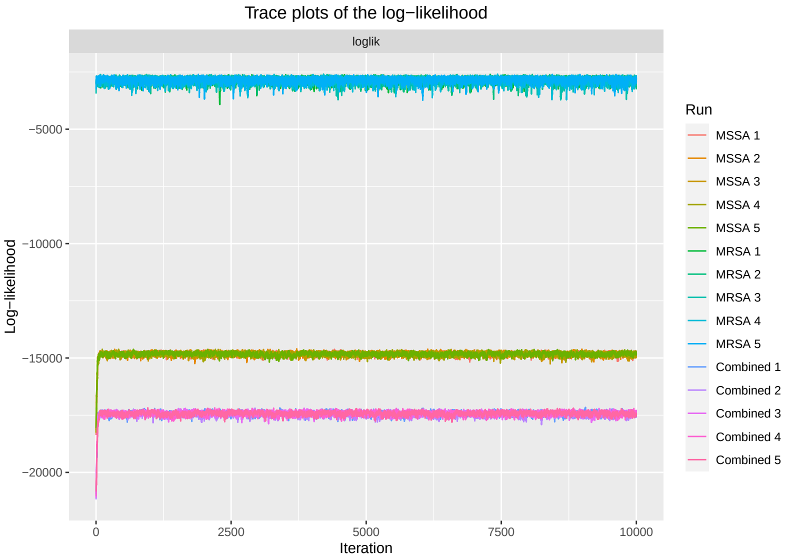


Figure B.8: Posterior trace plots for the log-likelihood when fitting the full model to the Brighton data set where the genetic data contains either MRSA, MSSA or both. The plot contains samples from the burn-in to demonstrate that the chain is initialised in regions of low likelihood and increases until convergence.

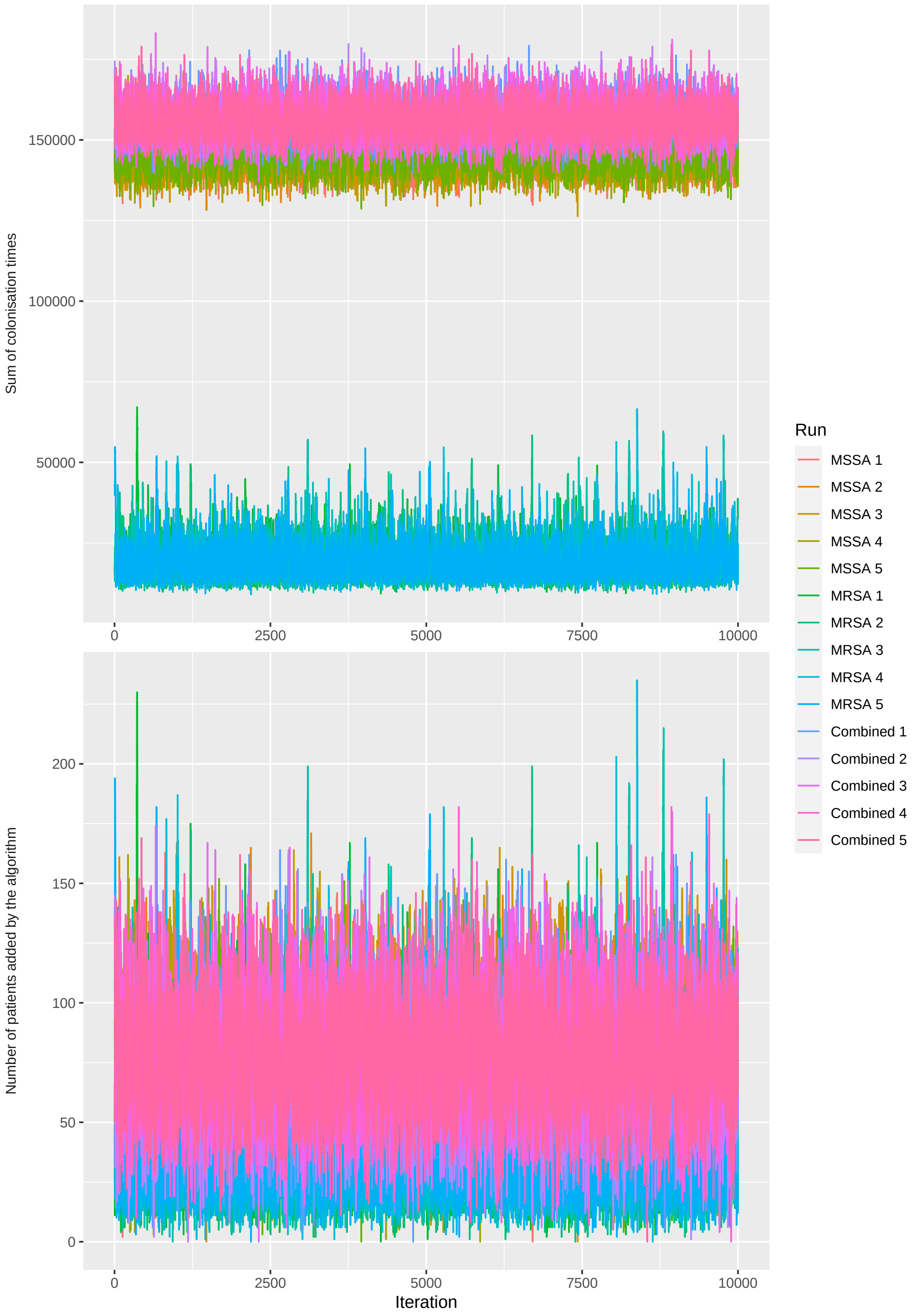


Figure B.9: Posterior trace plots for summary statistics of the augmented data which include the sum of colonisation times (above) and the number of patients added by the algorithm (below).

APPENDIX C

Simulation study results for FMD

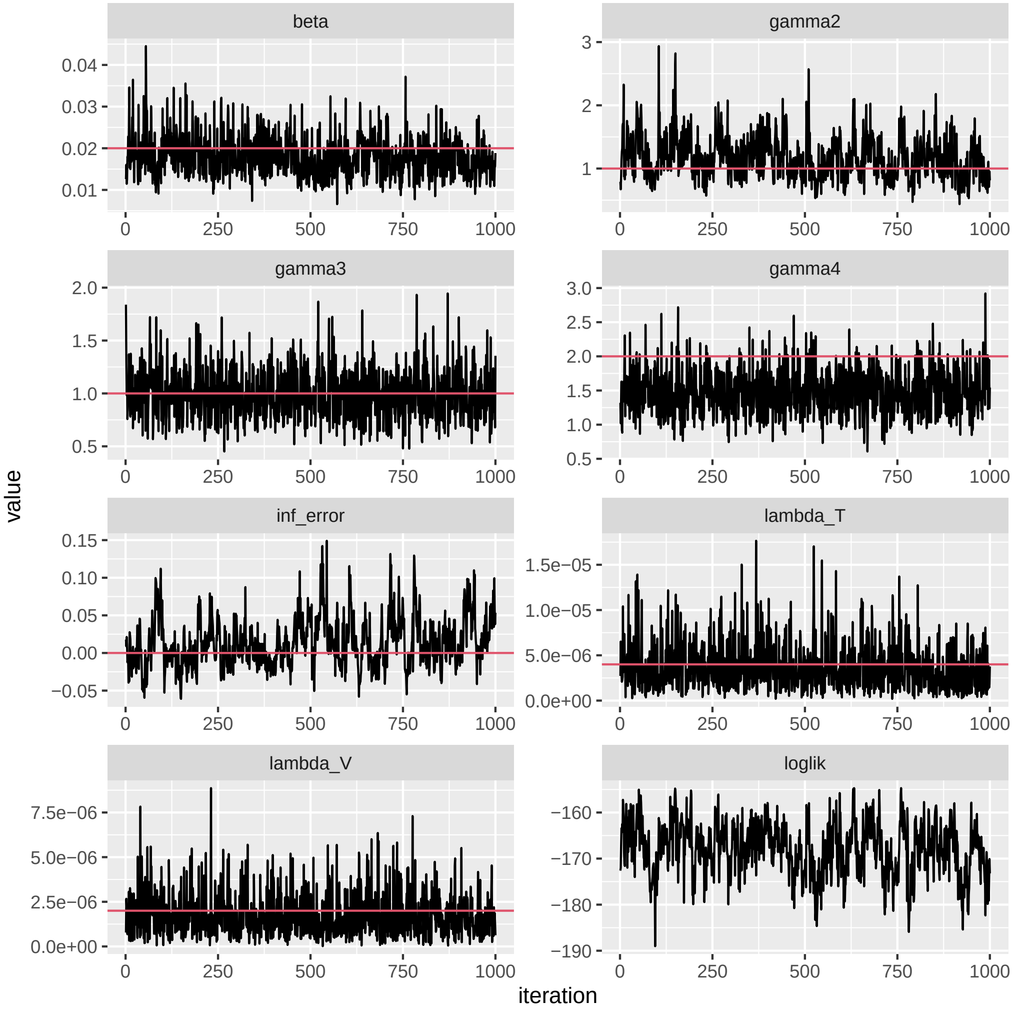


Figure C.1: Trace plots for all inferred model parameters and the log likelihood for simulated data. The lines in red indicate the true simulated parameter value. Each of the chains demonstrate good mixing and convergence to near the true value.

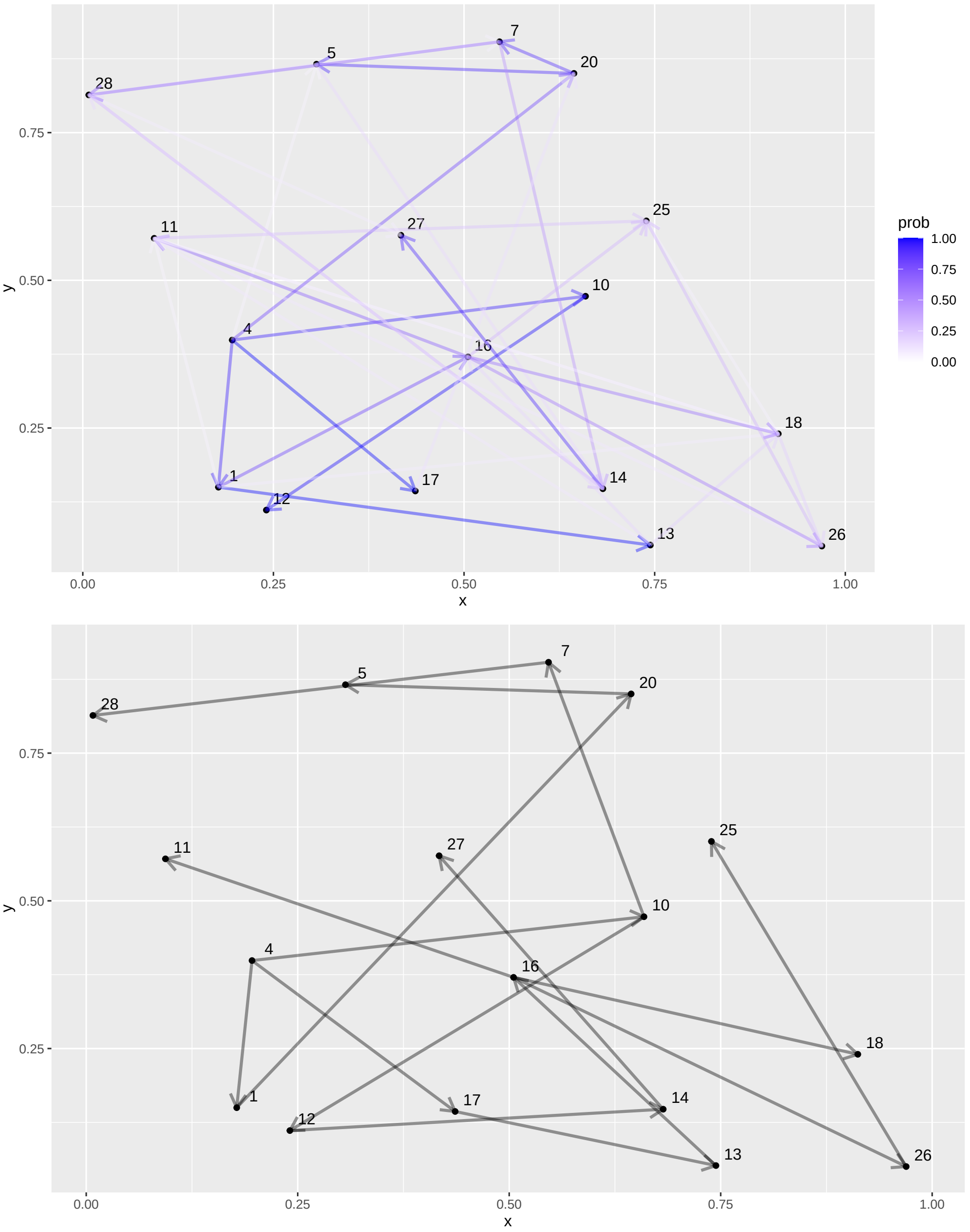


Figure C.2: Above: inferred posterior transmission network where the edge weights are the marginal posterior probability of that particular source infector pair. Below: the true simulated transmission network. The source accuracy for this data set was 64.7%.

APPENDIX D

Genetic model assessment for the Darlington data set

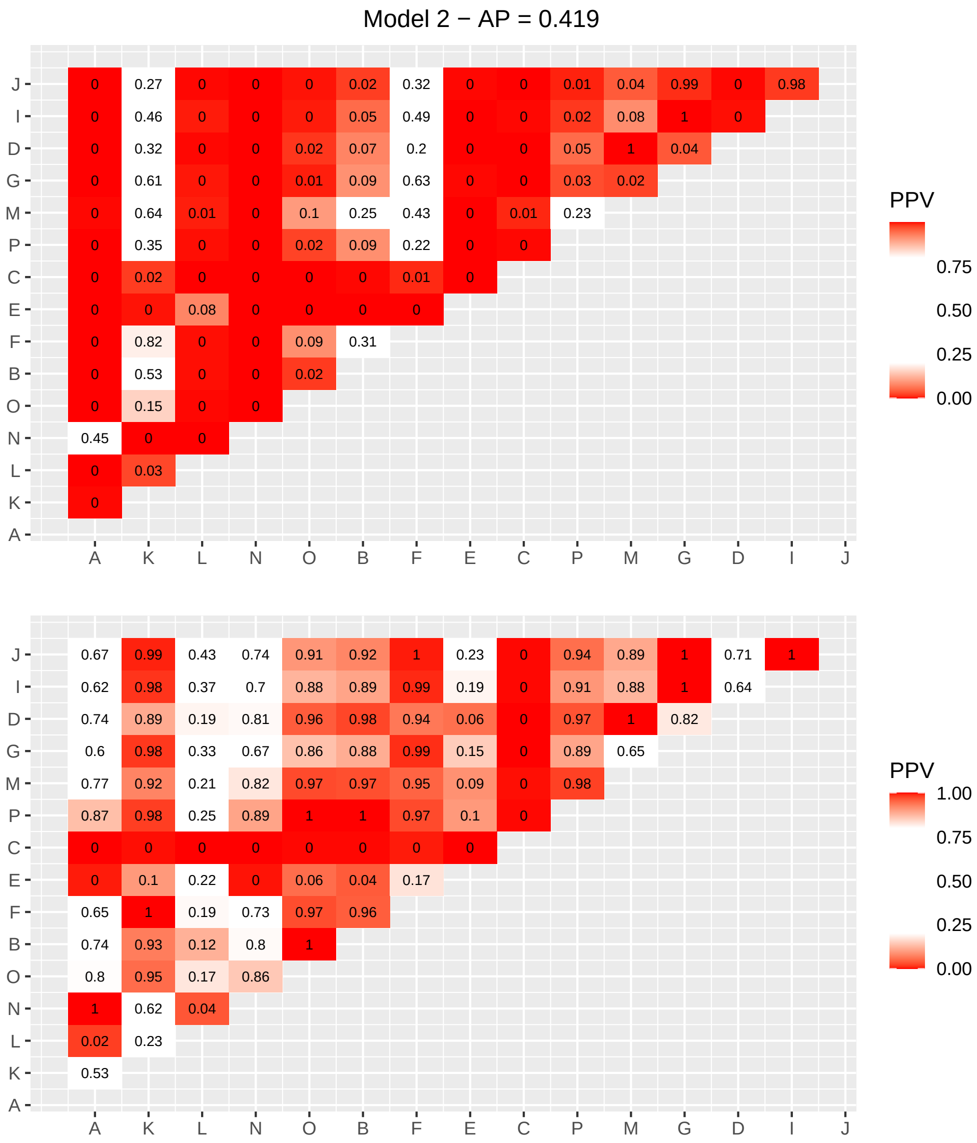

Model 2 − AP = 0.419
J 0 0.27 0 0 0 0.02 0.32 0 0 0.01 0.04 0.99 0 0.98
I 0 0.46 0 0 0 0.05 0.49 0 0 0.02 0.08 1 0
D 0 0.32 0 0 0.02 0.07 0.2 0 0 0.05 1 0.04
G 0 0.61 0 0 0.01 0.09 0.63 0 0 0.03 0.02
M 0 0.64 0.01 0 0.1 0.25 0.43 0 0.01 0.23
P 0 0.35 0 0 0.02 0.09 0.22 0 0
C 0 0.02 0 0 0 0 0.01 0
E 0 0 0.08 0 0 0 0
F 0 0.82 0 0 0.09 0.31
B 0 0.53 0 0 0.02
O 0 0.15 0 0
N 0.45 0 0
L 0 0.03
K 0
A
A K L N O B F E C P M G D I J
PPV
0.75
0.50
0.25
0.00
J 0.67 0.99 0.43 0.74 0.91 0.92 1 0.23 0 0.94 0.89 1 0.71 1
I 0.62 0.98 0.37 0.7 0.88 0.89 0.99 0.19 0 0.91 0.88 1 0.64
D 0.74 0.89 0.19 0.81 0.96 0.98 0.94 0.06 0 0.97 1 0.82
G 0.6 0.98 0.33 0.67 0.86 0.88 0.99 0.15 0 0.89 0.65
M 0.77 0.92 0.21 0.82 0.97 0.97 0.95 0.09 0 0.98
P 0.87 0.98 0.25 0.89 1 1 0.97 0.1 0
C 0 0 0 0 0 0 0 0
E 0 0.1 0.22 0 0.06 0.04 0.17
F 0.65 1 0.19 0.73 0.97 0.96
B 0.74 0.93 0.12 0.8 1
O 0.8 0.95 0.17 0.86
N 1 0.62 0.04
L 0.02 0.23
K 0.53
A
A K L N O B F E C P M G D I J
PPV
1.00
0.75
0.50
0.25
0.00

Figure D.1

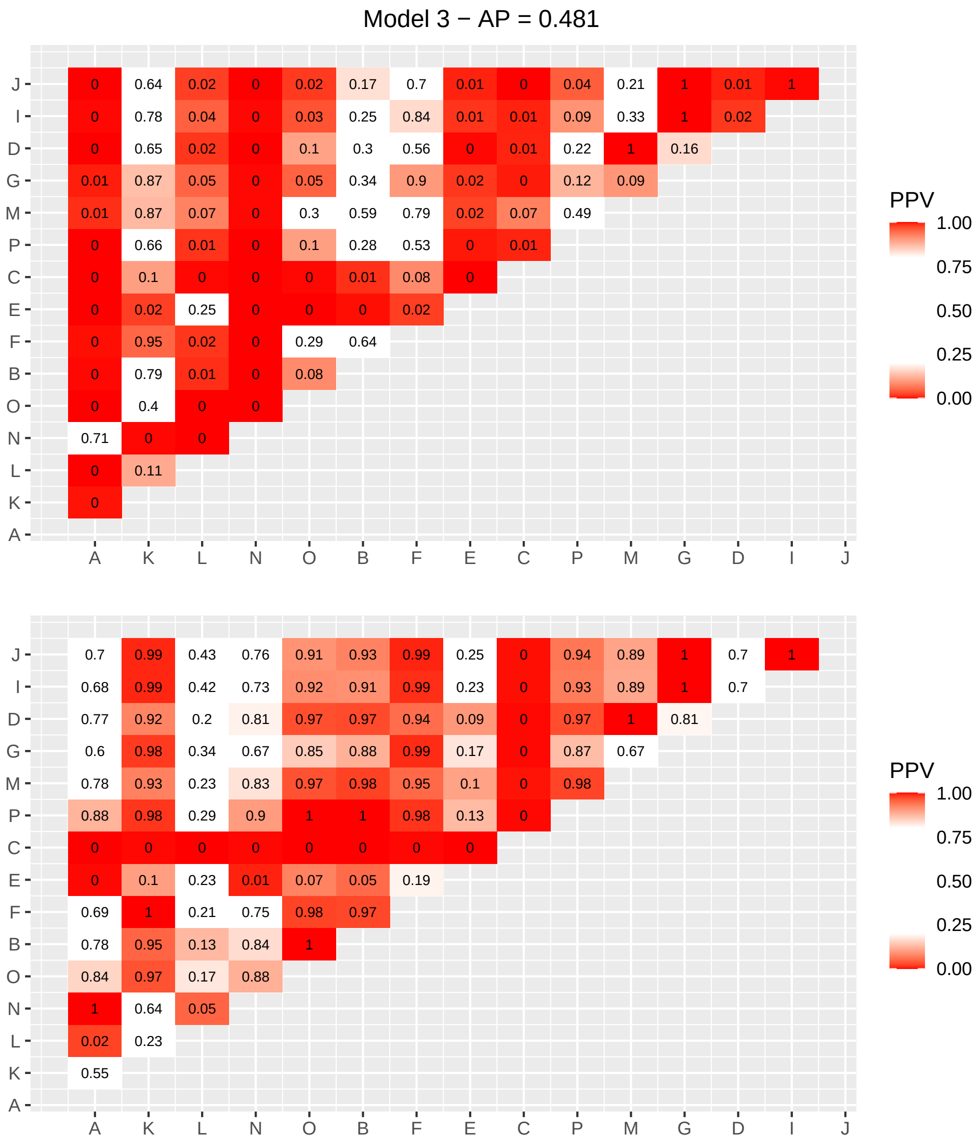
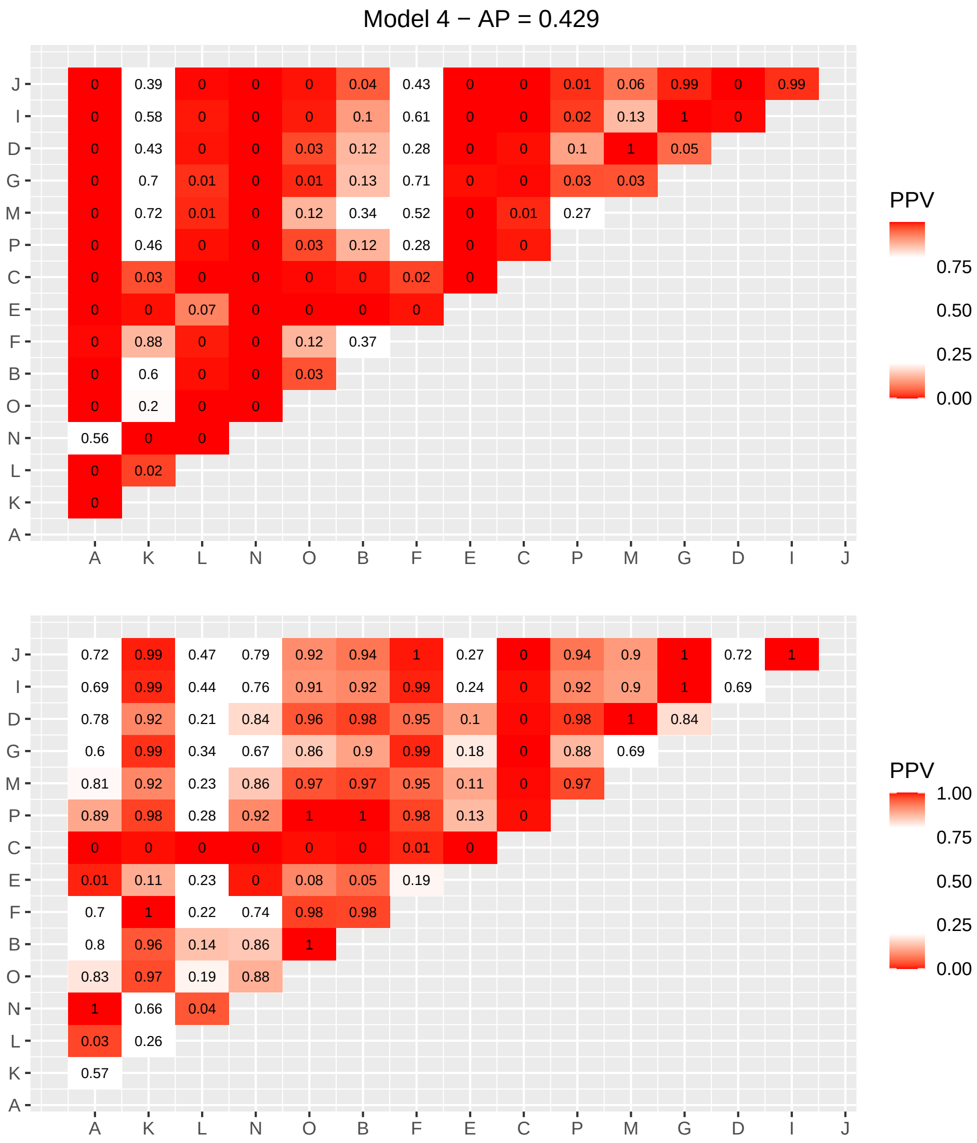

Model 3 − AP = 0.481
J 0 0.64 0.02 0 0.02 0.17 0.7 0.01 0 0.04 0.21 1 0.01 1
I 0 0.78 0.04 0 0.03 0.25 0.84 0.01 0.01 0.09 0.33 1 0.02
D 0 0.65 0.02 0 0.1 0.3 0.56 0 0.01 0.22 1 0.16
G 0.01 0.87 0.05 0 0.05 0.34 0.9 0.02 0 0.12 0.09
M 0.01 0.87 0.07 0 0.3 0.59 0.79 0.02 0.07 0.49
P 0 0.66 0.01 0 0.1 0.28 0.53 0 0.01
C 0 0.1 0 0 0 0.01 0.08 0
E 0 0.02 0.25 0 0 0 0.02
F 0 0.95 0.02 0 0.29 0.64
B 0 0.79 0.01 0 0.08
O 0 0.4 0 0
N 0.71 0 0
L 0 0.11
K 0
A
A K L N O B F E C P M G D I J
PPV
1.00
0.75
0.50
0.25
0.00
J 0.7 0.99 0.43 0.76 0.91 0.93 0.99 0.25 0 0.94 0.89 1 0.7 1
I 0.68 0.99 0.42 0.73 0.92 0.91 0.99 0.23 0 0.93 0.89 1 0.7
D 0.77 0.92 0.2 0.81 0.97 0.97 0.94 0.09 0 0.97 1 0.81
G 0.6 0.98 0.34 0.67 0.85 0.88 0.99 0.17 0 0.87 0.67
M 0.78 0.93 0.23 0.83 0.97 0.98 0.95 0.1 0 0.98
P 0.88 0.98 0.29 0.9 1 1 0.98 0.13 0
C 0 0 0 0 0 0 0 0
E 0 0.1 0.23 0.01 0.07 0.05 0.19
F 0.69 1 0.21 0.75 0.98 0.97
B 0.78 0.95 0.13 0.84 1
O 0.84 0.97 0.17 0.88
N 1 0.64 0.05
L 0.02 0.23
K 0.55
A
A K L N O B F E C P M G D I J
PPV
1.00
0.75
0.50
0.25
0.00

Figure D.2



Model 4 − AP = 0.429

Figure D.3

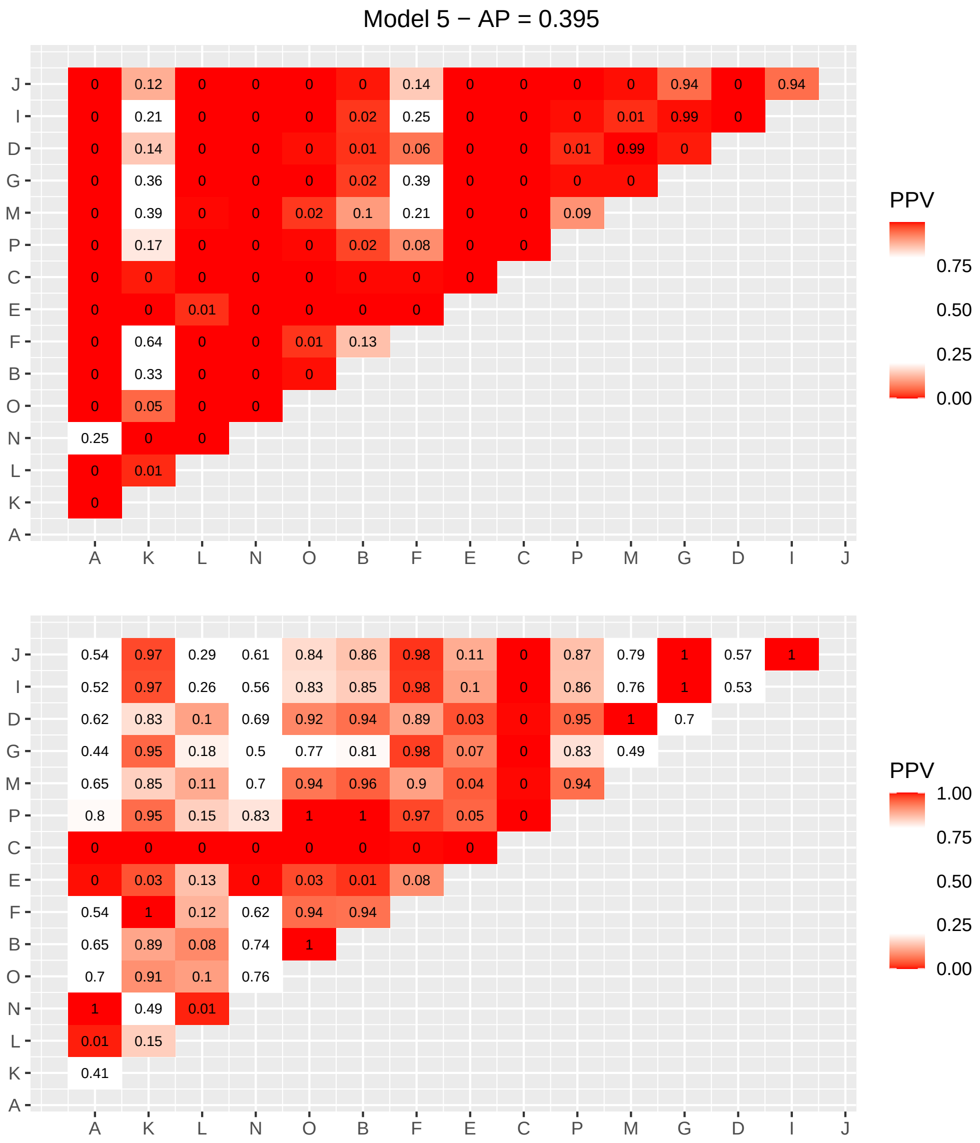
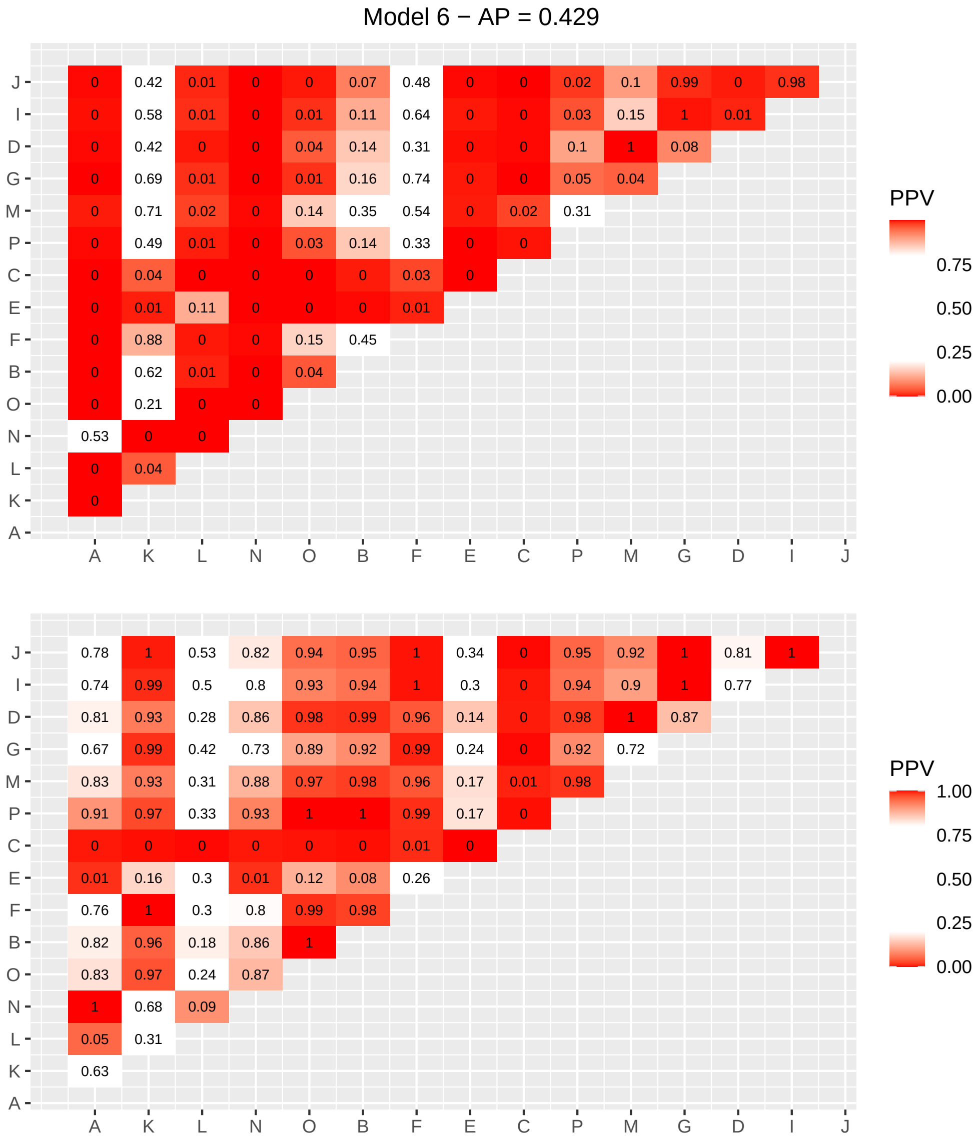

Model 5 − AP = 0.395

Figure D.4



Model 6 − AP = 0.429

Figure D.5

References

Albrich, W. C. and Harbarth, S. (2008). Health-care workers: source, vector, or victim of MRSA? *The Lancet Infectious Diseases*, 8(5):289–301.

Alexandersen, S., Zhang, Z., Donaldson, A., and Garland, A. (2003). The pathogenesis and diagnosis of foot-and-mouth disease. *Journal of Comparative Pathology*, 129(1):1–36.

Alharthi, M., Kypraios, T., and O'Neill, P. D. (2019). Bayes factors for partially observed stochastic epidemic models. *Bayesian Analysis*, 14(3).

Ali, M. S., Isa, N. M., Abedelrhman, F. M., Alyas, T. B., Mohammed, S. E., Ahmed, A. E., Ahmed, Z. S. A., Lau, N.-S., Garbi, M. I., Amirul, A. A.-A., Seed, A. O., Omer, R. A., and Mohamed, S. B. (2019). Genomic analysis of methicillin-resistant *Staphylococcus aureus* strain SO-1977 from sudan. *BMC Microbiology*, 19(1).

Andersson, H. and Britton, T. (2012). *Stochastic Epidemic Models and Their Statistical Analysis*. Springer London, Limited.

Andrade, J. and Duggan, J. (2020). An evaluation of hamiltonian monte carlo performance to calibrate age-structured compartmental seir models to incidence data. *Epidemics*, 33:100415.

Aristotelous, G. (2020). *Topics in Bayesian inference and model assessment for partially observed stochastic epidemic models*. PhD thesis, University of Nottingham.

Aristotelous, G., Kypraios, T., and O'Neill, P. D. (2022). Posterior predictive checking for partially observed stochastic epidemic models. *Bayesian Analysis*, 18(4).

Bailey, N. T. J. (1975). *The mathematical theory of infectious diseases and its applications*. Griffin, London, 2. ed. edition. Literaturverz. S. 383 - 403.

Bang, H. (2010). Statistical methods in molecular biology.

Barnes, S., Golden, B., and Wasil, E. (2010). MRSA transmission reduction using agent-based modeling and simulation. *INFORMS Journal on Computing*, 22(4):635–646.

Becker, N. G. (1993). Parametric inference for epidemic models. *Mathematical Biosciences*, 117(1-2):239–251.

Benzer, S. (1961). On the topography of the genetic fine structure. *Proceedings of the National Academy of Sciences*, 47(3):403–415.

Bondy, J. A. (1976). *Graph theory with applications*. Elsevier Science Publishing Co. Inc, New York.

Bratsun, D., Volfson, D., Tsimring, L. S., and Hasty, J. (2005). Delay-induced stochastic oscillations in gene regulation. *Proceedings of the National Academy of Sciences*, 102(41):14593–14598.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). Handbook of Markov chain Monte Carlo.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Brown, T. A. (2018). Genomes 4.

Cassidy, R. (2019). *Inference of transmission trees for epidemics using whole-genome sequence data*. PhD thesis, University of Nottingham.

Cassidy, R., Kypraios, T., and O'Neill, P. D. (2020). Modelling, Bayesian inference, and model assessment for nosocomial pathogens using whole-genome-sequence data. *Statistics in Medicine*, 39(12):1746–1765.

Chapman, L. A. C., Spencer, S. E. F., Pollington, T. M., Jewell, C. P., Mondal, D., Alvar, J., Hollingsworth, T. D., Cameron, M. M., Bern, C., and Medley, G. F. (2020). Inferring transmission trees to guide targeting of interventions against visceral leishmaniasis and post–kala-azar dermal leishmaniasis. *Proceedings of the National Academy of Sciences*, 117(41):25742–25750.

Charleston, B., Bankowski, B. M., Gubbins, S., Chase-Topping, M. E., Schley, D., Howey, R., Barnett, P. V., Gibson, D., Juleff, N. D., and Woolhouse, M. E. J. (2011). Relationship between clinical signs and transmission of an infectious disease and the implications for control. *Science*, 332(6030):726–729.

Choudhuri, S. (2014). Fundamentals of molecular evolution. In *Bioinformatics for Beginners*, pages 27–53. Elsevier.

Clancy, D. and O’Neill, P. D. (2008). Bayesian estimation of the basic reproduction number in stochastic epidemic models. *Bayesian Analysis*, 3(4).

Cottam, E. M., Haydon, D. T., Paton, D. J., Gloster, J., Wilesmith, J. W., Ferris, N. P., Hutchings, G. H., and King, D. P. (2006). Molecular epidemiology of the foot-and-mouth disease virus outbreak in the United Kingdom in 2001. *Journal of Virology*, 80(22):11274–11282.

Cottam, E. M., Thébaud, G., Wadsworth, J., Gloster, J., Mansley, L., Paton, D. J., King, D. P., and Haydon, D. T. (2008). Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proceedings of the Royal Society B: Biological Sciences*, 275(1637):887–895.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904.

Currie, A., Davis, L., Odrobina, E., Waldman, S., White, D., Tomassi, J., and Katz, K. C. (2008). Sensitivities of nasal and rectal swabs for detection of methicillin-resistant staphylococcus aureus colonization in an active surveillance program. *Journal of Clinical Microbiology*, 46(9):3101–3103.

Dagum, L. and Menon, R. (1998). OpenMP: an industry standard API for shared-memory programming. *IEEE Computational Science and Engineering*, 5(1):46–55.

Dani, A. (2014). Colonization and infection. *Central European Journal of Urology*, 67(01).

Davies, G. (2002). The foot and mouth disease (FMD) epidemic in the United Kingdom 2001. *Comparative Immunology, Microbiology and Infectious Diseases*, 25(5-6):331–343.

Deardon, R., Brooks, S. P., Grenfell, B. T., Keeling, M. J., Tildesley, M. J., Savill, N. J., Shaw, D. J., and Woolhouse, M. E. J. (2010). Inference for individual-level models of infectious diseases in large populations. *Statistica Sinica*, 20:239–261.

Deeth, L. E., Deardon, R., and Gillis, D. J. (2015). Model choice using the deviance information criterion for latent conditional individual-level models of infectious disease spread. *Epidemiologic Methods*, 4(1).

der Laan, M. V., Pollard, K., and Bryan, J. (2003). A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation*, 73(8):575–584.

Domingo, E., Baranowski, E., Escarmís, C., and Sobrino, F. (2002). Foot-and-mouth disease virus. *Comparative Immunology, Microbiology and Infectious Diseases*, 25(5-6):297–308.

Duan, L. L., Johndrow, J. E., and Dunson, D. B. (2018). Scaling up data augmentation MCMC via calibration. *Journal of Machine Learning Research*, 19(64):1–34.

Duault, H., Durand, B., and Canini, L. (2022). Methods combining genomic and epidemiological data in the reconstruction of transmission trees: A systematic review. *Pathogens*, 11(2):252.

Dunbar, M. R., Johnson, S. R., Rhyan, J. C., and McCollum, M. (2009). Use of infrared thermography to detect thermographic changes in mule deer (odocoileus hemionus) experimentally infected with foot-and-mouth disease. *Journal of Zoo and Wildlife Medicine*, 40(2):296–301.

Edwards, A. W. F. (2009). Statistical methods for evolutionary trees. *Genetics*, 183(1):5–12.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376.

Ferguson, N., Laydon, D., Nedjati Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunuba Perez, Z., Cuomo-Dannenburg, G., Dighe, A., Dorigatti, I., Fu, H., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Okell, L., Van Elsland, S., Thompson, H., Verity, R., Volz, E., Wang, H., Wang, Y., Walker, P., Winskill, P., Whittaker, C., Donnelly, C., Riley, S., and Ghani, A. (2020). Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand.

Ferguson, N. M., Donnelly, C. A., and Anderson, R. M. (2001). The foot-and-mouth epidemic in Great Britain: Pattern of spread and impact of interventions. *Science*, 292(5519):1155–1160.

Firestone, S. M., Hayama, Y., Lau, M. S. Y., Yamamoto, T., Nishi, T., Bradhurst, R. A., Demirhan, H., Stevenson, M. A., and Tsutsui, T. (2020). Transmission network reconstruction for foot-and-mouth disease outbreaks incorporating farm-level covariates. *PLOS ONE*, 15(7):e0235660.

Foxman, B. (2001). Molecular epidemiology: Focus on infection. *American Journal of Epidemiology*, 153(12):1135–1141.

Frénay, H. M., Theelen, J. P., Schouls, L. M., Vandenbroucke-Grauls, C. M., Verhoef, J., van Leeuwen, W. J., and Mooi, F. R. (1994). Discrimination of epidemic and nonepidemic methicillin-resistant *Staphylococcus aureus* strains on the basis of protein a gene polymorphism. *Journal of Clinical Microbiology*, 32(3):846–847.

Gaiarsa, S., Marco, L. D., Comandatore, F., Marone, P., Bandi, C., and Sassera, D. (2015). Bacterial genomic epidemiology, from local outbreak characterization to species-history reconstruction. *Pathogens and Global Health*, 109(7):319–327.

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall/CRC Texts in Statistical Science Ser. CRC Press LLC, Boca Raton, 2nd ed. edition. Description based on publisher supplied metadata and other sources.

Gelman, A. (2013). Two simple examples for understanding posterior p-values whose distributions are far from uniform. *Electronic Journal of Statistics*, 7(none).

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4).

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741.

Gibson, G. and Renshaw, E. (1998). Estimating parameters in stochastic compartmental models using Markov chain methods. *Mathematical Medicine and Biology*, 15(1):19–40.

Gilks, W. R., editor (1998). *Markov chain Monte Carlo in practice*. Interdisciplinary statistics. Chapman & Hall/CRC, Boca Raton, Fla., 1. ed., 1. crc press repr. edition.

Goldstein, N. D., Eppes, S. C., Mackley, A., Tuttle, D., and Paul, D. A. (2017). A network model of hand hygiene: How good is good enough to stop the spread of MRSA? *Infection Control & Hospital Epidemiology*, 38(8):945–952.

Grimmett, G. (2001). *Probability and random processes*. Oxford University Press.

Gross, J. L., Yellen, J., and Anderson, M. (2018). *Graph Theory and Its Applications*. Taylor and Francis Group.

Grubman, M. J. and Baxt, B. (2004). Foot-and-mouth disease. *Clinical Microbiology Reviews*, 17(2):465–493.

Guest, J. F., Keating, T., Gould, D., and Wigglesworth, N. (2020). Modelling the annual NHS costs and outcomes attributable to healthcare-associated infections in England. *BMJ Open*, 10(1):e033367.

Hall, M., Woolhouse, M., and Rambaut, A. (2015). Epidemic reconstruction in a phylogenetics framework: Transmission trees as partitions of the node set. *PLOS Computational Biology*, 11(12):e1004613.

Harris, S. R., Cartwright, E. J., Török, M. E., Holden, M. T., Brown, N. M., Ogilvy-Stuart, A. L., Ellington, M. J., Quail, M. A., Bentley, S. D., Parkhill, J., and Peacock, S. J. (2013). Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. *The Lancet Infectious Diseases*, 13(2):130–136.

Harris, S. R., Feil, E. J., Holden, M. T. G., Quail, M. A., Nickerson, E. K., Chantratita, N., Gardete, S., Tavares, A., Day, N., Lindsay, J. A., Edgeworth, J. D., de Lencastre, H., Parkhill, J., Peacock, S. J., and Bentley, S. D. (2010). Evolution of MRSA during hospital transmission and intercontinental spread. *Science*, 327(5964):469–474.

Hartigan, J. A. (1975). *Clustering Algorithms*, chapter The k-means algorithm, pages 84–87. John Wiley & Sons, 1st edition edition.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Hawkins, G., Stewart, S., Blatchford, O., and Reilly, J. (2011). Should healthcare workers be screened routinely for meticillin-resistant *Staphylococcus aureus*? a review of the evidence. *Journal of Hospital Infection*, 77(4):285–289.

Hayama, Y., Yamamoto, T., Kobayashi, S., Muroga, N., and Tsutsui, T. (2013). Mathematical model of the 2010 foot-and-mouth disease epidemic in Japan and evaluation of control measures. *Preventive Veterinary Medicine*, 112(3-4):183–193.

Higgs, P. G. and Derrida, B. (1992). Genetic distance and species formation in evolving populations. *Journal of Molecular Evolution*, 35(5).

Jewell, C. P., Keeling, M. J., and Roberts, G. O. (2008). Predicting undetected infections during the 2007 foot-and-mouth disease outbreak. *Journal of The Royal Society Interface*, 6(41):1145–1151.

Jombart, T., Cori, A., Didelot, X., Cauchemez, S., Fraser, C., and Ferguson, N. (2014). Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Computational Biology*, 10(1):e1003457.

Jombart, T., Eggo, R. M., Dodd, P. J., and Balloux, F. (2010). Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity*, 106(2):383–390.

Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism*, pages 21–132. Elsevier.

Kaufman, L. and Rousseeuw, P. J. (2005). *Finding Groups In Data: An Introduction to Cluster Analysis*. John Wiley and Sons.

Kaufmann, L. and Rousseeuw, P. (1987). Clustering by means of medoids. *Data Analysis based on the L1-Norm and Related Methods*, pages 405–416.

Keeling, M. J. (2005). Models of foot-and-mouth disease. *Proceedings of the Royal Society B: Biological Sciences*, 272(1569):1195–1202.

Keeling, M. J., Woolhouse, M. E. J., May, R. M., Davies, G., and Grenfell, B. T. (2002). Modelling vaccination strategies against foot-and-mouth disease. *Nature*, 421(6919):136–142.

Keeling, M. J., Woolhouse, M. E. J., Shaw, D. J., Matthews, L., Chase-Topping, M., Haydon, D. T., Cornell, S. J., Kappey, J., Wilesmith, J., and Grenfell, B. T. (2001). Dynamics of the 2001 UK foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape. *Science*, 294(5543):813–817.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772):700–721.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120.

Kimura, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences*, 78(1):454–458.

Klinkenberg, D., Backer, J. A., Didelot, X., Colijn, C., and Wallinga, J. (2017). Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLOS Computational Biology*, 13(5):e1005495.

Klinkenberg, D., Colijn, C., and Didelot, X. (2019). Methods for outbreaks using genomic data. In Held, L., Hens, N., O'Neill, P., and Wallinga, J., editors, *Handbook of Infectious Disease Data Analysis*, chapter 13. Chapman and Hall/CRC.

Kuhn, J. H., Bao, Y., Bavari, S., Becker, S., Bradfute, S., Brister, J. R., Bukreyev, A. A., Chandran, K., Davey, R. A., Dolnik, O., Dye, J. M., Enterlein, S., Hensley, L. E., Honko, A. N., Jahrling, P. B., Johnson, K. M., Kobinger, G., Leroy, E. M., Lever, M. S., Mühlberger, E., Netesov, S. V., Olinger, G. G., Palacios, G., Patterson, J. L., Paweska, J. T., Pitt, L., Radoshitzky, S. R., Saphire, E. O., Smither, S. J., Swanepoel, R., Towner, J. S., van der Groen, G., Volchkov, V. E., Wahl-Jensen, V., Warren, T. K., Weidmann, M., and Nichol, S. T. (2012). Virus nomenclature below the species level: a standardized nomenclature for natural variants of viruses assigned to the family Filoviridae. *Archives of Virology*, 158(1):301–311.

Kypraios, T., O'Neill, P. D., Huang, S. S., Rifas-Shiman, S. L., and Cooper, B. S. (2010). Assessing the role of undetected colonization and isolation precautions in reducing methicillin-resistant *Staphylococcus aureus* transmission in intensive care units. *BMC Infectious Diseases*, 10(1).

Lanave, C., Preparata, G., Sacone, C., and Serio, G. (1984). A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20(1):86–93.

Lau, M. S. Y., Marion, G., Streftaris, G., and Gibson, G. (2015). A systematic Bayesian integration of epidemiological and genetic data. *PLOS Computational Biology*, 11(11):e1004633.

Lau, M. S. Y., Marion, G., Streftaris, G., and Gibson, G. J. (2014). New model diagnostics for spatio-temporal systems in epidemiology and ecology. *Journal of The Royal Society Interface*, 11(93):20131093.

Lekone, P. E. and Finkenstädt, B. F. (2006). Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics*, 62(4):1170–1177.

Linderman, S. W., Johnson, M. J., and Adams, R. P. (2015). Dependent multinomial models made easy: Stick breaking with the Pólya-gamma augmentation.

Link, W. A. and Eaton, M. J. (2011). On thinning of chains in MCMC. *Methods in Ecology and Evolution*, 3(1):112–115.

Longe, J. (2015). *The Gale encyclopedia of medicine.* Gale/Cengage Learning, Farmington Hills, Mich, 5th edition edition.

Lyons, D. M. and Lauring, A. S. (2017). Evidence for the selective basis of transition-to-transversion substitution bias in two RNA viruses. *Molecular Biology and Evolution*, 34(12):3205–3215.

Mardones, F., Perez, A., Sanchez, J., Alkhamis, M., and Carpenter, T. (2010). Parameterization of the duration of infection stages of serotype O foot-and-mouth disease virus: an analytical review and meta-analysis with application to simulation models. *Veterinary Research*, 41(4):45.

Meyer, R. F. and Knudsen, R. C. (2001). Foot-and-mouth disease: a review of the virus and the symptoms. *Journal of environmental health*, 64:21–23.

Mohammadi-Kambs, M., Hölz, K., Somoza, M. M., and Ott, A. (2017). Hamming distance as a concept in DNA molecular recognition. *ACS Omega*, 2(4):1302–1308.

Mollentze, N., Nel, L. H., Townsend, S., le Roux, K., Hampson, K., Haydon, D. T., and Soubeyrand, S. (2014). A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proceedings of the Royal Society B: Biological Sciences*, 281(1782):20133251.

Mondal, B. and Choudhury, J. P. (2013). A comparative study on K means and PAM algorithm using physical characters of different varieties of mango in India. *International Journal of Computer Applications*, 78(5):21–24.

Morelli, M. J., Thébaud, G., Chadœuf, J., King, D. P., Haydon, D. T., and Soubeyrand, S. (2012). A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Computational Biology*, 8(11):e1002768.

Morris, R. S., Stern, M. W., Stevenson, M. A., Wilesmith, J. W., and Sanson, R. L. (2001). Predictive spatial modelling of alternative control strategies for the foot-and-mouth disease epidemic in Great Britain, 2001. *Veterinary Record*, 149(5):137–144.

Mutters, N. T., Heeg, K., Späth, I., Henny, N., and Günther, F. (2017). Improvement of infection control management by routine molecular evaluation of pathogen clusters. *Diagnostic Microbiology and Infectious Disease*, 88(1):82–87.

Nakhleh, L. (2013). Evolutionary trees. In *Brenner's Encyclopedia of Genetics*, pages 549–550. Elsevier.

Nascimento, F. F., dos Reis, M., and Yang, Z. (2017). A biologist's guide to Bayesian phylogenetic analysis. *Nature Ecology & Evolution*, 1(10):1446–1454.

Neher, R. A. and Bedford, T. (2018). Real-time analysis and visualization of pathogen sequence data. *Journal of Clinical Microbiology*, 56(11).

Nübel, U., Nachtnebel, M., Falkenhorst, G., Benzler, J., Hecht, J., Kube, M., Bröcker, F., Moelling, K., Bührer, C., Gastmeier, P., Piening, B., Behnke, M., Dehnert, M., Layer, F., Witte, W., and Eckmanns, T. (2013). MRSA transmission on a neonatal intensive care unit: Epidemiological and genome-based phylogenetic analyses. *PLoS ONE*, 8(1):e54898.

O'Neill, P. D. (2002). A tutorial introduction to Bayesian inference for stochastic epidemic models using markov chain monte carlo methods. *Mathematical Biosciences*, 180(1-2):103–114.

O'Neill, P. D. (2010). Introduction and snapshot review: Relating infectious disease transmission models to data. *Statistics in Medicine*, 29(20):2069–2077.

O'Neill, P. D. and Becker, N. G. (2001). Inference for an epidemic when susceptibility varies. *Biostatistics*, 2(1):99–108.

O'Neill, P. D. and Roberts, G. O. (1999). Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 162(1):121–129.

Paton, D. J., Gubbins, S., and King, D. P. (2018). Understanding the transmission of foot-and-mouth disease virus at different scales. *Current Opinion in Virology*, 28:85–91.

Pellis, L., Ball, F., Bansal, S., Eames, K., House, T., Isham, V., and Trapman, P. (2015). Eight challenges for network epidemic models. *Epidemics*, 10:58–62.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11.

Popovich, K. J., Green, S. J., Okamoto, K., Rhee, Y., Hayden, M. K., Schoeny, M., Snitkin, E. S., and Weinstein, R. A. (2020). MRSA transmission in intensive care units: Genomic analysis of patients, their environments, and healthcare workers. *Clinical Infectious Diseases*, 72(11):1879–1887.

Price, J., Didelot, X., Crook, D., Llewelyn, M., and Paul, J. (2013). Whole genome sequencing in the prevention and control of *Staphylococcus aureus* infection. *Journal of Hospital Infection*, 83(1):14–21.

Price, J. R. (2014). *Application of whole-genome sequencing to understand transmission of healthcare-associated Staphylococcus aureus*. PhD thesis, University of Sussex.

Price, J. R., Cole, K., Bexley, A., Kostiou, V., Eyre, D. W., Golubchik, T., Wilson, D. J., Crook, D. W., Walker, A. S., Peto, T. E. A., Llewelyn, M. J., and Paul, J. (2017). Transmission of *Staphylococcus aureus* between health-care workers, the environment, and patients in an intensive care unit: a longitudinal cohort study based on whole-genome sequencing. *The Lancet Infectious Diseases*, 17(2):207–214.

Revez, J., Espinosa, L., Albiger, B., Leitmeyer, K. C., and and, M. J. S. (2017). Survey on the use of whole-genome sequencing for infectious diseases surveillance: Rapid expansion of european national capacities, 2015–2016. *Frontiers in Public Health*, 5.

Robert, C. P. and Changye, W. (2020). Markov chain Monte Carlo methods, a survey with some frequent misunderstandings.

Roy, V. (2020). Convergence diagnostics for Markov chain Monte Carlo. *Annual Review of Statistics and Its Application*, 7(1):387–412.

Sainudiin, R., Clark, A. G., and Durrett, R. T. (2007). Simple models of genomic variation in human SNP density. *BMC Genomics*, 8(1):146.

Sassmannshausen, R., Deurenberg, R. H., Köck, R., Hendrix, R., Jurke, A., Rossen, J. W. A., and Friedrich, A. W. (2016). MRSA prevalence and associated risk factors among health-care workers in non-outbreak situations in the Dutch-German EUREGIO. *Frontiers in Microbiology*, 7.

Savill, N. J., Shaw, D. J., Deardon, R., Tildesley, M. J., Keeling, M. J., Woolhouse, M. E., Brooks, S. P., and Grenfell, B. T. (2006). Topographic determinants of foot and mouth disease transmission in the UK 2001 epidemic. *BMC Veterinary Research*, 2(1).

Scanvic, A., Denic, L., Gaillon, S., Giry, P., Andremont, A., and Lucet, J.-C. (2001). Duration of colonization by methicillin-resistant *Staphylococcus aureus* after hospital discharge and risk factors for prolonged carriage. *Clinical Infectious Diseases*, 32(10):1393–1398.

Senn, L., Basset, P., Nahimana, I., Zanetti, G., and Blanc, D. (2012). Which anatomical sites should be sampled for screening of methicillin-resistant *Staphylococcus aureus* carriage by culture or by rapid PCR test? *Clinical Microbiology and Infection*, 18(2):E31–E33.

Seymour, R. G., Kypraios, T., and O'Neill, P. D. (2022). Bayesian nonparametric inference for heterogeneously mixing infectious disease models. *Proceedings of the National Academy of Sciences*, 119(10).

Shopsin, B., Gomez, M., Montgomery, S. O., Smith, D. H., Waddington, M., Dodge, D. E., Bost, D. A., Riehman, M., Naidich, S., and Kreiswirth, B. N. (1999). Evaluation of protein a gene polymorphic region DNA sequencing for typing of *Staphylococcus aureus* strains. *Journal of Clinical Microbiology*, 37(11):3556–3563.

Shukla, S. K., Pantrang, M., Stahl, B., Briska, A. M., Stemper, M. E., Wagner, T. K., Zentz, E. B., Callister, S. M., Lovrich, S. D., Henkhaus, J. K., and Dykes, C. W.

(2012). Comparative whole-genome mapping to determine *Staphylococcus aureus* genome size, virulence motifs, and clonality. *Journal of Clinical Microbiology*, 50(11):3526–3533.

Snitkin, E. S., Zelazny, A. M., Thomas, P. J., Stock, F., Henderson, D. K., Palmore, T. N., and and, J. A. S. (2012). Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. *Science Translational Medicine*, 4(148):148ra116–148ra116.

Spencer, S. E. (2021). Accelerating adaptation in the adaptive Metropolis–Hastings random walk algorithm. *Australian & New Zealand Journal of Statistics*, 63(3):468–484.

Streftaris, G. and Gibson, G. J. (2012). Non-exponential tolerance to infection in epidemic systems – modeling, inference, and assessment. *Biostatistics*, 13(4):580–593.

Tang, P., Croxen, M. A., Hasan, M. R., Hsiao, W. W., and Hoang, L. M. (2017). Infection control in the new age of genomic epidemiology. *American Journal of Infection Control*, 45(2):170–179.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences.*

Thompson, D., Muriel, P., Russell, D., Osborne, P., Bromley, A., Rowland, M., Creigh-Tyte, S., and Brown, C. (2002). Economic costs of the foot and mouth disease outbreak in the United Kingdom in 2001. *Revue Scientifique et Technique de l'OIE*, 21(3):675–687.

Thompson, R. L. (1982). Epidemiology of nosocomial infections caused by methicillin-resistant *Staphylococcus aureus. Annals of Internal Medicine*, 97(3):309.

Thornley, J. H. and France, J. (2009). Modelling foot and mouth disease. *Preventive Veterinary Medicine*, 89(3-4):139–154.

Tildesley, M. J., Deardon, R., Savill, N. J., Bessell, P. R., Brooks, S. P., Woolhouse, M. E., Grenfell, B. T., and Keeling, M. J. (2008). Accuracy of models for the 2001 foot-and-mouth epidemic. *Proceedings of the Royal Society B: Biological Sciences*, 275(1641):1459–1468.

Toptas, M., Samanci, N. S., Akkoc, İ., Yucetas, E., Cebeci, E., Sen, O., Can, M. M., and Ozturk, S. (2018). Factors affecting the length of stay in the intensive care unit: Our clinical experience. *BioMed Research International*, 2018:1–4.

Touloupou, P., Retkute, R., Hollingsworth, T. D., and Spencer, S. E. (2022). Statistical methods for linking geostatistical maps and transmission models: Application to lymphatic filariasis in East Africa. *Spatial and Spatio-temporal Epidemiology*, 41:100391.

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The third revolution in sequencing technology. *Trends in Genetics*, 34(9):666–681.

van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50.

Vasylyeva, T. I., Friedman, S. R., Paraskevis, D., and Magiorkinis, G. (2016). Integrating molecular epidemiology and social network analysis to study infectious diseases: Towards a socio-molecular era for public health. *Infection, Genetics and Evolution*, 46:248–255.

Vergu, E., Busson, H., and Ezanno, P. (2010). Impact of the infection period distribution on the epidemic spread in a metapopulation model. *PLoS ONE*, 5(2):e9371.

Wallinga, J. (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology*, 160(6):509–516.

Watson, J. D. and Crick, F. H. C. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

Whittles, L. K. and Didelot, X. (2016). Epidemiological analysis of the Eyam plague outbreak of 1665–1666. *Proceedings of the Royal Society B: Biological Sciences*, 283(1830):20160618.

Woodbury, E. L. (1995). A review of the possible mechanisms for the persistence of foot-and-mouth disease virus. *Epidemiology and Infection*, 114(1):1–13.

Worby, C. J., Chang, H.-H., Hanage, W. P., and Lipsitch, M. (2014). The distribution of pairwise genetic distances: A tool for investigating disease transmission. *Genetics*, 198(4):1395–1404.

Worby, C. J., O'Neill, P. D., Kypraios, T., Robotham, J. V., Angelis, D. D., Cartwright, E. J. P., Peacock, S. J., and Cooper, B. S. (2016). Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *The Annals of Applied Statistics*, 10(1):395–417.

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

Yoon, S. H., Park, W., King, D. P., and Kim, H. (2011). Phylogenomics and molecular evolution of foot-and-mouth disease virus. *Molecules and Cells*, 31(5):413–421.

Young, B. C., Golubchik, T., Batty, E. M., Fung, R., Larner-Svensson, H., Votintseva, A. A., Miller, R. R., Godwin, H., Knox, K., Everitt, R. G., Iqbal, Z., Rimmer, A. J., Cule, M., Ip, C. L. C., Didelot, X., Harding, R. M., Donnelly, P., Peto, T. E., Crook, D. W., Bowden, R., and Wilson, D. J. (2012). Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. *Proceedings of the National Academy of Sciences*, 109(12):4550–4555.

Ypma, R. J. F., van Ballegooijen, W. M., and Wallinga, J. (2013). Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics*, 195(3):1055–1062.

Zou, Q., Lin, G., Jiang, X., Liu, X., and Zeng, X. (2018). Sequence clustering in bioinformatics: an empirical study. *Briefings in Bioinformatics*.