Imperial College London
Department of Infectious Disease Epidemiology

Integrated analysis of epidemiological and phylogenetic data to elucidate viral transmission dynamics

Lucy M Li

A thesis submitted for the degree of Doctor of Philosophy (2017)

Declaration of Originality

I declare that the work presented in this thesis is my own work. All other work included in this thesis has been appropriately referenced.

Lucy M Li, May 2017

Copyright Declaration

Acknowledgements

To my fellow students and researchers at DIDE, thank you for your advice, insightful comments, and feedback. Special thanks to Margarita Pons-Salort, George Shirreff, Nicholas Croucher, Isobel Blake, Anne Cori and Natsuko Imai for providing comments on various manuscripts and patiently explaining complex (or very obvious) concepts to me.

To Dr Erik Volz and Dr Deidre Hollingsworth, your invaluable input during my PhD has helped to keep me on track.

Finally I would like to express my gratitude for two supervisors, Professors Nick Grassly and Christophe Fraser. I have become a more confident and independent researcher as a result of your support and guidance.

Abstract

While infectious disease outbreaks are often summarised by population averages such as the reproductive number, variation between individuals in terms of onwards transmissions modulates the degree of unpredictability of an epidemic, and it needs to be accounted for in models of infection control. This heterogeneity among individuals can be quantified by the dispersion parameter k of the offspring distribution, a distribution that defines the number of secondary infections per infected individual. I have developed an inference framework to estimate k and other epidemiological parameters by fitting stochastic transmission models to both incidence time series and the pathogen phylogeny. Applying the framework to simulated data, I found that more accurate, less biased and more precise estimates of the reproductive number and k were obtained by combining epidemiologic and phylogenetic analyses. Accurately estimating k was necessary for unbiased estimates of the reproductive number, but it did not affect the accurate estimation of epidemic start date and the probability of sampling an infection. I further demonstrated that inference was possible in the presence of phylogenetic uncertainty by sampling from the posterior distribution of phylogenies. In addition to methodological contributions, I found that the inclusion of sequences in statistical inference for polio improved the precision of parameter estimates. Based on sequences collected from patients during a poliovirus outbreak, the estimated values of k were high regardless of the data used. On the other hand, the k estimates were low when a transmission model was fit to environmental sequences collected in Pakistan, which is still endemic for wild poliovirus. Furthermore, analysis of environmental sequences was informative of seasonality parameters whereas inference from incidence time series alone was not. This type of analysis using environmental sequences would be useful as polio eradication draws to a close as the number of symptomatic cases approaches zero.

Notation

A_t Number of lineages through time at time step t

Δt Size of simulation time step

$\mathbf{Epi}_t$ Number of reported cases at time step t

H_i Time step in which individual i becomes infected

I_t The number of infectious individuals at time step t

$I_{\mathbf{total}}$ Number of individuals infected during an outbreak

$\mathbf{Inc}_t$ Incidence, i.e. the number of reported cases at time step t

k Dispersion parameter of the negative binomial

λ_t Pairwise coalescent rate at time step t

N Population size (of infected individuals)

$N\mathbf{e}$ Effective population size (of infected individuals)

ϕ Offspring distribution

$\mathbf{Phy}_t$ Phylogenetic data at time step t

R_0 Basic reproductive number

R_t Effective reproductive number at time step t

σ^2 Variance of the offspring distribution

T_g Generation time

U_t Time intervals between events at time step t

Z_i Number of secondary infections caused by infected individual i

Abbreviations

AFP Acute Flaccid Paralysis

BSP Bayesian skyline plot

CNS Central nervous system

ESS Effective sample size of samples from the posterior distribution, e.g. using MCMC

FATA Federally Administered Tribal Areas in Pakistan

GPEI Global Polio Eradication Initiative

HIV Human Immunodeficiency Virus

HPD Highest posterior density

KP Khyber Pakhtunkhwa, a province in Pakistan

K-S Kolmogorov-Smirnov distance

MCC Maximum clade credibility

MCMC Markov Chain Monte Carlo

MERS-CoV Middle East Respiratory Syndrome coronavirus

$\mathbf{N_{eff}}$ Effective number of particles

OPV Oral polio vaccine

pH1N1 Pandemic H1N1

PMCMC Particle Markov Chain Monte Carlo

PV Poliovirus

RMSD root mean squared deviation

SARS Severe Acute Respiratory Syndrome

SEIR Susceptible-Exposed-Infected-Removed model of disease transmission

SIR Susceptible-Infected-Removed model of disease transmission

T_{MRCA} Time to the most recent common ancestor

VDPV Vaccine-Derived Poliovirus

VP Virion Protein of poliovirus

WPV Wild type poliovirus

List of Tables

List of Figures

Contents

Chapter 1

Introduction

In this Chapter I will introduce statistical inference methods that use epidemiological data or pathogen genetic data to estimate model parameters, and efforts to integrate the analyses for the purpose of epidemiological inference. I will then demonstrate how incorporating pathogen genetic data can help quantify heterogeneity in individual infectiousness and highlight the importance of estimating the offspring distribution for epidemiological inference. Despite advances in inference methods and the increasing use of pathogen genetics, analyses of polio data have not benefitted from these methods. Thus I will also provide background information on poliovirus and how an integrated inference framework that uses epidemiological and genetic data can benefit polio research. Parts of this chapter have been adapted from my review article (Li et al., 2014; attached in Appendix A).

1.1 Statistical inference in epidemiology

Transmission models are hypotheses expressed using mathematics, describing how an infection spreads through the population. Commonly used compartmental models track changes in the number of individuals in each disease state and movements between disease states (Anderson and May, 1991). For example, the simple susceptible-infected-removed (SIR) model tracks the number of susceptible, infected, and recovered individuals over time, and describes outbreak dynamics of infectious diseases (Kermack and McKendrick,

1927). The rate at which individuals move from one disease state to the next is determined by model parameters.

Inference of parameter values is possible by fitting mathematical models to data, which is also useful for model comparison and selection. Parameters estimated from data provide information on disease epidemiology. In the context of outbreaks, useful parameters to know include the reproductive number, and the generation time.

A simple approach to model fitting is by minimising the distance between model simulations and observed incidence time series, e.g. curve fitting using least squares. However, this assumes that the observed time series includes all infected individuals. For most infectious diseases, reporting of infected individuals is incomplete due to under-reporting or asymptomatic infections (Gamado et al., 2013). For infectious diseases such as poliovirus, less than 1% of infections result in symptoms that are reported in incidence time series. Inference in these cases requires an observation model in addition to the transmission model to describe the process by which data are generated.

Before discussing inference methods for incidence time series, I will briefly touch upon the two approaches to statistical inference, namely frequentist and Bayesian. The former approach regards data as a random realisation of the model, and thus parameter estimation involves finding the parameter values that maximises the frequency at which the data set could be observed. The goal of frequentist parameter estimation is to obtain point estimates of parameter values for a data set. Bayesian inference, on the other hand, characterises the posterior probability distribution of parameters rather than just a point estimates by updating prior belief about parameters (prior probability) with information from data (likelihood). In this thesis, I adopt the Bayesian inference method.

Regardless of the approach used to estimate parameters, a function is needed to describe the probability of observing data given a set of parameter values, i.e. the likelihood. For temporally correlated data, such as incidence time series, data points cannot simply be treated as independent observations. Like other state space models, the underlying transmission process captured by compartmental models is Markovian, and the observations are conditionally independent given the latent transmission model.

Because analytical solutions of likelihood functions are not usually available for nonlinear and stochastic models, such as those used in infectious disease epidemiology, Markov Chain Monte Carlo (MCMC) methods are frequently used to characterise the distribution of interest through sampling (Metropolis et al., 2004). For Bayesian analysis, the marginal posterior distribution of a parameter is of interest, which is the probability distribution of each parameter while integrating over all other parameters. Often the joint distribution of parameters is also of interest, especially for parameters that are highly correlated. At each iteration, new parameter values are proposed and are accepted with a probability proportional to its marginal posterior probability. By iteratively sampling parameter values and unobserved epidemic history, the MCMC method can be used to estimate parameters based on time-series data, while taking into account stochasticity in the transmission model (Lekone and Finkenstädt, 2006).

Separately proposing epidemic trajectory and parameter values in an MCMC framework is inefficient if there are many parameters and if the epidemic history is highly stochastic. This is because the epidemic history is highly correlated with parameter values. The development of particle MCMC (PMCMC) has enabled the calculation of the marginal parameter likelihood by integrating over possible stochastic epidemic histories (Andrieu et al., 2010). The cost of simulating multiple epidemic trajectories is offset by the more stable estimates of likelihood. Another advantage of using PMCMC is the co-estimation of the epidemic history at the same time as parameter estimation, which means no further simulations are needed to infer the epidemic history from the set of estimated parameter values.

Application of the PMCMC algorithm to infectious disease data has been limited. Examples include the analyses of the Middle East Respiratory Syndrome coronavirus (MERS-CoV) (Cauchemez et al., 2014), Ebolavirus (Camacho et al., 2015). In all these cases, only the time series data have been used to estimate parameter values. In the next section, I will discuss the role that pathogen phylogenies play in parameter inference, and how PMCMC can be used to estimate parameters from both epidemiological time series and phylogenetic data.

1.2 Phylodynamic inference

In parallel with developments in epidemiological inference, the field of phylodynamics arose to enable inference of epidemiological parameters from pathogen sequences. As sequencing technology has become more accessible and affordable, genetic analysis has played an increasingly important role in infectious disease research. Sequencing pathogens can confirm suspected cases of an infectious disease, discriminate between different strains, and classify novel pathogens. In addition to examining individual pathogen sequences, multiple sequences can be analysed together using phylogenetic methods to elucidate evolutionary history (Smith et al., 2009). If an outbreak is densely sampled, then the pathogen phylogeny is informative of the underlying transmission network and helps to uncover who infected whom (Cottam et al., 2008; Gardy et al., 2011; Jombart et al., 2014).

Just as mathematical modelling can be used to analyse surveillance data to reveal details of disease transmission (Section 1.1), analysis of pathogen genomes employs mathematical frameworks to link epidemiological, demographic and evolutionary processes to temporal changes in population-level genetic diversity. Whereas phylogenetics aims to delineate the relationship between individuals, population genetics aims to link population processes to observed patterns of genetic diversity. Inferring the pathogen population history is based on the genealogy of sampled sequences and is often carried out in a retrospective population genetics framework known as the coalescent (Kingman, 1982b). A genealogy describes the ancestry of sampled individuals. Going backward in time, pairs of lineages coalesce when they share a common ancestor until the last two lineages coalesce at the time to the most recent common ancestor for the entire sample (T_{MRCA}). Because the coalescent assumes a small sample compared to the population size, it is an especially useful method for analysing infectious diseases with mild or asymptomatic infections for which time series of reported cases severely underestimates prevalence.

Because of the simplistic assumptions of population genetics models, the population size inferred using coalescent-based methods cannot be directly interpreted as pathogen population size (prevalence of infection N). It is rather the effective population size Ne, which refers to the size of a Wright-Fisher population that would produce the same level

of genetic diversity as observed in the sample. In real populations, the variance of the offspring distribution is higher than expected in a Wright-Fisher population due to heterogeneity in host infectiousness, non-random mixing of the population, and migration events. The consequence of a large variance is that the discrepancy between the effective and census population sizes is greater (Magiorkinis et al., 2013). Accounting for the dispersion of the offspring distribution is especially important when analysing infectious disease data because of the widespread occurrence of transmission heterogeneity (Lloyd-Smith et al., 2005).

The coalescent method has been used to reconstruct N_e dynamics in the past, e.g. for dengue (Bennett et al., 2010), HIV-1 (Grassly et al., 1999; Volz et al., 2013; Faria et al., 2014), and influenza (Fraser et al., 2009). However, N_e is an abstract statistic that is not always proportional to the prevalence of infection N. Theoretical developments linking epidemiological model parameters to N_e (Volz et al., 2009, 2012; Koelle and Rasmussen, 2012) have enabled direct estimation of prevalence of infection N, as well as estimation of parameters relevant to public health such as the basic reproductie number R_0 and generation time T_g, which is related to the durations of latency and infectiousness.

An alternative population genetics approach is the birth-death model, which describes the probability distribution of a genealogy given rates of transmission, removal, and sampling (Stadler, 2010; Stadler and Bonhoeffer, 2013). These rates can be constant over time, as was the case for the original birth-death model, or can change in step-wise fashion (Stadler et al., 2013) or according to compartmental models of disease transmission (Kühnert et al., 2014; Leventhal et al., 2014). Unlike the coalescent framework, the birth-death model is still valid for densely sampled populations, which makes it more useful for studying smaller outbreaks. However, the birth-death model strongly depends on the sampling times of sequences, and thus accurate inference depends on correctly specifying the sampling process which is not always possible (Volz and Frost, 2014). The coalescent makes no assumptions about the sampling probability of sequences except that it is small, and is thus applicable to a greater range of epidemic situations.

Given the flexibility of the coalescent, I have adopted this model for statistical inference in this thesis. For the remainder of this thesis, I will assume a coalescent-based approach

to phylodynamic inference, i.e. estimation of model parameters given the underlying pathogen genealogy.

Packages such as BEAST (Drummond and Rambaut, 2007) and MrBayes (Ronquist et al., 2012) have integrated phylogenetic reconstruction with phylodynamic inference, thereby streamlining inference of epidemiological parameters from sequence data. However, the epidemiological models implemented in BEAST are simple models such as exponential or SIR. Furthermore, stochastic models can only be implemented in BEAST if the exact likelihood can be calculated, which is only possible for simple models. To integrate over both phylogenetic space and stochasticity in epidemiological dynamics, an inference can be repeated for a set of sampled phylogenies (Volz and Pond, 2014). This is also the approach adopted in the analyses presented in this thesis.

The contribution of genetic data to epidemiological studies is evident in recent epidemics of emerging and re-emerging infectious diseases. For example, despite the dearth of sequence data from the MERS-CoV outbreak (Centers for Disease Control and Prevention, 2014), coalescent-based analysis of only 10 genomic sequences produced estimates of time to most recent common ancestor (March 2012; 95% November 2011 June 2012), R_0 (1.21; 95% CI 1.08-1.40) and doubling time (43 days; 95%CI 23-104) (Cauchemez et al., 2014). Without further sequencing of the animal reservoirs, the authors could not infer whether these estimates applied to the animal reservoir or the human epidemic. The credible intervals around the estimates were unsurprisingly large given the small sample size.

Another example is the contribution of pH1N1 sequences during the swine flu pandemic. Analysing 11 haemagglutinin sequences collected over one month, the start date of the epidemic was estimated to be in late January 2009 (Fraser et al., 2009). Repeating the phylogenetic and molecular clock analyses with a further 12 sequences shifted the estimated start date 2 weeks earlier. Fitting an exponential growth model to the sequence data, an estimate of R_0 was estimated to be 1.22, slightly lower than inferred from epidemiological data but with overlapping confidence intervals. To determine at which point during the pandemic coalescent analysis would have provided accurate and precise estimates of evolutionary rate, R_0 and T_{MRCA}, real-time estimates of these parameters were obtained for genomic sequences collected in North America (Hedge et al., 2013). Accurate estimates could be obtained as early as May when 100 viral

genomes had been sequenced, with more precise estimates obtained by the end of June 2009 when 164 had been sequenced. However the inclusion of more sequences of longer length only slightly improved the accuracy of initial estimates (Hedge et al., 2013).

1.3 Integrated analysis

As both epidemiological data and pathogen genetics are informative of the underlying transmission process, there is a trend towards using both epidemiological and genetic data in an integrated inference framework. The two types of data contribute different information regarding the spread of an infectious disease. Phylodynamic methods using the coalescent do not require knowledge of the sampling probability of sequences. These methods are thus useful for quantifying the size of the infected population when the reporting probability is low and varies over time. On the other hand, uncertainty intervals surrounding parameter values estimated from genetic data tend to be larger than those estimated from epidemiological data (Rasmussen et al., 2011). This is because evolutionary processes as well as the transmission process shape the pathogen phylogeny, which reduces the information the pathogen phylogeny can provide on the spread of disease.

Informally, inference results from pathogen genetic data can be visually compared to those from epidemiological data. Skyline plots (Drummond et al., 2005) and skygrid plots (Gill et al., 2013) track changes in pathogen diversity (strictly $N_e T_g$) over time, and are often visually compared to reported incidence time series to determine similarity (Hedge et al., 2013; Bennett et al., 2010). Fitting parametric models of population growth to pathogen genetic data produce estimates of epidemic growth rate (Pybus et al., 2001). These estimates can then be used to infer the reproductive number if the generation time distribution is known. For pH1N1, coalescent analysis of early sequences produced estimates of R_0 that were slightly lower than values estimated from incidence time series though the confidence intervals still overlapped (Fraser et al., 2009).

Discrepancies in inference results from genetic data compared to those from epidemiological data suggest that some assumptions made in the population genetics model, epidemiological model, or both are wrong (Rasmussen et al., 2014a). An

integrated approach where the same transmission models are fit to both genetic and epidemiological data can help to identify the model that best describes both data sets, and find parameter estimates consistent with both sets of data.

For densely sampled outbreaks, methods exist to jointly infer phylogenies and transmission trees from sequences and epidemiological information (Ypma et al., 2013; Didelot et al., 2014). However reconstructing transmission trees is not feasible for large outbreaks or for under-reported infectious diseases. Instead, transmission models such as those fit to incidence time series can be fit to the pathogen phylogeny from a sparsely sampled outbreak. Definition the coalescent process using transmission model parameters has facilitated this type of inference by linking rates of coalescence to incidence and prevalence of infection (Volz et al., 2009; Volz, 2012; Koelle and Rasmussen, 2012). Combining these formulations of the coalescent with the PMCMC algorithm, Rasmussen et al. (2011) showed that a joint inference framework can be used to estimate epidemiological parameters from both pathogen phylogeny and incidence time series.

So far the joint inference framework has only been applied to a single set of simulated data so far (Rasmussen et al., 2011). A more recent paper (Rasmussen et al., 2014a) where the inference framework was applied to viral phylogenies for dengue produced estimates of prevalence over time, which was visually compared to incidence time series. No results were presented where models were fit to both incidence time series and phylogeny to estimate prevalence. Given that the PMCMC algorithm has been successfully used to analyse both epidemiological and phylogenetic data for a simulated data set, it would be interesting to test the inference framework on a wider range of parameter values with more simulations and apply the framework to real data to assess the value of an integrated inference approach. The real data that I analyse in this thesis are for polio, which will be discussed in more detail later in Section 1.5.

Furthermore, as the inference framework only works for a single phylogeny, the phylogeny is not estimated from genetic data at the same time as parameter estimation. Repeated inference from different phylogenies sampled from BEAST analysis of genetic data yields similar results (Rasmussen et al., 2014a,b), although the data in these papers were collected over a number of years and the uncertainty in branching times was small.

In this thesis, I will focus on the analysis of outbreak data. Integrating over phylogenetic uncertainty for outbreak data is important as the resolution of the viral phylogeny in terms of branching times is often poor.

Another extension that can be made to a joint inference framework is the inclusion of the offspring distribution in the stochastic epidemic model that is fit to data. This statistical distribution is important for accurate parameter estimation using the coalescent, and when fitting stochastic epidemic models to outbreak data. More details on this distribution and its relevance for epidemiological inference is given in the following section.

1.4 Overdispersed offspring distribution

1.4.1 Definition

The reproductive number is an important statistic for infectious diseases, and it is defined as the average number of infections caused by an infected individual during the course of their infection. This statistic is useful for determining whether an outbreak is dying out, or for estimating the critical vaccination threshold. It is a summary statistic for infection spreading in a population and is therefore dependent on properties of both the population at risk and of the infectious pathogen. However, there is often significant variation around this number between individuals due to heterogeneities in susceptibility and infectiousness between individuals (Becker and Britton, 1999). This variation can lead to unpredictable epidemic dynamics especially at the beginning of outbreaks when few individuals are infected.

The distribution of secondary infections per infected individual is known as the 'offspring distribution'. While not specified in typical compartmental models used in epidemiology, the offspring distribution is a concept integral to branching process models (Harris, 2002). Infectious disease outbreaks are more unpredictable when the variance in the offspring distribution is large. While the outbreaks are more likely to die out when the variance is large, the ones that do take off tend to be larger than outbreaks where the offspring numbers are more homogenous (Lloyd-Smith et al., 2005; Garske and Rhodes, 2008; de Silva et al., 2012a).

In terms of inference, the variance of the offspring distribution affects estimation of prevalence from pathogen phylogeny (Koelle and Rasmussen, 2012; Magiorkinis et al., 2013). This is because phylodynamic inference methods using the coalescent need to rescale the timescale according to the variance of the offspring distribution to use the original coalescent derivations for the simple Wright-Fisher model of population genetics (Kingman, 1982b). Thus characterising the offspring distribution is important for accurate estimation of epidemiological parameters when using pathogen phylogeny.

1.4.2 Dispersion parameter k of the negative binomial

To characterise the offspring distribution requires specification of a statistical distribution for the offspring numbers. If the sample variance is greater than the expected variance according to a model, the offspring distribution is considered overdispersed. The negative binomial is often used for biological count data (Bliss and Fisher, 1953; Shaw et al., 1998) including contact tracing data for infectious diseases (Lloyd-Smith et al., 2005; International Ebola Response Team et al., 2016). In the case of infectious disease spread, the negative binomial corresponds to a Gamma-Poisson mixture in which the infectious period is Gamma-distributed and the number of transmissions during that infectious period is Poisson distributed.

The negative binomial is usually used to describe the number of successful trials given r total trials and a probability of success p. For the purposes of infectious disease research it is more intuitive to parameterise the distribution by its mean $R = \frac{pr}{1-p}$ and dispersion $k = r$. The variance $\sigma^2 = R(1 + \frac{R}{k})$ is larger when k is small.

In a typical compartmental model, the offspring distribution is not directly specified but depends on the generation time distribution and the structure of the population. For an unstructured SIR model, the resulting offspring distribution is the geometric, which reflects the Poisson transmission process and exponentially distributed duration of infection. Heterogeneity can be introduced by including sub-populations modelled by different compartments with different levels of infectiousness.

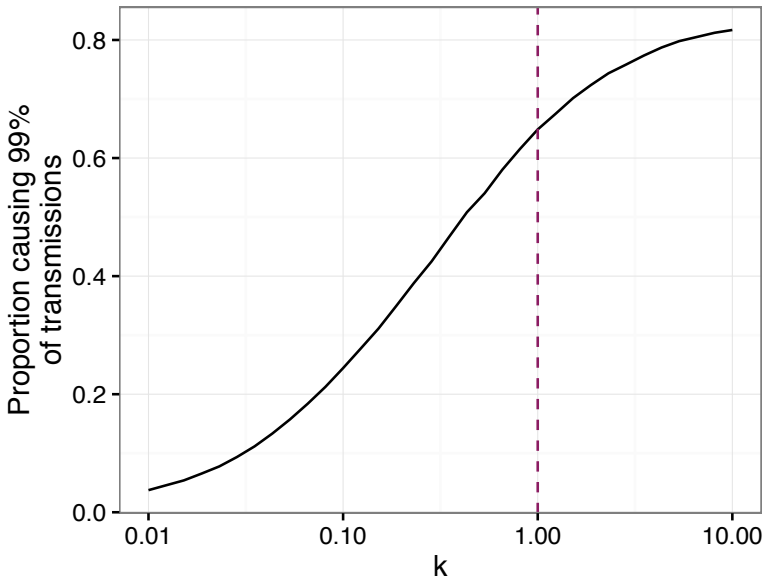


Figure 1.1: Effect of overdispersed offspring distribution (small k) on transmission heterogeneity. In an unstructured SIR model, the offspring distribution is geometric which means $k = 1$. At $k = 1$ (dashed line), the top 65% of individuals ranked by the number of onward transmissions cause 99% of infections. A reproductive number of $R = 2$ is assumed here.

1.4.3 Inferring k from data

Using contact tracing data, the value of k has been estimated for a range of directly transmitted infectious diseases with many examples of highly overdispersed offspring distribution. A classic example of an infectious disease with highly overdispersed offspring distribution is severe acute respiratory syndrome (SARS), for which a k of 0.16 (0.11, 0.64) was estimated for the outbreak in Singapore (Lloyd-Smith et al., 2005). More recently for the Ebola outbreak in West Africa, k was estimated to be between 0.03 and 0.52, depending on assumptions regarding sampling (International Ebola Response Team et al., 2016). The consequence of small k is that a small number of infected individuals contribute to the majority of transmissions (Figure 1.1). During the SARS epidemic in 2003, for example, there were many reports of extreme superspreading events (Riley et al., 2003). Similarly, there were superspreading events reported at funerals and in healthcare settings during the Ebola outbreak (Faye et al., 2015).

Obtaining contact tracing data is difficult as it is expensive and time-consuming, and

sampling is usually incomplete due to under-reporting. In the absence of contact tracing data, it is possible to estimate k from final size distributions (Garske and Rhodes, 2008; Blumberg and Lloyd-Smith, 2013). For a single outbreak, estimating k is difficult using epidemiological data alone. The pathogen phylogeny might be more informative of the dispersion parameter k because the times between coalescent events are shorter when k is small. For endemic settings, the rate of coalescence increases linearly with the variance of the offspring distribution (Kingman, 1982b). However, this is not the case in epidemic situations when the reproductive number changes over time (Koelle and Rasmussen, 2012). Adapting the original Wright-Fisher derivations for time-varying population sizes, Fraser and Li (2017) derived the coalescent for an arbitrary offspring distribution in a time-varying population (details provided in Chapter 2).

In the inference framework developed in this thesis, I focus on estimating epidemiological parameters concurrently with the dispersion parameter k using various transmission models. While branching process models contain offspring distribution parameters, typical compartmental models do not. A different formulation of the compartmental model is therefore needed to enable estimation of k by fitting such a model to data, and this is discussed in Chapter 3.

1.5 Polio

The inference methods discussed in Sections 1.1 to 1.3 have mainly focused on highly prevalent diseases (e.g. HIV, influenza, dengue) and emerging infectious diseases (SARS, Ebola). Thus, I applied the inference framework developed in this thesis to poliovirus data to highlight the value added by genetic data. Furthermore, there are not estimates of k for poliovirus in the literature, so inference of k from data should provide information on the level of overdispersion in the population and the extent to which superspreading contributes to infection spread.

1.5.1 Background

Poliomyelitis is a paralytic disease mainly affecting children. Its aetiological agent, poliovirus, is an enterovirus in the Picornaviridae family. The poliovirus genome is 7,500 nucleotides long with a single open reading frame. The resulting polyprotein is autocatalytic and yields 4 structural (Virion Protein 1-4; VP1-4) and 7 non-structural proteins (Racaniello and Baltimore, 1981). There are three wild poliovirus serotypes. Wild type 1 poliovirus (WPV1) causes the most morbidity and mortality, WPV2 has been eradicated, and WPV3 has not been observed since November 2012 (World Health Organization, 2016).

The virus is mainly transmitted faecal-orally although the oral-oral route is also possible where hygiene standards are high (Minor, 2004). Most infections are asymptomatic or result in only mild symptoms. However, poliovirus can occasionally become viraemic, and cross the blood-brain-barrier to infect the motor neurones in the central nervous system (CNS) causing paralysis (Centers for Disease Control and Prevention, 2012). Poliovirus can also enter the CNS via peripheral nerves in the muscle (Ren and Racaniello, 1992). It is not clear what the risk factors are for paralysis, although injections of non-poliovirus vaccines shortly before poliovirus infection have been shown to increase the risk of paralysis (Sutter et al., 1992). The case-fatality rate of poliomyelitis varies between different age groups, with adults at most risk of death during the acute phase of poliovirus infection (Nathanson and Kew, 2010).

As there is no effective cure, prevention via vaccination is crucial for poliovirus control. There are two classes of poliovirus vaccines, both protective against paralytic symptoms. The first is the inactivated polio vaccine (IPV) administered via injections, which provides long-term protection against paralysis but does not stop virus carriage and shedding (Salk et al., 1984). The oral polio vaccine (OPV) contains live attenuated viruses to induce gut immunity and thus prevent shedding of live virus (Sabin, 1957). The downside is that OPV replicates in the gut and mutates back into a virulent form for 1 in 1,000,000 vaccinated individuals (Fine and Carneiro, 1999). Furthermore, vaccine-derived poliovirus (VDPV) can spread to unvaccinated individuals and cause potentially large outbreaks.

The Global Polio Eradication Initiative (GPEI) was established in 1988 on the heel of

smallpox eradication. The annual incidence of poliomyelitis has decreased by > 99% from 350,000 in 1988 to just 74 in 2015 (World Health Organization, 2016). Polio is a good target for eradication for several reasons. First, vaccines that are cheap and effective means that immunisation alone can stop poliovirus circulation. Secondly, humans are the only host for poliovirus, which means there are no animal reservoirs to seed new infections after eradication (Dowdle and Birmingham, 1997). Thirdly, chronic infections have rarely been observed except in immunodeficient patients (Dowdle and Birmingham, 1997; de Silva et al., 2012b; Dunn et al., 2015), so there is unlikely to be a pool of latently infected individuals.

There are, nevertheless, many challenges to polio eradication. Political instability can cause gaps in vaccine coverage and allow re-emergence of polio in previously polio-free regions, as illustrated by the 2013 outbreak in Syria (Eichner and Brockmann, 2013), and the reappearance of polio in Nigeria after being undetected for 2 years (Centers for Disease Control and Prevention, 2016). Achieving high vaccine coverage is also hindered by attacks against health workers such as in Pakistan (Centers for Disease Control and Prevention, 2013). Both WPV and VDPV need to be eradicated to stop polio cases from occurring. The latter requires the withdrawal of OPV to prevent the spread of VDPV, a process that began in April 2016 with the removal of type-2 poliovirus in the OPV (World Health Organization, 2013).

1.5.2 Surveillance

Another challenge for polio eradication, and for studying polio epidemiology in general, is that most infections are asymptomatic.

As polio eradication approaches completion, it is vital to have a sensitive surveillance system to detect any remaining cases. The standard surveillance system for polio is detection of acute flaccid paralysis (AFP) cases (World Health Organization, 2004). However AFP is not a polio-specific symptom, and most AFP cases are not due to poliomyelitis as the eradication program approaches completion. Furthermore, the sensitivity of AFP surveillance is constrained by the large number of asymptomatic infections per paralytic case (Nathanson and Martin, 1979). Knowledge of the

case-to-infection ratio is important for quantifying the extent of PV infections in a population and acts as a marker for eradication progress.

There have been very few studies on the case-to-infection ratio; the most often quoted is the prospective study carried out in America in the 1950s (Melnick and Ledinko, 1953). In this study, 22,900 children under 15 were serologically tested before and after a polio epidemic. The overall case-to-infection ratio was 1:82 with 63 paralytic cases resulting from 5,200 infected children. The case-to-infection ratio increased with age, from 1:175 in children under 1 to 1:95 in 10-14 year olds. Because the total number of paralytic cases was quite low, there is great uncertainty surrounding these estimates. In a separate study in Finland during the 1954 polio epidemic, the ratio was found to be 1:250 before the epidemic and 1:110 during the epidemic (Penttinen and Patiala, 1961). Clearly the case-to-infection ratio varies between age-groups, populations and other environmental factors. However a temporally fixed ratio of 1:200 is often used for inference of infection prevalence (Wringe et al., 2008). Deviations from this fixed value could produce large differences in incidence and prevalence, which would impact other epidemiological estimates such as reproductive number and timing of the first infection.

1.5.3 Using pathogen genetics to characterise polio epidemiology

Existing work involving polio sequence data mainly focuses on the evolutionary relationships between isolates. The introduction of sequence data to poliovirus surveillance increased the resolution of strain identification from serotypes to genotypes that were associated with specific geographic locations (Kew et al., 1995). For example, phylogenetic analysis of environmental sequences collected in Gaza, Israel revealed two separate importation events when endemic circulation of poliovirus was suspected (Hovi et al., 2012). In endemic settings, phylogenetic analysis of environmental sequences have been used to identify transmission links between Afghanistan and Pakistan (Angez et al., 2012).

Despite the extensive use of genetic data in polio surveillance and research, there is no existing literature on demographic inference from polio sequence data. Given the

potential of genetic analysis to produce more informed estimates of case-to-infection ratio, a phylodynamic approach that incorporates coalescent-based genetic analysis with epidemiological investigation of polio could yield important estimates of infected population size that would help monitor the progress towards the eradication end goal.

Poliovirus is one of the fastest evolving human pathogens with substitution rates of 1×10^{-2} substitutions/site/year (Jorba et al., 2008). Therefore the genealogical relationship between samples can be inferred at greater resolution than would be possible if sequence divergence was small. Furthermore, the coalescent assumes that the sample size is much smaller than the population size, which is a characteristic of polio epidemiology due to the low probability of developing paralysis and under-reporting of cases (Centers for Disease Control and Prevention, 2012). The ratio between infections and reported cases was estimated to be 200:1 in a prospective study when serological tests were performed on a population before and after an epidemic (Melnick and Ledinko, 1953); no extensive studies of the infection-to-case ratio has since been carried out. Incorporating genealogical information based on sequence data would help to elucidate temporal and spatial variations in the proportion of infections resulting in reported cases.

Genetic analysis can shed light on the case-to-infection ratio as coalescent methods can be used to infer the prevalence of infection in the population. However this depends on the accurate estimation of the variance of the offspring distribution.

1.6 Aims and structure of the thesis

The two main aims of this thesis are

1. to develop and implement a statistical inference framework for integrated analysis of epidemiological and genetic data in an outbreak setting,
2. and to use the statistical inference framework to analyse poliovirus data.

For sparsely sampled diseases, there is currently only one publication so far that uses an integrated approach to epidemiological inference (Rasmussen et al., 2011). However the

paper only used an integrated approach for one simulated dataset, did not consider the offspring distribution, and only used a single pathogen phylogeny for inference. In this thesis, I develop an integrated inference approach that can be used to fit models with arbitrary offspring distributions and can integrate over estimates from multiple phylogenies (details in Chapter 2). I test the inference framework on a larger set of simulated data compared to Rasmussen et al. (2011) with results in shown in Chapter 3.

Because existing programs for PMCMC are not suitable for analysing phylogenetic data, I provide a parallelised implementation of the framework in C++ (github.com/lucymli/EpiGenMCMC) with an accompanying R package (github.com/lucymli/EpiGenR) to facilitate data input and output. The code is available at on GitHub and is attached in Appendix B.

Furthermore, I apply the inference framework to real data from a poliovirus outbreak (Chapter 4) to demonstrate the value added by including genetic data in the epidemiological analysis of poliovirus, which has not been done before. In Chapter 5, I demonstrate that the sequence data from environmental samples can also be used to estimate epidemiological parameters, which is useful for post-eradication surveillance of poliovirus.

Finally in Chapter 6 I will discuss limitations of the inference framework presented here, implications for polio endgame, and how it can be extended and applied to other infectious diseases.

Chapter 2

Inference framework

To estimate epidemiological parameters including k from both epidemiological data and pathogen phylogeny, I wrote an implementation of the particle Markov Chain Monte Carlo (PMCMC) algorithm (Andrieu et al., 2010) in C++. Before introducing PMCMC and providing details of my implementation, I will first introduce the coalescent and the Fraser and Li (2017) formulation of the coalescent that is used in my inference framework. Section 2.3.2 is part of a manuscript (Fraser and Li, 2017); equations were derived by Christophe Fraser, and both Christophe Fraser and I contributed to the writing of the manuscript.

2.1 Notation

The inference framework presented in this Chapter brings together methods and terminology from the coalescent, infectious disease epidemiology, and particle filtering. To avoid confusion, I will first define terms and notations used in this thesis.

The inference method here uses epidemic simulations to calculate the likelihood of parameter values. Epidemics are stochastically simulated in discrete time steps, where each time step lasts Δt. Individual i become infected at time step H_i, and infects Z_i other individuals ('offspring') during the course of their infection. Z_i is drawn from an offspring distribution ϕ_t where $t = H_i$. The mean and variance of ϕ_t are the reproductive number $R_t = E(Z_i|H_i = t)$ and σ^2, respectively. In this thesis, I use a

negative binomial offspring distribution parameterised by the mean R_t and dispersion parameter k, where $\sigma_t^2 = R_t(1 + \frac{R_t}{k})$.

At time step t, the simulated epidemic trajectory X_t comprises two numbers: incidence Inc_t and pairwise coalescent rate λ_t. The coalescent rate is calculated using Equation 2.15 based on the time-varying prevalence I_t and reproductive number R_t, and the generation time T_g and dispersion of the offspring distribution k which are invariant over time.

Raw epidemiological data are in the form of line lists, in which each row $i = \{1, ..., I_{\text{total}}\}$ contains information on individual i including the time step in which they become infected H_i. Line lists Epi_t are aggregated temporally into time series that record the reported number of individuals infected on a daily (or less frequent) basis. For infectious diseases with rapidly changing dynamics, size of the simulation time step could be less than the time unit of reporting. For example, if a simulation time step is 0.25 days and cases are reported on a daily basis, then the aggregated number of cases $\text{Inc}_{t:(t+3)} = \sum_{x=t}^{t+3} \text{Inc}_x$ in the simulation and the number of reported case on that day $\text{Epi}_{t:(t+3)}$ are used to calculate the likelihood. For simplicity, the derivations below assume that the simulation time steps correspond to the reporting time steps.

Phylogenetic data Phy_t comprise the number of lineages through time A_t and times between events U_t. Unlike epidemiological data where Epi_t is a number, both A_t and U_t are vectors including one or more numbers. More details will be provided in Section 2.3.2.

2.2 Overview of inference framework

The aim of the inference method presented in this chapter is to obtain the posterior distribution of parameter values given epidemiological and phylogenetic data (Figure 2.1). These parameters define the rates in a disease transmission model, which mathematically describes hypotheses regarding how an infectious disease spreads in the population. Because transmission models are often nonlinear and stochastic, it is not possible to directly calculate the probability of observing disease data given a set of parameter values. I therefore use a PMCMC approach to estimate parameter values by

comparing stochastic model simulations at various parameter values with the data. The coalescent provides the likelihood function linking the pathogen phylogeny to a simulated epidemic trajectory.

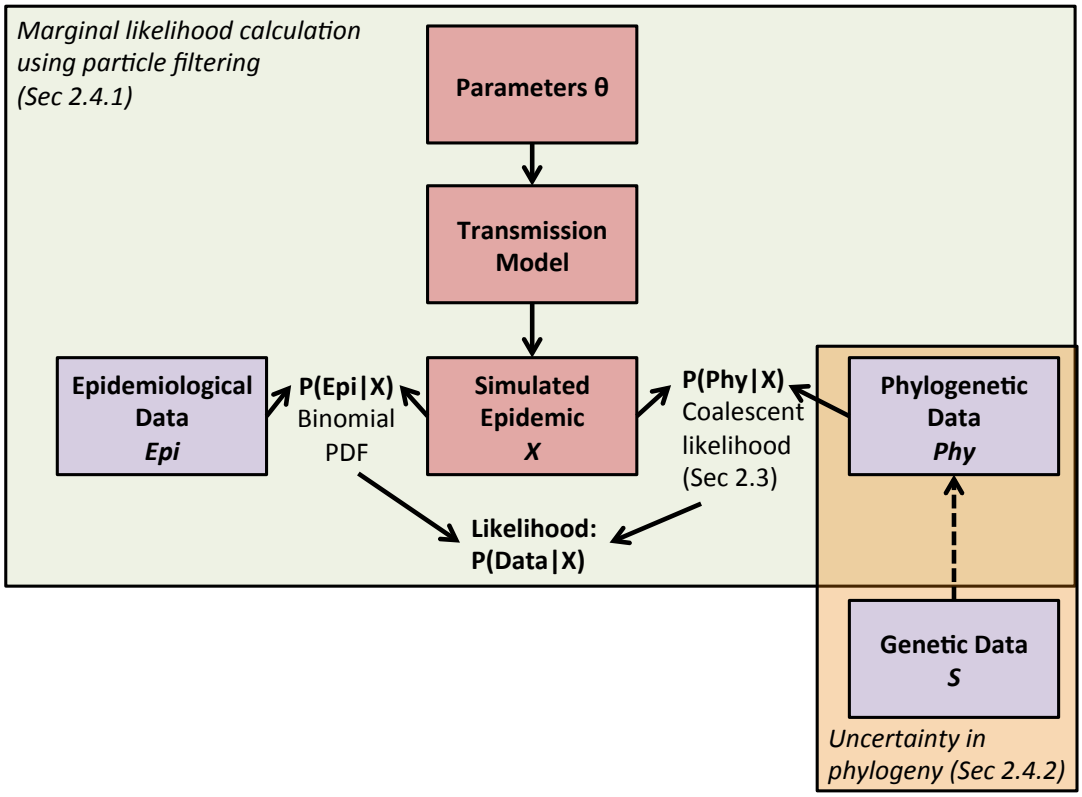


Figure 2.1: Overview of inference approach. To obtain a sample of parameter values distributed according to the posterior density, a Particle Markov Chain Monte Carlo (PMCMC) approach is used. The MCMC part of the algorithm samples parameter values from the posterior distribution, while the particle filtering part is used to calculate the marginal likelihood $P(D|X)$ while integrating over stochastic simulations. The coalescent provides the likelihood for the phylogenetic data, which are inferred from genetic (sequence) data using a phylogenetic reconstruction program such as MrBayes. Uncertainty in the reconstructed phylogeny needs to be accounted for and this is discussed in Section 2.4.2.

The method will be applied to simulated data 3 and polio data in Chapters 4 and 5.

2.3 Coalescent

2.3.1 Derivations of the original coalescent

For n individuals sampled from a population with N individuals, the sample genealogy can be constructed by tracing back the ancestry of sampled individuals until the most recent common ancestor is found. For infectious diseases, the total population from which the sample is taken refers the population of infected individuals, and the MRCA refers to the most recently infected individual who caused transmission chains that ultimately led to the infection of the sampled individuals. The coalescent provides the probability density function for times of coalescence in a sample genealogy going backward in time as a function of N. This allows the coalescent to be used in epidemiological inference within a PMCMC framework, as the likelihood of simulated epidemic trajectories can be calculated given a genealogy.

While the true topology and coalescent times of a sample genealogy are not usually known for infectious diseases, the genealogy can be approximated by reconstructing the dated phylogeny of pathogen sequences where branch lengths are in units of time. Neutral evolution, lack of within-host diversity, and lack of co-infections are assumed so that the disease spread is the only process shaping the pathogen phylogeny.

For the derivations in this and next sections, I assume that the dated phylogeny matches the true genealogy. Uncertainty in phylogeny and ways to incorporate this uncertainty are discussed later in this Chapter.

This original coalescent (Kingman, 1982b) was derived for a sample genealogy from a simple Wright-Fisher population model (Fisher, 1930; Wright, 1931). The main assumptions of Kingman's coalescent (1982b) are:

1. Discrete, non-overlapping generations. All infected individuals recover at the at the same time, and pass on their infections at the time of recovery.

2. Fixed population size. Infectious disease at endemic equilibrium so the prevalence of infection does not change over time.

3. Multinomial offspring distribution with equal probabilities. All individuals are similarly infectious.

4. No population structure or migration.

5. Small sample size compared to the population size N.

Despite the unrealistic and simplistic assumptions of Wright-Fisher, the coalescent derivations are robust to violations of assumptions 1-3 so long as time is re-scaled appropriately (Kingman, 1982a; Griffiths and Tavare, 1994; Möhle, 1998). In Section 2.3.2, I will present derivations of the coalescent (Fraser and Li, 2017) that allows time-varying population size and arbitrary offspring distribution. Extensions of the coalescent have been made to incorporate population and geographical structure, although these are ignored for the analyses carried out in this thesis (Notohara, 1990; Hudson, 1991; Volz, 2012). However, the assumption of small sample size is necessary for the coalescent derivations to hold, as the derivations are at the limit $N \to \infty$.

If two individuals are sampled at generation r, the total number of lineages is $n = 2$. The probability of no coalescent event in generation $r - 1$ is calculated as

$$1 - p_2 = \frac{N}{N}\frac{N-1}{N} = 1 - \frac{1}{N}. \tag{2.1}$$

Thus, the probability of coalescence p_2 between two lineages is

$$p_2 = \frac{1}{N}. \tag{2.2}$$

For $n > 2$ sampled lineages, there are $\binom{n}{2}$ number of pairs of lineages between which coalescence can occur. This assumes that the number of sampled lineages is small compared to N, and thus no more than one coalescent event is likely to occur within a single generation. The probability of a coalescence between 2 of n lineages p_n is

$$p_n = \frac{\binom{n}{2}}{N}. \tag{2.3}$$

The number of generations until a coalescence between 2 of n lineages $T^{(n)}$ follows a geometric distribution:

$$P(T^{(n)} = d) = \frac{\binom{n}{2}}{N}(1 - \frac{\binom{n}{2}}{N})^{d-1}. \tag{2.4}$$

As N tends to ∞ leading to a small probability of coalescence, and rescaling time to units of N generations, the geometric converges to an exponential distribution with rate $\frac{n}{2}$

$$P(\frac{T^{(n)}}{N} = m)\frac{n}{2}e^{-\frac{n}{2}m}, \tag{2.5}$$

where m is time measured in N generations.

Because time u is usually measured in continuous time units, and if we know the generation time T_g, then we can substitute $m = \frac{u}{NT_g}$ into Equation 2.5

$$P(\frac{T^{(n)}}{NT_g} = u) = \frac{\frac{A_r}{2}}{N}e^{-\frac{\frac{A_r}{2}}{N}\frac{u}{T_g}}. \tag{2.6}$$

The generation time for infectious diseases refers to the time interval between one individual becoming infected and that individual passing on the infection to another person.

For $n > 2$, the times to coalescence are a series $T_{\text{MRCA}} = \{T^{(n)}_{\text{MRCA}}, T^{(n-1)}_{\text{MRCA}}, ..., T^{(2)}_{\text{MRCA}}\}$.

The probabilities of observing this series is thus

$$P(T_{\text{MRCA}}) = \prod_i \frac{\binom{n}{2}}{N} e^{-\frac{\binom{n}{2}}{N}u}. \tag{2.7}$$

It can be concluded that for the Wright-Fisher model, rate of coalescence is inversely related to N. In terms of epidemiological inference, this means that when multiple coalescent events occur in a short period of time, the conclusion would be that the number of infected individuals is small. However, this relationship changes if the offspring distribution is more overdispersed than can be captured by a symmetric multinomial model. If the assumption of constant N still holds but the variance of the offspring distribution increases σ^2, the rate of coalescence for a given pair of sampled individuals is higher than in the original Wright-Fisher model (Figures 2.2A and B). In this case, N inferred based on the rate of coalescence in the genealogy would be lower than expected.

For populations where the variance of the offspring distribution σ^2 is greater than that for the Wright-Fisher model, the population size estimated based on the genealogy is the effective population size Ne. Although an abstract concept, Ne refers to the size of a Wright-Fisher population that would generate the same distribution of coalescent times as that observed in the sample genealogy. If the population size does not change over time, N is related to Ne via $N\text{e} = \frac{N}{\sigma^2}$ (Kingman, 1982b).

2.3.2 Coalescent with an arbitrary offspring distribution

While the relationship $N\text{e} = \frac{N}{\sigma^2}$ holds for endemic settings, prevalence of infection can change quickly during outbreaks. This means that the expected times to coalescence (Figure 2.2C). Furthermore, as discussed in Chapter 1, the offspring distribution can be highly overdispersed for many infectious diseases.

To incorporate time-varying population sizes and time-varying offspring distribution, Fraser and Li (2017) provided coalescent derivations based on a population with discrete non-overlapping generations.

In generation r there are N_r individuals indexed by $i = \{1, ..., N_r\}$. The probability mass

function $\phi_r(\nu)$ of an offspring distribution describes the probability of individual having ν offspring in the next generation. The offspring distribution at generation t has a mean of

$$R_r = \sum_{\nu=0}^{\infty} \nu \cdot \phi_r(\nu) \tag{2.8}$$

and a variance of

$$\sigma_r^2 = [\sum_{\nu=0}^{\infty} \nu^2 \phi_r(\nu)] - R_r^2. \tag{2.9}$$

Each individual i in generation r will have Z_i number of offspring in generation $r+1$, with probability $\phi_r(Z_i)$. The total number of offspring, and thus the number of individuals in the next generation, is

$$N_{r+1} = \sum_{i=1}^{N_r} Z_i \tag{2.10}$$

The total number of pairs of individuals in generation r is $\binom{N_{r+1}}{2}$. The proportion of these pairs that shared a common parent in generation r is

$$p_r = \frac{\sum_{i=1}^{N_r} \binom{Z_i}{2}}{\binom{N_{r+1}}{2}} = \frac{\sum_{i=1}^{N_r} Z_i^2 - \sum_{i=1}^{N_r} Z_i}{N_{r+1}^2 - N_{r+1}}. \tag{2.11}$$

which can be interpreted as the pairwise coalescent rate per generation.

Assuming a large population size N_r, then the observed mean and variance are equal to the expectation (Equation 2.8) and variance (Equation 2.9) of the offspring distribution, which means

$$R_r = \frac{\sum_{i=1}^{N_r} Z_i}{N_r} = \frac{N_{r+1}}{N_r} \tag{2.12}$$

and

$$\sigma_r^2 = \frac{\sum_{i=1}^{N_r} Z_i^2}{N_r} - R_r^2 \tag{2.13}$$

Re-arranging Equations 2.12 and 2.13 produces the following expressions: $N_r = \frac{N_{r+1}}{R_r}$ and $\sum_{i=1}^{N_r} Z_i^2 = (\sigma_r^2 + R_r^2)N_r$, respectively. Substituting these expressions and Equation 2.10 into Equation 2.11 results in a definition of p_r in terms of the N_{r+1}, and R_r and σ_r^2:

$$\begin{aligned} p_r &= \frac{(\sigma_r^2 + R_r^2)N_r - N_{r+1}}{N_{r+1}^2 - N_{r+1}} \\ &= \frac{(\sigma_r^2 + R_r^2)\frac{N_{r+1}}{R_r} - N_{r+1}}{(N_{r+1}^2 - N_{r+1})} \\ &= \frac{\frac{\sigma_r^2}{R_r} + R_r - 1}{(N_{r+1} - 1)} \end{aligned} \tag{2.14}$$

which can be approximated by the following expression when N_{r+1} is large

$$p_r \approx \frac{\frac{\sigma_r^2}{R_r} + R_r - 1}{N_{r+1}} \tag{2.15}$$

The expression in Equation 2.15 reduces to other formulations of the coalescent depending on assumptions of the offspring distribution. According to Wright-Fisher population dynamics, $R_r = 1$ and $\sigma_r^2 = R_r$. Thus

$$p_r = \frac{1}{N_{r+1}}. \tag{2.16}$$

In an endemic setting where $\sigma_r^2 \neq R_r$ and $R_r = 1$, Equation 2.15 becomes

$$p_r = \frac{\sigma_r^2}{N_{r+1}}. \tag{2.17}$$

These are the same as those derived in Kingman (1982b). Compared to the expression for the coalescent rate in endemic settings (Equation 2.17), Fraser and Li's (2017) formulation of the coalescent rate depends on the mean as well as the variance of the offspring distribution. If Equation 2.17 is used in an outbreak setting, this can lead to erroneous estimates of the prevalence of infection especially when the reproductive number is large.

These discrete time calculations can approximate continuous time dynamics when generation time is small, such as for acute infectious diseases. The coalescent rate in continuous time is given by $\lambda(u)\frac{p_r}{T_g}$ at time $u = rT_g$.

For compartmental models, Volz et al. (2009) derived the continuous-time coalescent rate for an SIR model

$$\lambda(u) = 2\frac{\text{Incidence}(u)}{\text{Prevalence}(u)^2}. \tag{2.18}$$

In an SIR model, the reproductive number $R(u) = \beta(u)S(u)T_g$, where $\beta(u)$ is the per-capita transmission rate and $S(u)$ is the number of susceptible individuals at time u. Incidence is calculate as $\text{Incidence}(u) = \beta(u)S(u)\text{Prevalence}(u)$, which means $R(u) = \frac{\text{Incidence}(u)}{\text{Prevalence}(u)}T_g$. The offspring distribution is geometric, which results from the mixture of a Poisson transmission process and an exponentially distributed duration of infectiousness T_g. The mean and variance of the geometric offspring distribution are $R(u)$ and $R(u)^2 +$

$R(u)$, respectively. This assumes that $R(u) = \frac{1-p}{p}$ is the mean number of failures, given a probability of success p. Equation 2.18 can be derived from Equation 2.15 by setting $N_{r+1} = \text{Prevalence}(u)$:

$$\begin{aligned}\lambda(u) &= \frac{2R}{\text{Prevalence}(u)T_g} \\ &= \frac{2\frac{\text{Incidence}(u)}{\text{Prevalence}(u)}T_g}{\text{Prevalence}(u)T_g} \\ &= 2\frac{\text{Incidence}(u)}{\text{Prevalence}(u)^2}.\end{aligned} \tag{2.19}$$

In the analyses presented in Chapters 3-5, I use the negative binomial to capture overdispersion in the offspring distribution. I assume that the reproductive number varies with small discrete time step t while the dispersion k does not change over time. Because variance of the negative binomial is $\sigma_t^2 = R_t + \frac{R_t^2}{k}$, the coalescent rate in a time step sized Δt is

$$\lambda_t = \frac{R_t(1+\frac{1}{k})}{N_{t+1}T_g}. \tag{2.20}$$

2.3.3 Coalescent likelihood

In the inference framework, likelihood given the genetic data is calculated as the probability of observing a dated phylogeny given a simulated epidemic trajectory. The coalescent rate given in Equation 2.15 is calculated for each simulated epidemic trajectory, and the rate parameterises an exponential distribution whose probability density function is used to calculate the probability of observing the time intervals between coalescent events.

The likelihood given a phylogeny is calculated in a piecewise fashion for small time intervals (see Figure 2.3 for example). The small time intervals are unequal in size, and bounded by the times for one of 3 'events': end of a simulation time step (total of T_{n_T} steps), coalescence (total of $n_{\text{tips}} - 1$ events), or sampling (total of n_{tips} events). The

length of each time interval between events is denoted by U_s where $s = \{1, ..., n_T + 2n_{\text{tips}} - 1\}$, and the number of lineages at the end of each interval is A_s. Let the function $g(t)$ return a vector of indices of time intervals the phylogeny corresponding to simulation time step t. The phylogenetic data at simulation time step t are summarised by $\text{Phy}_t = \{U_{g(t)}, A_{g(t)}\}$.

At each simulation time step t, I calculated the coalescent likelihood sequentially for each time interval $s \in g(t)$,

$$P(\text{Phy}_{t,s}|\lambda_{t,s}) = \begin{cases} \binom{A_s}{2}\lambda_t e^{-\binom{A_s}{2}\lambda_t U_s} \text{ if interval } s \text{ starts with a coalescent event} \\ e^{-\binom{A_s}{2}\lambda_t U_s} \text{ otherwise} \end{cases} \tag{2.21}$$

where λ_t is calculated from the simulated epidemic trajectory using Equation 2.15.

When data are sparsely sampled, the epidemiological and phylogenetic data can be considered to be independent. Thus, when inferring from both types of data, the overall likelihood is calculated as the product of epidemiological and phylogenetic likelihoods.

2.3.4 Average skyline

The coalescent likelihood is not only useful for calculating the probability of phylogenetic data, it can also be used to estimate the maximum likelihood value of the effective population size Ne, i.e. the classic skyline plot (Pybus et al., 2000). If there is prior information on the offspring distribution and generation time, the skyline plot can be converted to estimates of prevalence over time using Equation 2.15.

It is often useful to calculate the skyline which provides a non-parametric estimate of infection prevalence over time. A popular method for constructing the skyline in BEAST (Drummond et al., 2005), which uses the coalescent to calculate the prior probability of node times given a particular skyline. This requires specification of a prior for the skyline itself, and the choice of this prior might influence the resulting distribution of phylogenies and skyline especially for datasets with low diversity. When the number of substitutions

separating different sequences in a sample is small, the uncertainty in branching times is large, and the posterior distribution of prevalence estimates might be biased by the prior rather than informed by the data.

I took an alternative approach to estimate the skyline whereby I separated the processes of phylogeny and skyline reconstruction. Instead of a coalescent prior on branching times, I placed a uniform distribution on branching times which was an option in MrBayes (Ronquist et al., 2012) but not BEAST. The skyline was inferred after phylogeny reconstruction as detailed below. This approach circumvents the need to define a tree prior that depends on the effective population size, and does not rely on the coalescent to produce a posterior distribution of phylogenies.

MrBayes produces a posterior sample of phylogenies, which can be converted to dated phylogenies using the molecular clock rate estimated for each phylogeny. The classic skyline is estimated for each phylogeny in the posterior sample. The median and 95% highest posterior density (HPD) intervals of N_e estimated from the phylogenies are then calculated at each time point. While this does not capture the full uncertainty in estimating population history from the phylogeny using the coalescent, I assume that the maximum likelihood skyline has much higher probability than other skylines.

Derivations for the skyline are provided below. The skyline plot optimises the likelihood function for each time interval delimited by coalescent events indexed by b. Within each inter-coalescent interval b, there are s_b sampling events. Each interval b is divided into $s_b + 1$ sub-intervals indexed by a delimited by sampling events. Let δ_{ab} be an indicator function that is 1 if sub-interval a starts with a coalescent event, and is 0 otherwise. Let A_{ab} be the number of lineages in sub-interval a of inter-coalescent interval b, and U_{ab} be the length of the sub-interval a of inter-coalescent interval b. Using these definitions, the log-likelihood for a given inter-coalescent interval is given by

$$\log P(\mathrm{Phy}_b|N_{e_b}) = \sum_{a=1}^{s_b+1}[\delta_{ab}\log(\binom{A_{ab}}{2}\frac{1}{N_{e_b}}) - \binom{A_{ab}}{2}\frac{U_{ab}}{N_{e_b}}] \tag{2.22}$$

The effective population size N_e that maximises the log-likelihood of each inter-coalescent

interval b is obtained by differentiating the log-likelihood function (Equation 2.22) with respect to $N\text{e}_b$ and setting the derivative to 0. I am using the log-likelihood instead of the likelihood expression as the derivations are more straightforward, and the maximum log-likelihood estimate is the same as the maximum likelihood estimate of $N\text{e}$. Because time intervals are separated by coalescent events, there is only one coalescent-ended sub-interval within each inter-coalescent interval, and thus $\delta_{ab} = 1$ only when $a = s_b + 1$. Thus the expression for $\hat{N\text{e}}$, the maximum likelihood estimate of $N\text{e}$ is

$$\begin{aligned}
\frac{\partial}{\partial N\text{e}_b}[\log P(\text{Phy}_b|N\text{e}_b)] &= \frac{\partial}{\partial N\text{e}_b}(\log[\binom{A_{(s_b+1)b}}{2}\frac{1}{N\text{e}_b}]) - \frac{\partial}{\partial N\text{e}_b}\sum_{a=1}^{s_b+1}(\binom{A_{ab}}{2}\frac{U_{ab}}{N\text{e}_b}) = 0 \\
0 &= \frac{\partial}{\partial N\text{e}_b}(\log\binom{A_{(s_b+1)b}}{2}) - \frac{\partial}{\partial N\text{e}_b}(\log N\text{e}_b) - \sum_{a=1}^{s_b+1}(\binom{A_{ab}}{2}U_{ab})\frac{\partial}{\partial N\text{e}_b}(\frac{1}{N\text{e}_b}) \\
0 &= 0 - \frac{1}{N\text{e}_b} + \sum_{a=1}^{s_b+1}(\binom{A_{ab}}{2}U_{ab})\frac{1}{N\text{e}_b^2} \\
\hat{N\text{e}_b} &= \sum_{a=1}^{s_b+1}(\binom{A_{ab}}{2}U_{ab})
\end{aligned} \tag{2.23}$$

An R implementation of the average skyline is provided the `SmoothSkylines` function in my R package `EpiGenR` (github.com/lucymli/EpiGenR). More details on the R package is provided later in Section 2.4.7

2.4 Inference framework

2.4.1 Particle MCMC

The aim of the statistical inference framework presented here is to obtain the Bayesian posterior distribution $P(\theta|D) \propto P(D|\theta)P(\theta)$, where $\theta = (\theta_1, ..., \theta_{n_\theta})$ is a vector of n_θ parameters with parameter space $\Theta = (\Theta_1, ..., \Theta_{n_\theta})$. The prior probability $P(\theta)$ is updated with the likelihood of parameters given the data D. For compartmental transmission models, it is not usually possible to analytically solve the likelihood function. To solve this

problem, particle filtering has been implemented within an Markov chain Monte Carlo (MCMC) framework to estimate the likelihood by integrating over stochastic epidemic trajectories (Andrieu et al., 2010). A stochastic model simulation generates an epidemic trajectory $X_{0:n_T}$ from time step 0 to n_T describing the temporal changes in incidence, prevalence, and reproductive number. Initial model conditions are given by X_0. Data comprise incidence time series and phylogenetic data $D_{1:n_T} = \{\text{Epi}_{1:n_T}, \text{Phy}_{1:n_T}\}$.

The overall marginal likelihood is calculated sequentially for each discrete time step indexed by $t = \{1, ..., n_T\}$ (Equation 2.24; Figure 2.4).

$$\begin{aligned} P(D_{1:n_T}|\theta) &= \int P(D_T|X_{0:n_T}, \theta) P(X_{0:n_T}|\theta) dX_{0:n_T} \\ &= \int \prod_{t=1}^{n_T} \Big[P(D_t|X_t, \theta)\Big] P(X_0|\theta) \prod_{t=1}^{n_T} \Big[P(X_t|X_{t-1}, \theta)\Big] dX_{0:n_T} \end{aligned} \tag{2.24}$$

The proposal distribution of the MCMC needs to be tuned to optimally explore parameter space. The standard deviation σ_i of proposal distribution $q(\theta_i^*|\theta_i)$ is adjusted to optimise the acceptance probability of parameter θ_i. The acceptance probability a_i of parameter i is calculated periodically (e.g. every 200 iterations). If $a_i < a_{\text{lower}}$ or $a_i > a_{\text{upper}}$, the standard deviation of the proposal distribution σ_i is reduced or increased, respectively using

$$\sigma_i^* = \sigma_i e^{0.5*(a_\theta - a_{opt})} \tag{2.25}$$

where $a_{opt} = 0.234$ is the optimal acceptance probability for an MCMC chain (Roberts et al., 2001). Samples from MCMC were taken after x number of tuning steps (for the simulation data, $x = 50$).

2.4.2 Incorporation of phylogenetic uncertainty

Using a fixed phylogeny to infer parameters would not cause problems if confidence in the internal node times were high. However, low diversity among pathogen sequences reduces the resolution of the phylogeny and thus increases the uncertainties in parameter estimates. Phylogenetic reconstruction programs such as MrBayes (Ronquist et al., 2012) produce a posterior distribution $P(\text{Phy}|\text{Seq})$ of phylogenies Phy given the sequences Seq. To estimate marginal posterior probability $P(\theta|\text{Seq})$, I can integrate over the phylogenies Phy: $P(\theta|\text{Seq}) = \int_{\text{Phy}} P(\theta|\text{Phy})P(\text{Phy}|\text{Seq})d\text{Phy}$. The MCMC algorithm within MrBayes generates samples from $P(\text{Phy}|\text{Seq})$. Taking a random sample from the posterior distribution $P(\text{Phy}|\text{Seq})$, I can estimate the marginal posterior density $P(\theta|\text{Seq})$ using Equation 2.26.

$$P(\theta|\text{Seq}) = \frac{1}{M}\sum_{m=1}^{M} P(\theta|\text{Phy}^{(m)}). \tag{2.26}$$

Although this approach requires a large amount of computational resources, the ability to parallelise multiple MCMC chains makes this approach faster than methods that estimate phylogeny at the same time as implementing particle filtering.

A potential issue of pooling results for randomly sampled phylogenies is that the likelihoods given some phylogenies are lower than others, whereas in a joint inference framework the posterior phylogenies would agree both with the sequence data and the transmission model. Methods exist to correct for the potentially biased distribution of posterior phylogenies whereby the posterior distribution of phylogenies are sampled using an importance sampling scheme (Meligkotsidou and Fearnhead, 2007). This approach would provide a more accurate way of aggregating results of inference using different phylogenies, but might be hindered by the additional computational costs.

An alternative approach to integrate over phylogenetic uncertainty that I tried was to sample a new phylogeny from the posterior distribution of phylogenies generated using MrBayes. The proposal distribution was empirically constructed by constructing a distance matrix for the set of posterior phylogenies. Each pairwise distance was

calculated as the absolute difference in T_{MRCA}s of the two phylogenies. Applying this approach on simulated data led to no proposed phylogeny being accepted. Because the coalescent likelihood is dependent on the coalescent times and lineages-through-time, distance as calculated by the T_{MRCA} was probably not the best measure of difference between phylogenies. More sophisticated measures to quantify the distance between trees have recently been developed (Kendall and Colijn, 2015). However, a large sample of phylogenies would be required to ensure each phylogeny has a sufficient number of similar phylogenies. Otherwise, phylogeny proposals would never be accepted during MCMC.

One more approach that I tried was to integrate over all phylogenies during particle filtering. Instead of calculating the likelihood of each particle based on a single phylogeny, I calculated the average likelihood across all phylogenies and then averaged across those likelihoods to get an overall marginal likelihood that integrates over both stochasticity and phylogenetic uncertainty. However, this implementation was highly inefficient due to long computation times. Perhaps this approach can work for data sets where the phylogenetic uncertainty is small, and thus a small sample of phylogenies covers the range of possible phylogenies.

2.4.3 Pseudocode

Pseudocode of the inference procedure is provided below. I assume that both epidemiological and phylogenetic data are used.

1. Sample M phylogenies indexed by m from the posterior distribution of a Bayesian phylogenetic reconstruction program $\text{Phy}_{1:n_T}^{(m)}$.

   FOR each phylogeny m in 1 to M

   2 Calculate marginal likelihood $L := P(D|\theta^{init})$ using particle filtering and set $\theta := \theta^{init}$. (See particle filtering algorithm below)

   FOR iteration i in 1 to MCMC iterations

   3 Propose new parameter values $\theta^* := q(\theta^*|\theta)$ where $q(\cdot)$ is the proposal distribution.

4 Calculate the marginal likelihood $L^* = P(D|\theta^*)$ using the particle filtering algorithm below.

5 Calculate acceptance probability of new parameters $p_a = \frac{q(\theta|\theta^*)P(\theta^*)P(D|\theta^*)}{q(\theta^*|\theta)P(\theta)P(D|\theta)}$.

6 Draw a random number $z \sim$Unif$(0,1)$. IF $z < p_a$ THEN $\theta = \theta^*$ and $L := L^*$ ELSE $\theta := \theta$.

END LOOP

8 Remove first 50% of samples as burn-in and sample every x iterations from θ values accepted by MCMC.

9. Concatenate samples from all phylogenies, and calculate the median and 95% highest posterior density intervals.

The particle filtering algorithm used to calculate the marginal likelihood is given below. J is the number of particles, where each particle is associated with an epidemic trajectory $X^{(j)}_{0:n_T}$, and a particle weight $\omega^{(j)}$.

FOR time t in 1 to n_T

FOR particle j in 1 to J

1 Simulate $X^{(j)}_1$ according to the model.

2 Set the weight to the likelihood $\omega^{(j)} := P(D_t|X^{(j)}_t)$.

END LOOP

3 Calculate the mean weight $\bar{\omega}_t := \frac{1}{J}\sum_{j=1}^{J}\omega^{(j)}$.

4 Use a multinomial distribution with probabilities $\Omega^{(j)} = \frac{\omega^{(j)}}{\sum_{i=1}^{J}\omega^{(j)}}$ to resample J particles for the next time step.

END LOOP

5. Calculate the marginal likelihood $L(\theta|D_{1:T}) = \prod_{t=1}^{T}\bar{\omega}_t$

END LOOP

For incidence time series, the likelihood calculation uses the probability mass function of the binomial: $P(\text{Epi}_t|X_t^{(j)}) = \binom{X_t^j}{\text{Epi}_t}\rho^{\text{Epi}_t^{(j)}}(1-\rho)^{X_t^{(j)}-\text{Epi}_t}$, where ρ is the probability of a case being reported. If the simulation time step is smaller than the reporting period for incidence data, then I only calculate the likelihood every x number of simulation time steps, such that $x\Delta t$ is equal to or greater than the reporting period.

2.4.4 Summary statistics

Inference results are presented by calculating the median and 95% highest posterior density (HPD) interval of the posterior sample of parameter estimates, which is the smallest interval that captures 95% of parameter values in the posterior.

Because MCMC produces autocorrelated samples from the posterior, I used the effective sample size (ESS) to ensure that a chain has converged and sufficient numbers of samples have been generated to approximate the posterior density.

During particle filtering, an effective number of particles N_{eff} can be calculated after each resampling step to determine the how representative the particles are of the posterior distribution of epidemic trajectories. This is calculated using

$$N_{\text{eff}} = \frac{1}{\sum_{j=1}^{J}(\Omega^{(j)})^2}. \tag{2.27}$$

N_{eff} is small when only a few particles have significantly large weight. The number of particles should be chosen such that N_{eff} remains large enough during all steps of particle filtering. More discussion on choosing the number of particles is presented in Section 2.4.5.

2.4.5 Number of particles and resampling frequency

As with other Monte Carlo approaches, the accuracy of the marginal likelihood estimate increases with the number of particles. The marginal likelihood calculated using particle

filtering is unstable if not enough particles are used. This leads to low acceptance rates for parameters and the MCMC chain getting 'stuck' at certain parameter values.

To test how many particles are needed, I repeatedly estimated the likelihood of a set of parameter values using particle filtering (Figure 2.5). With 1,000 particles, the marginal likelihood can vary by 17.9 log likelihood units even though the same model, parameter values, and data are used. Increasing the number of particles to 10,000 reduced the range of the marginal likelihood estimates to 1.8 log likelihood units. When this was repeated for another set of parameter values with much lower likelihood, even 10,000 was not enough to reliably estimate the marginal likelihood (Figure 2.5A).

The number of particles required for stable estimation of the marginal likelihood depends on a number of factors, but in particular on the length of the time series. : for time series with 25, 50 and 100 time points, 75, 200 and 500 particles, respectively, are need to achieve 30% acceptance rate Andrieu et al. (2010).

Another factor that affects the number of particles is how far are the current parameter values from the maximum likelihood parameter values. Repeating particle filtering at parameter values with lower likelihood, I found that the marginal likelihood is unstable even when 10,000 particles are used (Figure 2.5B).

The consequence of unreliable estimates of the marginal likelihood is imprecision and possible bias in the posterior distribution of parameters. Estimating only the reproductive number, the posterior distribution estimated using 100 particles was not only broader but centred around a larger R_0 compared to the posterior estimates obtained using 5,000 particles (Figure 2.6). The posterior distribution of R_0 estimated using 1,000 particles was similar to that obtained using 5,000 particles, suggesting that 1,000 would be sufficient if only R_0 was estimated. More particles woudl scale with the number of parameters estimated.

When only 100 particles were used, simulations with lower R_0 were less likely to be captured in the sample of particles than when 5,000 particles were used. Although the particle filter should provide unbiased estimate of the marginal likelihood (Andrieu et al., 2010), epidemic simulations from a single infected individual are likely to die out within the first few generations and therefore particle depletion is an issue when the number of

particles is small relative to the extinction risk. Consequently, the marginal likelihood is only estimable when R_0 is high due to the reduced risk of extinction. The number of particles should therefore reflect the risk of particle depletion during particle filtering.

Although the particle filter produces an unbiased estimate of marginal likelihood, it is very computationally intensive. The number of particles required scales with the length of simulations, the number of transitions in the model, and the reporting probability of cases. As I simulated from the index case, the epidemic trajectories at the beginning of simulations were highly unpredictable. Datasets with overdispersed offspring distribution further increased the stochasticity of simulations, necessitating a large number of particles to obtain a stable estimate of the marginal likelihood. In our implementation, I needed 10,000 particles for $k = 0.1$ and at least 1,000 for $k = 1$. For simpler models, approximations such as the Kalman filter can be used. The strength of PMCMC, however, is the applicability to a wide range of models including high-dimensional ones (Sheinson et al., 2014).

By default, resampling takes place at set intervals, e.g. every 10 simulation time steps. Because resampling adds to the computation time of particle filtering, resampling can be limited to when the N_{eff} drops below a threshold. The N_{eff} is inversely correlated with the variance in particle weights, and calculated as $(\sum_{j=1}^{J}(\Omega^{(j)})^2)^{-1}$ for time step t. However, if the N_{eff} regularly drops below the threshold, then the gain in computational time might be insignificant.

The importance of resampling frequently is demonstrated in Figure 2.7. Before the epidemic simulation begins, there are 0 infected individuals and 0 reported cases in the data. Thus all particle weights are equal to 1. After the epidemic simulation begins but before the first data point, epidemics might take off in some particles but die out in others. At the time of the first reported case, the filtering step removes all particles with extinct epidemics, and resamples with replacement from the epidemics that have taken off. If there were not enough particles, there is a possibility that all particles have 0 weight ($N_{\text{eff}} = 0$). If only a few particles having significant weights, then the N_{eff} is very small and the resulting marginal likelihood might be unreliable.

Later in the time-series, the selection pressure on particles decreases as the simulations followed more deterministic paths. Although not implemented in this thesis, a particle

filtering scheme that adaptively changes the number of particles would be useful. For example, a large number of particles is usually needed at the beginning of an outbreak time series due to large amounts of stochasticity, but this number can be reduced once a threshold number of infected individuals is surpassed.

2.4.6 Performance

The time per MCMC iteration depends on the length of the time series data, the number of particles and the number of random number draws per simulation time step. For a simulated dataset with around 130 time steps, it took 0.77 seconds per MCMC iteration on a Linux cluster with 20 cores (Imperial College High Performance Computing Service, 2016) using 20,000 particles and both incidence time series and pathogen phylogeny for inference.

2.4.7 EpiGenMCMC program and EpiGenR package

Although there are existing software and packages to conduct MCMC, PMCMC, and phylodynamic inference, none was sufficient to perform the statistical inference described in this Chapter. The popular `R` package `pomp` (King et al., 2016a,b) provides an extensive set of inference methods including PMCMC for fitting state space models such as stochastic compartmental transmission models to time series data. However the package is not suitable for fitting to phylogenetic data, and it is not easily parallelisable to make use of multiple cores on high performance computing clusters. SSM (https://github.com/sballesteros/ssm) is a more efficient implementation of PMCMC that is a standalone program, and it has been used in the recent Ebola outbreak to estimate the real-time effective reproductive number. Again, this program does not make use of phylogenetic data in inference.

On the other hand, the main software used for parameter inference from genetic data are BEAST (Drummond and Rambaut, 2007) and BEAST2 (Bouckaert et al., 2014) packages. These packages have made phylodynamic inference more accessible by providing a graphical user interface. As of now, nevertheless, neither BEAST nor BEAST2 has

implemented a particle filter in estimating the likelihood. While the birth-death model implemented in BEAST2 enables inference of the reproductive number using stochastic SIR models, the model does not account for heterogeneous transmission and is currently limited to a small set of epidemiological models.

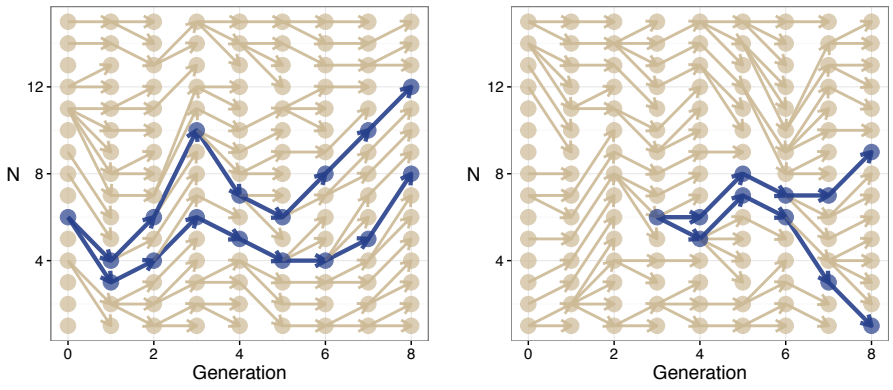
To use the PMCMC algorithm with both epidemiological and phylogenetic data, and to enable estimation with a large number of particles, I wrote my own implementation of the algorithm in C++. The code is available on Github: github.com/lucymli/EpiGenMCMC. Parallelisation of the particle filter was achieved through OpenMP. Particle simulations were split between cores (20 on Imperial College HPC).

I used thread-specific random number engines from the GSL library to ensure thread safety, which was important to ensure that particle simulations were independent. The Standard Library random number generators were not guaranteed to ensure thread safety, causing some particle simulations to be correlated.

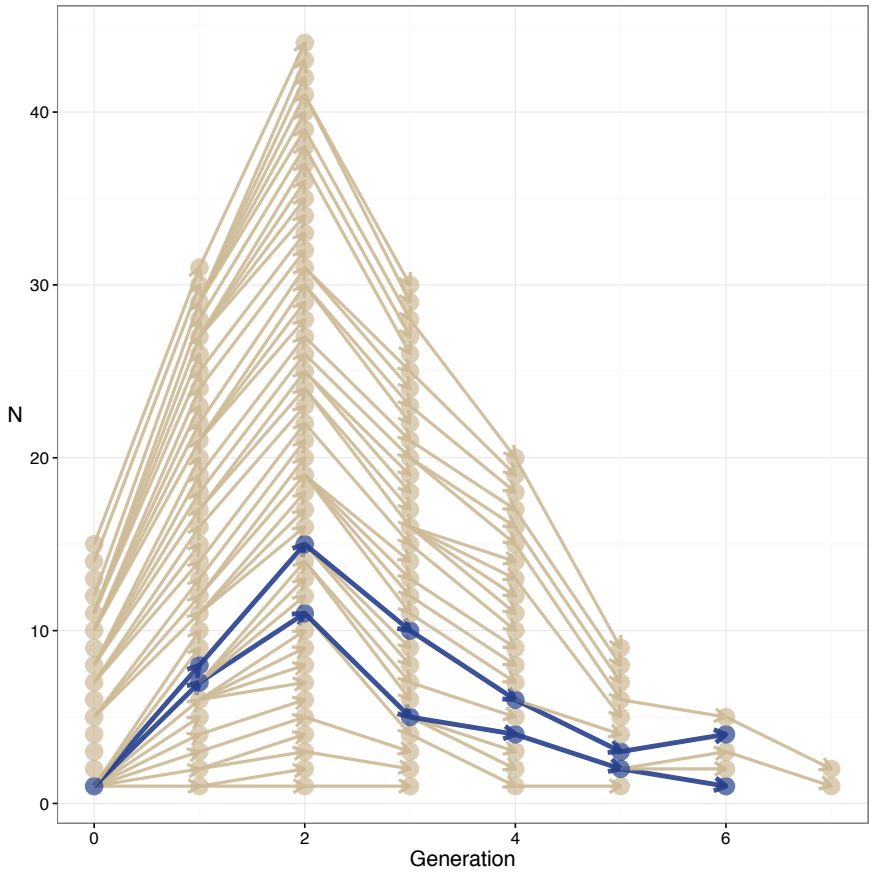
While the C++ program provides at least 2 orders of magnitude speed-up compared to an R implementation, parsing data especially phylogenetic data is not straightforward. I therefore developed a package to interface with the EpiGenMCMC program, with functions to parse phylogenetic data and line lists and generate input files necessary to run the C++ program. The R package can be downloaded using `devtools::install_github("lucymli/EpiGenR")`.

2.5 Summary

I developed a statistical inference framework to solve the problem of inference from both epidemiological and genetic data while incorporating phylogenetic uncertainty and estimating overdispersion of the offspring distribution. This implementation is parallelised and can be used on high performance computing clusters to reduce computation time. The implementation of Fraser and Li's (2017) formulation of coalescent likelihood enables the estimation of dispersion parameter k of the offspring distribution by model fitting. Such an analysis is presented in the next Chapter.



(A) Constant N, $k = \infty$ (B) Constant N, $k = 1$



(C) Time-varying N, $k = 1$

Figure 2.2: Expected time to coalescence for two sampled lineages depends on the underlying population dynamics and offspring distribution. A For Wright-Fisher populations, the expected time to coalescence is inversely proportional to the population size N. B When the offspring distribution follows a negative binomial with dispersion $k = 1$ with variance σ^2 instead of a Poisson distribution, the expected time to coalescence is proportional to $\frac{\sigma^2}{N}$. C When the population size varies over time, the expected time to coalescence also changes over time and is proportional to $\frac{R_r(1+\frac{1}{k})}{N_r}$, where R_r and N_r are the mean reproductive number and population size at generation r. Time is defined in discrete generations indexed by r.

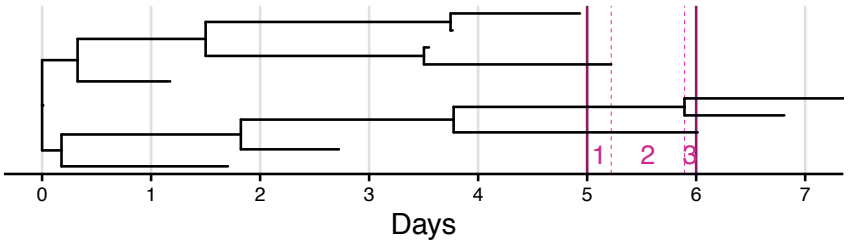


Figure 2.3: An example of using Equation 2.15 to calculate the coalescent likelihood. The overall likelihood is based on the product of likelihoods for intervals (delimited by solid vertical lines). And the likelihood for each interval is calculated as the product of the likelihood for each sub-interval. Sub-intervals are separated by coalescence, sampling event, or simulation time step (dotted vertical lines). Assume that the prevalence is $N_{t=5}$ and $N_{t=6}$ at time steps 5 and 6 (purple solid lines). The coalescent rate λ_t is calculated using Equation 2.15. The number of lineages is tracked by the variable $A_n = \{A_1, A_2, A_3\}$. During sub-interval 1, the probability of two lineages out of three coalescing after U_1 is $p_1 = \binom{A_1}{2}\lambda_t e^{-\binom{A_1}{2}\lambda_t U_1}$. The probability of no coalescent events during sub-interval 2 is $p_2 = e^{-\binom{A_2}{2}\lambda_t U_2}$, and during sub-interval 3 is $p_3 = e^{-\binom{A_3}{2}\lambda_t U_3}$. The overall likelihood for this interval is thus $p = p_1 p_2 p_3$. The time components of variables A_{nt} and U_{nt} are dropped for simplicity.

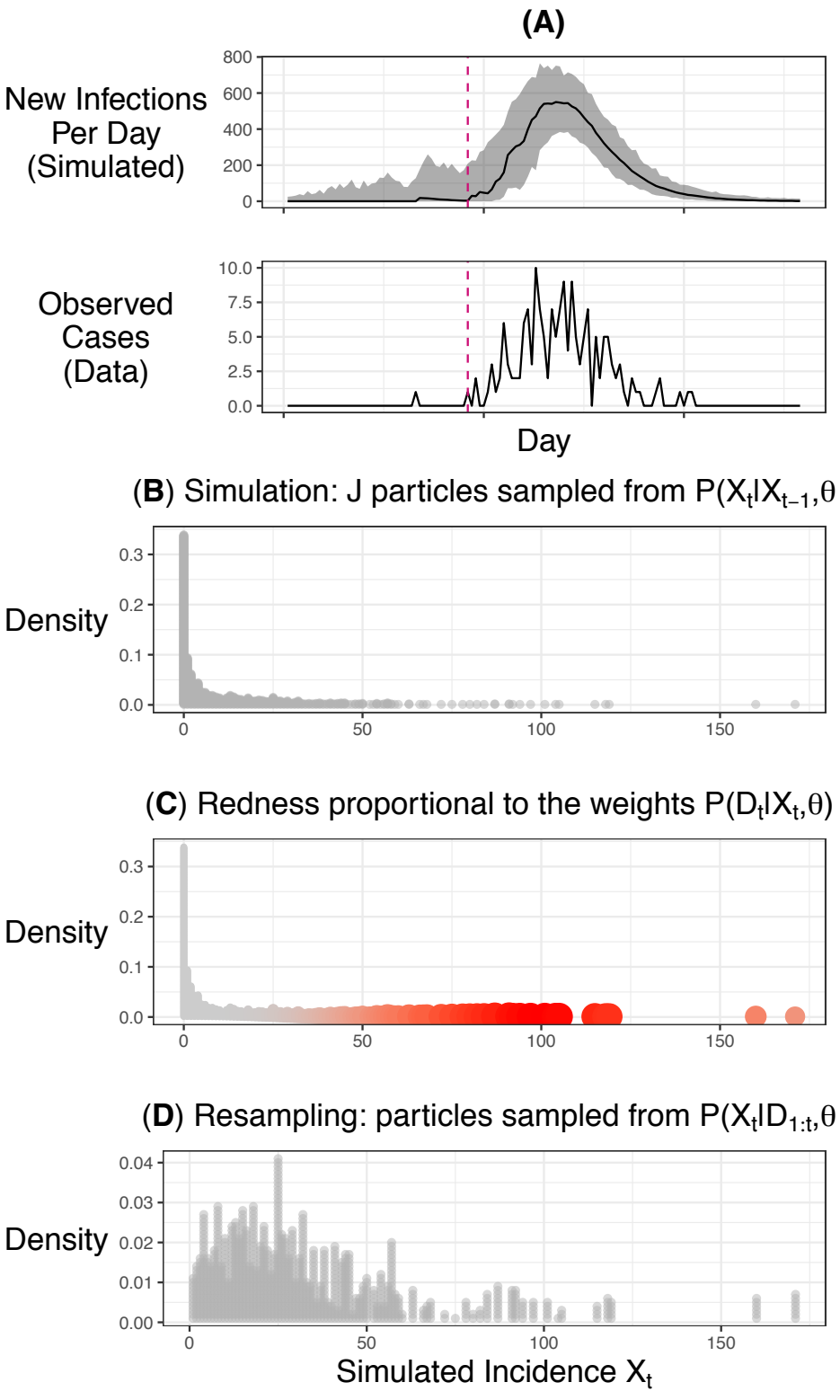
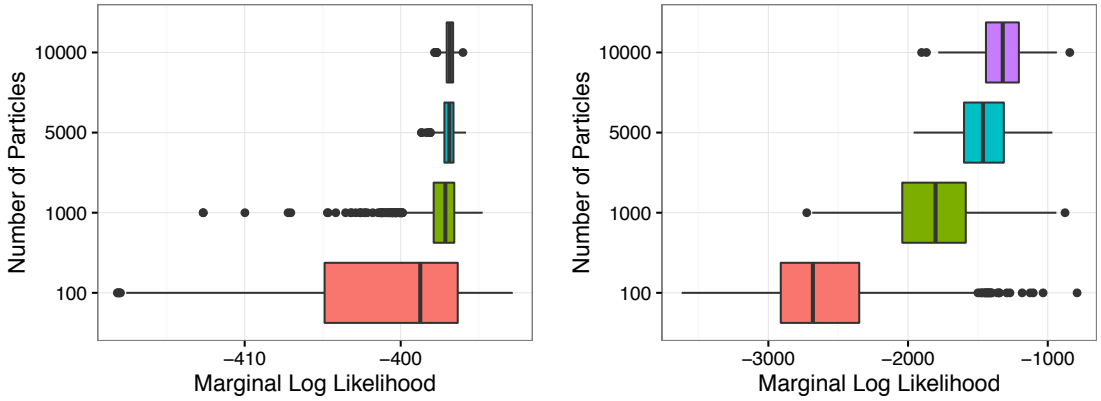


Figure 2.4: (A) The median and range of simulated epidemic trajectories during particle filtering (PF). (B)-(D) show the steps that occur during 1 iteration of PF. (B) J epidemics (particles) are simulated. The frequency distribution of the simulated X_t is proportional to the probability density $P(X_t|X_{t-1}, \theta)$. (C) The weight of each simulated epidemic (particle) is calculated according to the likelihood $P(D_t|X_t, \theta)$. (D) Particles are resampled with replacement according to multinomial distribution where probabilities are the normalized particle weights. Further details of the PF implementation are given as pseudocode and discussed in more detail in Section 2.4.



(A) High likelihood (B) Low likelihood

Figure 2.5: Increasing the number of particles increases the precision of the marginal likelihood calculated during particle filtering. Using the SEIR model that will be applied to polio data in Chapter 4, I set the parameter values to maximum likelihood values presented in (Blake et al., 2014). Particle filtering was repeated 1,000 times and the distribution of marginal likelihoods is shown here. With 1,000 replicates, the marginal likelihood calculated for the same set of parameter values varied across a range of 17.91 log likelihood units. This range reduced to 2.86 and 1.82 for 5,000 and 10,000 particles, respectively. Changing R_c to 4, the marginal likelihood estimates are lower and highly variable even if 10,000 particles are used.

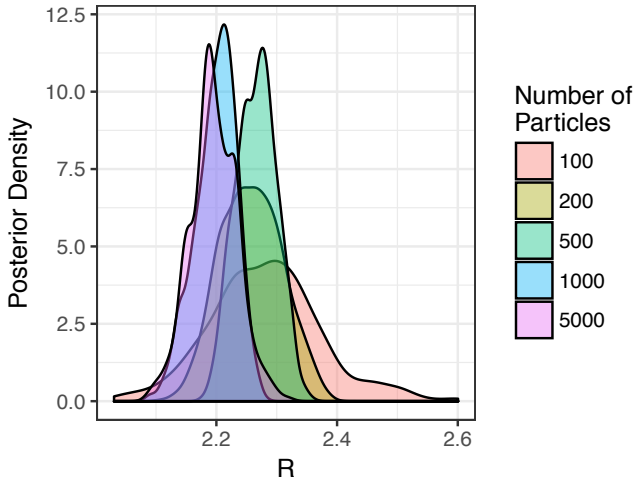


Figure 2.6: The posterior distributions of R_0 estimated using 100, 200, 500, 1,000, and 5,000 particles. All other parameters were fixed.

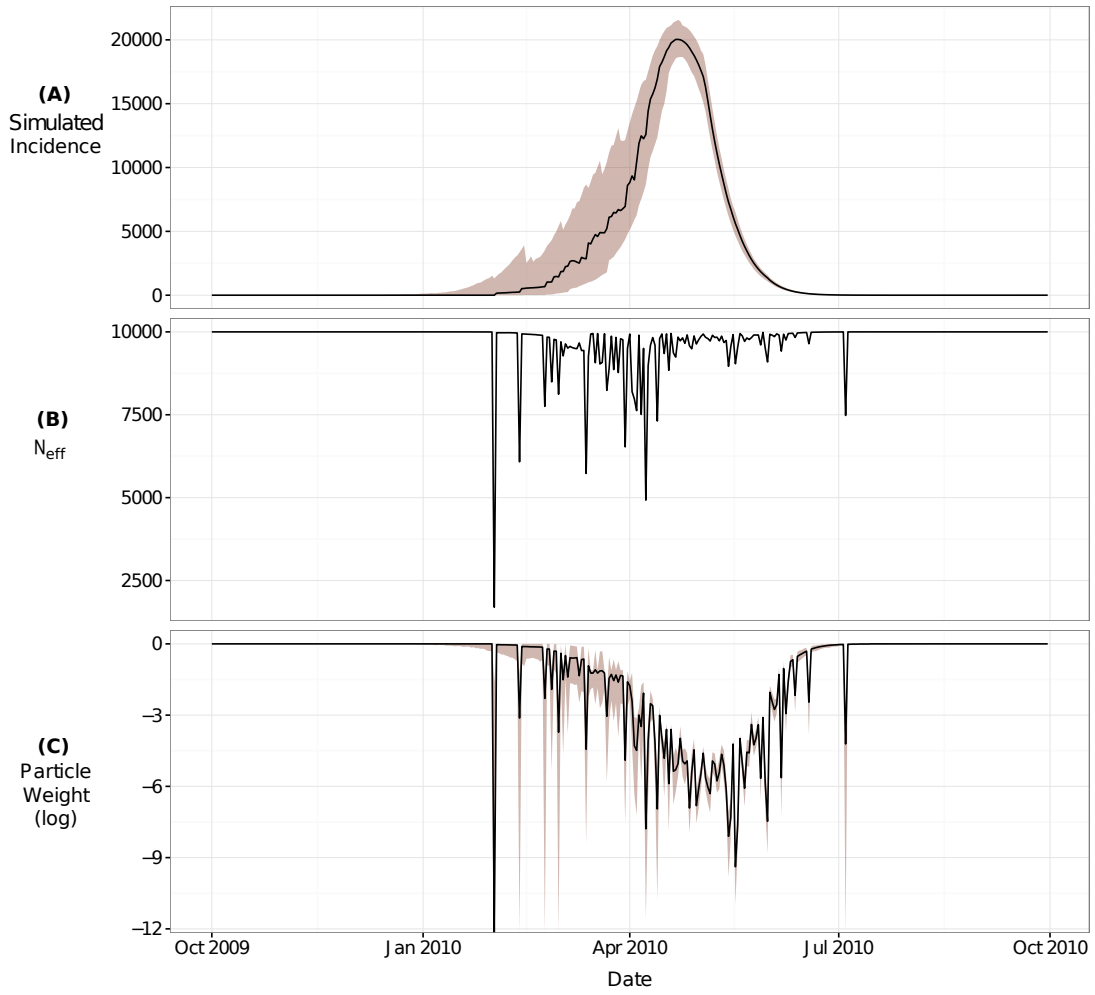


Figure 2.7: Effective sample size of particles during particle filtering. From top to bottom: **(A)** The median and range of simulated incidence across particles. **(B)** The effective sample size of particles N_{eff} after every 10 simulation step, calculated as $(\sum_{j=1}^{J}(W_t^j)^2)^{-1}$ for time step t. **(C)** The distribution of particle weights after each simulation step. An individual can become an incident case after an incubation period with a mean of 16.5 days. The probability of an incident case being reported is 1 in 200.

Chapter 3

Integrated analysis of phylogenetic and epidemiological data to estimate k

As discussed in Chapter 1, superspreading is a common phenomenon in infectious disease epidemiology. The offspring distribution captures the variation in the number of secondary infections per infected individual. Small values of the dispersion parameter k are indicative of superspreading. In this chapter, I use simulated data to demonstrate how k can be estimated from epidemiological and phylogenetic data, and explore the effects of phylogenetic uncertainty on parameter estimates.

This chapter is part of a manuscript that I submitted to Molecular Biology and Evolution titled "Quantifying transmission heterogeneity using both pathogen phylogenies and incidence time series".

3.1 Introduction

The intensity of epidemics is often summarised by the reproductive number R, the average number of secondary infections caused by a typical infectious individual over the course of their infectious period. This statistic is useful for determining whether an

epidemic can take off and if so the final size of the epidemic. However, large variation between individuals is frequently observed in outbreaks of directly transmitted acute infections leading to superspreading events such that a few individuals cause most of the infections (Lloyd-Smith et al., 2005). The offspring distribution captures the distribution of secondary infections per infectious individual and can be parameterised by a negative binomial with mean R and dispersion k. Small values of k, which lead to superspreading events, can affect the effectiveness of control strategies due to the presence of superspreaders (Garske and Rhodes, 2008).

Inferring the value of k from data is not straightforward, even in the presence of contact tracing data as many infections may be asymptomatic or not reported. The offspring distribution fit to incomplete transmission chain data has to be corrected for biased and under-reporting (International Ebola Response Team et al., 2016). Obtaining precise estimates of k from just incidence time series is usually not possible because k only affects the noisiness of the incidence time series at low numbers.

Besides epidemiological data in the form of incidence time series, pathogen population genetics are playing an increasingly important role in inferring epidemiological parameters (Volz et al., 2009; Koelle and Rasmussen, 2012; Volz, 2012; Stadler et al., 2013; Kühnert et al., 2014). For coalescent-based approaches, the offspring distribution is integral to the inference process as it affects the relationship between the underlying epidemic and the observed distribution of coalescent (branching) events in the pathogen phylogeny. When the offspring distribution is overdispersed, shorter intervals between coalescent times in the pathogen phylogeny are observed. This is expected as coalescent events correspond to transmission events during the epidemic and superspreaders can cause the aggregation of many transmission events within a short period of time.

Given that epidemiological parameters could be estimated either from epidemiological data or from phylogenetic data, combining the analysis of both types of data should provide more accurate and precise estimates. Rasmussen et al. (2011) found that estimating parameters jointly from both incidence time series and pathogen phylogeny reduced uncertainties in estimates of parameters and the prevalence over time. However, this work did not allow for uncertainty in the pathogen phylogeny and was limited to simple SIR models. While Rasmussen et al. (2014b) showed that parameter estimates

did not significantly change for phylogenies of sequences collected over many years, the uncertainty in pathogen phylogeny during outbreaks is generally greater and needs to be accounted for to ensure accurate estimation of transmission parameters.

Using the statistical inference framework developed in this thesis (Chapter 2), I fit a stochastic compartmental model with an explicit offspring distribution to estimate k and other epidemiological parameters from outbreak data. I simulated outbreaks to assess the accuracy, precision, and bias of parameter estimates, and in the presence of phylogenetic uncertainty.

3.2 Method

3.2.1 Details of simulations

I simulated data sets according to a Susceptible-Infected-Removed (SIR) model under 6 combinations of basic reproductive number $R_0 = \{2, 5\}$ and $k = \{0.1, 1, 10\}$. R_0 values of 2 and 5 are reasonable for directly transmitted viral infections (Ferguson et al., 2005; Fraser et al., 2004). Although they do not capture the full range of possible R_0 values for directly transmitted infectious diseases, I was more interested in the impact that different values of k has on parameter estimation. The values of k were selected because they capture the plausible range of values for k. The lowest of k estimated for a range of infectious diseases was 0.16 for SARS (Lloyd-Smith et al., 2005). Because the implicit assumption of classic SIR models is that the offspring distribution follows the geometric, the scenario of $k = 1$ is considered. The variance of the offspring distribution does not change significantly for $k > 10$, and thus I did not consider values of $k > 10$.

Assuming a constant population size $N = 20,000$, the S_t and I_t track the number of susceptible and infectious individuals at time t, while the number of recovered individuals is given by $N - S_t - I_t$. The duration of infection is exponentially distributed with rate $\gamma = 0.2$ day^{-1}. The reproductive number at time t was calculated as $R_t = R_0 \frac{S_t}{N}$. More details about the simulated datasets are found in Figures 3.2A and Figure 3.2B.

Epidemics were simulated with time steps of $\Delta t = 0.1$ days, starting with 1 infected

individual $I_0 = 1$ and $S_0 = 19,999$. At each time point t, the number of people recovering was drawn from a binomial distribution $Q_t^{I\to R} \sim Bin(I_t, \gamma dt)$. Another random number was drawn from the negative binomial to determine the number of new infections $Q_t^{S\to I} \sim NBin(R_t Q_t^{I\to R}, kQ_t^{I\to R})$, which translated to a variance of $Q_t^{I\to R} R_t(1 + \frac{R_t}{k})$. This meant that all secondary infections caused by an individual occurred at the end of the infectious period. This approximation was not an issue for these simulations due to the short infectious period; it would not be suitable for analysis of chronic infections.

All parameter values used to generate the parameters, as well as prior distributions used during inference, are listed in Table 3.1.

Table 3.1: Parameters of the SIR model used to generate the simulated outbreak data. The prior distribution of the epidemic start date T_0 is bounded by the time of the first reported case or the time of root in the phylogeny, whichever is earlier.

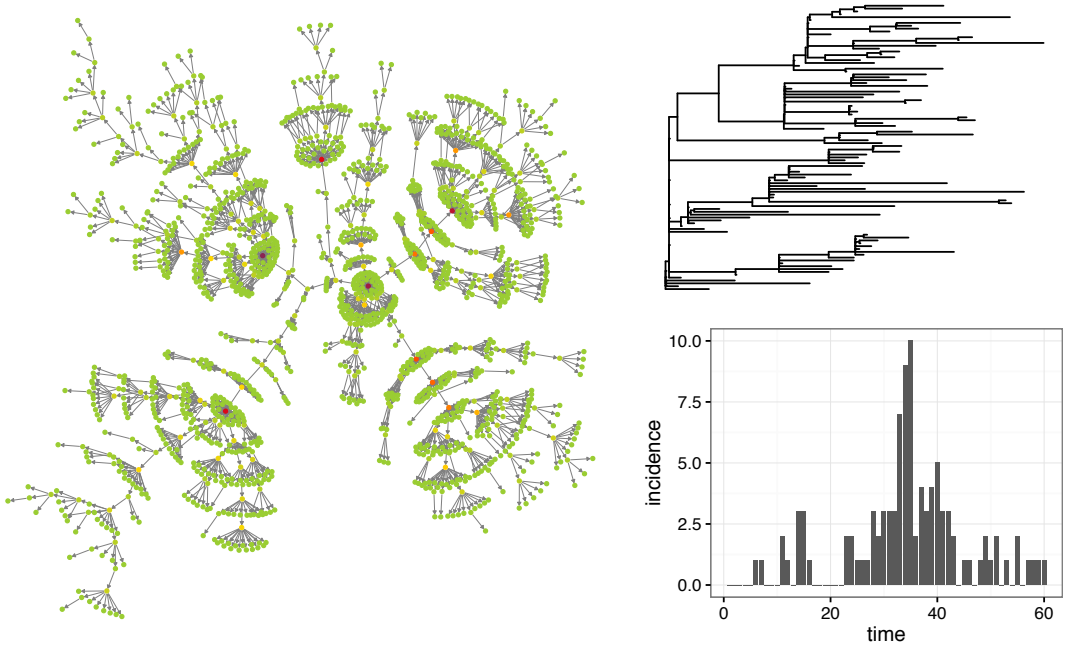
Parameter	**Value**	**Prior**
Population N_{Total}	20,000	-
Initial number of infected I_0	1	-
Duration of infection (days) $\frac{1}{\gamma}$	-	Uniform(3,7)
Basic reproductive number R_0	-	Uniform(1,20)
Offspring distribution dispersion parameter k	-	$\frac{1}{k} \sim$Uniform$(1 \times 10^{-4}, 1 \times 10^{4})$
Reporting probability ρ	-	Uniform(0.0,1.0)
Time of first infection T_0	-	Uniform(01-Jan-16,$\cdot$)

Simulations continued until the epidemic died out. I only kept the simulations with a final epidemic size of at least 10.

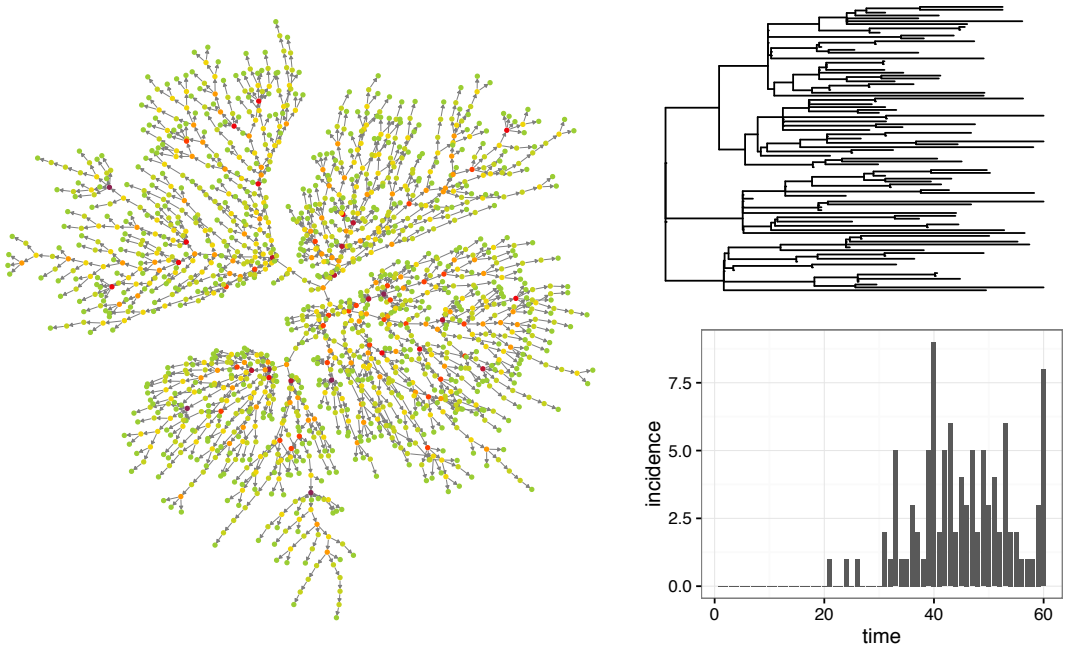
To generate the observation data I randomly sampled individuals with probability $\rho = \{0.01, 0.1\}$ at the time of recovery. As I tracked who-infected-whom, I reconstructed the dated phylogeny for the sampled individuals. The branching points corresponded to transmission times, and tip dates corresponded to sampling times, i.e. the times of recovery.

3.2.2 Details of MCMC

At the start of an MCMC chain, the initial parameter values were set to their true parameter values in the case of simulated data to reduce convergence time. I simulated



(A) $k = 0.1$



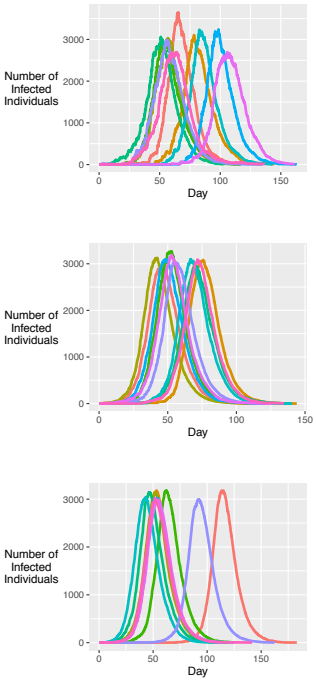
(B) $k = 10$

Figure 3.1: Pipeline to generate simulated data. A comparison is made between an example outbreak simulated with $k = 0.1$ (A) and with $k = 10$ (B). For each set of data, an outbreak is simulated according to a modified SIR model (see Section 3.2.1). The transmission network shown in the top panel of each figure. A proportion of infected individuals are sampled when they recover. The genealogy of the sampled individuals is constructed based on the transmission tree and represented as a dated phylogeny (bottom left plot of each figure). The sampled individuals are also aggregated on a daily basis to produce a time series of daily incidence (bottom right plot). The node colours in the transmission tree scale with the number of secondary infections, with the most infectious individuals coloured red, and the least infectious coloured yellow.

R0	k	Days to Epidemic Peak	Total Infected	Sample Size
2	0.1	65	16,284	163
		78	15,817	155
		57	15,534	136
		51	15,423	149
		51	16,354	151
		83	15,759	1586
		99	15,747	1544
		56	16,009	1559
		106	15,915	1577
		64	16,015	1587

R0	k	Days to Epidemic Peak	Total Infected	Sample Size
2	1	46	15,805	166
		76	15,998	145
		42	15,903	152
		53	16,243	139
		70	15,778	171
		66	15,917	1542
		48	16,176	1631
		55	15,808	1577
		53	16,045	1589
		72	15,745	1611

R0	k	Days to Epidemic Peak	Total Infected	Sample Size
2	10	115	15,970	164
		53	15,983	175
		52	15,907	155
		62	15,940	145
		46	16,072	159
		43	15,874	1593
		54	15,969	1633
		92	15,916	1576
		54	16,020	1548
		52	15,873	1551

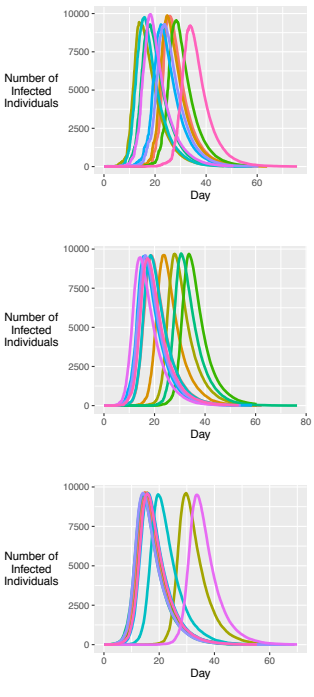


(A) $R_0 = 2$

R0	k	Days to Epidemic Peak	Total Infected	Sample Size
5	0.1	26	19,906	210
		25	19,862	186
		14	19,867	187
		28	19,862	199
		18	19,862	172
		16	19,892	2051
		22	19,863	1971
		24	19,830	1987
		18	19,877	2036
		34	19,849	2084

R0	k	Days to Epidemic Peak	Total Infected	Sample Size
5	1	19	19,867	204
		24	19,856	191
		28	19,862	211
		34	19,853	204
		30	19,845	179
		18	19,885	1958
		16	19,858	1964
		16	19,861	1996
		14	19,860	1972
		17	19,851	1952

R0	k	Days to Epidemic Peak	Total Infected	Sample Size
5	10	15	19,856	201
		16	19,857	190
		30	19,848	204
		16	19,860	205
		14	19,873	194
		20	19,860	1981
		16	19,856	1976
		14	19,868	1961
		34	19,867	2033
		16	19,854	1966



(B) $R_0 = 5$

Figure 3.2: Details of simulation data analysed in this chapter.

an additional 100 data sets from the stochastic SIR model with true parameter values $R_0 = 2$, $k = 1$, and $\rho = 1\%$, but starting the MCMC chains at parameter values far from their true values and using a heated chain at the beginning of the MCMC (multiplying p_a by a factor).

For the simulated data, up to 500,000 MCMC iterations were carried out, sampling parameter values every 100 iterations. Convergence was determined by calculating the effective sample size after removing the first 50% of samples as burn-in. Samples with an effective sample size less than 200 were removed from the final result plot.

At each iteration of the MCMC, I used a Gaussian distribution $q(\theta_i^*|\theta_i)$ to generate a new parameter value θ_i^* centred around the old parameter value θ_i, where $i = \{1, ..., n_\theta\}$ and n_θ is the total number of parameters to be estimated. For k, I estimated its reciprocal $\frac{1}{k}$ so the proposal distributions were centred around $\frac{1}{k}$ instead.

The standard deviation σ_i of proposal distribution $q(\theta_i^*|\theta_i)$ was adjusted to optimise the acceptance probability of parameter θ_i. During the first 20,000 proposals of a parameter θ_i, the acceptance probability of the parameter a_i was calculated every 200 proposals. If $a_i < 0.15$ or $a_i > 0.75$, the standard deviation of the proposal distribution σ_i was adjusted using Equation 2.25.

3.2.3 Incorporating phylogenetic uncertainty

For one of the outbreaks simulated using $R_0 = 2$ and $k = 0.1$ and sampled with 1% probability, I simulated sequence evolution down the sampled phylogeny using seq-gen (Rambaut and Grassly, 1997). In addition to the sampled sequences, I also simulated the sequence evolution of an outgroup so that the tree could be rooted. Each sequence was 1000 nucleotides in length with equal equilibrium frequencies of A, C, T, and G. I used the JC69+Γ model of substitution (Jukes and Cantor, 1969) with a rate of substitution of 0.15 per site per year. This value was at least 10 times higher than reported rates for viral evolution. Polioviruses, for example, evolve at a rate of ~ 0.01 substitutions per site per year (Jorba et al., 2008). I used an artificially high value to ensure that sufficient divergence between sequences sampled during the outbreak so that sufficiently resolved phylogenies could be reconstructed. There was still uncertainty around branching times

despite the high rate of substitution.

After obtaining the sequence data, I used MrBayes (Ronquist et al., 2012) to estimate the phylogeny assuming a strict molecular clock and a JC69+Γ model, which included a single rate parameter and gamma-distributed heterogeneity in rates among sites. I used the outgroup sequence to root the phylogenies, and the tip sampling dates to estimate the rate of nucleotide substitution. From the resulting posterior distribution, I sampled 100 dated phylogenies. I divided the branch lengths of phylogenies measured in substitutions per site by the estimated molecular clock rates to obtain dated phylogenies. For each dated phylogeny, I re-estimated the epidemiological parameters. An overall posterior distribution was obtained by concatenating samples of parameter values obtained for each phylogeny.

3.2.4 Assessing accuracy of estimates

I assessed the accuracy, bias and precision of parameter estimates. The accuracy was determined by the percentage of simulations for which the true parameter value was within the 95% HPD interval of estimates. Bias was the distance between the median parameter estimate and the true parameter value. Finally, the precision was determined by the root mean squared deviation (RMSD) using the formula $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\theta_i - \hat{\theta})^2)}$, where θ_i for $i = 1, .., n$ were the n values of parameter θ sampled from the posterior distribution of the parameter and $\hat{\theta}$ was the true parameter value.

3.3 Results

3.3.1 Overview of analyses

I tested the PMCMC inference framework on simulated data first to determine the accuracy of parameter estimates, assess the value of phylogenetic data in epidemiological inference, and to demonstrate the importance of estimating k. For the simulation study, I generated 60 simulated data sets using a stochastic SIR model under various combinations of R_0, k and reporting probability ρ (see Figures 3.2A and 3.2B). Each

data set comprised a phylogeny and an incidence time series for a sample of infected individuals. The phylogeny is a dated phylogeny representing the genealogy of the sampled individuals. For each data set, I performed 3 sets of inference: using incidence time series; using phylogenetic data; or using both. The following parameters were concurrently estimated: R_0, T_g, k, T_0, and ρ (only when incidence time series were used during inference). The results are shown in Figure 3.3 and Table 3.2.

To assess the consequences of not estimating k, I re-estimated all other parameter values except k for 30 of the simulated data sets (those with ρ=1% sampling) while fixing $k = 1$. Again, I conducted statistical inference 3 times using one or both sets of data. The posterior estimates are shown in Figure 3.4.

For one of the simulated data sets ($R_0 = 2, k = 0.1$), I simulated the evolution of pathogen sequences down the true phylogeny to obtain a sample of pathogen sequences. Using MrBayes (Ronquist et al., 2012) to reconstruct the phylogeny from the pathogen sequences, I obtained a posterior distribution of phylogenies given the simulated pathogen sequences. I then sampled 100 phylogenies from this posterior distribution and re-estimated the parameter values using each sampled phylogeny. The posterior estimates using all phylogenies are shown in Figure 3.5.

For all the parameter estimation above, I set the initial parameter values to be very close to the true parameter values (those used to simulate the data). To test that I can obtain the true parameter values in the absence of prior information, I generated 100 extra simulated data sets using $R_0 = 2, k = 1$ and $\rho = 1\%$. For each of these, the initial parameter values were randomly sampled from the prior distributions for the parameters. The precision, bias, and accuracy of estimates are presented in Table 3.4.

3.3.2 Phylogenetic data are more informative than epidemiological time series for estimating k given an accurately constructed phylogeny

Based on data simulated from an SIR model, pathogen phylogeny was needed to accurately estimate the dispersion parameter k of the offspring distribution when k was

Table 3.2: Precision (RMSD), bias, and accuracy (true value found in HPD) of parameter estimates when fitting models to either epidemiologic, genetic, or both types of data. The estimates of epidemic start dates T_0 were converted to the number of days after an arbitrary date. These statistics were evaluated for each set of simulated sequences and then averaged across all simulations that used both sets of data, only epidemiological data, or only phylogenetic data. For bias and precision, I normalised the statistics by the true parameter value.

Data	R_0**:RMSD**	R_0**:Bias**	R_0**: in HPD**	**k:RMSD**	**k:Bias**	**k: in HPD**
Both	0.15	-0.08	100%	13.70	-0.03	66.7%
Epi	0.17	-0.04	100%	673.94	93.43	100.0%
Phy	0.19	-0.12	100%	19.78	-0.19	100.0%
Data	T_0**:RMSD**	T_0**:Bias**	T_0**: in HPD**	ρ**:RMSD**	ρ**:Bias**	ρ**: in HPD**
Both	11.52	0.27	100%	0.106	-0.065	62.5
Epi	11.50	0.39	100%	0.127	-0.100	83.3
Phy	12.54	-1.35	100%	NA	NA	NA

small (Figure 3.3). This suggested that superspreading events left sufficient signal to allow inference of k in the phylogeny but not the incidence time series. There was insufficient signal in the epidemiological time series to determine the value of k. Although using both epidemiological and phylogenetic data produced the least biased and most precise estimates of k, the accuracy was lower than using each set of data alone. In fact, using phylogenetic data alone seemed to produce the most accurate estimates that were only slightly less precise and slightly more biased (Table 3.2).

I also estimated the basic reproductive number R_0, time of the first infection T_0, and probability of sampling an infectious individual in the incidence time series ρ. There were no noticeable differences between estimates of these parameters when only epidemiologic, only phylogenetic, or both data sets were used for inference (Figures 3.3 B-D). I did not estimate ρ when just using the phylogenetic data, as the reporting rate ρ referred to the probability that an infection appeared in the incidence time series.

Estimates of the R_0 and k were closer to the true value when both genetic and epidemiological data were used in inference, compared to fitting to each set of data individually (Table 3.2). The accuracy of estimates for T_0 and ρ while fitting to both sets data was similar to fitting to epidemiological or phylogenetic data alone.

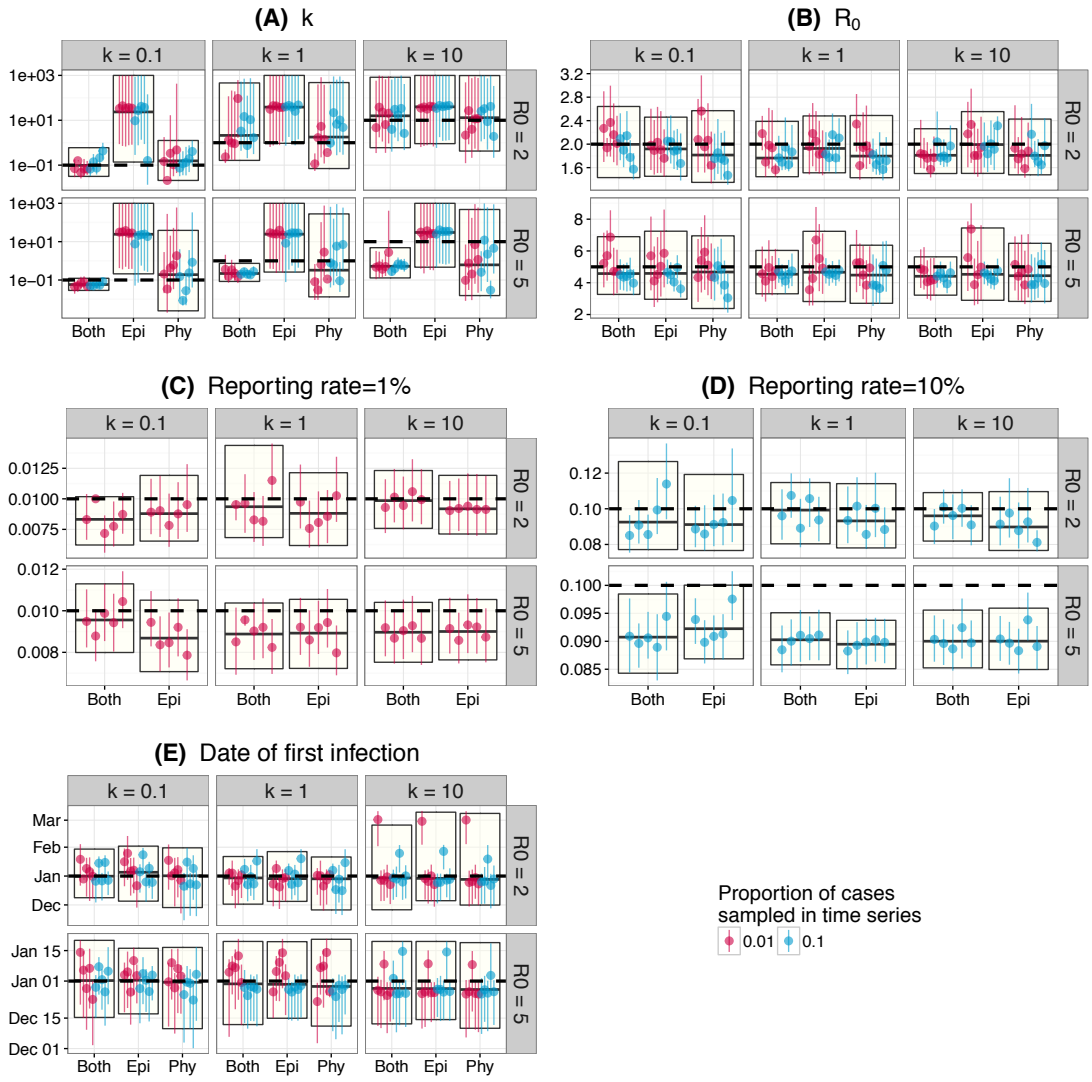


Figure 3.3: Parameter estimates from simulated data for **(A)** k, **(B)** R_0, **(C-D)** reporting rate when simulated data were sampled at 1% and 10%, and **(E)** date of first infection. Within each of the 6 panels of each subplot, the results of inference from both epidemiological and phylogenetic data (left), only epidemiological data (middle), and only phylogenetic data (right) are shown. The horizontal lines denote the true parameter value for that set of parameters, i.e. the parameter value used to simulate the data. The boxes with a horizontal line in the middle indicate the median and 95% HPD interval of parameter estimates pooled from all simulations for that parameter set. The vertical lines with a single red dot denote the median and 95% HPD interval of each individual simulation. I did not estimate the reporting rate when inferring just from phylogenetic data, as the reporting rate referred to the probability that an infection appeared in the incidence time series. A uniform prior distribution was used for all parameters, with bounds described in Table 3.1

3.3.3 Estimates of R_0 were biased if k was fixed at the incorrect value

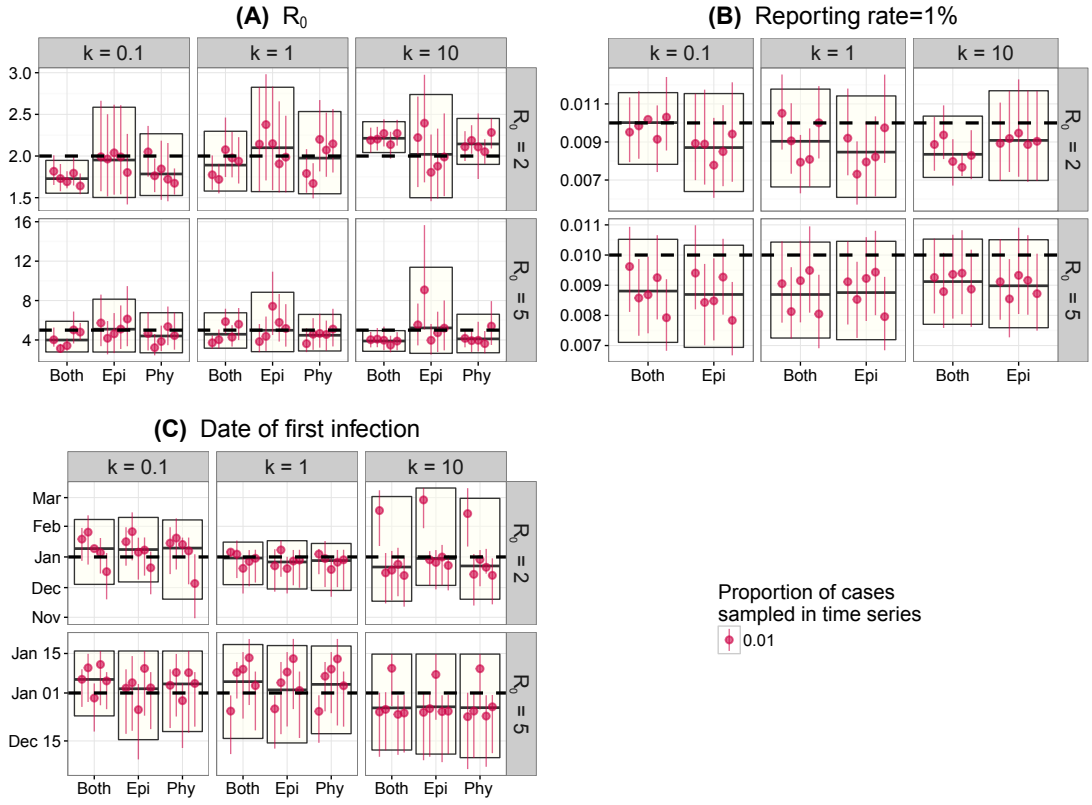


Figure 3.4: Parameter estimates when reporting rate was 1 in 100 and k was fixed to 1. Estimates are shown for for **(A)** R_0, **(B)** reporting rate, and **(C)** date of first infection. Within each of the 6 panels of each subplot, the results of inference from both epidemiological and phylogenetic data (left), only epidemiological data (middle), and only phylogenetic data (right) are shown. The horizontal dashed lines denote the true parameter value for that set of parameters, i.e. the parameter value used to simulate the data. The boxes indicate the median and 95% HPD interval of parameter estimates pooled from replicate simulations. The vertical lines with a single dot denote the median and 95% HPD interval of each individual simulation. Simulations where the MCMC chain did not converge were left out of the plot. Estimates of the reporting rate did not include inference from phylogenetic data, as the reporting rate refers to the probability that an infection appears in the epidemiological time series.

The dispersion parameter k of the offspring distribution is usually not estimated when fitting compartmental models. I investigated the effects of assuming the wrong value of k on parameter estimates, especially on R_0 estimates. The implicit assumption of compartmental transmission models is that the offspring distribution is geometrically distributed, which was equivalent to fixing $k = 1$ in the negative binomial. For a subset of

Table 3.3: The Kolmogorov-Smirnov (K-S) distance between the posterior distributions of R_0 estimated assuming a geometric offspring distribution (i.e. fixing $k = 1$) and those estimated while estimating k (see results in Table 3.2 and Figure 3.3). K-S values closer to 1 reflect larger discrepancies between the posterior distributions, whereas those close to 0 suggest no difference in posterior distributions. The numbers in the brackets denote the range of K-S distances from different sets of simulated data, and the number preceding the brackets denotes the median K-S distance.

k	**Both**	**Epi**	**Phy**
0.1	0.725 (0.078, 0.956)	0.000 (0.000, 0.125)	0.201 (0.000, 0.887)
1	0.325 (0.111, 0.979)	0.151 (0.078, 0.247)	0.422 (0.142, 0.777)
10	0.646 (0.135, 0.977)	0.066 (0, 0.101)	0.395 (0.067, 0.954)

simulated outbreaks (those where I sampled 1% of individuals), I re-estimated parameters with a fixed $k = 1$.

I compared these results (Figure 3.4) to those obtained when k was also estimated (Figure 3.3), and found significant differences in R_0 estimates when the true value of $k \neq 1$. This was evidenced by an increase in the Kolmogorov-Smirnov distances (Massey Jr, 1951), a measure of distance between two distributions, when the true value of $k \neq 1$ compared to when the true value of $k = 1$ (Table 3.3). However, this was only the case for inference using phylogenetic data. Those using just epidemiological time series were unaffected by assumptions of k.

Estimates of the reporting rate and the epidemic start date were not affected by assumptions of k, regardless of the data used during inference.

3.3.4 Estimation from multiple phylogenies

For a subset of data (those generated using $R_0 = 2$ and $k = 0.1$), I re-estimated the parameters for each phylogeny inferred from the simulated sequences (Figure 3.5). Estimates of R_0 and k obtained from inferred phylogenies instead of the true phylogeny reduced precision and increased bias, although estimates of k were still more precise than those estimated from epidemiological data (Table 3.2). Interestingly, estimates of the epidemic start date were less biased when using inferred phylogenies than the true phylogeny, although this might be due to the sample phylogeny randomly having a tree height further away from the epidemic start date.

3.3.5 Changing initial parameter values

At the start of an MCMC chain, the initial parameter values were set to their true parameter values in the case of simulated data to reduce convergence time. I simulated an additional 100 data sets from the stochastic SIR model with true parameter values $R_0 = 2$, $k = 1$, and $\rho = 1\%$, but starting the MCMC chains at parameter values far from their true values. Using a heated chain at the beginning of the MCMC (multiplying p_a by a factor), I found that the MCMC chain converged on the same posterior distributions as when the initial parameter values were close to the true parameter values (Table 3.4).

Table 3.4: The precision (RMSD), bias and accuracy (% in HPD) of parameter estimates when the initial parameter values were very far from the true parameter values, averaged across 100 simulations for the parameter combination $R_0 = 2$, $k = 0.1$, and $\rho = 1\%$. The RMSD and bias values presented here have not been normalised. The MCMC chain was initially heated to accept jumps to parameter sets with low posterior densities in order to escape local optima.

Data	R_0 : **RMSD**	R_0 : **Bias**	R_0 : **in HPD**	**k: RMSD**	**k: Bias**	**k: in HPD**
Both	0.3028	0.0702	100%	0.08970	-0.03504	100%
Epi	0.4018	0.0678	100%	0.09206	-0.07422	100%
Phy	0.3574	0.1344	100%	0.08810	-0.05097	100%
Data	T_0 : **RMSD**	T_0 : **Bias**	T_0 : **in HPD**	ρ**: RMSD**	ρ**: Bias**	ρ**: in HPD**
Both	12.9879	1.5898	100%	0.001551	-0.001024	100%
Epi	12.6207	-0.8961	100%	0.001795	-0.001174	100%
Phy	14.2418	-1.0382	100%	NA	NA	NA

3.4 Discussion

Building on methods that enable parameter inference for stochastic models and phylodynamic approaches integrating both epidemiological and phylogenetic data, I presented a framework for quantifying the offspring distribution dispersion k while inferring key epidemiological parameters from both types of data. The addition of pathogen phylogeny to epidemiological inference was necessary to accurately estimate the dispersion of the offspring distribution k. This would be useful for detecting superspreading dynamics in infectious disease outbreaks where data from densely sampled transmission networks are not available.

Existing approaches to epidemiological inference from pathogen phylogeny do not usually

account for overdispersion in the offspring distribution (Volz et al., 2009). While the variance of the offspring distribution can be increased by dividing the population into a limited number of infectious categories (Volz and Pond, 2014), the number of secondary infections per individual lies on a continuum in real populations. Also, discretisation of infectiousness requires a structured coalescent approach whereas estimating the offspring distribution parameters assumes homogeneous mixing.

The use of a single representative phylogeny to infer epidemiological parameters is sufficient for well-resolved phylogenies. Rasmussen et al. (2014b) found that parameters of an HIV transmission model were broadly consistent amongst 10 phylogenies sampled from Bayesian phylogeny reconstruction in BEAST. As these sequences were collected over a number of years, there was sufficient confidence in the branching times that different phylogenies sampled from the posterior distribution produced similar estimates. In an outbreak setting when transmission happens over a short time compared with viral evolution, greater uncertainty in branching times meant that I needed to use a larger number of phylogenies to be confident of the posterior distribution of parameters.

A potential source of bias in the k estimates is from the strong prior I placed on the value of k such that highly super-spreading scenarios would have greater prior density than more homogeneous settings. Random draws from the prior distribution I used would yield a mean of 0.00142 and a median of 0.0002, with interquartile bounds of 0.00013 and 0.0004. A more suitable prior would have been a uniform prior on k between 0 and 10 as increases in k above 10 would not affect the clustering of cases that much. However, despite the strong prior density, the posterior densities of k did not conform to the prior, which suggests that there was sufficient signal in the data to identify the most likely values of k.

The conclusion that the method presented in this chapter can be used to accurately estimate k needs to be caveated by the fact that this relies on accurate phylogenetic reconstruction. Despite the high mutation rate used to simulate the sequence data in Section 3.3.4, the uncertainty in k was much greater than when the true phylogeny was used for inference. The mutation rate was at least an order of magnitude faster than rates reported for viruses with short generation times (Duffy et al., 2008). The choice for the high mutation rate was because the simulated epidemics were over a short period

of time (around 100 days). The polio outbreak analysed in Chapter 4 lasted almost a year, for example. Sufficient numbers of mutations are necessary to resolve the tree. However, despite the high rate of nucleotide substitution in this case, there was still significant uncertainty in the branching times. This could be due to the phylogenetic reconstruction programs forcing bifurcations in the tree, even if the true phylogeny contained multifurcations due to superspreading events.

Further studies are required to more extensively test the inference method in the presence of phylogenetic uncertainty, perhaps with simulations lasting a longer period of time, and in a larger population but with the same number of sequence samples. The latter could improve the resolution of the phylogeny as multifurcations in the true phylogeny would be much less likely. Additionally, this needs to be repeated for a greater number of epidemiological parameter combinations rather than just one, which was the case in this chapter.

While the simulation parameters and model used in this chapter are different from those used to analyse poliovirus data in the next two chapters, they were chosen for their simplicity so that the accuracy of the inference framework could be assessed. These simulations shown here are comparable to other directly transmitted infections such as pandemic influenza the early stages of pandemic influenza where $R_0 = 1.5$ and $T_g = 1.91$ were estimated (Fraser et al., 2009). The scenarios with higher R_0 values are less likely to occur in emerging outbreaks, but illustrate the potential use of this inference framework in settings with higher transmission intensities.

Although the particle filter produces an unbiased estimate of marginal likelihood, it is very computationally intensive. The number of particles required scales with the length of simulations, the number of transitions in the model, and the reporting probability of cases. As I simulated from the index case, the epidemic trajectories at the beginning of simulations were highly unpredictable. Datasets with overdispersed offspring distribution further increased the stochasticity of simulations, necessitating a large number of particles to obtain a stable estimate of the marginal likelihood. In our implementation, I needed 10,000 particles for $k = 0.1$ and at least 1,000 for $k = 1$. For simpler models, approximations such as the Kalman filter can be used. The strength of PMCMC, however, is the applicability to a wide range of models including

high-dimensional ones (Sheinson et al., 2014).

An issue that I did not explore was the effect of sampling strategy on phylodynamic inference. In the analysis of simulated data, I assumed that a uniform sampling strategy in which every infected individual was equally likely to appear in the data. However, in an epidemic setting reporting rates may change over time due to many factors. During the Tajikistan poliovirus outbreak, for example, there were proportionally fewer pathogen isolates that were sequenced during the peak of the epidemic compared to early and late phases. On the other hand, surveillance systems may improve after an outbreak is declared, resulting in denser sampling. Coalescent analyses are particularly sensitive to the sampling strategy. Non-random sampling in which individuals from epidemiologically linked clusters were more likely to be sampled resulted in under-estimate of the effective population size (de Silva et al., 2012a).

Beyond state space models, PMCMC can be applied to a wider class of epidemic models as long as they can be simulated to produce trajectories of the number of infectious individuals over time, and that the mean and variance of the offspring distribution can be defined. Branching process models are often used to analyse data from the start of an epidemic as they assume no density-dependent effects. Although they do not satisfy the Markovian requirement, their likelihood can also be estimated using a PMCMC approach, as long as the timings of future 'offspring' are tracked in each particle.

The simulations presented in this Chapter demonstrate that the inference framework described in Chapter 2 can account for demographic stochasticity to quantify heterogeneity between individuals at the same time as estimating other epidemiological parameters. Including uncertainty in the phylogeny reduced confidence in estimated k and further work is necessary to explore the effects of phylogenetic uncertainty on the accuracy of epidemiological parameters. This inference method could be applied to rapidly evolving viral infections including polio (see Chapters 4 and 5).

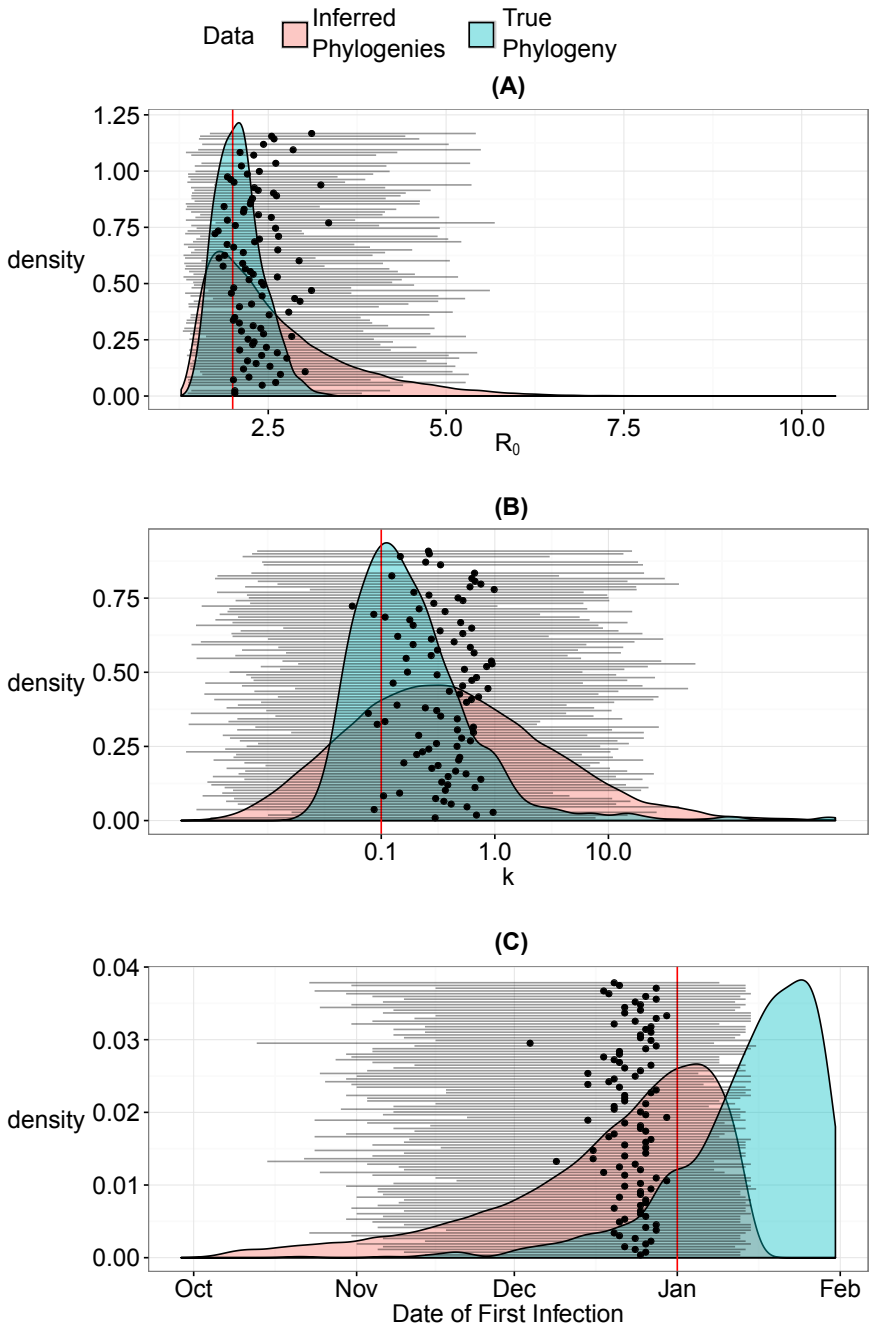


Figure 3.5: Parameter inference by integrating over multiple phylogenies. For one of the simulated data sets ($R_0 = 2, k = 1$), I simulated the evolution of a 1000-nucleotide gene using the HKY+Γ substitution model. Using MrBayes to reconstruct the phylogeny from the sequences of the sampled individuals, I obtained a posterior distribution of phylogenies given the sequences. Sampling 100 trees from this posterior distribution, I re-estimated the parameters using both incidence time series and pathogen phylogeny for each of the sampled phylogenies. The horizontal lines show the median and 95% HPD intervals of parameter estimates for each of the 100 trees. The red vertical lines denote the true parameter values. The red distributions are the posterior distributions obtained by pooling the parameter estimates from all 100 analyses, and the blue distributions are the posterior distributions obtained using the true genealogy of sampled individuals.

Chapter 4

Phylodynamic analysis of the 2010 Tajikistan poliovirus outbreak

I showed that the inference framework described in Chapter 3 can be used to recover the true parameter values of simulated outbreaks by fitting a compartmental transmission model to one or both of epidemiological data and pathogen phylogeny. In this Chapter, I will apply the inference framework to real data from a poliovirus outbreak.

This chapter is part of the manuscript that I submitted to Molecular Biology and Evolution titled "Quantifying transmission heterogeneity using both pathogen phylogenies and incidence time series".

4.1 Introduction

Despite the elimination of poliovirus from most of the world, poliovirus outbreaks can still occur as long as poliovirus remains endemic in a few countries. Gaps in immunity can lead to potentially large outbreaks of polio in non-endemic countries. In 2010, for example, importation of wild poliovirus type 1 (WPV1) occurred in 10 countries that were previously polio-free (Centers for Disease Control and Prevention, 2010).

The largest outbreak in 2010 occurred in Tajikistan, which resulted in 518 cases of poliomyelitis (Centers for Disease Control and Prevention, 2010). The build-up of a

large susceptible population combined with delayed response to the outbreak led to this large outbreak. Fitting a compartmental transmission model to incidence time series aggregated from acute flaccid paralysis (AFP) surveillance data, Blake et al. (2014) estimated the reproductive numbers for people in different age groups and the number of infections prevented by supplementary immunisation activity (SIA) rounds. The epidemic history was not co-estimated along side parameter values. Although simulating from the maximum likelihood parameter values helped to quantify stochasticity in the system, uncertainties in prevalence over time were not available.

Another difficulty in Blake et al. (2014) and other polio modelling papers was the need to assume a case-to-infection ratio to estimate epidemiological parameters. Polio cases are identified through AFP surveillance, but because most infections are asymptomatic, the case-to-infection ratio (i.e. reporting probability) is very low. Although a 1 in 200 case-to-infection ratio is often assumed for modelling purposes (Melnick and Ledinko, 1953; Grassly et al., 2006; Blake et al., 2014), this ratio can vary temporally, spatially, by age, and by serotype. Accurate estimation of the case-to-infection ratio is needed to quantify the extent of poliovirus spread. In an outbreak setting, this can help to estimate the true size of the infected population, and thus plan intervention strategies accordingly. Furthermore, in the long term, the case-to-infection ratio is needed to determine the length of time that needs to pass after the last reported case of polio before eradication can be declared (Eichner and Dietz, 1996; Kalkowska et al., 2015).

Using the statistical inference approach presented in Chapter 2, I aimed to

1. estimate case-to-infection ratio, with and without pathogen phylogenies,

2. compare parameter estimates obtained through maximum likelihood and those obtained using PMCMC,

3. and determine if sufficient information can be derived from pathogen phylogenies during an outbreak to estimate epidemiological parameters.

4.2 Method

4.2.1 Data

According to WHO guidelines, two stool samples should be taken within two weeks of an individual being diagnosed with acute flaccid paralysis (AFP). Stool specimens are tested for wild and vaccine-derived polioviruses, and viral isolates are sequences. On occasion, stool collection does not take place within 2 weeks of paralysis, or the patient is lost to follow-up. In such cases, the patient is designated a polio-compatible case if their symptoms are compatible with that of poliovirus.

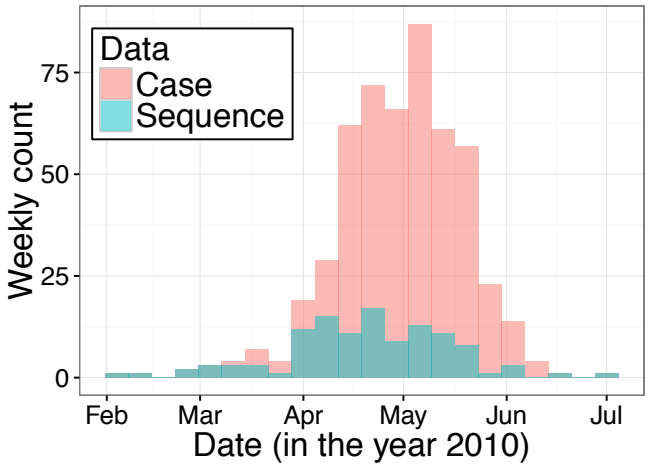
Data used for the analysis in this Chapter were collected during the 2010 outbreak of WPV1 in Tajikistan, which resulted in 463 laboratory-confirmed and 58 polio-compatible cases between February and July 2010, inclusive (Centers for Disease Control and Prevention, 2010). I constructed the incidence time series by aggregating cases on a daily basis according to their dates of paralysis and age group. This data set was divided into three age groups: 0-5, 6-14, and 15+ years, which corresponded to the target age groups of supplementary immunisation activities (Centers for Disease Control and Prevention, 2010).

A total of 116 Virion Protein 1 (VP1) sequences were obtained from the stool samples in Tajikistan with Genbank accession numbers KC880365-KC880521 (Yakovenko et al., 2014). Each sequence was associated with the date of collection, but age groups were ignored.

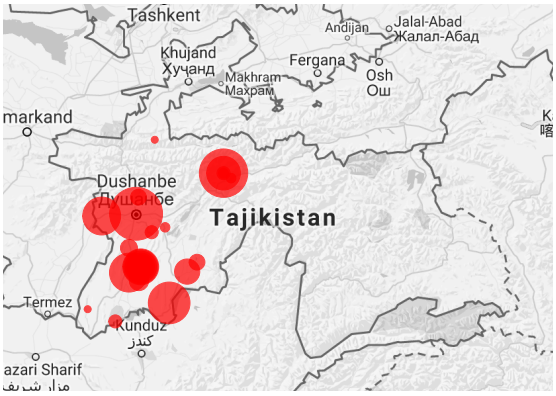
Not all isolates confirmed to be WPV1 were sequenced, nevertheless, due to lack of resources to process all samples (Figure 4.1). Changes in the sampling rate of sequences do not have an effect on inference from phylogeny data because the coalescent likelihood does not depend on the sampling rate.

4.2.2 Phylogenetics

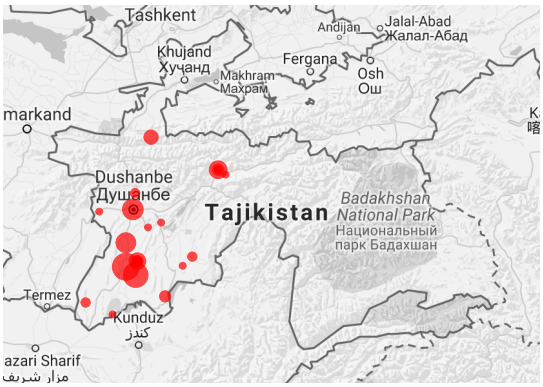
The K80+Γ model (Kimura, 1980) was selected as the substitution model as it returned the lowest Bayesian Information Criterion (BIC) score in jModelTest2 (Guindon and



(A) Temporal distribution of data



(B) Cases



(C) Sequences

Figure 4.1: Temporal and spatial distribution of the 2010 Tajikistan outbreak of wild poliovirus type 1 (WPV1). (A) Weekly incident case numbers (red bars) of WPV1 during the 2010 Tajikistan outbreak, and the number of cases for which a virus sequence was obtained (blue bars). The sampling dates are the dates of symptoms onset. Note that the bars are not stacked but superimposed. The spatial distribution of data is shown in (B) and (C). The sizes of circles are proportional to (B) the number of cases in a district (largest is 73), and (C) the number of cases for which sequences were obtained (largest is 19).

Gascuel, 2003; Darriba et al., 2012). This model of substitution included two rates for transitions and for transversions, and assumed equal base frequencies and gamma-distributed heterogeneity in rates among sites. Phylogenies were rooted using an outgroup sequence sampled in Uttar Pradesh, India in 2009 (GenBank: KC800662)

A maximum likelihood tree was constructed in RAxML (Stamatakis, 2014) using the GTR + Γ model. The K80 model was not available in RAxML and the GTR +Γ model had the lowest BIC score out of all available nucleotide substitution models in the program. A

root-to-tip plot was created from the resulting maximum likelihood tree by plotting the distance (number of substitutions) from the root node to each tip against the sampling times. A strong correlation as measured by the R^2 shows that the sequences evolve according to a strict molecular clock.

The posterior distribution of phylogenies was estimated using MrBayes (Ronquist et al., 2012), assuming a K80+Γ model of substitution and a strict molecular clock. I used 5 million generations with 1 cold and 3 heated chains and sampled every 1,000 generations. The first 50% of samples were removed as burn-in. The tip dates were fixed to the dates of sampling. As with the simulated data, dated phylogenies were obtained by dividing the branch lengths of reconstructed phylogenies by the molecular clock rate. I sampled 100 dated phylogenies from the MrBayes posterior and estimated parameter values based on each phylogeny.

An averaged skyline was created from the sampled phylogenies from MrBayes using the pipeline described in Section 2.3.4. To compare this to Bayesian skyline plots, I also analysed the sequences using BEAST (Drummond and Rambaut, 2007). I carried out 2 independent MCMC runs with 100 million generations, sampling every 5,000 generations, and with a 50% burn-in. After checking that the effective sample size was greater than 200 for all parameters in both chains, the chains were merged.

4.2.3 Statistical inference

I fit to the polio data a modified SEIR model similar to that used in Blake et al. (2014) but with an explicit offspring distribution. Let the transitions between states indexed by age group i be $Q_{t,i} = \{Q_{t,i}^{S_i \to E_i}, Q_{t,i}^{E_i \to I_i}, Q_{t,i}^{I_i \to R_i}, Q_{t,i}^{S_i \to R_i}\}$, where each transition corresponds respectively to a new infection, an infected person becoming infectious, recovery, and vaccination. The transitions are drawn from the following probability distributions

$$Q_{t,i}^{E_i \to I_i} \sim Bin(E_{t,i}, \gamma_1 dt) \tag{4.1}$$

$$Q_{t,i}^{I_i \to R_i} \sim Bin(I_{t,i}, \gamma_2 dt) \tag{4.2}$$

$$Q_{t,i}^{S_i \to R_i} \sim Bin(S_{t,i}, V_t \upsilon) \tag{4.3}$$

where γ_1 is the rate of becoming infectious, γ_2 is the rate of recovery, and V_t is an indicator function equal to 1 on the following dates and 0 otherwise: 6 May, 20 May, 3 June, 17 June 2010.

To determine the number of infections, I assumed that all secondary infections caused by each individual occurred at the same time as the generation time was short. The reproductive number of each age group at time t was $R_{t,i} = \frac{1}{\gamma_2} \sum_j \beta_{t,ij} S_{t,j}$. All density-dependent transmission rates $\beta_{t,ij}$ where $i \neq 1$ and $j \neq 1$ were assumed to be a proportion of the transmission rate amongst children of the youngest age group: $\beta_{1,1}\beta_p$. The number of transitions from susceptible to exposed individuals was drawn from a negative binomial distribution parameterised by its mean and dispersion parameters

$$Q_{t,i}^{S_i \to E_i} \sim NBin(R_{t,i} Q_{t,i}^{I_i \to R_i}, k Q_{t,i}^{I_i \to R_i}) \tag{4.4}$$

Whereas in the simulations I assumed the time of case reporting coincided with the time of recovery, here I could not make that assumption as symptoms may manifest even if an individual is no longer infectious. The incubation period refers to the time interval between infection ($S_i \to E_i$) and the onset of paralytic symptoms. Instead, I adopted the approach used in Blake et al. (2014) and modelled the incubation period as an Erlang distribution with mean $\xi = 16.5$ days and a shape parameter of $\alpha = 16$, which was equivalent to the sum of α independent exponential variables with rate $\frac{1}{16.5}$ per day. I obtained these values by fitting the Erlang distribution to observed data on incubation periods Casey (1942), and estimating the maximum likelihood values of the distribution (Figure 4.2). In practical terms, this meant having 16 compartments where the progression from one to the next was exponentially distributed. During simulations,

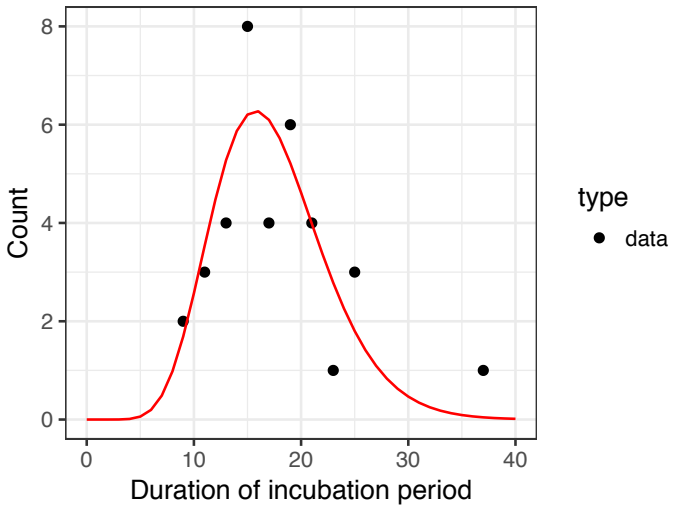


Figure 4.2: The best-fitting Erlang distribution (red line) for the incubation period data (black dots) from Casey (1942). The Erlang distribution with mean $\xi = 16.5$ days and a shape parameter of $\alpha = 16$ produced the best fit as assessed by least-squares.

individuals who became infected ($S_i \rightarrow E_i$) also moved from the S_i to the first incubation period compartment. Their progression in the SEIR compartments was independent of their progression down the 16 incubation period compartments.

The likelihood $P(\text{Phy}_t|X_t)$ is calculated as described in Section 2.3.3, but using $X_t = E_t + I_t$. The SEIR model used in this Chapter is akin to an SIR model with gamma-distributed generation time because I still simulated secondary infections at the end of an individual's infectious period. Because the latent period was very short compared to the infectious period, I approximated the coalescent likelihood with the same formula as above, replacing I_t with $E_t + I_t$ and setting T_g as the duration of infection.

Although I included age-structure in the simulation model, I ignored age-structure in the likelihood calculation based on the phylogeny. This is due to the low sampling probability (less than 1 in 200) of poliovirus sequences. Let $A = A_1 + A_2 + A_3$, where A_i be the number of lineages at time step t in age group i.. The time index t is dropped for simplicity According to derivations of the coalescent for structured SIR models (Volz, 2012), the overall coalescent rate at simulation time step t is

$$\lambda = \sum_i \lambda_i = \sum_i [\beta_{ii} S_i I_i \frac{A_i(A_i - 1)}{I_i^2} + \sum_{j \neq i} \beta_{ij} S_i I_j \frac{A_i A_j}{I_i I_j}] \tag{4.5}$$

where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. Because the low reporting probability of polio infections is less than 1 in 200 and not all reported cases were sequenced in the case of the Tajikistan outbreak, $I_j >> A_j$ and thus $\frac{A_j}{I_j} = \frac{A_i}{I_i} \approx \frac{A_i - 1}{I_i}$. Substituting this into Equation 4.5 yields

$$\lambda = \sum_i [\beta_{ii} S_i I_i \frac{A_i(A_i - 1)}{I_i^2} + \sum_{j \neq i} \beta_{ij} S_i I_j \frac{A_i(A_i - 1)}{I_i I_i}] \tag{4.6}$$

$$= \sum_i \sum_j \beta_{ij} S_i I_j \frac{A_i(A_i - 1)}{I_i^2} \tag{4.7}$$

$$= \beta S I \frac{A(A-1)}{I^2} \tag{4.8}$$

where the transmission rate $\beta S = \sum_i \sum_j \beta_{ij} S_i$. Thus, the overall coalescent rate of an age-structured population is approximately equal to the unstructured coalescent rate, when the probability of an infected individual appearing in the sample is low.

I fixed the initial susceptible population sizes to the maximum likelihood estimates obtained in Blake et al. (2014). I also placed a strong prior on the duration of infectiousness based on likelihood profile obtained in Blake et al. (2014). Fixed parameter values and prior distributions on estimated parameters are outlined in Table 4.1.

I used 10,000 particles and up to 150,000 MCMC iterations sampling every 20 iterations. The Markov chains were terminated earlier than 150,000 iterations if estimates of the marginal posterior density had an ESS of at least 100.

Parameter	Value	Estimated	Prior
Population sizes in thousands $N_{\text{Total},1}, N_{\text{Total},2}, N_{\text{Total},3}$	656, 1249, 372	No	-
Susceptible individuals at start (in thousands) $S_{0,1}, S_{0,2}, S_{0,3}$	109.6, 176.1, 104.2	No	-
Initial number of infected $I_{0,1}, I_{0,2}, I_{0,3}$	1,0,0	No	-
Mean duration of latency $T_l = \frac{1}{\gamma_1}$	4	No	-
Mean duration of infectiousness $T_i = \frac{1}{\gamma_2}$	-	Yes	Gam($\alpha = 5.12, \beta = 1.7$)
Initial reproduction number of children aged 0-5 years R_c	-	Yes* (proposal and prior on β)	Unif(0.00001, 0.1)
Initial reproduction number of people >5 years R_a	-	Yes* (proposal and prior on β_p)	Unif(0.00001, 1.0)
Offspring distribution dispersion parameter k	-	Yes	Unif(0.00001, 1000)
Infection:Case ratio (inverse of reporting fraction) $\frac{1}{\rho}$	-	Yes	Unif(1,1×10^6)
Time of first infection T_0	-	Yes	
Vaccine efficacy υ	-	Yes	Unif(0.0, 1.0)
Mean and shape parameters of the Erlang distributed incubation period ξ, α	16.5 days, 16	No	

Table 4.1: Model parameters of the transmission model fit to Tajikistan polio data. Values of fixed parameters are given in the column 'Value'. The population was divided into 3 age groups: 0-5 years, 6-14 years and 15+ years. The initial numbers of susceptibles were fixed to the maximum likelihood estimates from Blake et al. (2014). Vaccinations took place on the following dates: 06 May, 20 May, 03 Jun, 17 Jun and 17 Jun 2010. On these dates, individuals were moved from the susceptible to the recovered compartment with probability υ. Gamma distributions are parameterised by the shape and scale parameters. *The reproductive numbers R_c and R_a were calculated from the estimated transmission rate amongst young children β, the relative transmission rate between all other groups β_p, the duration of infectiousness $\frac{1}{\gamma_2}$, and numbers of susceptibles S_0.

4.3 Results

4.3.1 Phylogenetics

Although poliovirus is one of the fastest evolving viruses, the short duration of the outbreak combined with the short length of VP1 sequences meant that the uncertainties in branching times were large. Based on the maximum likelihood phylogeny, the rate of

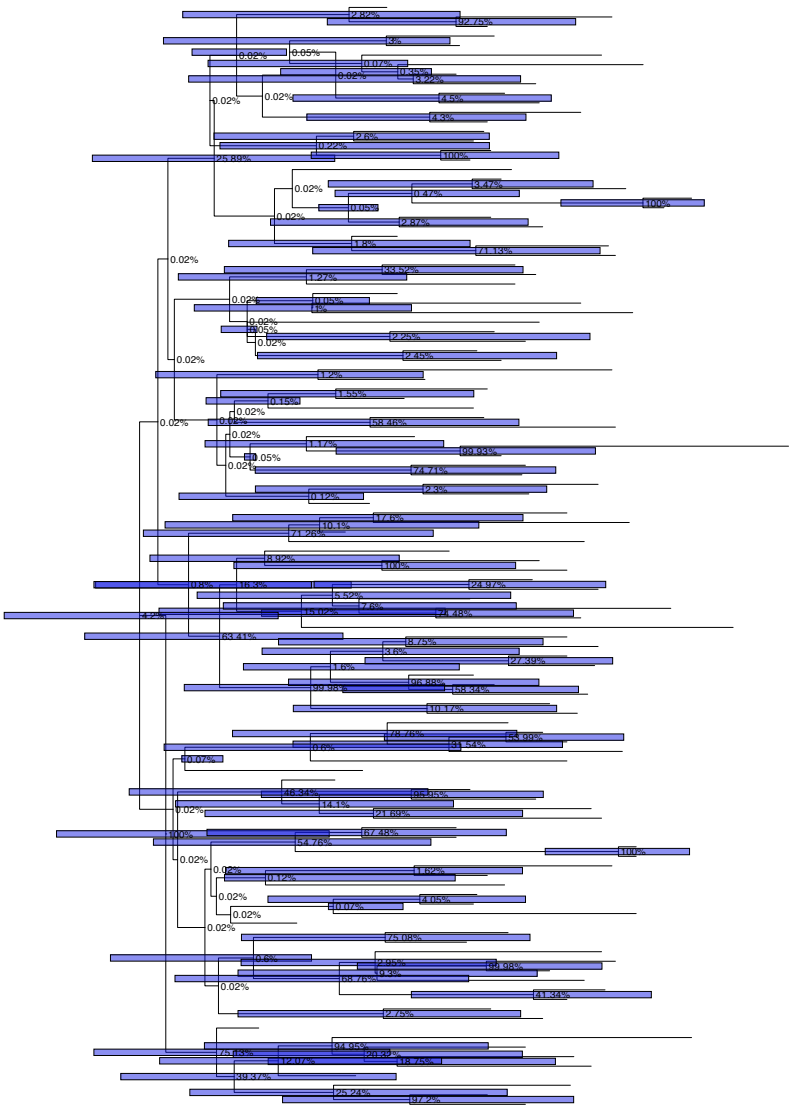


Figure 4.3: The maximum clade credibility (MCC) phylogeny for 116 wild poliovirus type 1 VP1 sequences from the 2010 Tajikistan outbreak. The MCC phylogeny was constructed from the posterior distribution of Bayesian phylogenetic analysis carried out in MrBayes 3.2 (Ronquist et al., 2012). The horizontal bars are the 95% highest posterior density intervals of branching time estimates.

substitution as estimated by the slope of the linear regression of divergence from root against time was $1.39(0.86 - 1.93) \times 10^{-2}$ substitutions per site per year (Figure 4.4). Usually poliovirus evolution follows a strict molecular clock (Jorba et al., 2008). However, the short duration of the outbreak meant that the sampling times did not correlate strongly with divergence from the root.

Bayesian analysis of the poliovirus sequences using MrBayes also yielded wide intervals

around estimates of molecular clock rate: $1.04(0.67 - 1.38) \times 10^{-2}$ substitutions per site per year. The Bayesian analysis yielded 95% highest posterior density interval estimates around branching times, which were all very wide (Figure 4.3).

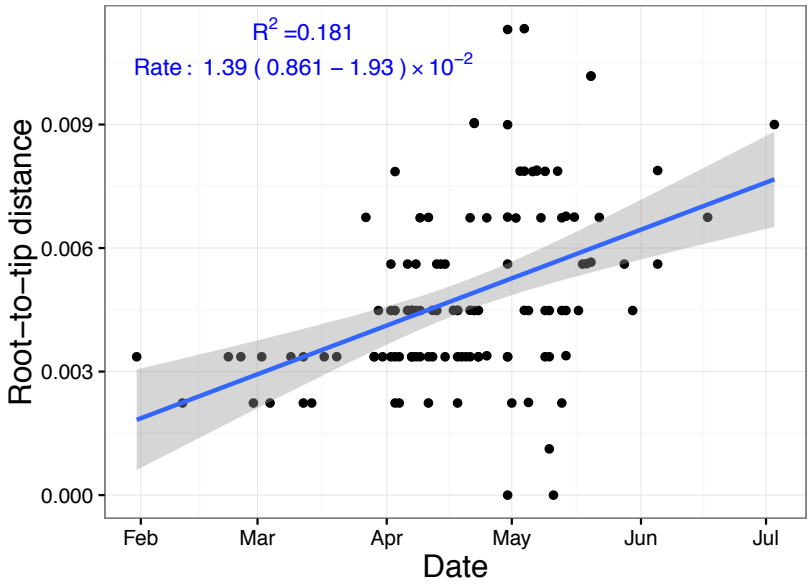


Figure 4.4: Root-to-tip distance as measured in substitutions per site for Tajikistan WPV1 sequences. The slope measures the substitution rate estimated to be $1.39(95\%CI : 0.86 - 1.93) \times 10^{-2}$ substitutions per site per year. Each dot represents a sequence. The line is the best-fitting linear model with the 95% confidence interval shown in the shaded region.

4.3.2 Non-parametric inference of N_e

To determine if there was sufficient signal in the pathogen phylogeny to infer parameter values, I estimated the effective population size N_e over time using the Bayesian skyline plot (BSP) approach in BEAST and using the average skyline pipeline outlined in Section 2.3.4. Both sets of N_e estimates yielded wide credible intervals (Figure 4.5). While the BSP suggested plateauing in infection numbers, the average skyline captured the rise and fall in prevalence of infection. Despite the uncertainty in N_e, the median N_e estimates suggested that there was sufficient information in the phylogeny to estimate model parameters.

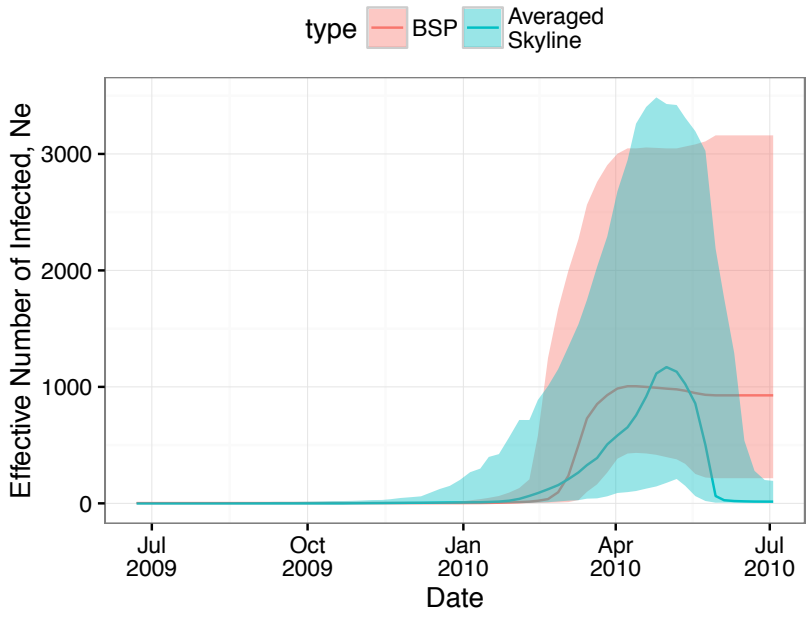


Figure 4.5: Estimates of effective population size based on poliovirus sequences collected during the 2010 Tajikistan outbreak. Two methods were used to estimate the effective population size over time: Bayesian skyline plot (BSP) as implemented in BEAST (Drummond and Rambaut, 2007) and the average skyline as described in Section 2.3.4. A generation time T_g of 10 days was used to convert the y-axis from $N_e T_g$ to N_e.

Parameters	**Both**	**Epi**	**Phy**	**Blake et al. 2014**
R_c	2.58 (2.23-2.98)	2.62 (2.07-3.32)	1.79 (1.46-2.11)	2.18 (2.06-2.45)
R_a	0.59 (0.48-0.7)	0.57 (0.44-0.69)	0.95 (0.15-1.47)	0.46 (0.42-0.52)
k	64.044 (1.842-518.102)	68.956 (0.003-728.782)	6.661 (0.119-349.728)	1
T_g	06 Dec 09 (06 Nov 09-25 Dec 09)	13 Jan 10 (13 Dec 09-27 Jan 10)	13 Jan 10 (13 Dec 09-27 Jan 10)	17 Dec 09 (21 Nov 09-6 Jan 10)
Infections per case	1 in 290 (212-369)	1 in 349 (213-496)	NA	1 in 200
Vaccine effectiveness	71.3% (53.9%-85.9%)	57.1% (24.3%-77.2%)	57.1% (24.3%-77.2%)	69% (55%-80%)

Table 4.2: Posterior parameter estimates for the 2010 Tajikistan poliovirus outbreak compared with the maximum likelihood estimates obtained in Blake et al. (2014)).

4.3.3 Phylodynamic inference

The posterior distributions of parameters are presented in Figure 4.6, and the median and 95% HPD intervals are in Table 4.2.

I obtained more precise estimates of the reporting rate when using both epidemiological and phylogenetic data. These estimates are dependent on the initial number of susceptibles, which I fixed to their maximum likelihood estimates obtained by fitting to epidemiological data only (Blake et al., 2014). The initial number of susceptibles was difficult to estimate without a very strong prior and necessitated a much longer MCMC chain to obtain sufficient samples from the posterior distribution. To obtain estimates of other parameters within a reasonable amount of time, I did not estimate the initial number of susceptibles.

The basic reproductive number of children aged 0-5 years R_c was estimated to be 2.58 (2.23-2.98) when both epidemiological and phylogenetic data were used in inference. These values were intermediate between estimates using just epidemiological data, and estimates using just phylogenetic data. The maximum likelihood estimate of 2.18 from Blake et al. (2014) was included within the 95% highest posterior density (HPD) interval except when only phylogenetic data were used for inference.

The posterior distributions of the reproductive number of older children and adults R_a all included the maximum likelihood estimate of 0.46. Adding the phylogenetic data did not significantly alter parameter estimates. This was not surprising as the credible interval surrounding estimates using just phylogenetic data was much wider.

The value of k was likely high, indicating the lack of superspreading dynamics. The estimated values were 6.7 (0.1-349.7) when only genetic data were used for inference. These values were much higher when epidemiological data were used: 69.0 (2.6×10^{-3}-729.8), and when both data sets were used at the same time: 64.0 (1.8-518.1).

Given the large credible intervals around estimates of vaccine effectiveness per campaign using phylogenetic data, only epidemiological data were informative of this parameter. The credible intervals using just epidemiological data included the maximum likelihood estimate from Blake et al. (2014) at 69% (55%-80%).

Finally, the estimated start date of the epidemic for analysis using both data sets, epidemiological data only, and phylogenetic data only all overlapped with each other, as well as with estimates from Blake et al. (2014).

Overall, it seemed like incidence time series were more informative for estimates of

reproductive number, k, and vaccine effectiveness than pathogen phylogeny alone. However, in all these cases, including the pathogen phylogeny improved the precision of estimates.

4.4 Discussion

While phylodynamic analyses have been used to characterise the epidemiological dynamics of other viral diseases such as influenza and HIV, such methods are not widely used for poliovirus analysis. Molecular surveillance through sequencing of poliovirus isolates has mainly been used for tracking the geographic spread of poliovirus in endemic countries (Angez et al., 2012), detecting orphan lineages which are indicative of long-term silent transmission (Gumede et al., 2014), and reconstructing the history of pathogen diversity (Burns et al., 2013). Although model-based parameter inference has been used to analyse epidemiological data for polio (Grassly et al., 2006; Mangal et al., 2013; Blake et al., 2014), it has not been used to analyse viral sequence data.

The gold standard of polio surveillance has traditionally been through Acute Flaccid Paralysis (AFP) surveillance, in which stool samples from patients with AFP symptoms are tested for the presence of poliovirus. As the number of poliovirus infections decreases, there might be too few symptomatic cases reported through AFP surveillance to provide sufficient data in terms of incidence time series and viral sequences. Environmental sampling of poliovirus shed by asymptomatically infected individuals will thus play an increasingly important role in monitoring poliovirus and quantifying its epidemiology as eradication gets closer.

In all cases, the confidence intervals obtained through profile likelihood were smaller than the 95% HPD intervals, even when only the incidence data are used for inference. This shows that integration over stochastic outcomes is important for fully describing the range of plausible parameter estimates.

Despite the uncertainty in phylogenetic topology and molecular clock rate, integrating over parameter estimates obtained for individual trees produced estimates of reproductive number, epidemic start date, and k. However, the 95% HPD interval

ranges were larger when using phylogenetic data compared to using incidence time series. In terms of vaccine effectiveness, most information came from the incidence time series rather than phylogenetic data. In this model, polio incidence dropped immediately after SIA campaigns but the prevalence of infection did not change as sharply. Perhaps these changes in prevalence of infection due to SIA campaigns did not leave noticeable patterns on the pathogen phylogeny, hence the lack of information on the per-campaign effectiveness of SIAs from the pathogen phylogeny.

Very large values of k were estimated for this outbreak, which appear implausible for a directly transmitted infectious disease (Lloyd-Smith et al., 2005). The wide distribution of posterior values for k could be due to the flat prior and lack of information in the phylogeny to precisely estimate k. Furthermore, the combination of short gene segment (906 bases) and rapid progression of the epidemic meant that the phylogeny could not be accurately characterised. As discussed in Chapter 3, more simulation studies are required to explore how phylogenetic uncertainty affects estimates of k, and further methodological developments may be required to fully integrate this uncertainty in epidemiological inference.

The prior on the fraction reported ρ could also have suffered from bias from the prior distribution, as the mean of the prior was 2×10^{-6}. However, the data seemed sufficiently informative to overcome this strong prior as the posterior values of ρ were much higher than that and in line with expected values.

As I did not estimate the initial number of susceptibles, I was able to obtain a broad estimate of the case-to-infection ratio using just the incidence time series. However, more precise estimates of the case-to-infection can be obtained by using both incidence time series and pathogen phylogeny.

To summarise, the addition of pathogen phylogeny to statistical inference improved the precision of case-to-infection ratio estimates, which were lower than the 1:200 value usually used in modelling of poliovirus transmission. Compared to the maximum likelihood parameter estimates based on incidence time series only (Blake et al., 2014), the Bayesian estimates here had broader credible intervals because of the incorporation of stochasticity in epidemic trajectories. Despite the increase in computation time, conducting a Bayesian analysis using the PMCMC approach is needed to fully capture

the uncertainty in parameter estimates. Finally, the analysis here shows that parameter estimates from pathogen phylogeny are mostly consistent with incidence time series. However, the pathogen phylogeny here was not as informative as in the simulated epidemics (Chapter 3) due to uncertainty in branching times. The rapidity of the Tajikistan outbreak reduced the reliability of phylogenies constructed from sequences. In the next Chapter, I will show how pathogen phylogeny constructed from sequences collected over a long period of time is more informative of certain parameters than incidence time series.

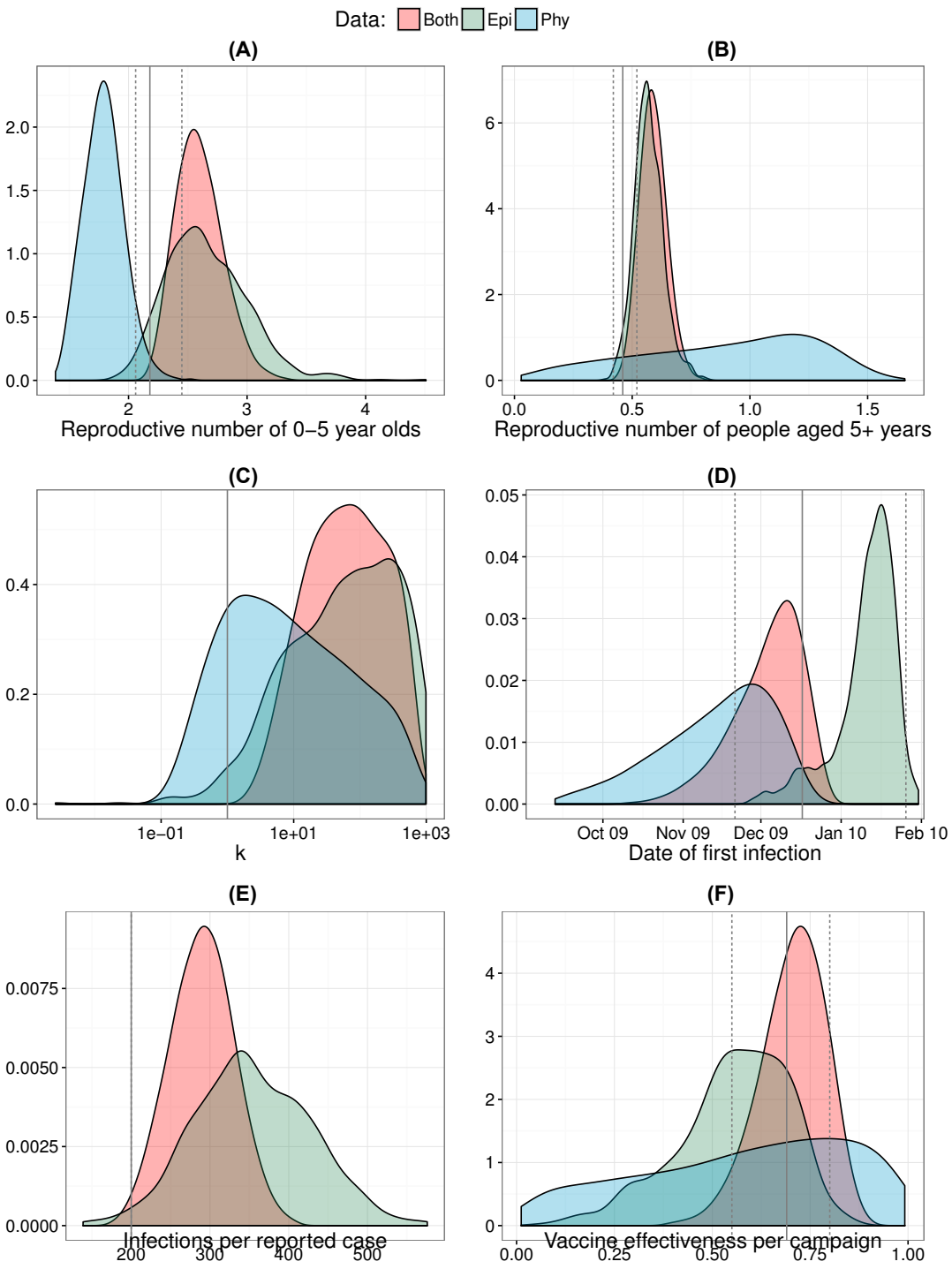


Figure 4.6: Posterior densities of epidemiological parameters for the 2010 WPV1 outbreak in Tajikistan. The estimated parameters include **(A)** the reproductive number of 0-5 year olds, **(B)** the reproductive number for 5+ year olds, **(C)** k, **(D)** the date of first infection, **(E)** infections per reported case, and **(F)** vaccine effectiveness per campaign.The solid and dashed vertical lines are the maximum likelihood estimates and 95% confidence intervals estimated in Blake et al. (2014). The solid vertical lines not accompanied by dashed lines correspond to parameter values that were fixed and not estimated.

Chapter 5

Phylodynamic analysis of environmental polio sequences

In this Chapter, I demonstrate that epidemiological models can be fit to sequence data from polioviruses isolated in wastewater and sewage samples. Environmental surveillance was set up in Pakistan in 2009 to supplement acute flaccid paralysis (AFP) surveillance of symptomatic polio cases. As the incidence of poliovirus infections decreases, environmental surveillance is playing an increasingly important role in monitoring the spread of poliovirus and quantifying infection numbers. Using the inference framework in Chapter 2, I estimate the average reproductive number of wild type-1 poliovirus (WPV1) each year between 2012 and 2015 in Pakistan based on just the environmental sequences, and compare the results to those obtained from analysing just incidence time series from AFP surveillance and highlight information added by analysing environmental sequences. The genetic data analysed in this chapter were sequenced by Sohail Zaidi and colleagues at the Pakistan National Institute of Health.

5.1 Introduction

The incidence of symptomatic cases of WPV1 has been on a downward trend for the last two decades. In 2015, there were only 75 confirmed cases of wild type-1 poliovirus (WPV1) globally. While the gold standard of polio surveillance is through Acute Flaccid

Paralysis (AFP) reporting, AFP surveillance can only detect symptomatic cases which occur in less than 1 in 200 infections. Because poliovirus infects the gut, viral particles are shed into the environment via faeces. By setting up sampling sites in sewage networks, poliovirus particles shed by infected individuals can be detected irrespective of symptoms. Environmental surveillance is thus particularly helpful at the end of the eradication programme as it is sensitive to infections even when there are no symptomatic infections. However, the sensitivity of environmental surveillance depends on a converging sewage system, distance of sampling sites from source (toilets), and sampling frequency (Hovi et al., 2012).

In countries already free of poliovirus, environmental surveillance data enable early detection of silent poliovirus transmission. In 2013, for example, the detection of WPV1 sequences in the environment in Israel led to mass immunisation campaigns with oral poliovirus vaccine (OPV) to boost the population immunity, so that infections can be halted before a paralytic case of polio occurs. In endemic settings such as Pakistan, environmental surveillance supplements AFP surveillance in detecting poliovirus infections. Furthermore, phylogenetic reconstruction of environmental sequences is informative of the spatial routes of poliovirus spread (Alam et al., 2014, 2016).

Environmental sampling sites were set up in Pakistan in 2009 (Angez et al., 2012). Since then, monthly samples have been taken from these sites. Thus there is sufficient number of sequences and sequence diversity to build phylogenies that are informative of epidemiological parameters. Estimates based on the sequences should therefore be more precise than those obtained for the Tajikistan outbreak analysis. However, analysis of data collected over several years from a large, endemic country presents additional challenges compared to outbreak analysis. First, using compartmental models that require specification of susceptible numbers was difficult because the proportion of susceptible changed due to routine immunisation, supplementary immunisation activity, births, deaths, as well as infections. To overcome this issue, I used a branching process model instead which tracked infected numbers only assuming no saturation effects. Secondly, annual variations in transmission rates have been observed in many temprate and sub-tropical countries including Pakistan with peak transmission occurring in late summer/autumn (Nathanson and Kew, 2010). I therefore included seasonality

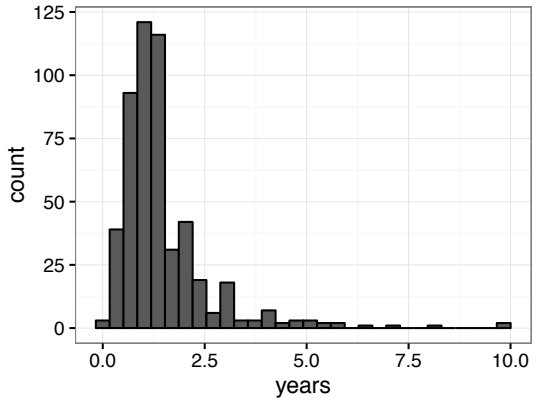


Figure 5.1: Age distribution of acute flaccid paralysis cases confirmed to be caused by wild type-1 poliovirus in Pakistan between 2012 and 2015, inclusive.

parameters in the transmission model to allow the reproductive number to vary according to the time of year.

Using data from 2012 to 2015, I estimated parameters of a transmission model using just environmental sequences, just incidence time series (from AFP surveillance), or both types of data. The results below highlight the epidemiological insight that can be gained from analysing environmental surveillance data, which will become increasingly important during the last stages of eradication.

5.2 Method

5.2.1 Data

Sequences analysed in this chapter were obtained from 38 environmental sampling sites in 13 cities across Pakistan (Alam et al., 2016). The viral isolates were sequenced in the 906-nucleotide VP1 region. Each sequence was associated with the date of sampling.

Between 2012 and 2015, inclusive, 369 environmental samples tested positive for WPV1, resulting in 683 sequences. More than one sequence per environmental sample appeared in the data set if these sequences were more than 1% divergent, suggesting the co-circulation of separate lineages in the population. As Sanger sequencing was used to obtain sequences, multiple sequences from a single environmental sample was only

possible if multiple plaques were sequenced.

Incidence time series were constructed from AFP cases confirmed to be caused by WPV1 and with an onset of paralysis between 2012 and 2015, inclusive. A total of 518 cases were reported in this period, with most reported cases of paralytic polio occurring in children under the age of 5 (Figure 5.1). This differed from the Tajikistan outbreak during which older children and adults were infected and reported as cases (Blake et al., 2014). Consequently, the model fit to data in this Chapter is not structured by age, and estimated sizes of the infected population correspond to children under 5.

5.2.2 Phylogenetics

I used jModelTest2 (Guindon and Gascuel, 2003; Darriba et al., 2012) to determine the best substitution model for the environmental WPV1 sequences in Pakistan. To reduce computation time, I sampled up to 2 sequences for each month between January 2012 and December 2015, inclusive. The resulting dataset comprised 294 sequences. The top 3 models with the lowest Bayesian Information Criterion (BIC) were SYM+Γ, K80+Γ and K80+Γ + I. Due to its simplicity, and agreement with the model used to analyse the Tajikistan sequences, I chose K80+Γ as the substitution model for phylogenetic reconstruction.

To estimate the rate of substitution, I used the smaller set of 294 sequences to construct a maximum likelihood phylogeny using RAxML (Stamatakis, 2014), again to reduce computation time. A root-to-tip plot was created from the resulting maximum likelihood tree.

For the Bayesian reconstruction of viral phylogeny in MrBayes (Ronquist et al., 2012), I analysed all 683 sequences. I used 3 million generations, sampling every 5000 and removing the first 50% of samples as burn-in.

5.2.3 Model

Unlike the Tajikistan analysis in Chapter 4, there are no reliable estimates of population immunity in Pakistan. Furthermore, the sampling dates of sequence data are over a longer

period of time: 4 years compared to 6 months in the case of Tajikistan. This means that a compartmental model should also include births and deaths. Assuming the size of the susceptible population does not significantly change due to infection events within each year, I used a continuous-time branching process model instead, which only tracks the number of infected individuals. In this model, the effective reproductive number R_t varies according to the time of year and undergoes step changes at the end of each year:

$$R_t = \bar{R}_{f(t)}(1 + \alpha cos(2\pi(t\Delta t + \tau))), \tag{5.1}$$

where $f(t)$ is a function that returns the year in which time step t occurs, $\bar{R}_{f(t)}$ is the average reproductive number in year $f(t)$, α is the amplitude of seasonal variation, and τ is the timing of peak reproductive number.

In a continuous-time branching process, each individual i gets infected at time step H_i and recovers at time step $H_i + W_i$ where $W_i\Delta t \sim$ discretised Gamma(a, b) is the generation time drawn from a gamma distribution with shape parameter a and scale parameter b, and Δt is the size of each simulation time step. I used a gamma generation time distribution with a mean of $ab = 10.8$ days and variance $ab^2 = 62.2$ days^2. These parameters are based on estimated values of generation time obtained in the Tajikistan analysis in Chapter 4. The mean duration of infectiousness estimated from the Tajikistan data was 6.8 days, and the assumed duration of latency was 4 days Paul (1955), thus resulting in a mean duration of infection of 10.8 days. Assuming exponentially distributed periods of latency and infectiousness, I adjusted the resulting gamma distribution parameters to match the mean and variance.

Secondary infections occur at the time of recovery. A number Z_i is drawn from a negative binomial distribution to determine the number of secondary infections caused by individual i at time step $H_i + W_i$. The negative binomial is parameterised by the mean and variance rather than the total number of trials and the probability of success. The negative binomial distribution has mean R_t and variance $R_t(1 + \frac{R_t}{k})$, where k is the dispersion parameter, and $t = H_i + W_i$.

5.2.4 Inference

Using the inference framework presented in Chapter 2, I fit the model to environmental sequences, to incidence time series, and to both at the same time. Although the two data sets were not necessarily collected from the same individuals, the inference approach I have developed in this thesis assumes that epidemiological data and genetic data are independently sampled from the same population of infected individuals. Thus inferred parameter values using either or both data sets should reflect the parameter values of the same population. The estimated parameters and their prior distributions are given in Table 5.1. Unlike in Chapter 4, I placed a beta prior on the case-to-infection ratio ρ (Figure 5.2). This prior placed less weight very low values of case-to-infection ratio.

To reduce the computation time, I drew 1000 random numbers from the gamma generation time distribution at the start of each MCMC iteration. On each core that the PMCMC was run, the program randomly selected a starting position in the array of 1000 numbers to obtain the generation time of the first infection. With each subsequently infected individual, the generation time of that individual was obtained by shifting the pointer along the matrix by 1. Each particle tracked the times of recovery for currently infected individuals, so this information was retained during particle filtering.

Because there is less stochasticity in transmission dynamics compared to the Tajikistan outbreak, I used 2,000 particles for the analyses as this was sufficient to obtain reliable estimates of the marginal likelihood. Each simulation time steps was $\Delta t = 0.25$ days.

I used up to 1 million MCMC iterations, sampling parameter values every 20 iterations. The first 25% of samples in each chain were removed as burn-in. Chains were terminated early if the effective sample size of the posterior exceeded 100 to reduce computation time. The median length of the chain was 105,840, and the shortest chain had 39,420 iterations.

5.2.5 Phylogeography

To determine the level of spatial structure in the population, I estimated the migration rate between provinces from the full set of 683 sequences using discrete phylogeographic methods in BEAST2 (Bouckaert et al., 2014). In this analysis, geographic location was

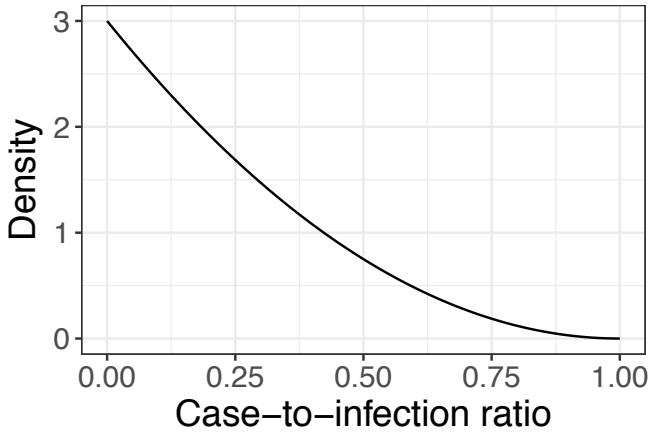


Figure 5.2: The beta prior placed on the case-to-infection ratio ρ with parameters $\alpha = 1$ and $\beta = 3$.

treated as a trait that switches between states (provinces) according to a Markov transition matrix, similar to nucleotide or amino acid evolution (Lemey et al., 2009).

In addition to migration rate parameters for pairs of provinces, indicator parameters were also estimated to determine if migration rates between pairs of provinces were significantly above 0. The migration rate matrix was assumed to be symmetric, which meant that the direction of migration was not inferred.

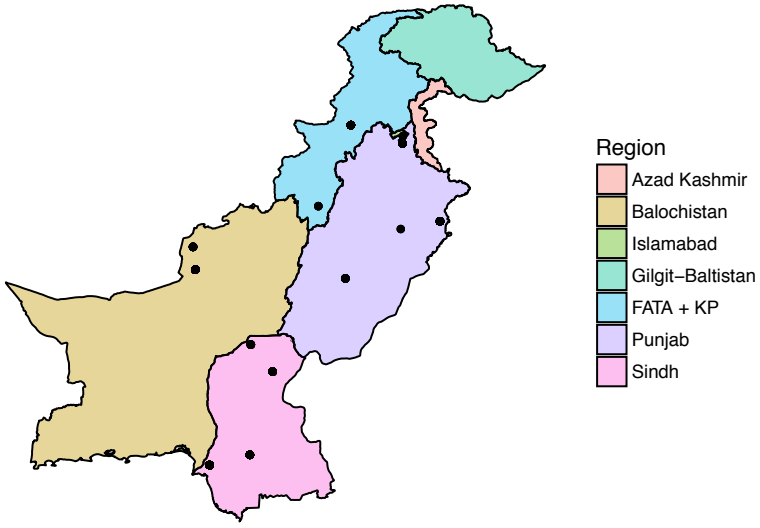
The discrete phylogeography analysis was carried out in BEAST2 with 60 million MCMC iterations, sampling every 25,000 iterations, and discarding the first 6 million (10%) iterations as burn-in.

5.3 Results

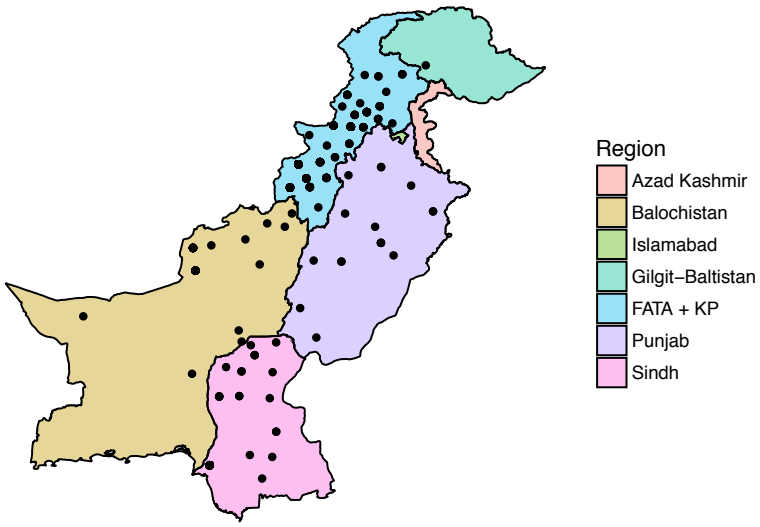
5.3.1 Spatiotemporal distribution

The environmental sampling sites were located in major cities (Figure 5.3A), however the spatial distribution of AFP cases was much wider (Figure 5.3B).

The overall number of environmental samples collected in Pakistan increased over time (Figure 5.4) as a reflection of the increasing number of sampling sites. The proportion of samples positive for WPV1 varied over time, perhaps reflecting seasonal changes in



(A) Environmental sampling sites



(B) Confirmed cases of WPV1

Figure 5.3: Cities in Pakistan with environmental sampling sites (A), and where confirmed wild type 1 polio cases have been detected between 2012 and 2015, inclusive (B).

Parameter name	Value	Estimated	Prior
$\bar{R}_{2012}$	-	Yes	unif(0.01, 10.0)
$\bar{R}_{2013}$	-	Yes	unif(0.01, 10.0)
$\bar{R}_{2014}$	-	Yes	unif(0.01, 10.0)
$\bar{R}_{2015}$	-	Yes	unif(0.01, 10.0)
Amplitude of seasonality α	-	Yes	unif(0.1, 1.0)
Timing of peak τ	-	Yes	unif(0.0, 2.0)
k	-	Yes*	unif(0, 1×105.0)
Generation time shape a	1.88	No	-
Generation time scale s	5.76	No	-
Initial infected I_0	-	Yes	unif(1.0, 1×10^6)
Case-to-infection ratio ρ	-	Yes	beta(1.0, 3.0)**

Table 5.1: Parameters of the branching process model fit to Pakistan data. For those that are estimated, the prior distributions are given. Fixed parameter values are given in column 2. The generation time distribution is gamma distributed with mean $as = 10.8$ days and variance $as^2 = 62.2$ days^2. The reproductive number at time t R_t is calculated using Equation 5.1. *k is not estimated when just using incidence time series. As shown in Chapter 2, epidemiological data in the form of incidence time series are not informative of the value of k. **The beta prior distribution is illustrated in Figure 5.2.

reproductive number.

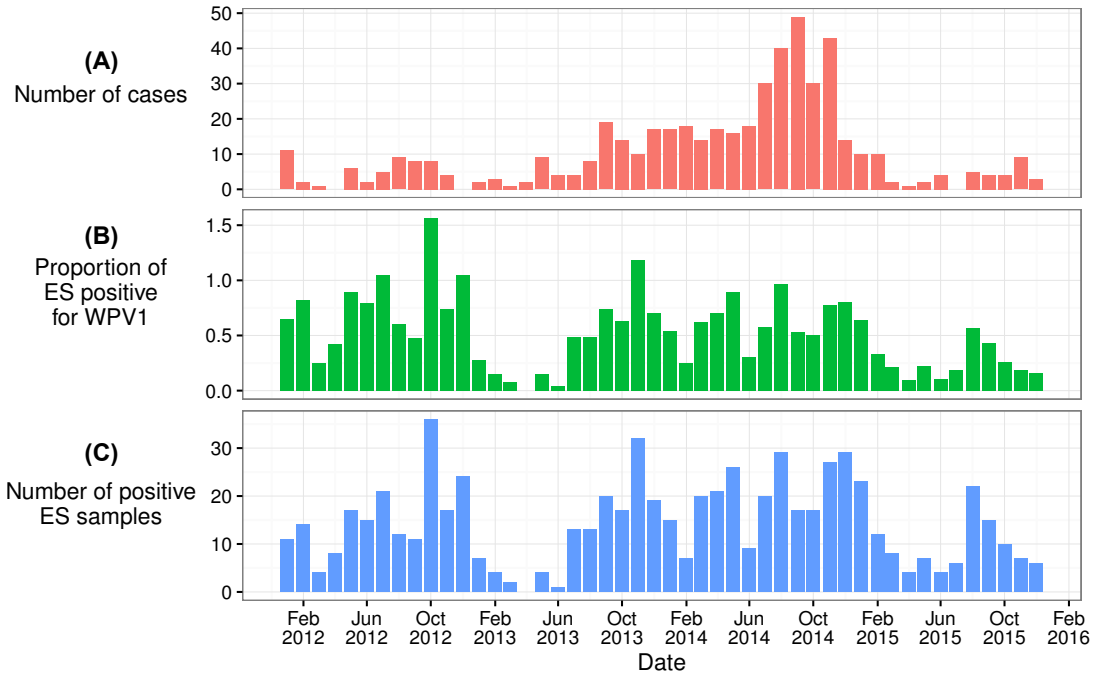


Figure 5.4: Monthly time-series of the numbers of (A) confirmed WPV1 cases, (B) environmental surveillance (ES) samples, and (C) proportion of ES samples that tested positive for wild type-1 poliovirus.

Although the majority of infections over the 2012-2015 period occurred in the Federally

Administered Tribal Areas (FATA), there were no environmental sampling sites in that region (Figure 5.5). The peak in case numbers in 2014 was also observed in the number and proportion of positive environmental samples for some of the provinces (Balochistan and Punjab). After removing seasonal trend, the spearman's rank correlation coefficient between the number of confirmed WPV1 cases (from AFP surveillance) and the number of positive environmental samples was 0.54. This suggested the two sources of data were moderately correlated in terms of counts.

5.3.2 Viral evolution

Root-to-tip analysis revealed substitution rates of 8.77×10^{-3} (7.81×10^{-3} - 9.73×10^{-3}, 95% confidence interval) substitutions per site per year (Figure 5.6). I obtained slightly higher estimates when I used a Bayesian approach implemented in MrBayes (Ronquist et al., 2012): 9.3×10^{-3} (8.45×10^{-3} - 1×10^{-2}, 95% HPD interval). The median estimates were lower than those estimated in the Tajikistan analysis, but the 95% HPD intervals here are also narrower. This is because there is greater diversity between these environmental samples compared to sequences collected during the Tajikistan outbreak, and thus the phylogeny is better resolved.

Assuming a constant rate of nucleotide substitution and neutral evolution, I used the posterior distribution of phylogenies obtained through MrBayes to infer the averaged skyline (as described in Chapter 2), i.e. the effective number of infectious individuals over time N_e (Figure 5.7). There was signal for seasonal patterns in transmission using all sequences across the country, and at province levels. The skyline inferred from trees of uniformly sampling sequences also displayed seasonal trends (not shown), indicating that these trends are not an artefact of sampling.

The annual cycles in transmission rates were more evident in the N_e estimated using the average skyline plot (Figure 5.7) than in the rate of polio AFP cases over time (Figure 5.4). This suggested that seasonal patterns in transmission potential left a more distinct signal on the viral phylogenies from environmental sequences than on the incidence time series.

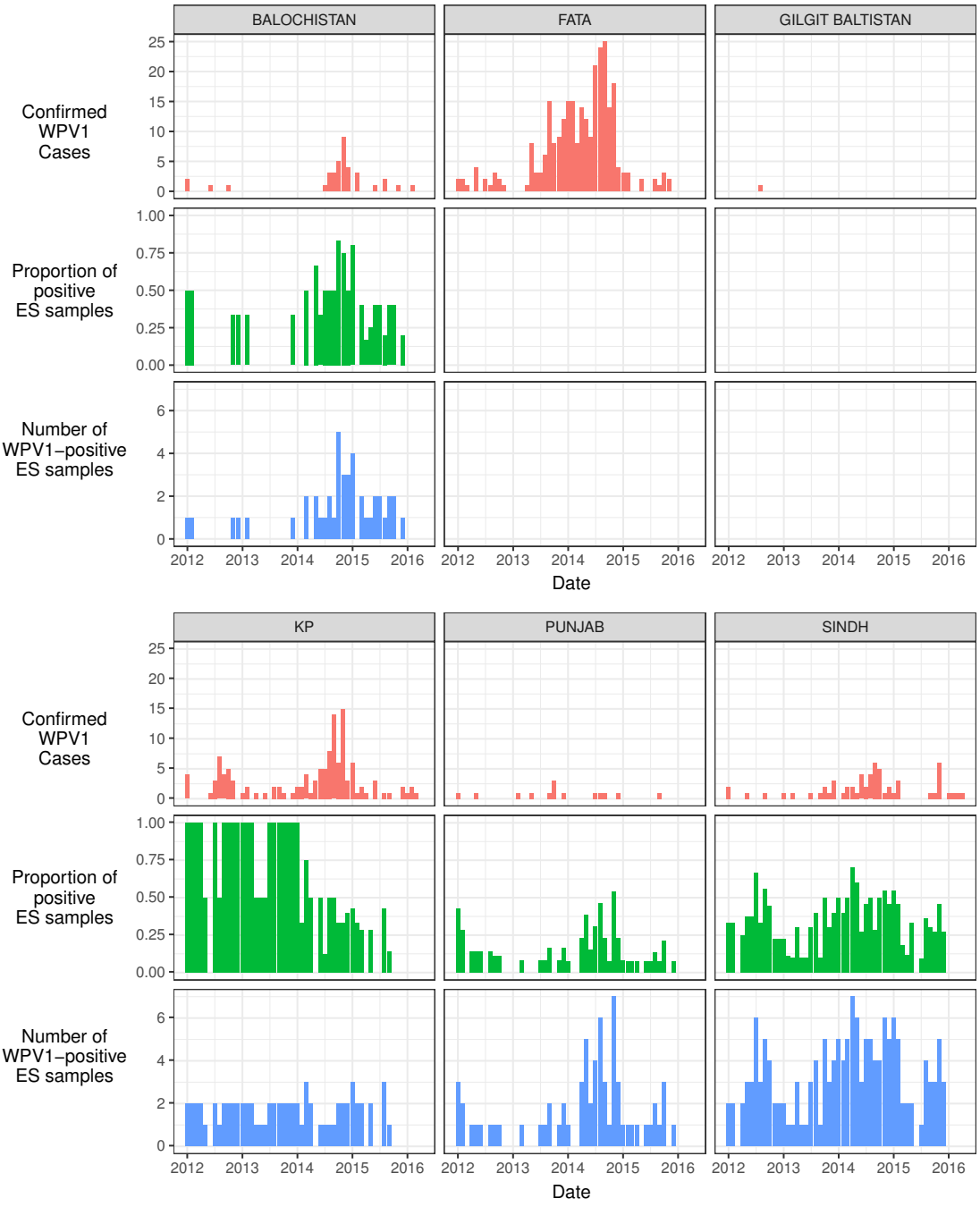


Figure 5.5: Monthly time-series of the numbers of confirmed WPV1 cases, environmental surveillance (ES) samples, and proportion of ES samples that tested positive for wild type-1 poliovirus in 6 provinces in Pakistan: Balochistan, FATA, Gilgit Baltistan, Khyber Pakhtunkhwa, Punjab, and Sindh.

5.3.3 Epidemiological parameters estimated using model fitting

The skyline plot suggested that there was sufficient signal in the pathogen phylogeny to quantify seasonal trends in transmission. Thus I fit a stochastic branching process

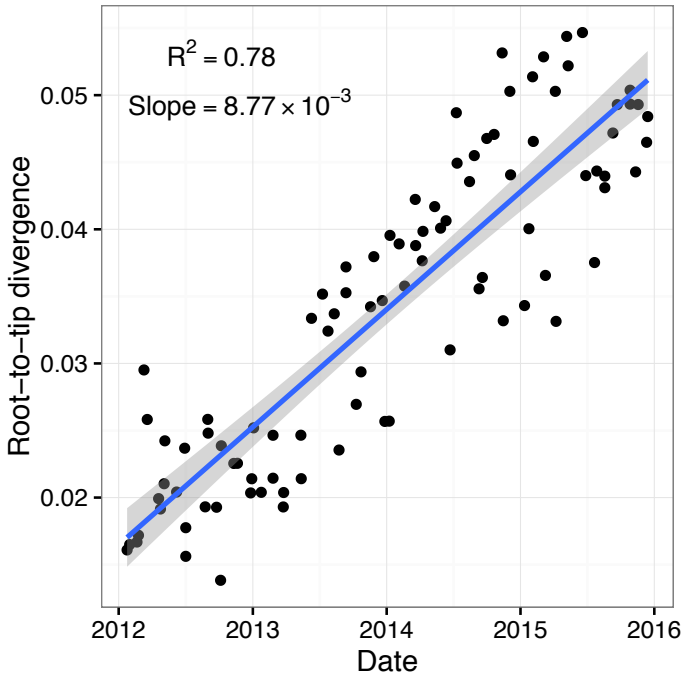


Figure 5.6: Maximum likelihood estimate of the rate of divergence. Based on the maximum likelihood phylogeny constructed in RAxML, I regressed the root-to-tip distance measured in substitutions per site against tip sampling times. I obtained estimates of $R^2 = 0.78$ and a divergence rate of $= 8.77(95\%CI : 7.82, 9.72) \times 10^{-3}$ substitutions per site per year. Each dot represents one sequence in the sample. The line is the best-fitting regression line. The shaded region shows the 95% confidence interval around the best-fitting regression line.

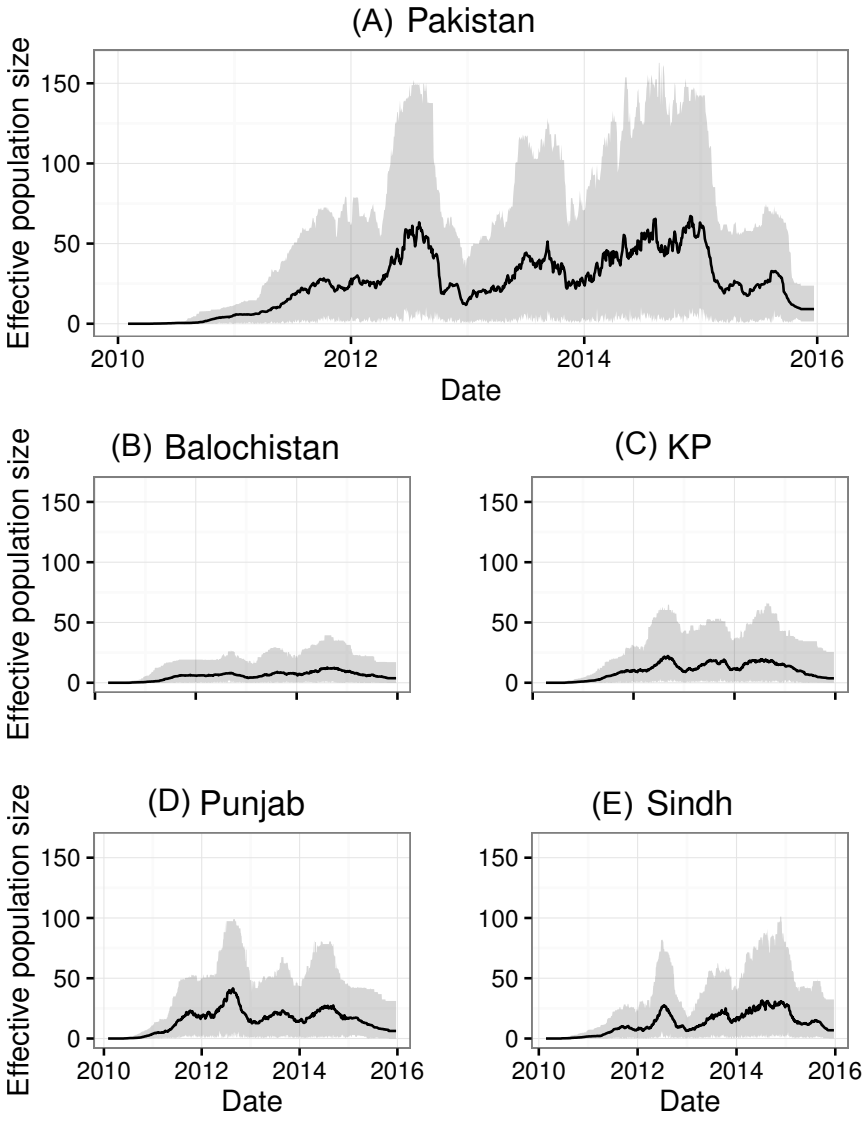


Figure 5.7: The average skyline plot for (A) all sequences in Pakistan, and for individual provinces: (B) Balochistan, (C) Khyber Pakhtunkhwa, (D) Punjab, and (E) Sindh. A generation time of $T_g = 10$ days was assumed to converge $N_e T_g$ to N_e. The shaded areas represent the 95% HPD intervals around N_e*estimates*

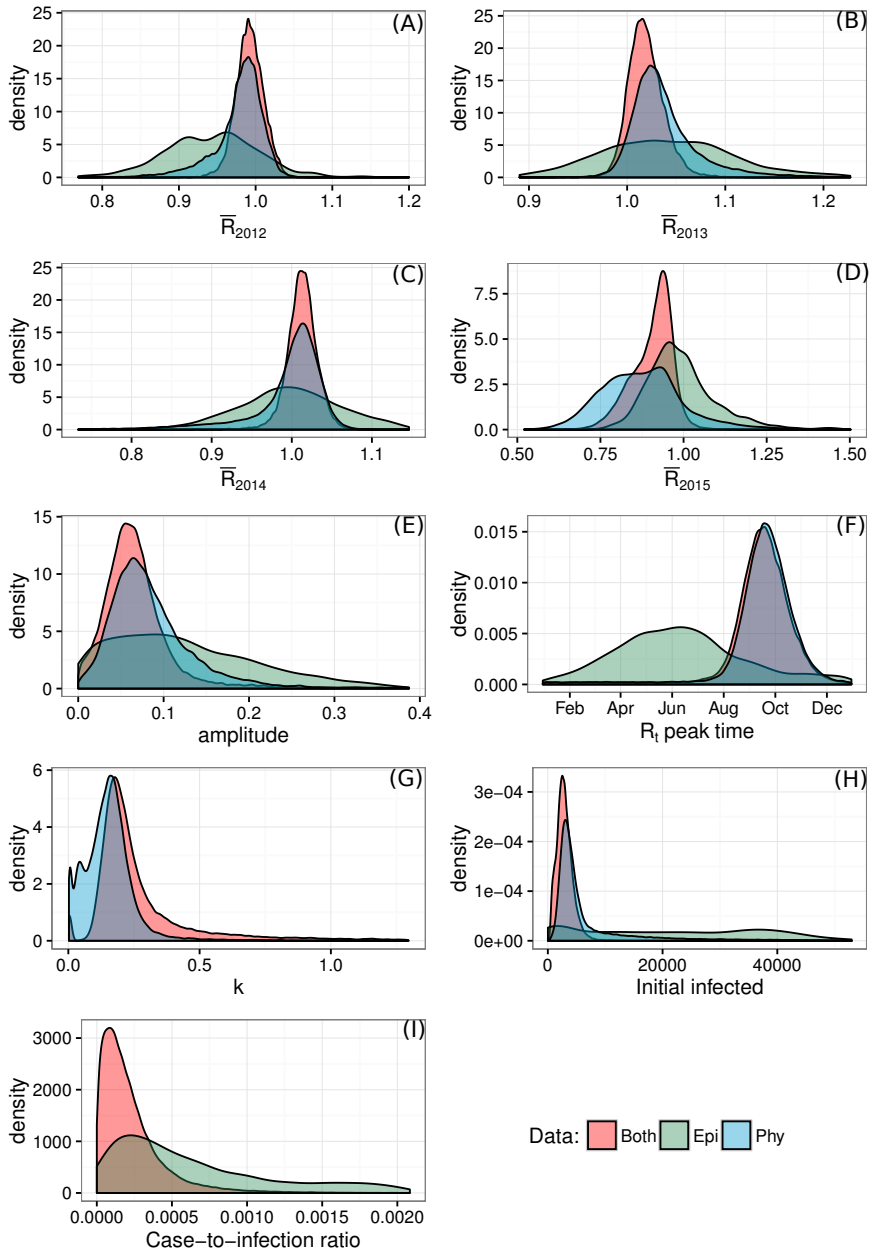


Figure 5.8: Posterior distributions of parameter values estimated for Pakistan data collected between 2012 and 2015. The estimated parameters include **(A-D)** the average annual reproductive numbers from 2012 to 2015, **(E)** amplitude of seasaonal variation, **(F)** peak reproductive number during a year, **(G)** k, **(H)** initial number of infected individuals, and **(I)** case-to-infection ratio.

model with a reproductive number that varied with time. The resulting estimates are listed in Table 5.2, with the full posterior distribution of each parameter visualised in Figure 5.8. The average reproductive number was estimated to be around 1 for all 4 years between 2012 and 2015, regardless of which source of data was used during inference.

Table 5.2: Median and 95% Highest Posterior Density bounds of parameter estimates inferred from wild poliovirus 1 sequences, from polio case reports, and from both at the same time.

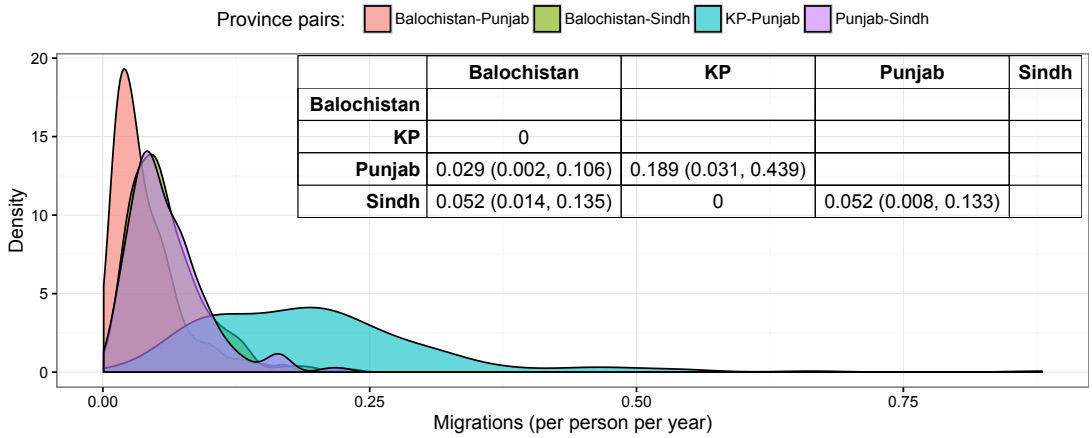
	Both	**AFP**	**ENV**
Amplitude	0.063 (0-0.166)	0.111 (0-0.293)	0.078 (0-0.387)
R_t peak time	19 Sep (16 Jul-03 Dec)	07 Jun (26 Jan-26 Nov)	22 Sep (11 Jul-15 Dec)
k	0.215 (0.004-1.296)	Not estimated	0.152 (0.002-0.306)
Initial number of infectious	2,776 (487-5,524)	22,366 (12-52,963)	3,992 (505-27,806)
Case:Infection	1:5,943 (458-52,195)	1:1,998 (216-17,254)	Not estimated
R_{2012}	0.992 (0.957-1.028)	0.945 (0.823-1.065)	0.987 (0.769-4.185)
R_{2013}	1.018 (0.988-1.057)	1.04 (0.89-1.207)	1.032 (0.949-1.227)
R_{2014}	1.013 (0.976-1.05)	0.999 (0.865-1.146)	1.009 (0.733-1.095)
R_{2015}	0.919 (0.773-1.006)	0.979 (0.813-1.236)	0.88 (0.521-1.502)

This is consistent with infectious disease dynamics at endemic equilibrium. Although there appeared to be a reduction in the reproductive number in 2015 when analysing environmental sequences, the uncertainty in parameter estimates was also larger in 2015. To confirm the apparent decrease in the reproductive number from 2014 to 2015, I would need to analyse more recent environmental sequence data collected in 2016.

When just environmental sequences were used, the estimated dispersion parameter $k \approx 0.15$ meant that when R = 1, just 1% of infected individuals caused 80% of infections. A similar estimate of k was obtained when both the environmental sequences and the incidence time series were used during inference. This value is consistent with other directly transmitted acute infectious diseases such as SARS and measles (Lloyd-Smith et al., 2005). I did not estimate k when inferring from just incidence time series, as incidence time series are not informative of the value of k (Chapter 3). Based on the median estimates of seasonal amplitude α, variation in reproductive number was likely present and small. However, I could not exclude zero seasonality in the reproductive number since all 3 credible intervals contained the amplitude parameter $\alpha = 0$.

The estimated case-to-infection ratio using just AFP time series was very low at 1:1,998 (216-17,254). The estimate was even lower when environmental sequences were also included in the analysis: 1:5,943 (458-52,195).

The posterior distribution of the initial number of infected individuals was very flat when only incidence time series data were used for inference. This was perhaps a reflection of



	Balochistan	KP	Punjab	Sindh
Balochistan				
KP	0			
Punjab	0.029 (0.002, 0.106)	0.189 (0.031, 0.439)		
Sindh	0.052 (0.014, 0.135)	0	0.052 (0.008, 0.133)	

Figure 5.9: Posterior distribution of migration rates between provinces. The median and 95% HPD intervals of migration rates between pairs of provinces are given in the table. Symmetric migration rates were assumed, hence the empty cells in the table.

the small number of reported cases at the beginning of 2012, which was not sufficiently informative of the number of infectious individuals at the start of the time series.

5.3.4 Phylogeographic analysis

To determine the relative contributions of infections from each province, separate skylines were calculated using sequences from each province (Figure 5.7). Punjab and Sindh appeared to have the largest contributors to infection. Seasonal trends in N_e were evident in the skyline plot for each province, suggesting that seasonal trends in transmission were present in all provinces.

N_e would be over-estimated if migration rates between provinces were low due to the strong spatial structure. To quantify the migration rate based on sequences alone, I estimated the migration rates between provinces in BEAST2 (Bouckaert et al., 2014) by treating provinces as discrete traits. The pair with the highest migration rate was between Khyber Pakhtunkhwa (KP) and Punjab (Figure 5.9). Other links were found between Balochistan and Punjab, Balochistan and Sindh, and Sindh and Punjab.

The low or zero migration rates estimated for most province pairs suggested that poliovirus transmission was highly spatially structured.

5.4 Discussion

Using just the VP1 gene sequences of poliovirus isolates collected in the environment, I was able to estimate model parameters and characterise the transmission dynamics of WPV1 in the population. These estimates were broadly consistent with those obtained by analysing incidence time series except the timing of peak transmission, although the estimates from environmental sequences were likely more reliable given the smaller uncertainty around the estimates. Thus, even in the absence of symptomatic cases, the epidemiological dynamics of poliovirus can be deduced by analysing the environmental sequences. While there are still AFP cases, nevertheless, both sequence data and incidence time series should be jointly analysed as this would produce the most precise parameter estimates.

Although the incidence time series provided higher point estimates (median) of the seasonality amplitude parameter than the viral phylogeny, the credible intervals were also much larger, indicating less clear signals of seasonality.

Besides providing more precise estimates of seasonality parameters, pathogen phylogeny was informative of the value of k. Because the pathogen phylogeny was more precisely estimated here compared to the Tajikistan analysis in Chapter 4, there was sufficient signal in the pathogen phylogeny to estimate k, whose value was similar to that of SARS which was characterised by superspreading events (Riley et al., 2003). The highly overdispersed offspring distribution means that a small group of individuals are contributing to poliovirus transmission, which are evident in environmental sequence data but are not picked up by traditional surveillance of symptomatic cases.

The estimated case-to-infection ratios ($\sim$ 1 in 6,000) were much lower than those estimated for the Tajikistan outbreak ($\sim$ 1 in 300). While this value is low, it is on the same order of magnitude as case-to-infection values estimated for poliovirus serotypes 2 and 3 (Nathanson and Kew, 2010). One possible reason for the low values might be due to areas in Pakistan with poorer surveillance than the rest of the country. Political instability due to the Taliban insurgency in the FATA region of Pakistan could have contributed to low reporting rates in the region (Hussain et al., 2016).

Unlike the prior used in the analysis of the Tajikistan poliovirus outbreak in Chapter 3

which was biased towards very low case-to-infection ratios ρ, the prior on ρ used in this chapter was a beta distributionwhere the mean was 0.25 and the interquartile interval was [0.09, 0.37]. This was a more reasonable distribution as values of $\rho < 0.1$ had similar prior densities but $\rho > 0.1$ had increasing smaller prior densities. The resulting posterior distribution did not conform to this beta prior, suggesting that it was informed by the data.

Results of the phylogeographic analyses were consistent with existing knowledge of epidemiological links between Balochistan and Sindh, and lack of movement between KP/FATA and Balochistan (Alam et al., 2016).

In terms of implementation, I only needed 2,000 particles for the analysis in this Chapter, unlike the analyses of simulated outbreak data (Chapter 3) and data from the Tajikistan outbreak of WPV1 (Chapter 4) that required 10,000 particles. This was because simulations did not start from a single infected individuals as polio is still endemic in Pakistan. Fewer particles were therefore needed to capture the stochasticity in epidemiological dynamics.

A major caveat in the analysis presented here is the assumption of a panmictic population. The low migration rates estimated using discrete phylogeography were indicative of strong population structure. Structured coalescent approaches (Rasmussen et al., 2014b) could be used to infer migration rates between provinces using both epidemiological and phylogenetic data, though the number of sequences per geographic region might not be sufficiently large to obtain precise estimates of migration rates. A potential workaround would be to divide the sequences not according to the province, but using a binary division e.g. North-South. Ongoing work on estimating connectivity at the district level using spatial models can help to determine the most suitable population structures (Molodecky, unpublished).

Besides the lack of spatial structure, another limitation of this study was the lack of data from Afghanistan. Because the border between Pakistan and Afghanistan is porous in many places leading to many cross-border transmissions (Angez et al., 2012), this might cause over-estimates of Ne and thus prevalence of infection in Pakistan.

In addition to data from a broader geographic range, analysis of more recent sequences

would improve the precision of reproductive number estimates during 2015. This is because including 2016 sequences would increase the number of coalescent events in the phylogeny during 2015, and thus increase the precision of parameter estimates for 2015. This would help to determine if there is a real decreasing trend in the reproductive number.

Another set of data that could be informative of epidemiological dynamics is the collection of poliovirus sequences sampled from AFP patients. Comparing parameter estimates from human and environmental sources of poliovirus sequences can further validate the parameter estimates, or could highlight differences in the processes generating the two sources of WPV1 sequences. I have just begun this analysis as the Pakistan team recently (Dec 2016) shared the WPV1 sequence data from AFP cases in Pakistan between 2012 and 2015.

Finally, more complex epidemiological models could be fit to the data to incorporate changes in the case-to-infection ratio, which I assumed to be constant over time in this Chapter. A more detailed analysis could be done by estimating a case-to-infection ratio for each year, and for each administrative region.

To summarise, the estimates based on environmental sequences were corroborated by parameters estimated from incidence time series alone. Furthermore, the inclusion of environmental sequences helped to quantify seasonality and heterogeneity in individual infectiousness. As polio eradication approaches completion, the number of AFP cases will continue to decline. At the same time, the number of environmental sampling sites is continuing to increase (Asghar et al., 2014). Using the approach presented here, environmental sequences of poliovirus can be used to quantify the remaining number of individuals infected with poliovirus even in the absence of symptomatic cases.

Chapter 6

Discussion

In this Chapter, I summarise the methodological developments and epidemiological findings presented in Chapters 2-5. Caveats of these studies are discussed in Section 6.2. Finally, extensions to the present work are presented in Section 6.3.

6.1 Summary of thesis contributions

The methodological problem addressed in this thesis was that of inference. While mathematical models have been used in epidemiology for many decades, inference frameworks that estimate parameters of nonlinear, stochastic models have only emerged over the last two decades. Estimating parameters from epidemiological data help to quantify underlying biological processes and transmission events, and can be useful for designing public health policies and predicting the future trends of an infectious disease. The inference framework presented in Chapter 2 addressed three challenges in epidemiological inference:

1. combined analyses of incidence time series and pathogen genetic sequence data;
2. estimation by integrating over model stochasticity and phylogenetic uncertainty;
3. quantifying heterogeneity in individual transmissibility by fitting transmission models that allow for arbitrary variance in the offspring distribution (number of secondary infections).

Various inference approaches have been previously been developed to address a subset of these problems. The inference framework described in this thesis is related to previous efforts to integrate inference from epidemiological and phylogenetic data (Rasmussen et al., 2011) in which the first challenge and, to a certain extent, the second challenge were addressed. The additional contributions made by this thesis include estimation of heterogeneity in individual transmissibility and integration over uncertainty in the underlying phylogeny. While in Rasmussen et al. (2014a) and Rasmussen et al. (2014b) the authors did estimate parameters using 10 different phylogenies, they did not pool together the estimates, and the phylogenies had relatively little uncertainty in branching times compared to phylogenies from outbreaks.

In terms of implementation, the PMCMC algorithm is available in a small number of existing program including SSM (Dureau et al., 2013), LibBi (Murray, 2013), and the `R` package `pomp` (King et al., 2016b). However, these implementations are not applicable to phylogenetic data because they were designed for time series data. Rasmussen et al. (2014b) provided a Java implementation of PMCMC that fits to both epidemiological and phylogenetic data. However the code is not parallelised and no longer maintained. The implementation provided here provides a parallelised C++ implementation of PMCMC that works on multi-core CPUs, and also an R package that simplifies the process of send data to and parsing the output of the C++ program (github.com/lucymli/EpiGenMCMC). This combination balances computational speed and ease of utilisation.

Developments in phylodynamic inference, on the other hand, have focused on using pathogen phylogenies to reconstruct epidemic history and estimate epidemiological parameters. Bayesian phylogenetic reconstruction programs such as BEAST (Drummond and Rambaut, 2007; Drummond et al., 2012), BEAST2 (Bouckaert et al., 2014), and MrBayes (Ronquist et al., 2012) have focused on simultaneous estimation of mutation model parameters and parameters of simple, deterministic population models such as exponential and logistic growth coalescent models (Griffiths and Tavare, 1994). Non-parametric estimation of demographic parameters is possible using skyline approaches (Drummond et al., 2006). Skyline approaches estimate $N\mathrm{e} \cdot T_g$, the product of effective population size and generation time. As discussed in Chapter 2, these values

are correlated with, but not necessarily proportional to the prevalence of infection.

More recently, an alternative population genetics approach using the birth-death model has been implemented in BEAST2 to directly estimate epidemiological parameters such as the reproductive number (Kühnert et al., 2014). The limitations of using a birth-death approach are the requirement to fix or estimate the sampling probability, and the large credible intervals surrounding parameter estimates (Boskova et al., 2014).

Application of the framework to simulation data highlighted the important role that phylogenies play in estimating the dispersion parameter k that determines the extent to which transmissions cluster in the population. Estimating k is not only useful for accurate inference from phylogenetic data using coalescent approaches, but also provides information on the likelihood of superspreading and the effectiveness of interventions. Although there have been developments in structured coalescent approaches, there are few phylodynamic methods that incorporate individual variation in transmission. One such method was used to estimate the variance of the offspring distribution for the 2014 Ebola outbreak by structuring the population into two infectiousness classes (Volz and Pond, 2014). However, discretising infectiousness might not capture extreme superspreading events, and definitions of high and low risk groups (as is done in studies of sexually transmitted diseases) are not always appropriate for viral infectious diseases.

In addition to methodological contributions, I also demonstrated that phylodynamic analyses of poliovirus sequences can shed light on outbreak dynamics as well as endemic patterns. In the case of the Tajikistan analysis, the addition of poliovirus sequences helped to narrow down the range of possible case-to-infection ratios. My analysis of environmental sequences from Pakistan demonstrated that viral sequences can help to quantify seasonality even though incidence time series did not reveal such patterns. When the reporting probability varies over time, seasonal patterns of transmission are not as evident in incidence time series as in phylogenetic patterns. Most importantly, this analysis showed that in the absence of reported cases, environmental sequences can be used to estimate the size of the infected population and the reproductive number. These are key indicators of progress towards eradication. After the disappearance of symptomatic polio cases or in areas where AFP surveillance is challenging, environmental surveillance will play a vital role in monitoring the decline of poliovirus

transmission in the population. Although environmental surveillance has been implemented in some countries for 2 decades, new countries are setting up environmental surveillance sites such as the UK (Public Health England, 2016).

Compared to our analysis of outbreak data from Tajikistan, the analysis of poliovirus data collected over the course of several years from Pakistan differed in terms of the transmission model and computational requirements. Unlike in an outbreak setting, compartmental models for long-term dynamics of an infectious disease need to include birth and death rates, in addition to quantification of immunity. To overcome this issue, I used a branching process model instead that produced estimates of the mean annual effective reproductive number. I assumed a seasonally changing reproductive number, with a step change in the mean at the end of each calendar year. For more complex dynamics, more breaks points would be needed to account for changes in transmission dynamics. However, this would increase the number of parameters that need to be estimated, which means longer computation time for PMCMC to reach convergence.

6.2 Limitations

6.2.1 Accuracy of phylogenetic inference

In all the analyses carried out in this thesis, I assumed that a dated phylogeny or a posterior distribution of phylogenies was accurately estimated. However, the rooting of the phylogeny and estimates of the molecular clock can both affect the accuracy of parameter estimation. For example, initial phylogenetic analysis during the 2014 outbreak of Ebolavirus suggested that these lineages diverged from viral lineages that caused Ebola outbreaks in the 70s (Baize et al., 2014). This would suggest a large Ne for the Ebolavirus population. However, Dudas and Rambaut (2014) noted that the divergence of Guinea sequences from those of previous outbreaks was because they were sequenced most recently and had accumulated the highest number of substitutions. Assuming that the Ebolavirus genome followed a molecular clock model, the tree was re-rooted at the lineage that caused an outbreak in 1976. According to the re-rooted phylogeny, Ebolavirus associated with the 2014 outbreak likely descended from the lineage that had previously caused outbreaks

in West Africa in the early 2000s, instead of silently circulating for decades.

6.2.2 Evolutionary assumptions

The use of pathogen phylogeny in the inference framework presented in this thesis was based on the assumption of neutral evolution and lack of recombination. In the case of poliovirus analyses, the VP1 sequence that was analysed is under strong negative selection as the VP1 protein is a viral surface protein that interacts with host cell receptors (Jorba et al., 2008). However, most substitutions are synonymous, which means they do not change the amino acid sequence and have little biological consequence. Neutral evolution can therefore be assumed in such cases as there is no selection pressure on these synonymous mutations.

Recombination is highly prevalent amongst poliovirus strains (Lukashev, 2005), and analysis of whole genome sequences from the 2010 Tajikistan outbreak revealed multiple recombination events (Yakovenko et al., 2014). However, recombination with the VP1 gene is rare, which means the standard coalescent approach can be used for inference (Jorba et al., 2008).

Extending the method to other pathogens or to the whole poliovirus genome would require considerations for these evolutionary complexities. Possible extensions to the method to incorporate selection and recombination events are discussed in Section 6.3.

6.2.3 Assumptions of the coalescent with arbitrary offspring distribution

In Fraser and Li's (2017) formulation of the coalescent, the mean and variance of the offspring distribution were assumed to be constant during each discrete generation of disease transmission. This discrete generation scheme can approximate the overlapping generations in real disease epidemics if the generation time is short. This was the case for the simulated and real infectious disease data analysed in this thesis. However, if the reproductive number changed rapidly relative to the generation time, the coalescent rate

might change significantly during a single generation of disease transmission and lead to biased parameter estimates.

6.2.4 Population structure in the transmission model

Although I have considered individual-level heterogeneity, I did not consider population structure except for the Tajikistan analysis in Chapter 4. For that analysis, I structured the population by age but ignored this age structure when calculating the coalescent likelihood due to the low sampling rate (see Section 4.2.3 for more details). For the Pakistan analysis, I did not structure the population by age because almost all infections occurred in children under 5. Phylogeographic analysis suggested strong spatial structure given the low rates of migration between regions. The lack of spatial structure in the model fitted to the Pakistan could have affected inference results. For example the estimate reproductive numbers might not be reflective of each province. The results in Chapter 5 highlighted the information available in environmental sequences, but further work involving structured models is necessary to produce accurate estimates of parameter values at the provincial level.

6.2.5 Variable reporting probability

Incomplete reporting is a common issue in infectious disease research due to asymptomatic cases and under-reporting. The wrong assumption about the reporting probability can lead to biased estimates of infection prevalence, final epidemic sizes and reproductive numbers (Gamado et al., 2013).

Often when estimating epidemiological parameters from incidence time series, the reporting proportion is estimated as a constant (Blake et al., 2014; Camacho et al., 2015). In the analysis of simulated data in Chapter 3, I demonstrated that reporting probability can be estimated from incidence time series if the initial state variable values are fixed, but more accurate and precise estimates were obtained by incorporating phylogenetic data in statistical inference. Thus, pathogen phylogeny can be analysed along other epidemiological data to improve estimates of the reporting probability. A

limitation was that I assumed a constant reporting probability ρ for all the analyses in this thesis. However, the reporting probability can change over time, for example when an outbreak is covered in the media or as a result of changes in the investments made in surveillance systems. In addition to temporal changes in reporting probability, there might be overdispersion in the reporting probability. This can be modelled using a negative binomial distribution, which was used for the analysis of the recent Ebola outbreak for example (Camacho et al., 2015).

6.2.6 Initial parameter values

A major issue of PMCMC and MCMC algorithms, in general, is the choice of initial parameter values. If the starting values are too far from the most probable values, then the MCMC chain will get stuck in regions of low likelihood. In Chapter 3, I used chain heating at the start of the MCMC to allow large jumps in parameter space. If prior information on the parameter values is available, for example from field studies or using values estimated using alternative inference methods, then these values can be used to narrow down the choice of initial parameter values.

Regardless of the method of choosing the initial parameter values, there is always the risk of convergence at a local optimum. Ideally, multiple chains of MCMC are run with different starting values. Due to computational and time constraints, I did not do this for all the simulation results presented in Chapter 3. For the poliovirus analyses, I ran two chains for the inferences using only epidemiological data, but not when phylogenetic data were used.

6.2.7 Optimisation

Beyond optimisations of the computational implementation, the actual algorithm for PMCMC can be modified to reduce computation time. The number of particles does not need to be fixed for the duration of particle filtering. For datasets with less overdispersion, or when the number of infections is large, the number of particles could be reduced as the set of plausible epidemic trajectories is less variable. This would

require an automated system of determining when to reduce particles, which I have not so far implemented.

In Chapter 5, I showed that fewer particles were needed if epidemic simulations began with a few hundred infected individuals. However, this was not done for simulations in Chapters 3 and 4 as these were focused in outbreak settings. In these settings, the first few generations of infections are highly stochastic as the initial number of infected individuals is only one. Not incorporating this stochasticity could lead to under-estimates of k and the reproductive number. Because low values of k correlates with high rates of extinction, simulating from a single infection penalises against values of k that are too low.

Besides the intensive marginal likelihood calculation using particle filtering, MCMC algorithms are generally difficult to implement because of low acceptance rates. There exist variations of the particle filter that forego MCMC altogether. The SMC^2 algorithm (Chopin et al., 2013) uses particle filtering not only to calculate the marginal likelihood but also to update the parameter values. Instead of an MCMC algorithm that carries out a random walk across parameter space, the SMC^2 algorithm uses particle filtering to generate samples of θ from $P(\theta)$, and then uses particle filtering as described in Chapter 2 to calculate the marginal likelihood of each sample of θ. The sample of θ is then updated by resampling with probabilities proportional to the marginal likelihoods $P(D|\theta)$. The resulting sample of θ is distributed according to $P(\theta|D)$. On the one hand, this approach is more parallelisable than PMCMC as the marginal likelihood calculation for each set of parameter values can be sent to a different node, and the marginal likelihood calculation itself can be parallelised across different cores of that node. However, this approach requires a sufficiently large sample of θ to cover the possible range of parameter values, which might be difficult if there are many parameters.

Both PMCMC and SMC^2 are computationally intensive. For simple epidemiological models with few parameters and data from a small number of time steps, Approximate Bayesian Computation (ABC) methods offer an alternative that requires less computational time. Sequential versions of ABC use particle filtering to estimate the posterior distribution of parameters $P(\theta|D)$ (Toni et al., 2009). In such algorithms, various parameter values sampled from the prior distribution $P(\theta)$ are used to generate epidemic simulations, which are compared to real data (time series or sequence data)

and the prior distribution accordingly updated. These steps are repeated until the prior distribution no longer changes. The resulting distribution should correspond to the posterior distribution $P(\theta|D)$.

6.3 Future directions

While the coalescent results are robust to changes in the assumptions about discrete generations and offspring distribution, violations of the evolutionary assumptions have more consequences with regards to parameter estimates. Various methods have been and are being developed to incorporate more complex evolutionary processes and population dynamics. In this section I present extensions that can be made to the inference framework presented in this thesis.

6.3.1 Population structure

In the environmental polio sequence analysis (Chapter 5), I fit an unstructured transmission model to incidence and phylogenetic data. Given the strong spatial structure in Pakistan, it would be interesting to co-estimate reproductive numbers for each province while taking into account migration between different locations. For epidemiological data, gravity and radiation models have been used to capture the spread of individuals in continuous two-dimensional space (Simini et al., 2012), and there are numerous compartmental models that capture movements between discrete locations. Similarly for genetic analysis, continuous phylogeography methods have been developed to reconstruct locations of ancestral lineages (Lemey et al., 2010; Guindon et al., 2016; Bouckaert, 2016). Discrete phylogeographic methods also exist (De Maio et al., 2015; Kühnert et al., 2016; Mueller et al., 2016) such as the implementation in BEAST2 (Lemey et al., 2009) that was used to analyse environmental polio sequences (Chapter 5). These rely on coalescent methods for structured populations (Volz, 2012) that estimate transition rates between different locations. While structured coalescent models have been fit to pathogen phylogeny (Rasmussen et al., 2014a,b), there is no study yet that simultaneously fits a structured transmission model to both epidemiological and

genetic data. Just as an integrated method can improve the accuracy and precision of parameter estimates, an inference framework that fits the same spatial model to both epidemiological and phylogenetic data could improve estimates of migration rates and epidemiological parameters for each discrete location.

6.3.2 Recombination

Another evolutionary force affecting the coalescent approach is frequent recombination, which is a common feature of a wide range of viral and bacterial pathogens (Awadalla, 2003). In the presence of recombination, phylogenies reconstructed from sequences are no longer meaningful and can lead to wrong parameter estimates. The analyses of poliovirus sequences were based on the assumption of no recombination. Even though many recombination hotspots have been detected in the polio genome, the VP1 gene is highly conserved and recombination events within this region are rare (Jorba et al., 2008). However, if the analyses were extended to the whole genome, then recombination events would need to be accounted for. The ancestral recombination graph (ARG) was developed to jointly estimate coalescent and recombination parameters (Griffths and Marjoram, 1997). Computational optimisations allowed application of ARG to large-scale genomic sequences (Rasmussen et al., 2014c). However, the use of ARG in infectious disease research is still limited. Thus, further developments in this area would allow the extension of phylodynamic inference methods to more complex data sets including bacterial data sets.

6.3.3 Selection

Besides recombination, positive and negative selection pressures can impact the relationship between demographic processes such as disease transmission and the pathogen phylogeny. This can lead to biased estimates of the prevalence of infection (Bedford et al., 2011). For other viruses where there is directional selection, e.g. influenza lineages adapting to escape the immune system, it is vital to account for these selection pressures to accurately recover the epidemiological history.

6.3.4 Within-host evolution

In addition to performing analyses with longer sequences, there is also a need to develop methods that exploit as many sequences as possible. For population studies, available sequences are often subsampled to remove individuals from the same household or in the same close contact network to have a representative sample of the population. Furthermore, sequences from the same individuals are often discarded, though these may be informative for within-host evolution. Although some effort has been made to link within-host to between-host evolution (Didelot et al., 2014; Vrancken et al., 2014), the effect of within-host evolution on population genetic inference is still not well-studied. Combining analyses across different scales could improve the accuracy of epidemiological predictions and provide better mechanistic explanations of observed trends.

6.3.5 Real-time estimation

While I showed that parameter estimation is possible using data from the whole outbreak, it would also be interesting to determine the smallest amount of data necessary to recover the parameter estimates. Using genomic data collected in North America during the swine flue outbreak, Hedge et al. (2013) found that accurate estimates of R_0 and T_{MRCA} of the tree could be obtained as early as May 2010, by which point 100 viral genomes had been sequenced. At the beginning of an outbreak, stochasticity plays a large role in determining the outcome of the emerging outbreak. Inference from data collected during these early stages might produce parameter estimates with greater uncertainty bounds as few cases are sampled. Joint analysis of epidemiological and genetic data would be useful in this case to produce more precise and accurate estimates of parameters, therefore it would be interesting to conduct a study using simulated data to see when joint analysis becomes useful.

6.3.6 Application to other pathogens

Besides wild-type polioviruses, vaccine-derived poliovirus (VDPV) can also cause paralysis due to reversion of the live attenuated virus in the oral poliovirus vaccine.

Although infections caused by VDPVs usually only spread to a few individuals before dying out, outbreaks can occur if VDPVs start circulating (cVPDV) in under-immunised populations. In Nigeria, for example, cVDPV2 has continuously been detected since 2005 (Burns et al., 2013). A joint analysis of cVDPV2 sequences with incidence time series from AFP surveillance of cVDPV2 cases could help to estimate the remaining pool of individuals infected with cVDPV2.

Most phylodynamic methods have been applied to viruses because of the rate of evolution is on a similar time-scale to outbreaks. However, whole-genome sequencing of bacterial isolates is becoming more widespread and can help to uncover genetic determinants of clinical severity, elucidate pathogen-host interactions and quantify evolutionary rates at within- and between-host levels Wilson (2012). Epidemiological investigations using bacterial genomes have also been possible even though bacteria acquire point mutations at a lower rate per base than viruses because longer bacterial genomes should provide sufficient genetic resolution for phylogenetic analysis. For example, whole-genome sequencing has been used to refine the tuberculosis transmission network built using contact information Gardy et al. (2011), an outbreak of MRSA in a hospital and surrounding community in near real-time Harris et al. (2013).

Regardless of the pathogen that is analysed, there is an increasing trend in epidemiological research to integrate different sources of information for inference. As demonstrated in this thesis and previous work, combining the analyses of different types of data such as epidemiological and genetic data can improve the accuracy of and reduce uncertainty in parameter estimates.

References

Alam, M. M., Sharif, S., Shaukat, S., Angez, M., Khurshid, A., Rehman, L., Zaidi, S. S. Z., 2016. Genomic surveillance elucidates persistent wild poliovirus transmission during 2013-2015 in major reservoir areas of Pakistan. Clinical Infectious Diseases, civ831.

Alam, M. M., Shaukat, S., Sharif, S., Angez, M., Khurshid, A., Malik, F., Rehman, L., Zaidi, S. S. Z., 2014. Detection of multiple cocirculating wild poliovirus type 1 lineages through environmental surveillance: impact and progress during 2011–2013 in Pakistan. Journal of Infectious Diseases 210 (suppl 1), S324–S332.

Anderson, R. M., May, R. M., 1991. Infectious diseases of humans: dynamics and control. Vol. 28. Oxford University Press, Oxford.

Andrieu, C., Doucet, A., Holenstein, R., 2010. Particle markov chain monte carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72 (3), 269–342.

Angez, M., Shaukat, S., Alam, M. M., Sharif, S., Khurshid, A., Zaidi, S. S. Z., 2012. Genetic relationships and epidemiological links between wild type 1 poliovirus isolates in Pakistan and Afghanistan. Virology Journal 9 (1), 1.

Asghar, H., Diop, O. M., Weldegebriel, G., Malik, F., Shetty, S., El Bassioni, L., Akande, A. O., Al Maamoun, E., Zaidi, S., Adeniji, A. J., et al., 2014. Environmental surveillance for polioviruses in the Global Polio Eradication Initiative. Journal of Infectious Diseases 210 (suppl 1), S294–S303.

Awadalla, P., 2003. The evolutionary genomics of pathogen recombination. Nature Reviews Genetics 4 (1), 50–60.

Baize, S., Pannetier, D., Oestereich, L., Rieger, T., Koivogui, L., Magassouba, N., Soropogui, B., Sow, M. S., Keïta, S., De Clerck, H., et al., 2014. Emergence of Zaire Ebola virus disease in Guinea—preliminary report. New England Journal of Medicine.

Becker, N. G., Britton, T., 1999. Statistical studies of infectious disease incidence. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61 (2), 287–307.

Bedford, T., Cobey, S., Pascual, M., 2011. Strength and tempo of selection revealed in viral gene genealogies. BMC Evolutionary Biology 11 (1), 1.

Bennett, S., Drummond, A., Kapan, D., Suchard, M., Munoz-Jordan, J., Pybus, O., Holmes, E., Gubler, D., 2010. Epidemic dynamics revealed in dengue evolution. Molecular Biology and Evolution 27 (4), 811–818.

Blake, I. M., Martin, R., Goel, A., Khetsuriani, N., Everts, J., Wolff, C., Wassilak, S., Aylward, R. B., Grassly, N. C., 2014. The role of older children and adults in wild poliovirus transmission. Proceedings of the National Academy of Sciences 111 (29), 10604–10609.

Bliss, C. I., Fisher, R. A., 1953. Fitting the negative binomial distribution to biological data. Biometrics 9 (2), 176–200.

Blumberg, S., Lloyd-Smith, J. O., 2013. Inference of R0 and transmission heterogeneity from the size distribution of stuttering chains. PLoS Computational Biology 9 (5), e1002993.

Boskova, V., Bonhoeffer, S., Stadler, T., 2014. Inference of epidemiological dynamics based on simulated phylogenies using birth-death and coalescent models. PLoS Computational Biology.

Bouckaert, R., 2016. Phylogeography by diffusion on a sphere: whole world phylogeography. PeerJ 4, e2406.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., Drummond, A. J., 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Computational Biology 10 (4), e1003537.

Burns, C. C., Shaw, J., Jorba, J., Bukbuk, D., Adu, F., Gumede, N., Pate, M. A., Abanida, E. A., Gasasira, A., Iber, J., et al., 2013. Multiple independent emergences of type 2 vaccine-derived polioviruses during a large outbreak in northern Nigeria. Journal of Virology 87 (9), 4907–4922.

Camacho, A., Kucharski, A., Aki-Sawyerr, Y., White, M. A., Flasche, S., Baguelin, M., Pollington, T., Carney, J. R., Glover, R., Smout, E., et al., 2015. Temporal changes in Ebola transmission in Sierra Leone and implications for control requirements: a real-time modelling study. PLoS Currents Outbreaks 7.

Casey, A. E., 1942. The incubation period in epidemic poliomyelitis. Journal of the American Medical Association 120 (11), 805–807.

Cauchemez, S., Fraser, C., Van Kerkhove, M. D., Donnelly, C. A., Riley, S., Rambaut, A., Enouf, V., van der Werf, S., Ferguson, N. M., 2014. Middle East respiratory syndrome coronavirus: quantification of the extent of the epidemic, surveillance biases, and transmissibility. The Lancet Infectious Diseases 14 (1), 50–56.

Centers for Disease Control and Prevention, 2010. Outbreaks following wild poliovirus importations–Europe, Africa and Asia, January 2009-September 2010. Morbidity and Mortality Weekly Report 59 (43), 1393–1399.

Centers for Disease Control and Prevention, 2012. Poliomyelitis, 12th Edition. Epidemiology and prevention of vaccine-preventable diseases. Public Health Foundation, Ch. 17, pp. 249–262.

Centers for Disease Control and Prevention, 2013. Progress toward eradication of polio-worldwide, January 2011-March 2013. MMWR. Morbidity and mortality weekly report 62 (17), 335.

Centers for Disease Control and Prevention, July 2014. Middle East Respiratory Virus (MERS).
URL `http://www.cdc.gov/coronavirus/mers/`

Centers for Disease Control and Prevention, 2016. Two cases of polio detected in Nigeria.

Chopin, N., Jacob, P. E., Papaspiliopoulos, O., 2013. SMC^2: an efficient algorithm for sequential analysis of state space models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75 (3), 397–426.

Cottam, E. M., Thébaud, G., Wadsworth, J., Gloster, J., Mansley, L., Paton, D. J., King, D. P., 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.

Darriba, D., Taboada, G. L., Doallo, R., Posada, D., 2012. jModelTest 2: more models, new heuristics and parallel computing. Nature Methods 9 (8), 772–772.

De Maio, N., Wu, C.-H., O'Reilly, K. M., Wilson, D., 2015. New routes to phylogeography: A Bayesian structured coalescent approximation. PLoS Genetics 11 (8), e1005421.

de Silva, E., Ferguson, N. M., Fraser, C., 2012a. Inferring pandemic growth rates from sequence data. Journal of The Royal Society Interface, rsif20110850.

de Silva, R., Gunasena, S., Ratnayake, D., Wickremesinghe, G., Kumarasiri, C., Pushpakumara, B., Deshpande, J., Kahn, A. L., Sutter, R. W., 2012b. Prevalence of prolonged and chronic poliovirus excretion among persons with primary immune deficiency disorders in Sri Lanka. Vaccine 30 (52), 7561–7565.

Didelot, X., Gardy, J., Colijn, C., 2014. Bayesian inference of infectious disease transmission from whole genome sequence data. Molecular Biology and Evolution, msu121.

Dowdle, W. R., Birmingham, M. E., 1997. The biologic principles of poliovirus eradication. Journal of Infectious Diseases 175 (Supplement 1), S286–S292.

Drummond, A. J., Ho, S. Y., Phillips, M. J., Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. PLoS Biology 4 (5), e88.

Drummond, A. J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology 7 (1), 214.

Drummond, A. J., Rambaut, A., Shapiro, B., Pybus, O. G., 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Molecular Biology and Evolution 22 (5), 1185–1192.

Drummond, A. J., Suchard, M. A., Xie, D., Rambaut, A., 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution 29 (8), 1969–1973.

Dudas, G., Rambaut, A., 2014. Phylogenetic analysis of Guinea 2014 EBOV Ebolavirus outbreak. PLoS Currents 6.

Duffy, S., Shackelton, L. A., Holmes, E. C., 2008. Rates of evolutionary change in viruses: patterns and determinants. Nature Reviews Genetics 9 (4), 267–276.

Dunn, G., Klapsa, D., Wilton, T., Stone, L., Minor, P. D., Martin, J., 2015. Twenty-eight years of poliovirus replication in an immunodeficient individual: Impact on the global polio eradication initiative. PLoS Pathog 11 (8), e1005114.

Dureau, J., Ballesteros, S., Bogich, T., 2013. SSM: inference for time series analysis with State Space Models. arXiv preprint arXiv:1307.5626.

Eichner, M., Brockmann, S. O., 2013. Polio emergence in syria and israel endangers europe. Lancet 382 (1777), 62220–5.

Eichner, M., Dietz, K., 1996. Eradication of poliomyelitis: when can one be sure that polio virus transmission has been terminated? American Journal of Epidemiology 143 (8), 816–822.

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., Tatem, A. J., Sousa, J. D., Arinaminpathy, N., Pépin, J., et al., 2014. The early spread and epidemic ignition of HIV-1 in human populations. Science 346 (6205), 56–61.

Faye, O., Boëlle, P.-Y., Heleze, E., Faye, O., Loucoubar, C., Magassouba, N., Soropogui, B., Keita, S., Gakou, T., Koivogui, L., et al., 2015. Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study. The Lancet Infectious Diseases 15 (3), 320–326.

Ferguson, N. M., Cummings, D. A., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., Iamsirithaworn, S., Burke, D. S., 2005. Strategies for containing an emerging influenza pandemic in Southeast Asia. Nature 437 (7056), 209–214.

Fine, P. E., Carneiro, I. A., 1999. Transmissibility and persistence of oral polio vaccine viruses: implications for the global poliomyelitis eradication initiative. American Journal of Epidemiology 150 (10), 1001–1021.

Fisher, R. A., 1930. The genetical theory of natural selection. Clarendon, Oxford.

Fraser, C., Donnelly, C. A., Cauchemez, S., Hanage, W. P., Van Kerkhove, M. D., Hollingsworth, T. D., Griffin, J., Baggaley, R. F., Jenkins, H. E., Lyons, E. J., et al., 2009. Pandemic potential of a strain of influenza A (H1N1): early findings. Science 324 (5934), 1557–1561.

Fraser, C., Li, L. M., 2017. Coalescent models for populations of with time-varying population sizes and arbitrary offspring distributions. bioRxiv doi: 10.1101/131730.

Fraser, C., Riley, S., Anderson, R. M., Ferguson, N. M., 2004. Factors that make an infectious disease outbreak controllable. Proceedings of the National Academy of Sciences of the United States of America 101 (16), 6146–6151.

Gamado, K. M., Streftaris, G., Zachary, S., 2013. Modelling under-reporting in epidemics. Journal of Mathematical Biology, 1–29.

Gardy, J. L., Johnston, J. C., Sui, S. J. H., Cook, V. J., Shah, L., Brodkin, E., Rempel, S., Moore, R., Zhao, Y., Holt, R., et al., 2011. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. New England Journal of Medicine 364 (8), 730–739.

Garske, T., Rhodes, C., 2008. The effect of superspreading on epidemic outbreak size distributions. Journal of Theoretical Biology 253 (2), 228–237.

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B., Suchard, M. A., 2013. Improving bayesian population dynamics inference: a coalescent-based model for multiple loci. Molecular biology and evolution 30 (3), 713–724.

Grassly, N. C., Fraser, C., Wenger, J., Deshpande, J. M., Sutter, R. W., Heymann, D. L., Aylward, R. B., 2006. New strategies for the elimination of polio from India. Science 314 (5802), 1150–1153.

Grassly, N. C., Harvey, P. H., Holmes, E. C., 1999. Population dynamics of HIV-1 inferred from gene sequences. Genetics 151 (2), 427–438.

Griffiths, R. C., Tavare, S., 1994. Sampling theory for neutral alleles in a varying environment. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 344 (1310), 403–410.

Griffths, R., Marjoram, P., 1997. An ancestral recombination graph. In: Donnelly, P., Tavaré, S. (Eds.), Progress in population genetics and human evolution. Vol. 87. Springer.

Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52 (5), 696–704.

Guindon, S., Guo, H., Welch, D., 2016. Demographic inference under the coalescent in a spatial continuum. Theoretical Population Biology 111, 43–50.

Gumede, N., Jorba, J., Deshpande, J., Pallansch, M., Yogolelo, R., Muyembe-Tamfum, J. J., Kew, O., Venter, M., Burns, C. C., 2014. Phylogeny of imported and reestablished wild polioviruses in the Democratic Republic of the Congo from 2006 to 2011. Journal of Infectious Diseases 210 (suppl 1), S361–S367.

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., et al., 2013. Whole-genome sequencing for analysis of an outbreak of methicillin-resistant *Staphylococcus aureus*: a descriptive study. The Lancet Infectious Diseases 13 (2), 130–136.

Harris, T. E., 2002. The theory of branching processes. Courier Corporation.

Hedge, J., Lycett, S., Rambaut, A., 2013. Real-time characterization of the molecular epidemiology of an influenza pandemic. Biology Letters 9 (5), 20130331.

Hovi, T., Shulman, L., Van der Avoort, H., Deshpande, J., Roivainen, M., De Gourville, E., et al., 2012. Role of environmental poliovirus surveillance in global polio eradication and beyond. Epidemiology and infection 140 (1), 1.

Hudson, R. R., 1991. Gene genealogies and the coalescent process. Oxford Surveys in Evolutionary Biology 7 (1-44).

Hussain, S. F., Boyle, P., Patel, P., Sullivan, R., 2016. Eradicating polio in pakistan: an analysis of the challenges and solutions to this security and health issue. Globalization and Health 12 (1), 63.

Imperial College High Performance Computing Service, 2016. http://www.imperial.ac.uk/admin-services/ict/self-service/research-support/hpc/.

International Ebola Response Team, Agua-Agum, J., Ariyarajah, A., Aylward, B., Bawo, L., Bilivogui, P., Blake, I. M., Brennan, R. J., Cawthorne, A., Cleary, E., Clement, P., Conteh, R., Cori, A., Dafae, F., Dahl, B., Dangou, J.-M., Diallo, B., Donnelly, C. A., Dorigatti, I., Dye, C., Eckmanns, T., Fallah, M., Ferguson, N. M., Fiebig, L., Fraser, C., Garske, T., Gonzalez, L., Hamblion, E., Hamid, N., Hersey, S., Hinsley, W., Jambei, A., Jombart, T., Kargbo, D., Keita, S., Kinzer, M., George, F. K., Godefroy, B., Gutierrez, G., Kannangarage, N., Mills, H. L., Moller, T., Meijers, S., Mohamed, Y., Morgan, O., Nedjati-Gilani, G., Newton, E., Nouvellet, P., Nyenswah, T., Perea, W., Perkins, D., Riley, S., Rodier, G., Rondy, M., Sagrado, M., Savulescu, C., Schafer, I. J., Schumacher, D., Seyler, T., Shah, A., Van Kerkhove, M. D., Wesseh, C. S., Yoti, Z., 11 2016. Exposure patterns driving Ebola transmission in West Africa: A retrospective observational study. PLOS Medicine 13 (11), 1–23.

Jombart, T., Cori, A., Didelot, X., Cauchemez, S., Fraser, C., Ferguson, N., 2014. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLoS Computational Biology 10 (1), e1003457.

Jorba, J., Campagnoli, R., De, L., Kew, O., 2008. Calibration of multiple poliovirus molecular clocks covering an extended evolutionary range. Journal of Virology 82 (9), 4429–4440.

Jukes, T. H., Cantor, C. R., 1969. Evolution of protein molecules. Mammalian Protein Metabolism 3 (21), 132.

Kalkowska, D. A., Tebbens, R. J. D., Pallansch, M. A., Cochi, S. L., Wassilak, S. G., Thompson, K. M., 2015. Modeling undetected live poliovirus circulation after apparent interruption of transmission: implications for surveillance and vaccination. BMC Infectious Diseases 15 (1), 66.

Kendall, M., Colijn, C., 2015. A tree metric using structure and length to capture distinct phylogenetic signals. arXiv preprint arXiv:1507.05211.

Kermack, W. O., McKendrick, A. G., 1927. A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences 115 (772), 700–721.

Kew, O. M., Mulders, M. N., Lipskaya, G. Y., da Silva, E. E., Patlansch, M. A., 1995. Molecular epidemiology of polioviruses. In: Seminars in Virology. Vol. 6. Elsevier, pp. 401–414.

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.

King, A. A., Ionides, E. L., Bretó, C. M., Ellner, S. P., Ferrari, M. J., Kendall, B. E., Lavine, M., Nguyen, D., Reuman, D. C., Wearing, H., Wood, S. N., 2016a. pomp: Statistical Inference for Partially Observed Markov Processes. R package, version 1.10. URL `http://kingaa.github.io/pomp`

King, A. A., Nguyen, D., Ionides, E. L., 2016b. Statistical inference for Partially Observed Markov processes via the **R** package `pomp`. Journal of Statistical Software 69 (i12).

Kingman, J., 1982a. Exchangeability and the evolution of large populations.

Kingman, J. F., 1982b. On the genealogy of large populations. Journal of Applied Probability, 27–43.

Koelle, K., Rasmussen, D. A., 2012. Rates of coalescence for common epidemiological models at equilibrium. Journal of The Royal Society Interface 9 (70), 997–1007.

Kühnert, D., Stadler, T., Vaughan, T. G., Drummond, A. J., 2014. Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth–death SIR model. Journal of the Royal Society Interface 11 (94), 20131106.

Kühnert, D., Stadler, T., Vaughan, T. G., Drummond, A. J., 2016. Phylodynamics with migration: A computational framework to quantify population structure from genomic data. Molecular biology and evolution, msw064.

Lekone, P. E., 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.

Lemey, P., Rambaut, A., Drummond, A. J., Suchard, M. A., 2009. Bayesian phylogeography finds its roots. PLoS Comput Biol 5 (9), e1000520.

Lemey, P., Rambaut, A., Welch, J. J., Suchard, M. A., 2010. Phylogeography takes a relaxed random walk in continuous space and time. Molecular biology and evolution 27 (8), 1877–1885.

Leventhal, G. E., Günthard, H. F., Bonhoeffer, S., Stadler, T., 2014. Using an epidemiological model for phylogenetic inference reveals density dependence in HIV transmission. Molecular Biology and Evolution 31 (1), 6–17.

Li, L. M., Grassly, N. C., Fraser, C., 2014. Genomic analysis of emerging pathogens: methods, application and future trends. Genome biology 15 (11), 1.

Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., Getz, W. M., 2005. Superspreading and the effect of individual variation on disease emergence. Nature 438 (7066), 355–359.

Lukashev, A. N., 2005. Role of recombination in evolution of enteroviruses. Reviews in medical virology 15 (3), 157–167.

Magiorkinis, G., Sypsa, V., Magiorkinis, E., Paraskevis, D., Katsoulidou, A., Belshaw, R., Fraser, C., Pybus, O. G., Hatzakis, A., 2013. Integrating phylodynamics and epidemiology to estimate transmission diversity in viral epidemics. PLoS Computational Biology 9 (1), e1002876.

Mangal, T. D., Aylward, R. B., Grassly, N. C., 2013. The potential impact of routine immunization with inactivated poliovirus vaccine on wild-type or vaccine-derived poliovirus outbreaks in a posteradication setting. American Journal of Epidemiology, kwt203.

Massey Jr, F. J., 1951. The Kolmogorov-Smirnov test for goodness of fit. Journal of the American statistical Association 46 (253), 68–78.

Meligkotsidou, L., Fearnhead, P., 2007. Postprocessing of genealogical trees. Genetics 177 (1), 347–358.

Melnick, J. L., Ledinko, N., 1953. Development of neutralizing antibodies against the three types of poliomyelitis virus during an epidemic period: The ratio of inapparent infection to clinical poliomyelitis. American Journal of Epidemiology 58 (2), 207–222.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E., 2004. Equation of state calculations by fast computing machines. The Journal of Chemical Physics 21 (6), 1087–1092.

Minor, P. D., 2004. Polio eradication, cessation of vaccination and re-emergence of disease. Nature Reviews Microbiology 2 (6), 473–482.

Möhle, M., 1998. Robustness results for the coalescent. Journal of Applied Probability, 438–447.

Mueller, N. F., Rasmussen, D. A., Stadler, T., 2016. The structured coalescent and its approximations. bioRxiv, 091058.

Murray, L. M., 2013. Bayesian state-space modelling on high-performance hardware using libbi. arXiv preprint arXiv:1306.3277.

Nathanson, N., Kew, O. M., 2010. From emergence to eradication: the epidemiology of poliomyelitis deconstructed. American journal of Epidemiology 172 (11), 1213–1229.

Nathanson, N., Martin, J. R., 1979. The epidemiology of poliomyelitis: enigmas surrounding its appearance epidemicity and disappearance. American Journal of Epidemiology 110 (6), 672–92.

Notohara, M., 1990. The coalescent and the genealogical process in geographically structured population. Journal of mathematical biology 29 (1), 59–75.

Paul, J. R., 1955. Epidemiology of poliomyelitis. Monograph Series. World Health Organization 26, 9.

Penttinen, K., Patiala, R., 1961. The paralytic/infected ratio in a susceptible population during a polio type I epidemic. In: Annales medicinae experimentalis et biologiae Fenniae. Vol. 39. p. 195.

Public Health England, 2016. PHE national polio guidelines: Local and regional services. Public Health England.

Pybus, O. G., Charleston, M. A., Gupta, S., Rambaut, A., Holmes, E. C., Harvey, P. H., 2001. The epidemic behavior of the hepatitis C virus. Science 292 (5525), 2323–2325.

Pybus, O. G., Rambaut, A., Harvey, P. H., 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics 155 (3), 1429–1437.

Racaniello, V. R., Baltimore, D., 1981. Molecular cloning of poliovirus cDNA and determination of the complete nucleotide sequence of the viral genome. Proceedings of the National Academy of Sciences 78 (8), 4887–4891.

Rambaut, A., Grassly, N. C., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Computer Applications In The Biosciences 13 (3), 235–238.

Rasmussen, D. A., Boni, M. F., Koelle, K., 2014a. Reconciling phylodynamics with epidemiology: the case of dengue virus in southern Vietnam. Molecular Biology and Evolution 31 (2), 258–271.

Rasmussen, D. A., Ratmann, O., Koelle, K., 2011. Inference for nonlinear epidemiological models using genealogies and time series. PLoS Computational Biology 7 (8), e1002136–e1002136.

Rasmussen, D. A., Volz, E. M., Koelle, K., 2014b. Phylodynamic inference for structured epidemiological models. PLoS Computional Biology 10 (4), e1003570.

Rasmussen, M. D., Hubisz, M. J., Gronau, I., Siepel, A., 2014c. Genome-wide inference of ancestral recombination graphs. PLoS Genetics 10 (5), e1004342.

Ren, R., Racaniello, V. R., 1992. Poliovirus spreads from muscle to the central nervous system by neural pathways. Journal of Infectious Diseases 166 (4), 747–752.

Riley, S., Fraser, C., Donnelly, C. A., Ghani, A. C., Abu-Raddad, L. J., Hedley, A. J., Leung, G. M., Ho, L.-M., Lam, T.-H., Thach, T. Q., et al., 2003. Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. Science 300 (5627), 1961–1966.

Roberts, G. O., Rosenthal, J. S., et al., 2001. Optimal scaling for various Metropolis-Hastings algorithms. Statistical Science 16 (4), 351–367.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., Huelsenbeck, J. P., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61 (3), 539–542.

Sabin, A. B., 1957. Properties and behavior of orally administered attenuated poliovirus vaccine. Journal of the American Medical Association 164 (11), 1216–1223.

Salk, D., Van Wezel, A., Salk, J., 1984. Induction of long-term immunity to paralytic poliomyelitis by use of non-infectious vaccine. The Lancet 324 (8415), 1317–1321.

Shaw, D., Grenfell, B., Dobson, A., 1998. Patterns of macroparasite aggregation in wildlife host populations. Parasitology 117 (06), 597–610.

Sheinson, D. M., Niemi, J., Meiring, W., 2014. Comparison of the performance of particle filter algorithms applied to tracking of a disease epidemic. Mathematical Biosciences 255, 21–32.

Simini, F., González, M. C., Maritan, A., Barabási, A.-L., 2012. A universal model for mobility and migration patterns. Nature 484 (7392), 96–100.

Smith, G. J., Vijaykrishna, D., Bahl, J., Lycett, S. J., Worobey, M., Pybus, O. G., Ma, S. K., Cheung, C. L., Raghwani, J., Bhatt, S., et al., 2009. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature 459 (7250), 1122–1125.

Stadler, T., 2010. Sampling-through-time in birth–death trees. Journal of Theoretical Biology 267 (3), 396–404.

Stadler, T., Bonhoeffer, S., 2013. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. Philosophical Transactions of the Royal Society B: Biological Sciences 368 (1614), 20120198.

Stadler, T., Kühnert, D., Bonhoeffer, S., Drummond, A. J., 2013. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proceedings of the National Academy of Sciences 110 (1), 228–233.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30 (9), 1312–1313.

Sutter, R. W., Patriarca, P. A., Suleiman, A. J. M., Brogan, S., Malankar, P. G., Cochi, S. L., Al-Ghassani, A. A., El-Bualy, M. S., 1992. Attributable risk of DTP (diphtheria and tetanus toxoids and pertussis vaccine) injection in provoking paralytic poliomyelitis during a large outbreak in Oman. Journal of Infectious Diseases 165 (3), 444–449.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., Stumpf, M. P., 2009. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of the Royal Society Interface 6 (31), 187–202.

Volz, E. M., 2012. Complex population dynamics and the coalescent under neutrality. Genetics 190 (1), 187–201.

Volz, E. M., Frost, S. D., 2014. Sampling through time and phylodynamic inference with coalescent and birth–death models. Journal of The Royal Society Interface 11 (101), 20140945.

Volz, E. M., Ionides, E., Romero-Severson, E. O., Brandt, M.-G., Mokotoff, E., Koopman, J. S., 2013. HIV-1 transmission during early infection in men who have sex with men: a phylodynamic analysis. PLoS Medicine 10 (12), e1001568.

Volz, E. M., Koopman, J. S., Ward, M. J., Brown, A. L., Frost, S. D., 2012. Simple epidemiological dynamics explain phylogenetic clustering of HIV from patients with recent infection. PLoS Computational Biology 8 (6), e1002552.

Volz, E. M., Pond, S., 2014. Phylodynamic analysis of Ebola virus in the 2014 Sierra Leone epidemic. PLoS Currents Outbreaks.

Volz, E. M., Pond, S. L. K., Ward, M. J., Brown, A. J. L., Frost, S. D., 2009. Phylodynamics of infectious disease epidemics. Genetics 183 (4), 1421–1430.

Vrancken, B., Rambaut, A., Suchard, M. A., Drummond, A., Baele, G., Derdelinckx, I., Van Wijngaerden, E., Vandamme, A.-M., Van Laethem, K., Lemey, P., 2014. The genealogical population dynamics of HIV-1 in a large transmission chain: Bridging within and among host evolutionary rates. PLoS Computational Biology 10 (4), e1003505.

Wilson, D. J., 2012. Insights from genomics into bacterial pathogen populations. PLoS Pathogens 8 (9), e1002874.

World Health Organization, 2004. Polio laboratory manual.

World Health Organization, 2013. Polio Eradication & Endgame Strategic Plan, 2013–2018. WHO Press.

World Health Organization, 2016. Poliomyelitis.
URL `http://www.who.int/mediacentre/factsheets/fs114/en/`

Wright, S., 1931. Evolution in Mendelian populations. Genetics 16 (2), 97.

Wringe, A., Fine, P. E., Sutter, R. W., Kew, O. M., 2008. Estimating the extent of vaccine-derived poliovirus infection. PLoS One 3 (10), e3433.

Yakovenko, M., Gmyl, A., Ivanova, O., Eremeeva, T., Ivanov, A., Prostova, M., Baykova, O., Isaeva, O., Lipskaya, G., Shakaryan, A., et al., 2014. The 2010 outbreak of poliomyelitis in Tajikistan: epidemiology and lessons learnt. Euro Surveillance 19 (7), 20706.

Ypma, R. J., van Ballegooijen, W. M., Wallinga, J., 2013. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. Genetics 195 (3), 1055–1062.

Appendices

Appendix A

Review paper in Genome Biology (Li et al. 2014)

Li et al. Genome Biology 2014, 15:541
http://genomebiology.com/2014/15/11/541

Genome Biology

REVIEW

Genomic analysis of emerging pathogens: methods, application and future trends

Lucy M Li*, Nicholas C Grassly and Christophe Fraser

Abstract

The number of emerging infectious diseases is increasing. Characterizing novel or re-emerging infections is aided by the availability of pathogen genomes. In this review, we evaluate methods that exploit pathogen sequences and the contribution of genomic analysis to understand the epidemiology of recently emerged infectious diseases.

Introduction

When a pathogen crosses over from animals to humans, or an existing human disease suddenly increases in incidence, the infectious disease is said to be 'emerging'. The number of emerging infectious diseases (EIDs) has increased over the last few decades, driven by both anthropogenic and environmental factors [1]. These include the expansion of agricultural land, which increases the exposure of livestock and humans to infections in wildlife [2]; a greater volume of air traffic, enabling EIDs to rapidly spread across the world [3,4]; and climate change, which alters the ecology and density of animal vectors, thereby introducing diseases to new geographic locations [5]. Novel strains of existing pathogens also have the potential to cause large epidemics. The over- and misuse of antimicrobial drugs have contributed to the growing number of drug-resistant pathogen strains [6,7].

Detecting, characterizing and responding to an EID requires co-ordination and collaboration between multiple sectors and disciplines. Laboratory-based research helps to characterize the pathogen and its interactions with host cells, but is less useful for quantitative understanding of population-level disease dynamics. Modeling approaches enable a large number of hypotheses to be tested, which might not be logistically or ethically feasible in laboratory and field experiments. In addition to characterizing past disease dynamics, modeling future trends informs decisions regarding outbreak response and resource allocation [8]. Modeling plays an especially important role in epidemiological studies of infectious disease spread, because the transmission of infectious disease between individuals is not directly observable. At the individual level, transmission times and who infected whom are typically unknown. And at the population level, disease burden needs to be inferred from observable data. Important public health questions such as how quickly an epidemic spreads and how many people will be infected are hard to quantify without a mechanistic understanding of underlying factors driving disease transmission. By expressing disease spread in mathematical terms, statistical properties of epidemics can be estimated to help address specific questions regarding disease spread and control efforts [9].

Another discipline contributing to the study of EIDs is pathogen genomics. As sequencing technology has become more accessible and affordable, genetic analysis has played an increasingly important role in infectious disease research. Sequencing pathogens can confirm suspected cases of an infectious disease, discriminate between different strains, and classify novel pathogens. In addition to examining individual pathogen sequences, multiple sequences can be analyzed together using phylogenetic methods to elucidate evolutionary [10] and transmission [11] history. Just as mathematical models of disease transmission help to capture the epidemiological properties of an infectious disease, modeling the molecular evolution of pathogen genomes is important for phylogenetic methods.

Besides characterizing the genetics and evolution of a pathogen, mathematical models used in population genetics link demographic and evolutionary processes to temporal changes in population-level genetic diversity. The coalescent population genetics framework was developed so that demographic history could be inferred from the shape of the genealogy linking sampled individuals [12,13]. More recently, the birth-death model has been applied to infectious diseases to infer epidemiological history from a genealogy [14,15]. Given the link between pathogen

* Correspondence: mengqi.li09@imperial.ac.uk
Department of Infectious Disease Epidemiology, Imperial College London, St Mary's Campus, London W2 1PG, UK

evolution and disease transmission, there is a trend towards integrating both epidemiologic and genetic data in the same analytical framework [16-18].

In this review, we provide an overview of recent developments in genomic methods in the context of infectious diseases, evaluate integrative methods that incorporate genetic data in epidemiological analysis, and discuss the application of these methods to EIDs.

Role of genetics in studying infectious diseases

Over the last two decades, sequence data have increased in quality, length and volume due to improvements in the underlying technology and decreasing costs. As a result, pathogen sequences are regularly collected during routine surveillance and clinical studies. Just as mathematical modeling can be used to analyze surveillance data to reveal details of disease transmission (Box 1), analysis of pathogen genomes employs mathematical frameworks to elucidate pathogen biology, evolution and ecology (Figure 1).

At the most basic level, mathematical models are used to find the optimal alignment of pathogen sequences. Multiple sequence alignment is useful for finding highly conserved or variable regions, shedding light on the molecular biology of the pathogen. Furthermore, coupling sequences with clinical information can help identify the contribution of polymorphic sites to disease. Revealing the evolutionary history of a pathogen requires a quantitative description of relatedness. Based on polymorphic sites in the sequence alignment, a model of sequence evolution is then used to reconstruct the phylogeny [19]. Often, there is insufficient genetic diversity in the sample to fully infer the phylogeny without ambiguity. In such a case, it is useful to consider a tree as an unknown set of

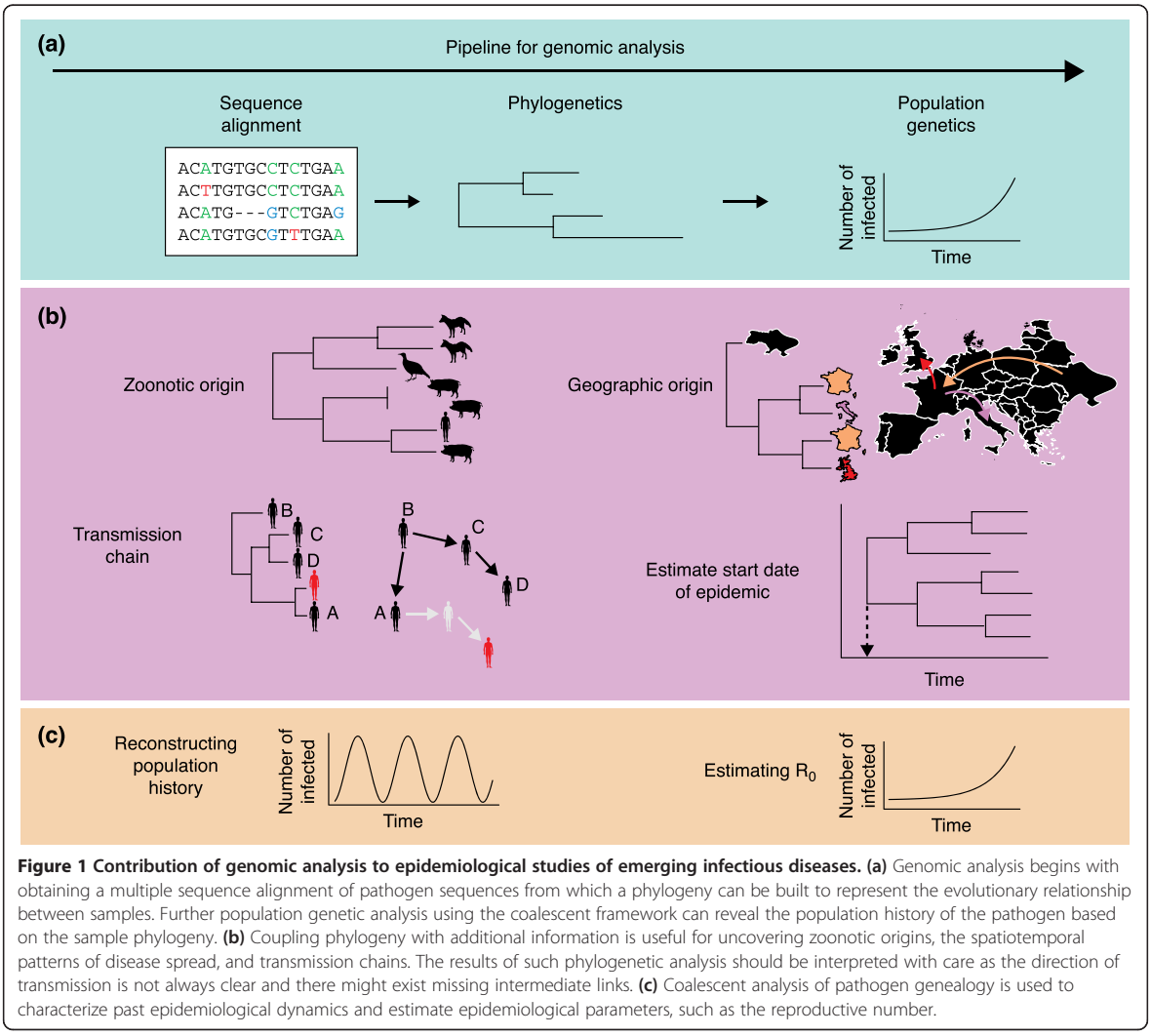


Figure 1 Contribution of genomic analysis to epidemiological studies of emerging infectious diseases. (a) Genomic analysis begins with obtaining a multiple sequence alignment of pathogen sequences from which a phylogeny can be built to represent the evolutionary relationship between samples. Further population genetic analysis using the coalescent framework can reveal the population history of the pathogen based on the sample phylogeny. **(b)** Coupling phylogeny with additional information is useful for uncovering zoonotic origins, the spatiotemporal patterns of disease spread, and transmission chains. The results of such phylogenetic analysis should be interpreted with care as the direction of transmission is not always clear and there might exist missing intermediate links. **(c)** Coalescent analysis of pathogen genealogy is used to characterize past epidemiological dynamics and estimate epidemiological parameters, such as the reproductive number.

parameters and obtain its posterior probability distribution using a Bayesian framework, such as the Markov Chain Monte Carlo (MCMC) approaches [20,21].

Biological samples from which pathogen genetic material is sequenced are usually associated with geographic or temporal information (Figure 1b). When this additional information is available, phylogenetic methods can reveal the spatiotemporal spread of the pathogen in the population. If an outbreak is densely sampled, then the pathogen phylogeny provides information about the underlying transmission network and helps to uncover who infected whom [22,23], though phylogenetic clustering alone is usually not sufficient to prove direct transmission or direction of infection (Figure 1b).

Incorporating sampling times helps to convert a phylogeny specified in units of nucleotide substitutions to a phylogeny specified in units of time [24]. The conversion is straightforward if sequence evolution follows a strict molecular clock, whereby the rate of substitution remains constant over time. However, selection pressure and population bottlenecks can lead to changes in the rate of substitution [25]. More flexible models have been developed to incorporate time-varying rates of evolution [26,27]. With branch lengths in units of real time, the start date of an epidemic can be estimated. Whereas phylogenetics aims to delineate the relationship between individuals, population genetics aims to link population processes to observed patterns of genetic diversity. Inferences regarding pathogen population history are based on the genealogy, or ancestry, of sequences from sampled individuals, and often carried out in a retrospective population genetics framework known as the coalescent [12] (Box 2). A genealogy describes the ancestry of sampled individuals. Going backwards in time, pairs of lineages coalesce when they share a common ancestor, until the last two lineages coalesce at the time of the most recent common ancestor (TMRCA) for the entire sample.

Since the turn of the century, the coalescent has been increasingly applied to infectious disease research to infer epidemic history from pathogen sequences, thereby linking pathogen evolutionary history to disease epidemiology (Figure 1c). The method is especially useful for analyzing infectious diseases with mild or asymptomatic infections, for which case-based surveillance data severely underestimate prevalence, because the coalescent assumes a small sample compared to the population size [28-30].

Other approaches have been developed to make epidemiological inferences from genetic data. Of particular note is the birth-death model [31], which describes the rates of transmissions, recoveries and deaths, and sampling events in terms of the sample genealogy [14]. Just as there are coalescent methods incorporating population structure [32-34] and compartmental models [35-37], similar methods exist in the birth-death framework [38,39]. Unlike the coalescent framework, the birth-death model is still valid for densely sampled populations, which makes it more useful for studying small outbreaks. However, accurately inferring epidemiological parameters depends on correctly specified sampling proportions [40]. Although the two approaches are methodologically different, both aim to reconstruct pathogen population history and produce estimates of epidemiological parameters, such as the reproductive number (R_0). The focus on the coalescent framework in this review is due to its more pervasive use in the literature and its greater versatility when integrated with epidemiological models compared to birth-death models.

Because of the simplistic assumptions of population genetics models, the population size inferred using coalescent-based methods cannot be directly interpreted as pathogen population size (prevalence of infection). It is rather the effective population size, N_e (Box 2), which refers to the size of a Wright-Fisher population that would produce the same level of genetic diversity as observed in the sample. In real populations, the variance of the offspring distribution (Box 1) is higher than expected in a Wright-Fisher population due to heterogeneity in host infectiousness, non-random mixing of the population, and migration events. The consequence of a large variance is that there is a greater discrepancy between the effective and census population sizes [41]. Accounting for the dispersion of the offspring distribution is especially important when analyzing infectious disease data because of the widespread occurrence of transmission heterogeneity [42].

Another statistical property of epidemics affecting the results of modeling studies is the generation time distribution, which describes the time between infection of the primary case and of secondary cases. Obtaining an estimate of the generation time is important for two reasons. First, estimates of R_0 from the initial growth rate of an epidemic depend on the generation time distribution [43]. As R_0 is the mean of the offspring distribution, its value affects the relationship between the effective population size, N_e, and the census population size, N. Second, the coalescent model was originally specified in units of generations, and so estimates in this framework need to be converted to natural units using the generation time, T_g.

Because transmission events are rarely observed, the generation time distribution is often approximated by the distribution of the serial interval, which is the time between onset of symptoms in the primary and secondary cases. The two distributions generally share the same mean but might have different variances [44]. Furthermore, the observed generation time decreases as the epidemic grows but increases again after the epidemic peak due to right censoring [45].

Integrating genetics with other data

As both sequence and surveillance data contain information regarding the transmission process, simultaneously analyzing both datasets should yield more accurate estimates of epidemiological parameters than separate analyses [17]. The recently established discipline of phylodynamics takes an interdisciplinary approach to understand the pathogen phylogenetics and epidemiology in terms of disease transmission.

Most efforts thus far have focused on enhancing phylogenetic and population genetic analyses by incorporating spatial and temporal information about the sequences. The molecular clock model assumes a constant rate of evolution and thus helps to estimate the time of the most recent common ancestor of the sample, which approximates the start date of an epidemic. Molecular clock analysis has been used to date the emergence of a range of emerging pathogens from HIV [46] to multidrug-resistant *Streptococcus pneumoniae* [47].

Linking geographic information with sequences can reveal the spatial spread of infectious disease. Phylogenetic reconstruction of seasonal influenza (H3N2) sequences has revealed the contribution of viral circulation in temperate regions to the global genetic diversity of influenza, and determined that not all epidemics in temperate regions are seeded by strains from South East Asia [48,49]. Also using global sequences, hepatitis C virus (HCV) subtypes were shown to spread from developed to developing countries [50]. Finally, phylogeographic analysis of methicillin-resistant *Staphylococcus aureus* samples identified England as the source of the EMRSA-15 lineage [51].

By contrast, there have been relatively few studies incorporating genetic data into epidemiological frameworks. Although genetic analysis plays an important role in elucidating transmission links in disease outbreaks [20,21,52], its integration with epidemiological models to understand population-level disease dynamics has been more limited. In one of the first papers to link coalescent inference to mathematical models in epidemiology, the effective population sizes of HIV-1 subtypes A and B were estimated from the maximum likelihood trees of viral sequences [53]. In addition to revealing population sizes, Pybus *et al*. [54] estimated the R_0 values of HCV subtypes (1a, 1b, 4 and 6) by inferring the epidemic growth rate from viral genealogy. Taking integration a step further, the coalescent process has been described for compartmental epidemiological models such as the Susceptible-Infected-Recovered (SIR) model, thereby enabling epidemiological parameters to be inferred from the genealogy [35]. To infer demographic history from both pathogen genomes and epidemiological data, Rasmussen *et al*. [17] developed a Markovian framework in which the population size at each time step was estimated by taking into account both the surveillance data and the genealogy. The epidemic history reconstructed using both datasets was more accurate than when analyzing each type of data separately.

In all the above methods, the genealogy of the sampled sequences was fixed. However, there might be great uncertainty regarding the order and the timing of coalescence, especially if the sequences are sampled within a short time period. While genealogical reconstruction using Bayesian MCMC approaches allows phylogenetic uncertainty to be incorporated into estimates of population size [13,31], an integrative model is lacking in which uncertainties arising from both genetic and epidemiological data are incorporated during demographic reconstruction.

Application to emerging pathogens

Models of pathogen evolution and mechanistic models of disease spread have increased in complexity. There is also greater computational power to test these models with data. However, these sophisticated models have mostly been applied to infectious diseases for which abundant data are available. For example, new methods are most often tested on the HIV-1 pandemic [15,34,35,55], for which data have been extensively collected from various settings and sources since the virus was first characterized three decades ago. It is worthwhile to evaluate how genomic methods have been applied to other diseases that have emerged more recently. In this section, we will present three case studies of recently emerged infectious diseases to illustrate the power and shortcomings of genomic methods discussed in this review.

Ebola virus emergence in West Africa

Since emerging in Guinea in March 2014, Ebola virus (EBOV) has spread to other countries in Western Africa, resulting in the largest outbreak of Ebola since it was first identified in 1976. The first viral genomes were made available just a month after alarm was raised about a new Ebola outbreak in Guinea [56], with further sequences collected in Sierra Leone [57]. By aligning all the genomes, a number of polymorphic sites were identified, including eight in highly conserved regions of the genome. Further association studies are needed to clarify the role of these genetic variants in determining disease outcome. Using the sampling dates of the sequences and a molecular clock model, phylogenetic analysis of 81 EBOV sequences revealed a start date of February 2014 in Guinea, spreading to Sierra Leone by April 2014 [57].

Uncovering the relationship between the 2014 EBOV lineage and previous EBOV outbreaks has proved trickier than understanding the disease dynamics during the 2014 outbreak. Initial phylogenetic analysis suggested that lineages causing the present outbreak did not cluster with EBOV strains that caused earlier outbreaks in Central Africa [56]. However, Dudas and Rambaut [58]

noted that the divergence of Guinea sequences from those of previous outbreaks was because they were sequenced most recently and had accumulated the highest number of substitutions. Assuming that the EBOV genome followed a molecular clock model, the authors re-rooted the tree to a lineage that caused an outbreak in 1976 [58]. Instead of silently circulating in West Africa, the EBOV lineage causing the current outbreak likely descended from a lineage that previously caused outbreaks in the Democratic Republic of Congo.

These studies highlight two issues. First, correct rooting of a phylogeny is important for accurate inference of past epidemic history. Correct rooting can be achieved by using an out-group, but one was not available in the case of this EBOV strain. This leads onto the second issue. Without sequences from animal hosts, the mechanism by which EBOV was sustained between outbreaks remains unknown.

Middle East respiratory syndrome coronavirus

Middle East respiratory syndrome coronavirus (MERS-CoV) first appeared in Saudi Arabia in 2012, and has since been reported in several neighboring countries in the Arabian Peninsula and on other continents [59].

Despite the dearth of sequence data, coalescent-based analysis of 10 genomic sequences produced estimates of the TMRCA (March 2012; 95% confidence interval (CI): November 2011 to June 2012), R_0 (1.21; 95% CI: 1.08, 1.40), and doubling time (43 days; 95% CI: 23, 104 days) [60]. Without further sequencing of the animal reservoirs, the authors could not infer whether these estimates applied to the animal reservoir or the human epidemic, because the methods are agnostic as to where transmission and evolution occur. The credible intervals around the estimates were unsurprisingly large given the small sample size.

Unlike the 2014 EBOV outbreak, which is sustained by human-to-human transmission [57], there appears to have been multiple introductions of MERS-CoV into the human population. Identification of the animal reservoir is therefore crucial for establishing risk factors of infection and planning appropriate interventions to control the disease. Since bats are reservoirs for other coronaviruses, their being a reservoir host is possible. A 182-nucleotide-long region of the RNA-dependent RNA polymerase gene was found to be 100% identical between a viral sample from a patient in Saudi Arabia and from a bat nearby, though the region is known to be highly conserved [61]. However, antibodies against human MERS-CoV have been detected in dromedary camels [62], the camel MERS-CoV genome is similar to human MERS-CoV [62], and there are reports of close contact between patients and camels [63]. Phylogenetic analysis of coronavirus sequences from bats, dromedaries and humans indicate a bat origin, with dromedary camel as an intermediate host [64]. It is possible that there are other animal reservoirs not yet sampled, which highlights the need to carry out extensive animal surveillance to characterize the emergence of an infection in humans.

Unraveling the complex evolutionary history of pandemic H1N1 influenza

With sequences collected over three decades from humans, pigs and birds, the origin of the pandemic H1N1 influenza A strain (pdmH1N1 or 'swine flu') was elucidated soon after emergence. Within two months of the first reported case of swine flu in humans, genomic analysis of the novel influenza strain had been carried out. A phylogeny was constructed for each of the eight genomic segments with sequences from humans, swine and birds. Comparison of these eight phylogenies revealed a complex history of reassortment with a mixture of gene segments from all three groups. The start of the pandemic was estimated to be the end of 2008 or early 2009, and the dates of the reassortment events leading to pdmH1N1 were also obtained [10]. Without good surveillance of influenza in the animal reservoir, the origin of the novel strain would have been difficult to uncover.

By analyzing 11 hemagglutinin sequences collected over a one-month period, the start date of the epidemic was estimated to be in late January 2009 [65]. Repeating the phylogenetic and molecular clock analyses with a further 12 sequences shifted the estimated start date two weeks earlier. Fitting an exponential growth model to the sequence data, R_0 was estimated to be 1.22, slightly lower than inferred from epidemiological data but with overlapping confidence intervals.

To determine at which point during the pandemic coalescent analysis would have provided accurate and precise estimates of evolutionary rate, R_0 and TMRCA, real-time estimates of these parameters were obtained for genomic sequences collected in North America [66]. Accurate estimates could have been obtained as early as May, when 100 viral genomes had been sequenced. More precise estimates could have been obtained by the end of June, when 164 had been sequenced. However, inclusion of more sequences of longer length only slightly improved the accuracy of initial estimates [66].

Future directions

Most statistical models in population genetics have focused on the application of such methods to viruses, although this bias is perhaps unsurprising given the large proportion of EIDs caused by viruses [1]. Whole-genome sequencing of bacterial isolates is becoming more widespread, and can help to uncover genetic determinants of clinical severity, elucidate pathogen-host interactions, and

quantify evolutionary rates at within- and between-host levels [67]. Epidemiological investigations using bacterial genomes have also been possible. Even though bacteria acquire point mutations at a lower rate per base than viruses, longer bacterial genomes have provided sufficient genetic resolution for phylogenetic analysis. For example, whole-genome sequencing has been used to refine the tuberculosis transmission network built using contact information [21], and to investigate an outbreak of methicillin-resistant *Staphylococcus aureus* in a hospital and surrounding community in near real-time [68]. The need for longer sequences when conducting epidemiological studies of bacterial infections adds to the per-sample cost of sequencing, and more computational resources are required for coalescent-based inference of pathogen history. However, this latter limitation may be overcome by only analyzing polymorphic sites if samples are similar.

Demographic reconstruction of emerging bacterial pathogens using coalescent-based approaches has been limited compared to work on viral pathogens. In one such study, the temporal changes in genetic diversity of *Streptococcus pneumoniae* in Iceland were estimated based on the coalescent model [47]. This study was limited to a single multidrug-resistant lineage in a single location, with data collected over decades. Over longer evolutionary time-scales, the accumulation of diversity through recombination can obscure phylogenetic relationships. More complex evolutionary models would be required to taken into account these genomic changes, increasing the uncertainty surrounding demographic estimates from genomic data.

In addition to performing analyses with longer sequences, there is also a need to develop methods that exploit as many sequences in the sample as possible. For population studies, available sequences are often subsampled to remove individuals from the same household or in the same close contact network to have a representative sample of the population. Furthermore, sequences from the same individuals are often discarded, though these may be informative for within-host evolution. Although some effort has been made to link within-host to between-host evolution [52,69], the effect of within-host evolution on population genetic inference is still not well studied. Combining analyses across different scales could improve the accuracy of epidemiological predictions and provide better mechanistic explanations of observed trends.

Conclusion

Genomic studies have contributed to better understanding of EIDs and their spatiotemporal spread. Sophisticated statistical methods have been developed to uncover the epidemiological features of infectious diseases based on the genealogy of their sequences. There is also growing

Box 1. Key concepts in mathematical modeling of infectious disease transmission

Representing infectious disease transmission in a mathematical framework requires distilling complex observations into simple but informative expressions. Perhaps the most important statistical property of interest to an epidemiologist is the basic reproductive number, R_0, which represents the mean number of secondary infections caused by each infected individual in a wholly susceptible population. An epidemic can only occur if $R_0 > 1$. As an epidemic progresses, or if there is pre-existing immunity in a population, R_0 is no longer appropriate for describing the number of secondary infections per primary infection. Instead the effective reproductive number, R, is used. Another important statistical property of an epidemic is the generation time, T_g, which is the mean time between when an individual becomes infected and when they infect others. The combination of R_0 and T_g provides an indication of how quickly an epidemic will spread.

The most common type of model used in infectious disease research is the compartmental model. Given a set of parameters, a compartmental model tracks the temporal dynamics of subpopulations that are characterized by disease status. For example, a Susceptible-Infected-Recovered (SIR) model describes the changes in the number of susceptible, infected and recovered (and immune) individuals. R_0 can be calculated by inferring the set of model parameters that can generate the epidemiological dynamics most similar to those observed in the data. Increasingly, model parameters are inferred in a Bayesian framework. Bayesian inference finds the posterior probability distribution of parameters, given prior information and the data. Exploring all possible parameter combinations is intractable. The use of Markov Chain Monte Carlo (MCMC) for Bayesian statistical inference has enabled efficient estimation of the posterior probability distribution when the distribution cannot be computed analytically [70].

Obtaining estimates of R_0 and T_g is not always sufficient to predict epidemic trajectory if there is significant heterogeneity between individuals. The offspring distribution with mean R and variance σ^2 describes the probability distribution of the number of secondary infections caused by each infected individual. In compartmental models, the offspring distribution is not explicitly specified but follows from the specification of the model - in the case of the SIR model it follows a geometric distribution. For certain diseases, the offspring distribution is more dispersed than captured by the geometric distribution [42]. In other words, most individuals cause no further infections whereas a few individuals are super-spreaders who cause the majority of infections. Accurate estimate of σ^2 is important for predicting epidemic outcome and assessing control measures.

Box 2. Coalescent inference from genetic data

Just as compartmental models can be fitted to surveillance data to infer the epidemiological dynamics of an infectious disease (Box 1), the coalescent framework allows inference of population history from pathogen sequences. The coalescent model describes the statistical properties of the genealogy underlying a small sample of individuals from a large population. In the simplest case, the forward-time dynamics of the population is assumed to follow the Wright-Fisher model, in which the haploid population has discrete, non-overlapping generations, undergoes neutral evolution, and remains the same size [71,72]. Extensions to the coalescent have assumed more complex population dynamics described by deterministic population equations [73], compartmental disease models [35], or non-parametric approaches [13,55,74,75].

Within this framework, going backwards in time, individuals in the current generation are randomly assigned to parents in the previous generation. If two individuals have the same parent, then a coalescent event has occurred. Eventually, all lineages in the sample coalesce to a single individual known as the most recent common ancestor of the sample.

The rate of coalescence is inversely related to population size. If the population follows the Wright-Fisher model, evolutionary changes are selectively neutral, so the shape of the genealogy reflects only demographic changes.

effort to integrate genomic analysis with analysis of epidemiological data. In recent cases of EIDs, genomic data have helped to classify and characterize the pathogen, uncover the population history of the disease, and produce estimates of epidemiological parameters.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We would like to thank Nick Croucher for discussions on bacterial genomics. LL is funded by a Medical Research Council Doctoral Training Partnership Studentship.

Published online: 22 November 2014

References

1. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, Daszak P: **Global trends in emerging infectious diseases.** *Nature* 2008, **451:**990–993.
2. Pulliam JR, Epstein JH, Dushoff J, Rahman SA, Bunning M, Jamaluddin AA, Hyatt AD, Field HE, Dobson AP, Daszak P: **Agricultural intensification, priming for persistence and the emergence of Nipah virus: a lethal bat-borne zoonosis.** *J R Soc Interface* 2012, **9:**89–101.
3. Wilder-Smith A, Gubler DJ: **Geographic expansion of dengue: the impact of international travel.** *Med Clin North Am* 2008, **92:**1377–1390.
4. Khan K, Arino J, Hu W, Raposo P, Sears J, Calderon F, Heidebrecht C, Macdonald M, Liauw J, Chan A, Gardam M: **Spread of a novel influenza A (H1N1) virus via global airline transportation.** *New Engl J Med* 2009, **361:**212–214.
5. Le Guenno B, Camprasse MA, Guilbaut JC, Lanoux P, Hoen B: **Hantavirus epidemic in Europe, 1993.** *Lancet* 1994, **343:**114–115.
6. Velayati AA, Masjedi MR, Farnia P, Tabarsi P, Ghanavi J, Ziazarifi AH, Hoffner SE: **Emergence of new forms of totally drug-resistant tuberculosis bacilli: super extensively drug-resistant tuberculosis or totally drug-resistant strains in Iran.** *Chest* 2009, **136:**420–425.
7. Ohnishi M, Golparian D, Shimuta K, Saika T, Hoshina S, Iwasaku K, Nakayama S, Kitawaki J, Unemo M: **Is *Neisseria gonorrhoeae* initiating a future era of untreatable gonorrhea?: Detailed characterization of the first strain with high-level resistance to ceftriaxone.** *Antimicrob Agents Chemother* 2011, **55:**3538–3545.
8. Anderson RM, May RM: *Infectious Diseases of Humans: Dynamics and Control*, Volume 28. Oxford: Oxford University Press; 1991.
9. Grassly NC, Fraser C: **Mathematical models of infectious disease transmission.** *Nat Rev Microbiol* 2008, **6:**477–487.
10. Smith GJ, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, Ma SK, Cheung CL, Raghwani J, Bhatt S, Peiris JS, Guan Y, Rambaut A: **Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic.** *Nature* 2009, **459:**1122–1125.
11. Cottam EM, Haydon DT, Paton DJ, Gloster J, Wilesmith JW, Ferris NP, Hutchings GH, King DP: **Molecular epidemiology of the foot-and-mouth disease virus outbreak in the United Kingdom in 2001.** *J Virology* 2006, **80:**11274–11282.
12. Kingman JF: **On the genealogy of large populations.** *J Appl Probability* 1982, **19:**27–43.
13. Drummond AJ, Rambaut A, Shapiro B, Pybus OG: **Bayesian coalescent inference of past population dynamics from molecular sequences.** *Mol Biol Evol* 2005, **22:**1185–1192.
14. Stadler T: **Sampling-through-time in birth-death trees.** *J Theor Biol* 2010, **267:**396–404.
15. Stadler T, Kouyos R, von Wyl V, Yerly S, Böni J, Bürgisser P, Klimkait T, Joos B, Rieder P, Xie D, Günthard HF, Drummond AJ, Bonhoeffer S, Swiss HIV Cohort Study: **Estimating the basic reproductive number from viral sequence data.** *Mol Biol Evol* 2012, **29:**347–357.
16. Grenfell BT, Pybus OG, Gog JR, Wood JL, Daly JM, Mumford JA, Holmes EC: **Unifying the epidemiological and evolutionary dynamics of pathogens.** *Science* 2004, **303:**327–332.
17. Rasmussen DA, Ratmann O, Koelle K: **Inference for nonlinear epidemiological models using genealogies and time series.** *PLoS Comput Biol* 2011, **7:**1002136.
18. Ypma RJ, van Ballegooijen WM, Wallinga J: **Relating phylogenetic trees to transmission trees of infectious disease outbreaks.** *Genetics* 2013, **195:**1055–1062.
19. Felsenstein J: *Inferring Phylogenies.* Sunderland, MA: Sinauer Associates; 2004.
20. Cottam EM, Thébaud G, Wadsworth J, Gloster J, Mansley L, Paton DJ, King DP, Haydon DT: **Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus.** *Proc Biol Sci* 2008, **275:**887–895.
21. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJ, Brinkman FS, Brunham RC, Tang P: **Whole-genome sequencing and social-network analysis of a tuberculosis outbreak.** *New Engl J Med* 2011, **364:**730–739.
22. Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG: **Measurably evolving populations.** *Trends Ecol Evol* 2003, **18:**481–488.
23. Ho SY, Lanfear R, Bromham L, Phillips MJ, Soubrier J, Rodrigo AG, Cooper A: **Time-dependent rates of molecular evolution.** *Mol Ecol* 2011, **20:**3087–3101.
24. Thorne JL, Kishino H, Painter IS: **Estimating the rate of evolution of the rate of molecular evolution.** *Mol Biol Evol* 1998, **15:**1647–1657.
25. Yoder AD, Yang Z: **Estimation of primate speciation dates using local molecular clocks.** *Mol Biol Evol* 2000, **17:**1081–1090.
26. Drummond AJ, Rambaut A: **BEAST: Bayesian evolutionary analysis by sampling trees.** *BMC Evol Biol* 2007, **7:**214.
27. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP: **MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space.** *Syst Biol* 2012, **61:**539–542.

28. Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC: **The genomic and epidemiological dynamics of human influenza A virus.** *Nature* 2008, **453:**615–619.
29. Volz EM, Ionides E, Romero-Severson EO, Brandt M-G, Mokotoff E, Koopman JS: **HIV-1 transmission during early infection in men who have sex with men: a phylodynamic analysis.** *PLoS Med* 2013, **10:**1001568.
30. Rasmussen DA, Boni MF, Koelle K: **Reconciling phylodynamics with epidemiology: the case of dengue virus in southern Vietnam.** *Mol Biol Evol* 2014, **31:**258–271.
31. Kühnert D, Stadler T, Vaughan TG, Drummond AJ: **Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death SIR model.** *J R Soc Interface* 2014, **11:**20131106.
32. Volz EM: **Complex population dynamics and the coalescent under neutrality.** *Genetics* 2012, **190:**187–201.
33. Frost SD, Volz EM: **Modelling tree shape and structure in viral phylodynamics.** *Philos Trans R Soc London B Biol Sci* 2013, **368:**20120208.
34. Rasmussen DA, Volz EM, Koelle K: **Phylodynamic inference for structured epidemiological models.** *PLoS Comput Biol* 2014, **10:**1003570.
35. Volz EM, Pond SLK, Ward MJ, Brown AJL, Frost SD: **Phylodynamics of infectious disease epidemics.** *Genetics* 2009, **183:**1421–1430.
36. Volz EM, Koopman JS, Ward MJ, Brown AL, Frost SD: **Simple epidemiological dynamics explain phylogenetic clustering of HIV from patients with recent infection.** *PLoS Comput Biol* 2012, **8:**1002552.
37. Koelle K, Rasmussen DA: **Rates of coalescence for common epidemiological models at equilibrium.** *J R Soc Interface* 2012, **9:**997–1007.
38. Stadler T, Bonhoeffer S: **Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods.** *Philos Trans R Soc Lond B Biol Sci* 2013, **368:**20120198.
39. Stadler T, Yang Z: **Dating phylogenies with sequentially sampled tips.** *Syst Biol* 2013, **62:**674–688.
40. Stadler T, Ku¨hnert D, Bonhoeffer S, Drummond AJ: **Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV).** *Proc Natl Acad Sci U S A* 2013, **110:**228–233.
41. Magiorkinis G, Sypsa V, Magiorkinis E, Paraskevis D, Katsoulidou A, Belshaw R, Fraser C, Pybus OG, Hatzakis A: **Integrating phylodynamics and epidemiology to estimate transmission diversity in viral epidemics.** *PLoS Comput Biol* 2013, **9:**1002876.
42. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz W: **Superspreading and the effect of individual variation on disease emergence.** *Nature* 2005, **438:**355–359.
43. Wallinga J, Lipsitch M: **How generation intervals shape the relationship between growth rates and reproductive numbers.** *Proc R Soc Biol Sci* 2007, **274:**599–604.
44. Svensson Å: **A note on generation times in epidemic models.** *Math Biosci* 2007, **208:**300–311.
45. Kenah E, Lipsitch M, Robins JM: **Generation interval contraction and epidemic data analysis.** *Math Biosci* 2008, **213:**71–79.
46. Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, Tatem AJ, Sousa JD, Arinaminpathy N, Pépin J, Posada D, Peeters M, Pybus OG, Lemey P: **HIV epidemiology. The early spread and epidemic ignition of hiv-1 in human populations.** *Science* 2014, **346:**56–61.
47. Croucher NJ, Hanage WP, Harris SR, McGee L, van der Linden M, de Lencastre H, Sá-Leão R, Song JH, Ko KS, Beall B, Klugman KP, Parkhill J, Tomasz A, Kristinsson KG, Bentley SD: **Variable recombination dynamics during the emergence, transmission and 'disarming' of a multidrug-resistant pneumococcal clone.** *BMC Biol* 2014, **12:**49.
48. Bedford T, Cobey S, Beerli P, Pascual M: **Global migration dynamics underlie evolution and persistence of human influenza A (H3N2).** *PLoS Pathog* 2010, **6:**1000918.
49. Bahl J, Nelson MI, Chan KH, Chen R, Vijaykrishna D, Halpin RA, Stockwell TB, Lin X, Wentworth DE, Ghedin E, Guan Y, Peiris JS, Riley S, Rambaut A, Holmes EC, Smith GJ: **Temporally structured metapopulation dynamics and persistence of influenza A H3N2 virus in humans.** *Proc Natl Acad Sci U S A* 2011, **108:**19359–19364.
50. Magiorkinis G, Magiorkinis E, Paraskevis D, Ho SY, Shapiro B, Pybus OG, Allain J-P, Hatzakis A: **The global spread of hepatitis C virus 1a and 1b: a phylodynamic and phylogeographic analysis.** *PLoS Med* 2009, **6:**1000198.
51. Holden MT, Hsu LY, Kurt K, Weinert LA, Mather AE, Harris SR, Strommenger B, Layer F, Witte W, de Lencastre H, Skov R, Westh H, Zemlicková H, Coombs G, Kearns AM, Hill RL, Edgeworth J, Gould I, Gant V, Cooke J, Edwards GF, McAdam PR, Templeton KE, McCann A, Zhou Z, Castillo-Ramírez S, Feil EJ, Hudson LO, Enright MC, Balloux F, *et al*: **A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic.** *Genome Res* 2013, **23:**653–664.
52. Didelot X, Gardy J, Colijn C: **Bayesian inference of infectious disease transmission from whole genome sequence data.** *Mol Biol Evol* 2014, **31:**1869–1879.
53. Grassly NC, Harvey PH, Holmes EC: **Population dynamics of hiv-1 inferred from gene sequences.** *Genetics* 1999, **151:**427–438.
54. Pybus OG, Charleston MA, Gupta S, Rambaut A, Holmes EC, Harvey PH: **The epidemic behavior of the hepatitis C virus.** *Science* 2001, **292:**2323–2325.
55. Strimmer K, Pybus OG: **Exploring the demographic history of DNA sequences using the generalized skyline plot.** *Mol Biol Evol* 2001, **18:**2298–2305.
56. Baize S, Pannetier D, Oestereich L, Rieger T, Koivogui L, Magassouba N, Soropogui B, Sow MS, Keïta S, De Clerck H, Tiffany A, Dominguez G, Loua M, Traoré A, Kolié M, Malano ER, Heleze E, Bocquin A, Mély S, Raoul H, Caro V, Cadar D, Gabriel M, Pahlmann M, Tappe D, Schmidt-Chanasit J, Impouma B, Diallo AK, Formenty P, Van Herp M, *et al*: **Emergence of Zaire Ebola virus disease in Guinea–preliminary report.** *New Engl J Med* 2014, **371:**1418–1425.
57. Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G, Wohl S, Moses LM, Yozwiak NL, Winnicki S, Matranga CB, Malboeuf CM, Qu J, Gladden AD, Schaffner SF, Yang X, Jiang P-P, Nekoui M, Colubri A, Coomber MR, Fonnie M, Moigboi A, Gbakie M, Kamara FK, Tucker V, Konuwa E, *et al*: **Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak.** *Science* 2014, **345:**1369–1372.
58. Dudas G, Rambaut A: **Phylogenetic analysis of Guinea 2014 EBOV Ebolavirus outbreak.** *PLoS Curr* 2014, **6:**.
59. Center for Disease Control and Prevention: **Middle East Respiratory Virus (MERS).** 2014, [http://www.cdc.gov/coronavirus/mers/]
60. Cauchemez S, Fraser C, Van Kerkhove MD, Donnelly CA, Riley S, Rambaut A, Enou V, van der Werf S, Ferguson NM: **Middle East respiratory syndrome coronavirus: quantification of the extent of the epidemic, surveillance biases, and transmissibility.** *Lancet Infect Dis* 2014, **14:**50–56.
61. Memish ZA, Mishra N, Olival KJ, Fagbo SF, Kapoor V, Epstein JH, Alhakeem R, Durosinloun A, Al Asmari M, Islam A, Kapoor A, Briese T, Daszak P, Al Rabeeah AA, Lipkin WI: **Middle East respiratory syndrome coronavirus in bats, Saudi Arabia.** *Emerg Infect Dis* 2013, **19:**1819–1823.
62. Haagmans BL, Al Dhahiry SH, Reusken CB, Raj VS, Galiano M, Myers R, Godeke GJ, Jonges M, Farag E, Diab A, Ghobashy H, Alhajri F, Al-Thani M, Al-Marri SA, Al Romaihi HE, Al Khal A, Bermingham A, Osterhaus AD, AlHajri MM, Koopmans MP: **Middle East respiratory syndrome coronavirus in dromedary camels: an outbreak investigation.** *Lancet Infect Dis* 2014, **14:**140–145.
63. Azhar EI, Hashem AM, El-Kafrawy SA, Sohrab SS, Aburizaiza AS, Farraj SA, Hassan AM, Al-Saeed MS, Jamjoom GA, Madani TA: **Detection of the Middle East respiratory syndrome coronavirus genome in an air sample originating from a camel barn owned by an infected patient.** *Mbio* 2014, **5:**e01450–14.
64. Corman VM, Ithete NL, Richards LR, Schoeman MC, Preiser W, Drosten C, Drexler JF: **Rooting the phylogenetic tree of Middle East respiratory syndrome coronavirus by characterization of a conspecific virus from an African bat.** *J Virol* 2014, **88:**11297–11303.
65. Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, Griffin J, Baggaley RF, Jenkins HE, Lyons EJ, Jombart T, Hinsley WR, Grassly NC, Balloux F, Ghani AC, Ferguson NM, Rambaut A, Pybus OG, Lopez-Gatell H, Alpuche-Aranda CM, Chapela IB, Zavala EP, Guevara DM, Checchi F, Garcia E, Hugonnet S, Roth C, WHO Rapid Pandemic Assessment Collaboration: **Pandemic potential of a strain of influenza A (H1N1): early findings.** *Science* 2009, **324:**1557–1561.
66. Hedge J, Lycett S, Rambaut A: **Real-time characterization of the molecular epidemiology of an influenza pandemic.** *Biol Lett* 2013, **9:**20130331.
67. Wilson DJ: **Insights from genomics into bacterial pathogen populations.** *PLoS Pathog* 2012, **8:**1002874.
68. Harris SR, Cartwright EJ, Török ME, Holden MT, Brown NM, Ogilvy-Stuart AL, Ellington MJ, Quail MA, Bentley SD, Parkhill J, Peacock SJ: **Whole-genome sequencing for analysis of an outbreak of methicillin-resistant *Staphylococcus aureus*: a descriptive study.** *Lancet Infect Dis* 2013, **13:**130–136.
69. Vrancken B, Rambaut A, Suchard MA, Drummond A, Baele G, Derdelinckx I, Van Wijngaerden E, Vandamme A-M, Van Laethem K, Lemey P: **The**

genealogical population dynamics of HIV-1 in a large transmission chain: Bridging within and among host evolutionary rates. *PLoS Comput Biol* 2014, **10:**1003505.
70. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E: **Equation of state calculations by fast computing machines.** *J Chem Phys* 2004, **21:**1087–1092.
71. Fisher RA: *The Genetical Theory of Natural Selection.* Oxford: Clarendon; 1930.
72. Wright S: **Evolution in Mendelian populations.** *Genetics* 1931, **16:**97–159.
73. Griffiths RC, Tavare S: **Sampling theory for neutral alleles in a varying environment.** *Phil Trans R Soc Lond Biol Sci* 1994, **344:**403–410.
74. Pybus OG, Rambaut A, Harvey PH: **An integrated framework for the inference of viral population history from reconstructed genealogies.** *Genetics* 2000, **155:**1429–1437.
75. Minin VN, Bloomquist EW, Suchard MA: **Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics.** *Mol Biol Evol* 2008, **25:**1459–1471.

doi:10.1186/s13059-014-0541-9
Cite this article as: Li *et al.*: Genomic analysis of emerging pathogens: methods, application and future trends. *Genome Biology* 2014 **15**:541.

Appendix B

Vignette for generating simulated data in EpiGenR and interfacing with EpiGenMCMC

Epidemic simulation and inference from the simulated data

Lucy M Li

2017-05-08

```
library(EpiGenR)
library(ape)
library(ggplot2)
library(grid)
library(gridExtra)
fig.counter <- list()
knitr::opts_chunk$set(warning=FALSE, error=TRUE, message=FALSE, echo=TRUE)
```

1. Simulate epidemic data
2. Convert line list data and pathogen phylogeny into list objects
3. Construct input objects for inference
4. Call the EpiGenMCMC program to estimate parameters

1. Simulate epidemic data

In the example here, I simulate an epidemic according to a stochastic SIR model, which is a state space model with 3 state variables: Susceptible, Infected, and Removed. Two events can occur to change the state variable values: infection and recovery. Simulation takes in discrete steps indexed by t, where each step size is dt. During each small time interval $[(t-1)\cdot dt, t\cdot dt]$, the number of recovery events given I_t infected individuals and a recovery rate of γ is approximate binomial $\text{recoveries}_t \sim Bin(I_t, \gamma\Delta t)$. Assuming all onward transmissions occur at recovery, the number of infection events during a time interval $[(t-1)\cdot dt, t\cdot dt]$ follows the offspring distribution which I model using the negative binomial $\text{infections}_t \sim NBin(\text{recoveries}_t \times R_t, \text{recoveries}_t \times k)$.

we assume an S->I->R model of disease progression in which susceptible individuals become infected and capable of infecting others, and later recover and stop being infectious. The time to recovery is exponentially distributed with rate γ. Upon recovery, an infector infects b_0 number of individuals. The number of onward infections, i.e. `offspring', caused by each infected individual is a random variable drawn from a negative binomial offspring distribution $B \sim NBin(R, k)$ with mean $R = \frac{\sum_{i=0}^{N-1} b_i}{N}$, dispersion parameter k, and variance $\sigma^2 = R(1+\frac{R}{k})$. The mean of the offspring distribution is the reproductive number of the infectious disease, and is related to the basic reproduction number R_0 via the proportion of susceptible individuals in the population: $R = R_0\frac{S}{N}$. The parameter k determines the level of overdispersion in the population. At smaller values of k, most individuals do not cause any further infections while a few contribute to most of the transmission events.

Setting N = 5,000, R_0 = 2, k = 0.5, and duration of infectiousness $\frac{1}{\gamma}$ = 5 days, we can simulate the outbreak using

```
seed.num <- 1010113
set.seed(seed.num)
sim.outbreak <- simulate_sir(params, dt, total_dt, min_epi_size, max_attempts, TRUE)
```

The offspring distribution of the simulated epidemic follows a negative binomial distribution

```
par(mar=c(5.1, 4.1, 0.25, 0.25))
offspring <- rnbinom(10000, mu=R0, size=k)
hist(offspring, xlab="Number of onward infections", main="")
legend("topright", legend=paste0("R0=", round(R0, 2), " and k=", k))
```

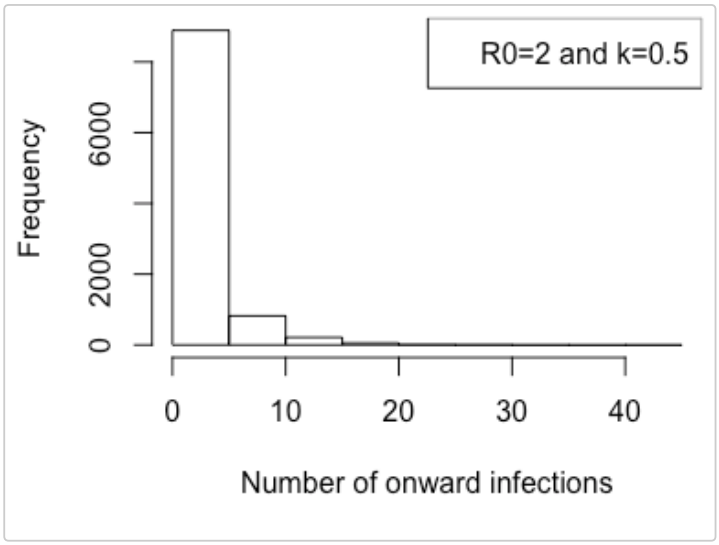


Figure 1. The offspring distribution of the simulated epidemic assuming R0=2 and k=0.5.

The final epidemic size was 4,052.

Simulated epidemic trajectories

The epidemic trajectories denoted by the incidence and prevalence curves are shown in the Figure 2 below. Assuming that infectious individuals are reported at the time of recovery, the incidence curve shows the daily number of reported cases.

```
P1 <- ggplot(data.frame(time_series_from_line_list(sim.outbreak))) + theme_bw() +
  geom_bar(aes(x=time, y=incidence), stat="identity") +
  xlab("Days since start of epidemic") +
  ylab("Incidence per day") + ggtitle("A")
P2 <- ggplot(data.frame(x=(1:sim.outbreak$total_dt)*dt, Prevalence=sim.outbreak$prevalence)) +
  theme_bw() + geom_line(aes(x=x, y=Prevalence)) +
  xlab("Days since start of epidemic") + ggtitle("B")
grid.arrange(P1, P2, ncol=1)
```

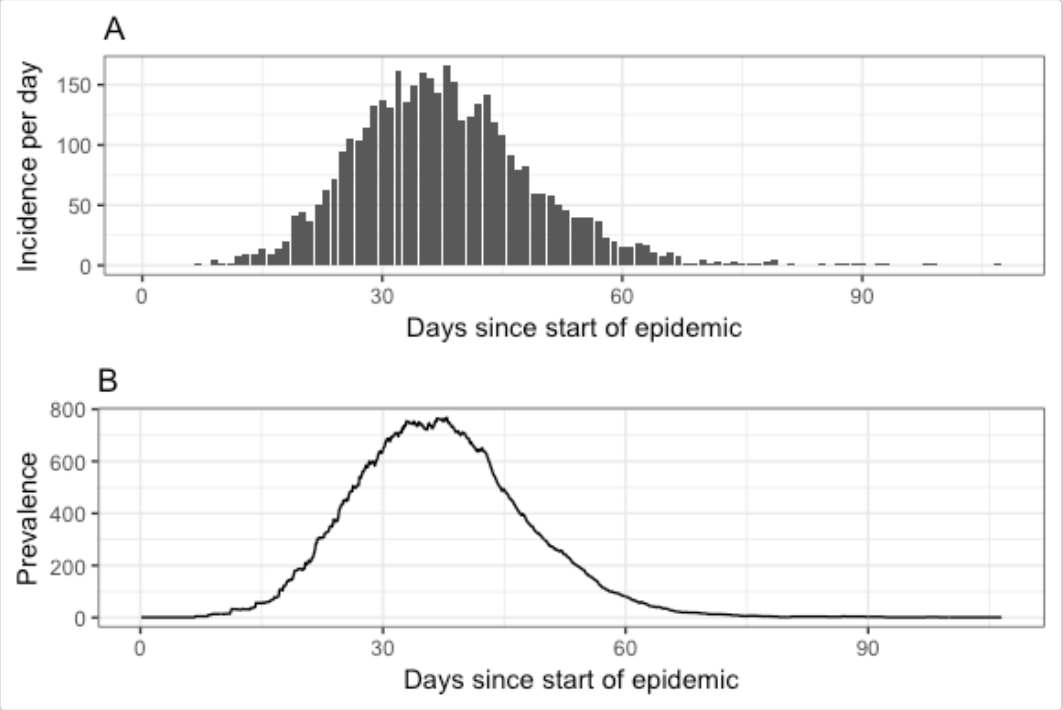


Figure 2. The daily incidence (A) and prevalence (B) of the simulated epidemic.

2. Convert line list data and pathogen phylogeny into list objects

Transmission Tree

By setting to we can track who infected whom in the outbreak and thus reconstruct the transmission tree. From the transmission tree, we can infer the pathogen phylogeny which describes the ancestral relationship between pathogen isolates from infected individuals.

```
sim.transmission.tree <- as.data.frame(get_transmission_tree(sim.outbreak$infected))
sim.transmission.tree$from <- as.factor(sim.transmission.tree$from)
sim.transmission.tree$to <- as.factor(sim.transmission.tree$to)
fig.counter.sim.graph <- fig.counter
```

We can visualise the transmission network using the function. Below is the transmission network of the first 100 infected people.

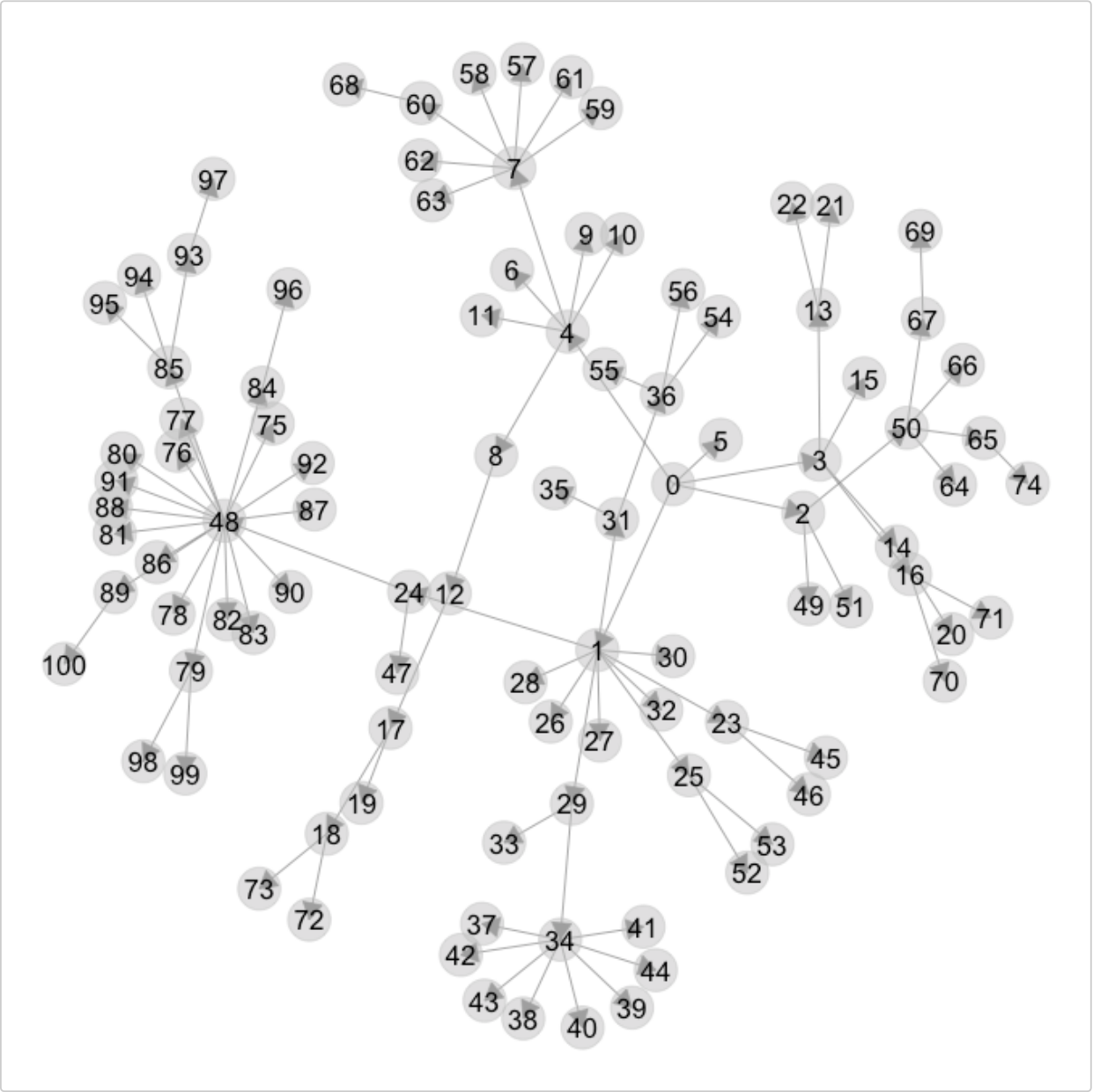


Figure 3. Transmission tree.

Phylogeny

The phylogenetic tree is related to the transmission tree. In the case of the latter, parents are represented by internal nodes whereas in the case of phylogenies, parents are represented by an external node (tip). The function produces the phylogenetic tree for a given outbreak. Figure 4 is the phylogenetic tree of the first 100 individuals to be infected during the epidemic.

```
tree <- get_phylo(sim.outbreak$infected)
```

```
par(mar=c(0.5, 0.5, 0.5, 0.5))
not.sampled.tips <- 101:length(tree$tip.label)
subtree <- drop.tip(tree, not.sampled.tips)
plot(subtree)
```

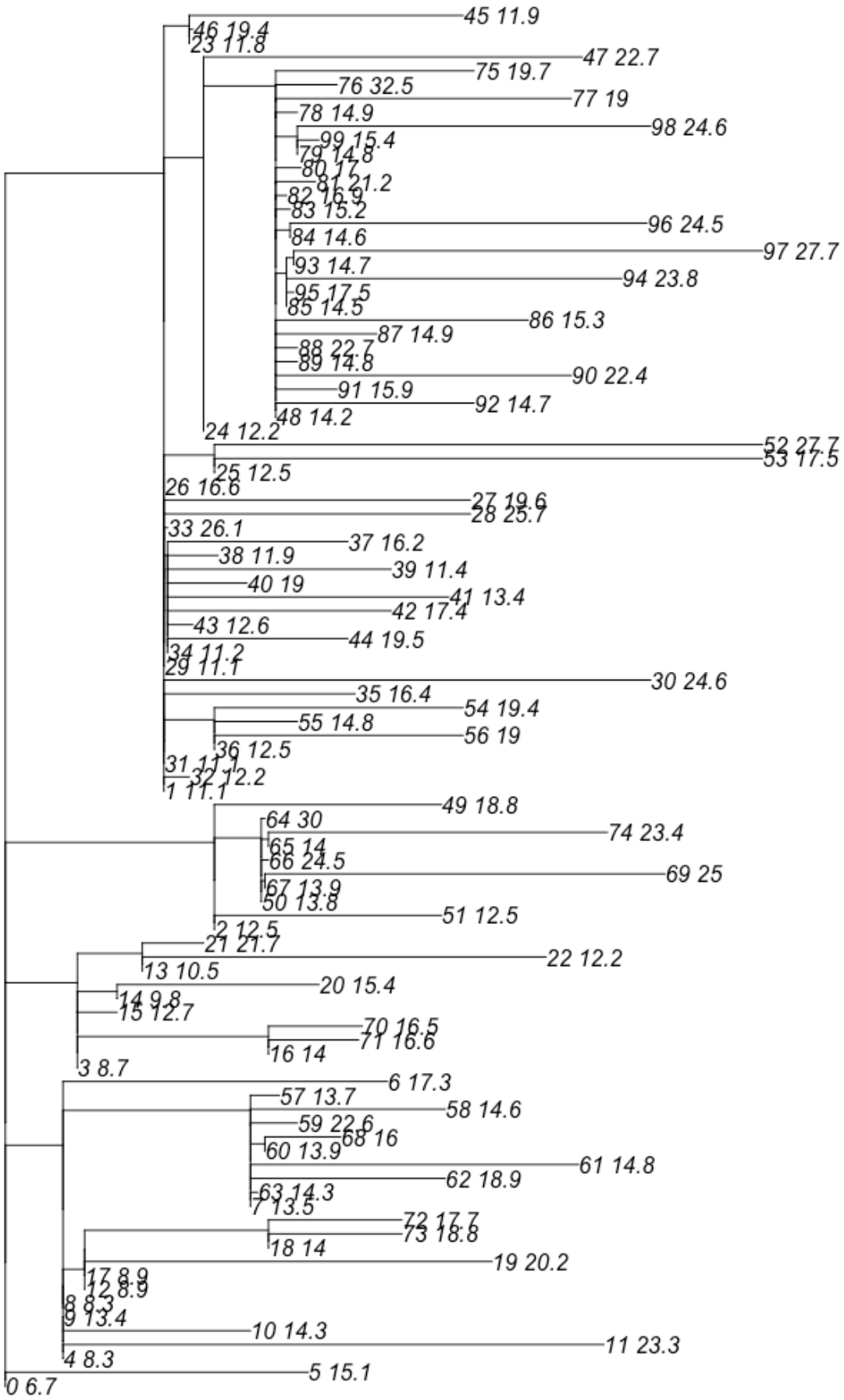

45 11.9
46 19.4
23 11.8
47 22.7
75 19.7
76 32.5
77 19
78 14.9
98 24.6
99 15.4
79 14.8
80 17
81 21.2
82 16.9
83 15.2
96 24.5
84 14.6
97 27.7
93 14.7
94 23.8
95 17.5
85 14.5
86 15.3
87 14.9
88 22.7
89 14.8
90 22.4
91 15.9
92 14.7
48 14.2
24 12.2
52 27.7
53 17.5
25 12.5
26 16.6
27 19.6
28 25.7
33 26.1
37 16.2
38 11.9
39 11.4
40 19
41 13.4
42 17.4
43 12.6
44 19.5
34 11.2
29 11.1
30 24.6
35 16.4
54 19.4
55 14.8
56 19
36 12.5
31 11.1
32 12.2
1 11.1
49 18.8
64 30
74 23.4
65 14
66 24.5
69 25
67 13.9
50 13.8
51 12.5
2 12.5
21 21.7
22 12.2
13 10.5
20 15.4
14 9.8
15 12.7
70 16.5
71 16.6
16 14
3 8.7
6 17.3
57 13.7
58 14.6
59 22.6
68 16
60 13.9
61 14.8
62 18.9
63 14.3
7 13.5
72 17.7
73 18.8
18 14
19 20.2
17 8.9
12 8.9
8 8.3
9 13.4
10 14.3
11 23.3
4 8.3
5 15.1
0 6.7

Figure 4. Phylogeny of the first 100 individuals to be infected during the simulated epidemic, out of a total of $N=$4052.

Producing time-series data from simulation

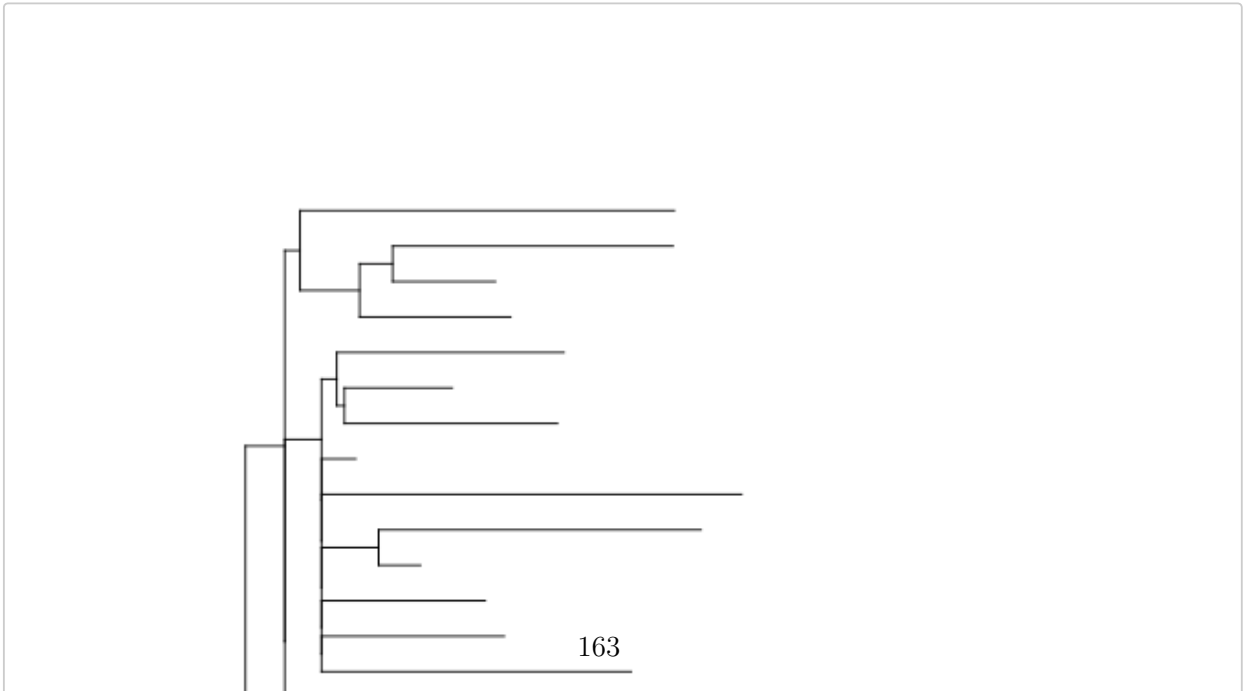
Inferring parameters of dynamic disease models such as the SIR require data to be in time-series format, i.e. a quantity per time step. For epidemiologic data, this could be the number of reported cases per day. If data collected during an outbreak is in the form of a line list where each line contains information about an infected individual, this can be converted to time-series format using the function . The first column should contain the ID of the infected individual and the second column the time of reporting. Here we assumed that an individual was reported upon recovery.

```
sampling.prob <- 0.01
data.dt <- 1
set.seed(seed.num)
sampled.sim.outbreak <- downsample(sim.outbreak, strategy="proportional", prob=sampling.prob)
epi_data <- time_series_from_line_list(sampled.sim.outbreak, step_size=data.dt)
head(epi_data)
```

```
##      time incidence
## [1,]    1         0
## [2,]    2         0
## [3,]    3         0
## [4,]    4         0
## [5,]    5         0
## [6,]    6         0
```

And the phylogeny of the randomly sampled individuals is given in Figure 5.

```
subtree <- drop.tip(reorder.phylo(tree, "postorder"), which(!(1:length(tree$tip.label) %in%
sampled.sim.outbreak$sampled_individuals)))
plot(subtree, show.tip.label=FALSE)
```



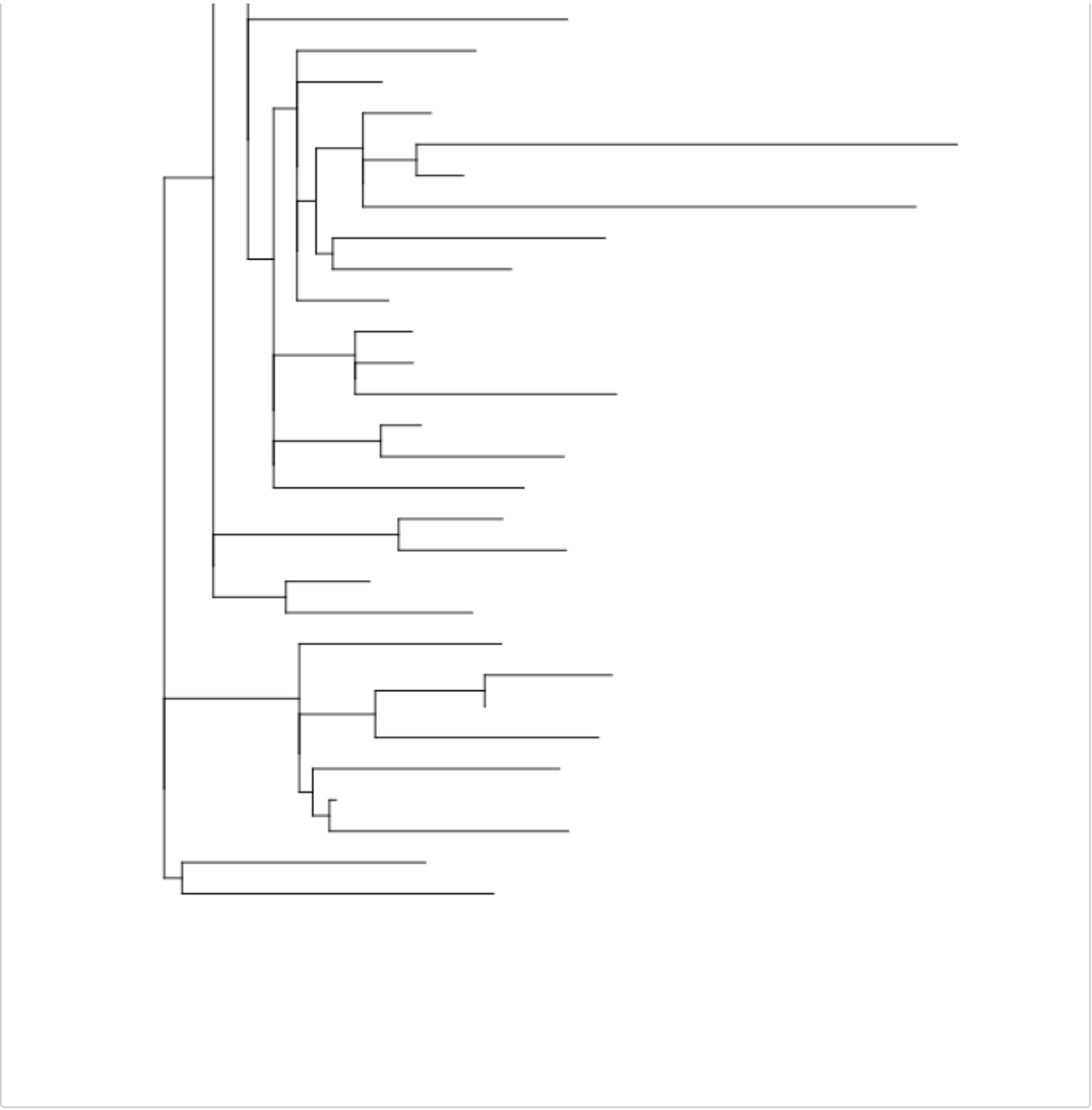


Figure 5. Phylogeny of 43 randomly sampled individuals. This is a subtree of the full phylogeny of N=4052 individuals.

```
gen_data <- time_series_from_tree(subtree, step_size=data.dt)
```

We can also obtain time-series data for both at the same time:

```
all_data <- get_data(epi=sampled.sim.outbreak, phy=subtree, dt=data.dt)
str(all_data)
```

```
## List of 2
##  $ epi: num [1:156, 1:2] -77 -76 -75 -74 -73 -72 -71 -70 -69 -68 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : chr [1:2] "time" "incidence"
##  $ gen: num [1:78, 1:2] 1 2 3 4 5 6 7 8 9 10
```

```
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : chr [1:2] "time" "incidence"
```

3. Construct input objects for inference

Create input files for EpiGenMCMC program

```
param_list <- create_params_list(
  param_names=c("R0", "k", "rateI2R", "N", "S", "reporting", "time_before_data"), # All parameter
values
  init_param_values=c(R0, k, 1/Tg, N, S, sampling.prob, 10), # Initial parameter values
  params_to_estimate=c("R0", "k", "rateI2R", "reporting", "time_before_data"), # Names of
parameters to be estimated
  transform=c(NA, "inverse", "inverse", NA, NA), # The algorithm will estimate the value of the
transformed parameter
  prior=c("unif", "unif", "unif", "beta", "unif"), # Prior distribution
  prior_params=list(c(1.0, 100.0), c(1.0, 10000.0), c(1.0, 30.0), c(1.0, 3.0), c(0.0, 300.0)), #
Parameters for the prior distribution
  proposal_params=list(c(0.5, 1.0, 100.0), c(1.0, 1.0, 10000.0), c(1.0, 1.0, 30.0), c(0.05, 0.0,
1.0), c(20.0, 0.0, 300.0))
# SD of proposal distribution, and the range of parameter values to be explored
)
mcmc_options <- create_mcmc_options (particles=1000, iterations=1000, log_every=1,
pfilter_every=20,
                                     which_likelihood=0,
        # 0= use both epi and genetic data, 1=use only epi data, 2=use only genetic data
                                     pfilter_threshold=1.0,
                                  log_filename="log.txt", traj_filename="traj.txt")

input_files <- EpiGenR::generate_cpp_input_files(dt=dt, params=param_list,
mcmc_options=mcmc_options,
                                                 initial_states=unlist(init.states),
data=all_data)
```

4. Call the EpiGenMCMC program to estimate parameters

```
EpiGenR::run_pMCMC("path/to/pmcmc", input_files, wait=FALSEs)
```