Phylodynamic Methods for Infectious Disease Epidemiology

by

David A. Rasmussen

Department of Biology
Duke University

Date: _____

Approved:

_____
Katia Koelle, Supervisor

_____
William Morris

_____
Sayan Mukherjee

_____
Allen Rodrigo

_____
Marcy Uyenoyama

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Biology
in the Graduate School of Duke University
2014

Abstract

Phylodynamic Methods for Infectious Disease Epidemiology

by

David A. Rasmussen

Department of Biology
Duke University

Date: _____

Approved:

_____
Katia Koelle, Supervisor

_____
William Morris

_____
Sayan Mukherjee

_____
Allen Rodrigo

_____
Marcy Uyenoyama

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biology in the Graduate School of Duke University
2014

Abstract

In this dissertation, I present a general statistical framework for phylodynamic inference that can be used to estimate epidemiological parameters and reconstruct disease dynamics from pathogen genealogies. This framework can be used to fit a broad class of epidemiological models, including nonlinear stochastic models, to genealogies by relating the population dynamics of a pathogen to its genealogy using coalescent theory. By combining Markov chain Monte Carlo and particle filtering methods, efficient Bayesian inference of all parameters and unobserved latent variables is possible even when analytical likelihood expressions are not available under the epidemiological model. Through extensive simulations, I show that this method can be used to reliably estimate epidemiological parameters of interest as well as reconstruct past disease dynamics from genealogies, or jointly from genealogies and other common sources of epidemiological data like time series. I then extend this basic framework to include different types of host population structure, including models with spatial structure, multiple-hosts or vectors, and different stages of infection. The later is demonstrated by using a multistage model of HIV infection to estimate stage-specific transmission rates and incidence from HIV sequence data collected in Detroit, Michigan. Finally, to demonstrate how the approach can be used more generally, I consider the case of dengue virus in southern Vietnam. I show how earlier phylodynamic inference methods fail to reliably reconstruct the dynamics of dengue observed in hospitalization data, but by deriving coalescent models that

take into consideration ecological complexities like seasonality, vector dynamics and spatial structure, accurate dynamics can be reconstructed from genealogies. In sum, by extending phylodynamics to include more ecologically realistic and mechanistic models, this framework can provide more accurate estimates and give deeper insight into the processes driving infectious disease dynamics.

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Abbreviations

CI	Credible interval.
DENV	Dengue virus.
MCMC	Markov chain Monte Carlo.
PMMH	Pseudo-marginal Metropolis-Hastings.
SIR	Susceptible-Infected-Recovered (model).
SSM	State-space model.

Acknowledgements

First and foremost, I would like to thank Katia Koelle for being an excellent advisor. You were always helpful, encouraging and very generous in supporting me through graduate school and beyond. I am truly grateful to have had the opportunity to work with you!

I would also like to give special thanks to those who have collaborated with me on my disseration work, especially Oliver Ratmann, Erik Volz and Maciej Boni. Thank you also to the many people who have provided me with support and encouragement over the years: my committee members Bill Morris, Sayan Mukherjee, Allen Rodrigo and Marcy Uyenoyama; the many members of the Koelle group for advice and friendship; the Biology Dept. support staff; Cameron Simmons and the entire dengue group at the Oxford University Clinical Research Group in Vietnam; and the U.S. National Science Foundation Graduate Research Program for funding. Finally, I am very grateful to all the great friends I have made here in Durham, especially Marissa Lee for putting up with me the most.

1

Introduction

Mathematical models play an important role in our understanding of the processes driving infectious disease dynamics. Simple but elegant models like the well known Susceptible-Infected-Recovered (SIR) class of models are capable of describing the epidemic dynamics of pathogens and can easily be extended to capture a wide range of more complex population dynamics (Anderson et al., 1979; Anderson and May, 1991). While simple, these models can provide insight into which processes are fundamental to the epidemiology of pathogens and help remove unnecessary complexity from our understanding of their dynamics. Historically, mathematical epidemiology has also greatly benefited from the availability of high quality data generated from surveillance. For example, classic childhood diseases like measles, rubella and pertussis have provided epidemiology with very detailed time series data to which model predictions can be compared and tested against. This interplay between modeling and empirical data has provided many basic insights and a better understanding of how factors such as human demography, population structure, demographic stochasticity, climatic fluctuations, and ecological competition among interacting pathogens influence disease dynamics (Andreasen et al., 1997; Bolker and Grenfell, 1995; Earn

et al., 2000; Pascual et al., 2000; Rohani et al., 1998).

In spite of the rich theoretical foundations of mathematical epidemiology, the practice of formally fitting models to empirical data using rigorous statistical methods has lagged behind other theoretical developments. Simulations from models are often only qualitatively compared against observational data. Even when models are directly fit to empirical data, there is often a discrepancy between the dynamical models used in the theoretical literature and the generally much more simplistic models actually fit to data. More broadly, developing statistical methods for fitting epidemiological models to data remains difficult because the processes underlying disease dynamics like transmission and recovery are rarely directly observed and must be inferred from partial observations. Only relatively recently have new methods been developed for fitting the types of nonlinear, and often stochastic, population dynamic models used in mathematical epidemiology to empirical data (Finkenstädt and Grenfell, 2000; Ionides et al., 2006; O'Neill and Roberts, 1999; Sisson et al., 2007).

Along with these methodological issues, epidemiologists often confront limitations in the quality and availability of data. Epidemiological data are often noisy, aggregated across spatial or temporal scales different than the ones of primary interest, and may contain systematic biases due to reporting practices (Rohani and King, 2010). Moreover, long-term time series data are often incomplete or even completely missing. Detailed time series like those for classic childhood diseases are generally the exception rather than the rule. This is especially true for the vast majority of emerging or reemerging infectious diseases where surveillance systems do not yet exist, as well as for most wildlife diseases (Woolhouse, 2002; Morens et al., 2004). Epidemiology is thus very often rich in theory but poor in data to test theory against.

Nontraditional sources of data may offer epidemiologists new ways of studying disease dynamics, especially the increasing abundance of genetic and molecular se-

quence data collected from pathogens. Traditionally, epidemiologists have used genetic data to identify pathogens, classify them taxonomically, and establish genetic relationships among isolates from different hosts (Maslow et al., 1993). Genetic data thus aided in the identification and description of pathogens but played little role in our more general understanding of their epidemiological dynamics. But as modern phylogenetic and population genetic methods for analyzing genetic data have developed, it has became clear that phylogenies reconstructed from pathogen sequence data can also offer a wealth of information about the ecological and evolutionary dynamics of pathogens. This realization has spawned considerable interest in what has been dubbed "phylodynamics" — the field that aims to quantitatively understand how ecological and evolutionary processes act, or interact, to shape pathogen phylogenies and patterns of genetic variation (Grenfell et al., 2004; Pybus and Rambaut, 2009; Volz et al., 2013b).

1.1 The rise of phylodynamics

Before the term "phylodynamics" was even coined, researchers working on the phylogenetics of infectious pathogens began to realize that sequence data could provide a window into the historical population dynamics of these pathogens (Holmes et al., 1995; Rodrigo and Felsenstein, 1999; Zanotto et al., 1996). Population geneticists had long recognized that genealogies of random samples collected from a population will contain information about the demographic history of that population (Kingman, 1982; Hudson et al., 1990; Donnelly and Tavaré, 1995). For example, historical changes in population sizes can shift the distribution of coalescent (i.e. branching) events over a genealogy, making it possible to identify periods of population growth or decline (Slatkin and Hudson, 1991; Nee et al., 1995). Other demographic features of a population, such as population structure and reproductive variability can likewise shape trees and therefore, at least in theory, be inferred from genealogies.

To exploit this information, several new statistical methods based on coalescent theory, which probabilistically relates genealogies to a population's demographic history, were developed to infer demographic parameters such as population sizes and migration rates from genealogies (Beerli and Felsenstein, 1999; Kuhner et al., 1998; Tavare et al., 1997). Around the same time, molecular sequence data from well-studied human pathogens such as HIV and hepatitis C were becoming increasingly abundant. Unlike most population genetic data, viral sequence data are often sampled serially over time so that samples are available at sequential time points. With such serially sampled data, it becomes possible to simultaneously estimate mutation rates and demographic parameters such as population sizes from genealogies (Pybus et al., 2000; Rodrigo and Felsenstein, 1999). Coupled with new methods for inferring time-calibrated phylogenies (Drummond et al., 2002), it was then possible to infer how pathogen population dynamics change in real calendar time from genealogies (Drummond et al., 2002, 2005; Pybus et al., 2001; Strimmer and Pybus, 2001).

These new phylodynamic inference methods became extremely popular among researchers working on infectious diseases, especially rapidly evolving RNA viruses (Lemey et al., 2003; Carrington et al., 2005; Biek et al., 2007; Rambaut et al., 2008). Early phylodynamic analyses showed that it was possible to reconstruct populations dynamics as well as to infer the timing and initial growth rates of epidemics (Pybus et al., 2001). More general methods such as the popular Bayesian Skyline and Skyride approaches also made it possible to reconstruct increasingly complex population dynamics, such as rapid fluctuations in population size over time (Drummond et al., 2005; Minin et al., 2008). However, these coalescent-based methods were generally based on very simple demographic models borrowed from traditional population genetics, such as the Wright-Fisher and Moran models, and were very different types of population dynamic models used in mathematical epidemiology. It was therefore not possible to apply the type of models typically used in epidemiology

to analyze time series to molecular sequence data, and thus not possible estimate many epidemiological parameters directly from genealogies.

As phylodynamic methods were applied to more pathogens, it also became increasingly apparent that coalescent-based methods using simple demographic models did not always reconstruct epidemiological dynamics that were consistent with other observational data (de Silva et al., 2012; Bennett et al., 2010; Siebenga et al., 2010). The appropriateness of standard coalescent models for infectious pathogens was therefore brought into question. Under standard coalescent models from population genetics, the rate of coalescence is simply inversely proportional to the effective population size, N_e. Thus, it is N_e, and not necessarily the absolute population size that is inferred from genealogies using most phylodynamic methods. For an infectious pathogen, it was generally assumed that N_e would be proportional to the number of infected hosts when estimated from a genealogy of samples taken from different infected hosts (Pybus et al., 2000; Drummond et al., 2002). However, it was not entirely clear what N_e represented for a pathogen, especially one undergoing nonlinear and complex population dynamics. This lead several authors to suggest that phylodynamic reconstructions of N_e should not be interpreted as reflecting the actual number of infections but as a measure of relative genetic diversity, which had long been recognized to be influenced by factors other than absolute population size, such as population structure and reproductive variance (Carrington et al., 2005; Griffiths and Tavare, 1994; Pybus et al., 2001; Rambaut et al., 2008). Yet it remained unclear why phylodynamic reconstructions of N_e so closely tracked the expected population dynamics of some pathogens but yet so poorly reproduced these patterns for others.

Important theoretical work by Volz et al. (2009) and Frost and Volz (2010) helped elucidate the relationship between the coalescent events observed in a genealogy and the population dynamics of a pathogen. If one assumed that each lineage in a pathogen genealogy was represented by a single infected host, so that the within-

host coalescent process was ignored, the coalescent events in the genealogy will correspond to the transmission events if both lineages are included in the sample. Using this reasoning, under continuous-time epidemiological models the rate of coalescence must at least in part depend on the rate at which transmission events occur in the host population (i.e the incidence), and not just the number of infected hosts (i.e the prevalence). Furthermore, under standard epidemiological models with random mixing, incidence is proportional to the product of the transmission rate, the number of susceptible individuals and the number of infected individuals (βSI in standard SIR model notation). Thus, the coalescent rate not only depends on prevalence, but also on the number of susceptible individuals and the transmission rate, which may vary nonlinearly over time. As pointed out by Frost and Volz (2010), there is generally no linear rescaling of prevalence into an effective population size that would be adequate to describe the dynamics of the coalescent process, and therefore there is generally no such thing as an effective population size for an infectious pathogen. Only under certain conditions, like at endemic equilibrium, can the rate of coalescence be linearly related to the prevalence of the disease and therefore an effective population size (Frost and Volz, 2010; Koelle and Rasmussen, 2012).

1.2 A new statistical framework

These developments in coalescent theory for infectious pathogens opened the door for more mechanistic epidemiological models to be fit to pathogen genealogies. Indeed, using the assumption that each lineage in the genealogy corresponds to a single infected host, it is possible to derive coalescent models for many different epidemiological models with various assumptions about the natural history of infection and transmission (Koelle and Rasmussen, 2012; Volz, 2012). Yet existing statistical methods for coalescent-based inference still only allowed for very simple demographic models. I therefore decided at the outset of my dissertation work to develop a more

general statistical framework for phylodynamic inference that could be used to fit the types of mechanistic, and often nonlinear, population dynamic models used in epidemiology to genealogies. With the help of Katia Koelle and Oliver Ratmann, I developed a framework for phylodynamic inference based on state-space models (SSMs) that is the subject of Chapter 2.

SSMs are widely used in the natural sciences to model dynamical systems when only partial observations of the system are available. In general, a SSM is composed of two interacting models: a process model and an observation model. The process model describes the stochastic dynamics of the process under study in terms of one or more variables (i.e. the "state-space"). However, we may only have direct observations on a subset of these variables, and these observations may contain errors or noise. Thus, we also require a probabilistic model to relate the observations back to the latent state variables. Analogously, in phylodynamics we generally only observe a partial genealogy containing a fraction of all infections in the population, so that the tree contains only the lineages that were directly sampled or have descendants that were sampled. Our insight was that we could use a coalescent model instead of a normal observation model in a SSM in order to relate the observed genealogy back to the unobserved disease dynamics. A SSM framework therefore seemed like a natural choice for phylodynamic inference because we could estimate epidemiological parameters from genealogies while at the same time using the coalescent model to simultaneously infer the unobserved latent variables, such as the number of susceptible or infected hosts over time.

However, many of the same challenges encountered when fitting epidemiological models to traditional data like time series of case reports are also present when trying to fit SSMs to genealogies. In a SSM, the latent state variables are treated as random variables that also need to be inferred, which can cause the models to become very high-dimensional. For nonlinear stochastic models where we lack direct

observations on the actual process underlying the observed dynamics, it is also often not possible to analytically compute the transition densities describing the time evolution of the state variables and therefore not possible to compute the likelihood of the model in closed-form. For example, in epidemiology we typically do not observe the actual transmission and recovery events driving the disease dynamics so we must integrate over all possible events when computing the transition densities involved in the likelihood. Standard likelihood-based inference methods therefore cannot be used for fitting stochastic, nonlinear models to observational data, nor genealogies.

Inspired by recent work done on fitting nonlinear SSMs to time series data (Bretó et al., 2009; Ionides et al., 2006; He et al., 2010), we decided to use particle filtering methods (as known as sequential Monte Carlo) to fit SSMs to genealogies. In essence, particle filtering methods provide a computational means of approximating high-dimensional distributions, as well as sequences of distributions, by using importance sampling methods (Doucet et al., 2001). In the context of inference for SSMs, particle filters can be used to obtain samples (i.e. particles) from the posterior density of latent state variables given the observed data. One can then average over these samples to integrate out the latent state variables and therefore compute an estimate of the marginal likelihood of the data given the model. Furthermore, an especially nice feature of the particle filtering methods we use for SSMs is that we only have to be able to simulate from the model to obtain samples from the posterior density of the latent state variables. This means that it is not necessary to compute the transition densities of the state variables explicitly, which opens they way for doing likelihood-based inference for a large class of nonlinear and non-Gaussian SSMs where other likelihood-based methods cannot be applied.

To perform full Bayesian inference of all epidemiological parameters and latent state variables, we combined our particle filtering algorithm for genealogies with an MCMC sampler using the particle MCMC framework of Andrieu et al. (2010). While

in theory it is possible to use standard MCMC approaches for SSMs, in practice it can be very difficult to design proposal mechanisms to sample from the joint density of the model parameters and latent state variables. We therefore modified the pseudo-marginal Metropolis-Hastings (PMMH) algorithm of Andrieu et al. (2010) to fit SSMs to genealogies. In this algorithm, a Metropolis-Hastings step is used to either accept or reject a new set of parameters each MCMC iteration while using a particle filter to compute a numerical approximation of the marginal likelihood of the genealogy given the proposed parameters. Because we marginalize out the latent state variables when running the particle filter, we do not have to design a separate proposal distribution for the latent state variables. The idea of using importance sampling methods like particle filtering within a MCMC algorithm to construct an estimate of some marginal density was not new in of itself. In fact, importance sampling methods had already been used in population genetics to target the marginal density of demographic parameters while integrating over the unknown genealogy of the samples (Beaumont, 2003). However, until the theoretical work of Andrieu et al. (2010), it remained unclear if it was legitimate to use particle filtering methods within MCMC. By demonstrating that the PMMH algorithm is a special case of a MCMC algorithm that operates on the expanded state space of all the random variables in the posterior density as well as the random variables used to generate the particle samples, it was shown that these methods were exact and will give an unbiased estimate of the target posterior density (Andrieu et al., 2010). While this has far reaching consequences for statistical inference with dynamical systems in general, it allowed us to efficiently perform phylodynamic inference with far more complex stochastic population dynamic models than what had previously been done before. In Chapter 2, I describe the PMMH algorithm for phylodynamic inference in detail and demonstrate how it can be used to estimate epidemiological parameters and dynamics from genealogies, or jointly from genealogies and other sources of data

such as time series.

1.3 The problem of structure

While the inference methods described in Chapter 2 allow for stochastic and possibly nonlinear population dynamics, they were built under the assumption that all lineages in the genealogy are in the same population of infected hosts. However, in searching for empirical data sets to apply these methods to, it soon became obvious that most real world pathogen populations are structured in such a way that lineages in different subpopulations do not necessarily have the same probability of coalescing. While earlier theoretical work had already extended coalescent models and methods to consider different forms of population structure (Beerli and Felsenstein, 1999; Notohara, 1990; Takahata and Slatkin, 1990), these earlier approaches generally assumed that populations are at equilibrium and that the rates at which lineages transition between populations are constant over time—assumptions that are not generally valid for infectious pathogens. I therefore started working on structured coalescent models for pathogens that could handle non-equilibrium population dynamics and changing migration rates. However, at the same time, Erik Volz independently developed a similar structured coalescent framework for pathogens (Volz, 2012). Moreover, it was very evident that his approach was more elegant and general in terms of the different forms of population structure it could accommodate, although his framework initially only considered deterministic population dynamics. I therefore began working with Erik Volz and Katia Koelle on extending the original PMMH framework for phylodynamic inference to incorporate the structured coalescent models of Volz (2012).

Unfortunately, extending the PMMH methods to accommodate structured models turned out to be rather difficult. Under structured coalescent models, the likelihood of a genealogy depends on the demographic history of the population as well

as the internal states of the lineages in the genealogy. Moreover, the probability that a lineage is in a certain state (i.e. population) at a given time can only be computed retrospectively conditional upon the state of the lineage at the time of sampling and the demographic history of the population over the time period spanned by the genealogy (Volz, 2012). As I show in Chapter 3, these backward-time dependencies can cause forward-in-time particle filtering methods to become quite inefficient. Initially, it seemed best to try to break down the problem into smaller, more manageable pieces. For example, I tried many different algorithms that iteratively sampled from the conditional densities of the lineage states and the epidemiological parameters in the model using a Gibbs sampling approach, so that the likelihood of the genealogy could be computed conditional on the current mapping of lineage states onto the genealogy and then a new mapping of lineage states could be sampled conditional on the current epidemiological parameters. However, these types of approaches did not work in practice because of strong correlations among the lineage states and the epidemiological parameters, which made it virtually impossible to achieve good MCMC mixing. Instead, I developed a modified version of the earlier PMMH algorithm that uses a particle filter to simultaneously integrate over the population dynamic variables and the unobserved lineage states, which allowed for far more efficient inference under structured models.

In Chapter 3, I describe the modified PMMH algorithm for structured models and apply it to HIV sequence data from Detroit, Michigan. The HIV application demonstrates how the algorithm can be used to infer stage-specific transmission rates for HIV as well as to reconstruct the changing patters of HIV incidence over the epidemic. I then explore more broadly how much information genealogies contain about population structure and parameters such as migration rates. During the course of testing the method, I found that there may be inherent limits to what can be inferred from genealogies about population structure irrespective of the algorithmic

issues discussed above. I found that the ability to precisely estimate parameters like migration rates from genealogies depends on how much uncertainty there is in the probable state of lineages along the internal branches of the genealogy. Although we may know the state of the lineage at the time of sampling, information about the probable state of a lineage generally decays as we move into the past. In Chapter 3, I show that the rate at which information about the lineage states decays depends in turn on the rate at which lineages transition between different populations. Critically, if information about the probable state of lineages decays much faster than lineages coalesce in the genealogy, then it is generally not possible to precisely infer migration rates or other parameters relating to population structure from genealogies.

1.4 Phylodynamic reconciliation

The methods described in Chapters 2 and 3 open the way for a wide variety of epidemiological models to be fit to genealogies. Using these methods, it becomes possible to revisit cases where phylodynamic reconstructions of population dynamics have differed from what was expected or observed in other data, and to then explore how these discrepancies arose. For example, several authors have suggested that discrepancies between phylodynamic estimates and other observational data may be due to an inappropriate or misspecified coalescent model that ignores important features of a population's demography or ecology (Carrington et al., 2005; Pybus and Rambaut, 2009; Siebenga et al., 2010). With the ability to directly fit coalescent models that include additional demographic and ecological complexities, it becomes possible to compare the population dynamics reconstructed under different coalescent models and test whether or not using more realistic coalescent models leads to more accurate estimates.

In Chapter 4, I use dengue serotype 1 (DENV-1) in southern Vietnam as a case study to better understand how discrepancies between dynamics observed in time

series data and those reconstructed from genealogies arise. DENV-1 provided a good case study because reliable time series of hospitalization case reports are available from hospitals in Ho Chi Minh City, to which phylodynamic estimates can be directly compared. A large amount of whole-genome sequence data for DENV-1 was also made available to me by Cameron Simmons' group at the Oxford University Clinical Research Unit in Vietnam. Using this sequence data, it was possible to reconstruct genealogies with relatively low amounts of phylogenetic uncertainty, ruling out the possibility that any inability to reconstruct dengue's dynamics was due to a lack of phylogenetic signal in the sequence data. Yet despite adequate sequence data, I was unable to reconstruct the highly seasonal incidence pattens observed in the dengue hospitalization data using standard coalescent-based approaches like the Bayesian skyline. Moreover, I was unable to reconstruct seasonal dynamics even when using a SIR model that allowed for dengue transmission dynamics to seasonally vary.

Working with Maciej Boni and Katia Koelle, I therefore explored how including other ecological and demographic complexities in the coalescent model used for inference affected phylodynamic estimates. I show in Chapter 4 that including vector-borne transmission by mosquitoes and spatial structure in the host population allowed us to reconcile phylodynamic estimates of DENV-1's dynamics with the hospitalization data. Models that include the vector and spatial structure fit the DENV-1 genealogies far better than models without vectors or population structure in terms of formal model comparisons. Moreover, using these more realistic coalescent models, I was able to reconstruct dengue population dynamics that were highly consistent with the hospitalization data, demonstrating the utility of using more mechanistic epidemiological models in phylodynamics that can incorporate additional ecological complexities.

1.5 A final note

The next two chapters describe the phylodynamic inference methods introduced above and then the application of these methods to DENV-1 in southern Vietnam is presented in Chapter 4. All three of these chapters were adapted from manuscripts that have been published in peer reviewed journals (Rasmussen et al., 2011, 2014a,b). Each of these manuscripts benefited greatly from the contributions of my collaborators and coauthors. In particular, I would like to acknowledge Oliver Ratmann for first recommending particle MCMC methods to me and providing statistical guidance with Chapter 2, Erik Volz for his theoretical contributions to the structured coalescent models presented in Chapter 3, Maciej Boni for hosting me in Vietnam and guidance on presenting the findings of Chapter 4, and Katia Koelle for first proposing the idea of developing new phylodynamic methods and her invaluable help and guidance on all three of of these chapters. It should be noted that whenever I use the word "we" in the following chapters, it refers to the coauthors of the published version of these manuscripts.

2

Inference for Nonlinear Epidemiological Models using Time Series and Genealogies

2.1 Introduction

Epidemiologists increasingly rely on the ability to fit mechanistic models of disease transmission to data in order to estimate key parameters and elucidate the underlying processes driving disease dynamics. However, the nature of epidemiological data makes model fitting statistically challenging. Case report data such as time series of disease incidence are often incomplete or subject to severe biases like underreporting. Moreover, disease dynamics are generally only partially observed in that the exact times at which infection and recovery events occur are rarely, if ever, directly observed (Cauchemez and Ferguson, 2008; O'Neill and Roberts, 1999; O'Neill, 2010). Researchers have therefore turned to the large amounts of molecular sequence data becoming available when case report data are insufficient. Gene genealogies can be reconstructed from sequence data and the times of coalescence events (i.e. branching events) can be used as a proxy for the timing of a subset of transmission events in the population. Using coalescent-based phylodynamic methods, it is then possible

to infer the past dynamics of a disease from the lineages present in the genealogy, opening up the possibility of fitting models directly to genealogies.

Several coalescent-based methods for inferring past population dynamics from genealogies have already been developed (Drummond et al., 2005; Kuhner et al., 1998; Minin et al., 2008; Strimmer and Pybus, 2001). These methods employ the basic result of coalescent theory that the rate of coalescence is inversely proportional to the effective population size, N_e (Hudson et al., 1990). Given the distribution of coalescence times over a genealogy, it is then possible to infer N_e, which for an infectious disease is generally interpreted as an estimate of the number of infected hosts (Pybus et al., 2001). Past population dynamics can also be inferred by specifying a demographic model and fitting it to a genealogy (Griffiths and Tavare, 1994; Nee et al., 1995). Most often, these demographic models are phenomenological and use simple parametric functions (e.g. constant size, exponential growth or logistic growth) or nonparametric functions that constrain population sizes to change smoothly or only at certain points in time (Drummond et al., 2005; Minin et al., 2008; Strimmer and Pybus, 2001). Fitting simple parametric models like exponential growth to genealogies can provide insight into the epidemic dynamics of pathogens and provide estimates of epidemic growth rates and times of emergence (Carrington et al., 2005; Fraser et al., 2009; Lemey et al., 2004). Phylodynamic methods have also been applied to systems with far more complex endemic disease dynamics where the prevalence of the disease can fluctuate rapidly or undergo complex periodic oscillations. Remarkably, phylodynamic analyses of RNA viruses can sometimes recover features of their complex population dynamics due to a fast rate of sequence evolution and sampling of sequences over time (Bennett et al., 2010; Rambaut et al., 2008; Siebenga et al., 2010).

While the vast majority of phylodynamic studies have inferred past dynamics by fitting phenomenological models to genealogies, a smaller body of work has in-

vestigated fitting mechanistic population dynamic models such as the well-known Susceptible-Infected-Recovered (SIR) class of models for the transmission dynamics of an infectious disease (Pybus et al., 2001; Volz et al., 2009). Using mechanistic population dynamic models in place of phenomenological models may have major benefits. First, biologically important parameter values can be estimated along with past population dynamics, which can provide insights into the underlying processes driving historical population dynamics. For example, Pybus et al. (2001) were able to estimate the basic reproductive number R_0 from viral genealogies for subtypes of Hepatitis C virus. Second, using these types of models should also improve our ability to correctly infer complex population dynamics, as they are constrained by population size trajectories that are dynamically feasible, rather than only biologically sensible (e.g., by being temporally continuous).

While the field of phylodynamics has made tremendous progress in recent years, methodological constraints limit the use of phylodynamic methods in epidemiological modeling more generally. First, only relatively simple epidemiological models where the number of infected hosts is a deterministic function of time can currently be fit using standard coalescent-based methods (Drummond et al., 2005; Volz et al., 2009; Pybus et al., 2000). However, epidemiological dynamics are inherently stochastic and both demographic and environmental stochasticity can play important roles in disease dynamics (Coulson et al., 2004; Earn et al., 2000; Rohani et al., 2002). Stochastic models are also essential for statistical inference since the variability, or over-dispersion, observed in real data can only be described statistically if stochasticity is included in the model (Bretó et al., 2009). This is especially true when fitting models to long-term data where the effects of stochasticity can accrue over time and cause the observed disease dynamics to deviate widely from the expectations of a deterministic model.

Current phylodynamic methods are also limited in that inference can only be

conducted using genealogies. While phylodynamic methods will generally be used in the absence of historical data, other sources of data such as time series may be available alongside of sequence samples. This is especially the case for well-studied RNA viruses, where time series of case report data are collected as part of epidemiological surveillance programs. A number of statistical methods already exist for fitting mechanistic population dynamic models to time series data (Cauchemez and Ferguson, 2008; Finkenstädt and Grenfell, 2000; Ionides et al., 2006). Generalizing such methods to fit mechanistic population models to genealogies as well would allow for inferences to be drawn from both time series and genealogies. Such an approach would allow for direct comparison between estimates derived from genealogies with estimates derived from time series data. Moreover, inference could then be conducted using both genealogical and population incidence data, potentially leading to more robust results.

The field of phylodynamics could therefore greatly benefit from having more flexible methods for genealogical-based inference. To this end, we have developed a general framework for phylodynamic inference that accommodates stochastic, mechanistic population dynamic models and can be integrated with other sources of data such as time series. In our framework, state-space models (SSMs) are used to model underlying biological processes mechanistically. While SSMs are already commonly fit to time series, we show how SSMs can also be fit to genealogies using coalescent methods. This allows for the model parameters and past population dynamics to be inferred from genealogies with or without accompanying time series data. Full Bayesian inference of all model parameters and past dynamics is performed using a method known as particle Markov Chain Monte Carlo (particle MCMC) (Andrieu et al., 2010), which uses particle filtering methods to fit SSMs to data without requiring an analytical likelihood function (Bretó et al., 2009; Ionides et al., 2006). This makes it possible to use a wide-class of SSMs for phylodynamic inference, including

the stochastic, continuous-time dynamic models commonly used in epidemiology and population biology.

We present our approach by first briefly reviewing the fundamentals of SSMs and the particle MCMC method. We then present a stochastic SIR model for the dynamics of an infectious disease that we use throughout the paper as our SSM. For conceptual clarity, we first show how particle MCMC can be used to fit a SSM to time series data without a genealogy since this is a familiar problem in statistical inference. We then go on to show how the SSM framework can be expanded to include genealogies and how particle MCMC can be used to infer model parameters and past population dynamics from genealogies with or without accompanying time series data. Finally, we test our particle MCMC approach on simulated time series and genealogies. We find that reliable estimates of model parameters and past population dynamics can be obtained from time series data, a genealogy, or both. Moreover, we find that estimates obtained from genealogies approach the accuracy of estimates obtained from time series data when a large number of samples are collected serially over time.

2.2 Methods

The general statistical framework we use to fit population dynamic models to either genealogies or time series data is based on state-space modeling. Structurally, state-space models (SSMs) consist of a process model and an observation model. The process model describes the underlying dynamics of the state variables x_t as a Markov process with model parameters θ for all time points t in $\{1, \ldots, T\}$:

$$x_t \sim p(x_t|x_{t-1}, \theta). \tag{2.1}$$

Below, we use a SIR compartmental model (Anderson and May, 1991) as the process model for the transmission dynamics of an infectious disease, with state variables

being the number of susceptible (S), infected (I), and recovered (R) individuals. The exact state of the population at any given time (e.g., S_t, I_t, R_t) is generally not observable. The state variables therefore remain unknown latent variables that must be inferred from available data. We therefore need an observation model to relate the observed data z_t to the underlying process model:

$$z_t \sim p(z_t|x_t, \theta). \tag{2.2}$$

For example, we will use an observation model that accounts for normally distributed observation noise in time series observations. While SSMs are typically used with time series data, here we use a more general approach where a coalescent model can be used in place of an observation model to relate a genealogy to the state variables in the process model.

To fit state space models to genealogical and/or time series data $z_{1:T}$, we use a Bayesian approach. Our primary goal is to find the posterior density of parameters θ and latent state variables $x_{1:T}$,

$$p(\theta, x_{1:T}|z_{1:T}) = \frac{p(z_{1:T}, x_{1:T}|\theta)p(\theta)}{p(z_{1:T})}. \tag{2.3}$$

From the posterior density, point estimates of model parameters as well as measures of uncertainty can be easily derived. However, for the models we consider here, the posterior density is analytically intractable. We therefore use an MCMC algorithm to sample from $p(\theta, x_{1:T}|z_{1:T})$ (for background on MCMC methods, see Gilks et al. (1996)). For illustrative purposes, we first present the following simple MCMC algorithm. Given current values of θ and $x_{1:T}$, we:

1. Propose new values of θ and $x_{1:T}$ by sampling from the proposal density $q(\theta^*, x^*_{1:T}|\theta, x_{1:T})$.

2. Evaluate the posterior probability of θ^* and $x^*_{1:T}$ given $z_{1:T}$ by computing $p(z_{1:T}, x^*_{1:T}|\theta^*)p(\theta^*)$.

3. With probability

$$\min\left(\frac{p(\theta^*, x^*_{1:T}|z_{1:T})}{p(\theta, x_{1:T}|z_{1:T})}\frac{q(\theta, x_{1:T}|\theta^*, x^*_{1:T})}{q(\theta^*, x^*_{1:T}|\theta, x_{1:T})}, 1\right), \tag{2.4}$$

set $\theta = \theta^*$ and $x_{1:T} = x^*_{1:T}$; otherwise set $\theta = \theta$ and $x_{1:T} = x_{1:T}$.

In practice, there are two major problems with using this naive MCMC approach. First, choosing an efficient proposal density for nonlinear and high-dimensional models is challenging (O'Neill, 2002). Second, it is often difficult or impossible to evaluate the likelihood needed in step 2 when the disease dynamics are only partially observed through temporally aggregated data and the exact infection times are unknown (O'Neill and Roberts, 1999; Becker and Britton, 1999). In our case, there is no analytical expression to impute over all unobserved infection times for continuous time, stochastic population models. We therefore use an approach known as particle MCMC (Andrieu et al., 2010), which employs a particle filtering algorithm to numerically construct an efficient proposal density without requiring that the likelihood be computed analytically.

The particle MCMC algorithm is essentially a particular version of the MCMC sampler presented above. While new values of θ and $x_{1:T}$ can be proposed together in Step 1, in particle MCMC new values for θ are first sampled from the proposal density $q(\theta^*|\theta)$ and then $x^*_{1:T}$ is independently proposed by sampling sequentially from $p(x_{1:T}|\theta^*, z_{1:T})$, so that the proposal density has the form

$$q(\theta^*, x^*_{1:T}|\theta, x_{1:T}) = q(\theta^*|\theta)\hat{p}(x^*_{1:T}|\theta^*, z_{1:T}), \tag{2.5}$$

where $\hat{p}(x^*_{1:T}|\theta^*, z_{1:T})$ is a Monte Carlo estimate of $p(x_{1:T}|\theta^*, z_{1:T})$ that must be obtained with a particle filtering algorithm (see below). The proposed $x^*_{1:T}$ is therefore "adapted" to the data, which in our case, greatly improves MCMC efficiency (Andrieu et al., 2010). As shown by Andrieu et al. (2010), the acceptance probability in

Step 3 is exactly given by

$$\min\left(\frac{\hat{p}(z_{1:T}|\theta^*)p(\theta^*)}{\hat{p}(z_{1:T}|\theta)p(\theta)}\frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}, 1\right), \tag{2.6}$$

where the Monte Carlo estimate $\hat{p}(z_{1:T}|\theta^*)$ to the marginal likelihood is a byproduct of the particle filtering algorithm (see below). The full justification for using this acceptance probability is non-trivial, and we refer to Andrieu et al. (2010). We can therefore approximate the joint posterior density of θ and $x_{1:T}$ using particle MCMC, which would otherwise be very difficult or impossible using standard MCMC methods.

The particle filtering algorithm used in particle MCMC allows us to numerically approximate $p(x_{1:T}|\theta, z_{1:T})$ by simulating the unknown trajectories of the state variables from the process model (for reviews, see Cappe et al. (2007) and Doucet et al. (2001)). The key idea behind particle filtering is to update particles sequentially through time so that at any time t, the weighted particles provide an approximation to the density $p(x_{1:t}|\theta, z_{1:t})$. This is done by propagating particles forward from time $t-1$ to t at each observation point in a two-step process. First, the state of each particle is updated by sampling new values from an importance density $q(x_t^j|x_{t-1}^j, z_t, \theta)$, where x_t^j refers to the state of the jth particle at time t. Second, after the state of the particles has been updated, each particle is filtered according to the observation model and assigned a weight w_t^j. In general, the unnormalized particle weights are calculated as

$$w_t^j = \frac{p(x_t^j|x_{t-1}^j, \theta)p(z_t|x_t^j, \theta)}{q(x_t^j|x_{t-1}^j, z_t, \theta)}. \tag{2.7}$$

In our case, there is no ideal importance density to sample from and particles are propagated by simulating directly from the process model (Ionides et al., 2006; Cappe et al., 2007; Gordon et al., 1993), so that equation (2.7) simplifies to:

$$w_t^j = p(z_t|x_t^j, \theta). \tag{2.8}$$

In other words, the unnormalized weight assigned to a particle is simply the probability of observing the data z_t given the state of the particle as specified by the observation model. The unnormalized weights can then be summed to approximate the conditional marginal likelihood $p(z_t|z_{1:t-1}, \theta)$,

$$\hat{p}(z_t|z_{1:t-1}, \theta) = \bar{w}_t = \frac{1}{N}\sum_{j=1}^{N} w_t^j. \tag{2.9}$$

By the law of total probability, an approximation to the marginal likelihood of the entire series of observations given θ is simply

$$\hat{p}(z_{1:T}|\theta) = \prod_{t=1}^{T} \bar{w}_t. \tag{2.10}$$

This numerical approximation to the marginal likelihood is exactly the term that is required to evaluate the acceptance probability in (2.6) needed to perform MCMC sampling.

A common problem with particle filtering algorithms is that particle weights degenerate over time, meaning that most particles will carry little weight while a few will carry most of the weight (Cappe et al., 2007; Doucet et al., 2001). If this occurs, the particle system will not provide a good approximation to the density $p(x_{1:t}|\theta, z_{1:t})$. For long time series, this becomes a serious problem. The standard way of dealing with this issue is to resample particles from the population so that unpromising particles with low weights are not propagated forwards through time while promising particles are used to replenish the particle population (Chopin, 2004). We therefore calculate the normalized weights of each of the particles:

$$W_t^j = \frac{w_t^j}{\sum_{j=1}^{N} w_t^j}, \tag{2.11}$$

and then resample particles according to their weights by multinomial sampling with replacement so that the total number of particles remains constant. Resampling is

performed after every time step, after which particle weights are reset to $1/N$. This particular particle filtering algorithm is known as bootstrap filtering and has the nice property that particle weights are independent of the particle's past trajectory (Doucet et al., 2001; Gordon et al., 1993). Note that without resampling, a proposal for $x_{1:T}$ in each step of the particle MCMC algorithm can be obtained simply by sampling a single particle trajectory $x_{1:T}^j$ from the particle filter approximation to $p(x_{1:T}|\theta, z_{1:T})$. However, because particles are resampled at each time step in the particle filter, we have to track the ancestry of particles so that a single trajectory representing the path of a single particle through state space can be sampled. Pseudo code for the full particle filtering algorithm with resampling is given in Appendix A.

2.2.1 Inference with time series data

We first consider fitting a SSM to time series data $y_{1:T}$ using particle MCMC. As our process model, we use a Susceptible-Infected-Recovered (SIR) epidemiological model with noise arising from variability in the transmission rate due to environmental factors. Using the Euler-Maruyama method, we can simulate this model forward in time with equations:

$$S_{t+dt} = S_t + \mu N dt - \mu S_t dt - (1 + F\xi)\beta(t)\frac{S_t}{N} I_t dt \tag{2.12a}$$

$$I_{t+dt} = I_t + (1 + F\xi)\beta(t)\frac{S_t}{N} I_t dt - \gamma I_t dt - \mu I_t dt \tag{2.12b}$$

$$R_{t+dt} = R_t + \gamma I_t dt - \mu R_t dt, \tag{2.12c}$$

where μ is the host birth/death rate, γ is the rate of recovery, $\beta(t)$ is the seasonally varying transmission rate, and N is the constant population size of the host, which we assume is known. We let the transmission rate vary sinusoidally with strength of seasonality α, so that $\beta(t) = \bar{\beta}(1 + \alpha \sin(2\pi t))$ and the mean transmission rate is given by $\bar{\beta} = R_0(\gamma + \mu)$, where R_0 is the basic reproduction number. The noise term ξ is given by $\frac{W}{\sqrt{dt}}$, where W is a normal random variate with mean equal to zero

and variance equal to one (Keeling and Rohani, 2008). The constant F scales the magnitude of environmental noise. Along with the equations in (2.12), we simulate a compartment, C, that tracks the cumulative number of infected individuals over time (i.e. the cumulative incidence):

$$C_{t+dt} = C_t + (1 + F\xi)\beta(t)\frac{S_t}{N}I_t dt. \tag{2.13}$$

From C, we can compute the number of new infections occurring between any two time points $t-1$ and t: $c_t = C_t - C_{t-1}$. Assuming that only a fraction ρ of these new cases are observed, and that observation error is normally distributed, the likelihood of observing y cases at time t is given by the observation model:

$$p(y_t|c_t) = Norm(y_t|\rho c_t, \tau\rho c_t), \tag{2.14}$$

where the mean is given by $\mu = \rho c_t$ and the observation variance is given by $\sigma^2 = \tau\rho c_t$, which depends on a scaling parameter τ, as in Ionides et al. (2006).

Adapting the particle MCMC algorithm described above to fit the SIR model to time series data is straightforward. In the particle filtering algorithm, particle trajectories are simulated from equations (2.12) and (2.13) with process noise, so that each particle has a simulated incidence value c_t^j. Particle weights are assigned using the observation model given in (2.15), so that unnormalized particle weights are calculated as

$$w_t^j = Norm(y_t|\rho c_t^j, \tau\rho c_t^j). \tag{2.15}$$

Particle MCMC can then be used to sample from the posterior density $p(\theta, x_{1:T}|z_{1:T})$. Here, θ contains all the parameters in the SIR model as well as the observation model parameters ρ and τ. We can infer the trajectory of any of the state variables but we limit ourselves to inferring I so that $x_{1:T}$ stands in for the number of infected hosts from $t = 1$ to T. Likewise, the initial conditions for all the state variables could also

be inferred but we do not estimate them here since they are known values in the mock data we use to test the algorithm.

2.2.2 Inference with a genealogy

We now turn to using particle MCMC to infer model parameters and latent variables from a genealogy. For illustrative purposes, we will use the same epidemiological model as detailed above. To see how a genealogy can be used to reconstruct the past population dynamics of a disease, first imagine that every infected individual is included in an infection tree with branching times that correspond to transmission events and tip times that correspond to recovery events. In this hypothetical case, the past prevalence of the disease at any time would simply be the number of lineages present in the genealogy at that time and the likelihood of the genealogy under a given population dynamic model could easily be computed since the times at which infection and recovery events occur would be known. In reality, we cannot observe the complete genealogy but we can reconstruct a partial genealogy from sequences sampled randomly from infected individuals over time. Coalescent theory provides us with the necessary probabilistic relationship between an incomplete genealogy and the underlying population dynamics $x_{1:T}$ needed to fit a SSM to a genealogy of randomly sampled individuals. Specifically, the coalescent model will allow us to calculate the likelihood of observing a certain genealogy given the population dynamics $x_{1:T}$, just as the observation model allowed us to calculate the likelihood of time series observations given $x_{1:T}$.

Under the standard neutral coalescent model, the times between coalescent events in a genealogy are exponentially distributed so the probability of observing a coalescent event after time t is

$$f(t) = \lambda e^{-\lambda t}, \tag{2.16}$$

where λ is the rate of coalescence. For many population models, the rate of coa-

lescence depends on the number of lineages present in the genealogy i, the effective population size N_e, and a factor τ that rescales generation time into calendar time, so that

$$\lambda = \frac{\binom{i}{2}}{N_e \tau}. \tag{2.17}$$

For an infectious disease, N_e depends on the number of infected hosts I and the variance in the number of secondary infections an infected individual causes (Rodrigo and Felsenstein, 1999). Genealogical time has generally been rescaled into calendar time by defining the generation time scaling factor τ as the duration of infection (Drummond et al., 2005; Pybus et al., 2000). However, for an epidemiological model like the SIR model, the generation time of the disease is more appropriately defined as the average length of time it takes an infected individual to infect a susceptible host. Under our SIR model, the generation time is not constant over time since it depends on the rate at which infections occur, which is equal to $\beta(t)\frac{S_t}{N}I_t$. There is therefore no linear relationship that can be used to rescale genealogical time into calendar time. We therefore follow Volz et al. (2009), and write the rate of coalescence under our SIR model as:

$$\lambda_t = \frac{\binom{i}{2}}{\binom{I_t}{2}}\beta(t)\frac{S_t}{N}I_t. \tag{2.18}$$

Thus, (2.18) has the intuitive interpretation that the rate of coalescence is equal to the overall rate of transmission in the population multiplied by the probability of observing a transmission event in the sample fraction, which is given by the ratio of the two binomial coefficients in the leading term. In practice, we round I_t to its nearest integer value when computing the rate of coalescence so that $\binom{I_t}{2}$ is always computable.

The exponential probability density function given in (2.16) can be combined with the expression for the rate of coalescence for an SIR model given in (2.18) to

calculate the likelihood of the waiting time between any two coalescence events as a function of the state variables in the SIR model. The total likelihood of a genealogy can therefore be obtained by dividing the genealogy into coalescent intervals and taking the product of the likelihoods over all coalescent intervals. However, to enable comparison with inference using time series, the genealogy must also be partitioned at intervals that correspond to the observation times $1:T$ in the time series. Each of these time intervals is further divided into subintervals of size dt, where dt is the time step used in the simulation of the process model, given by equations (2.12) above. We assume that these dt subintervals are sufficiently small so that the number of infected and susceptible individuals does not significantly change within a subinterval. This assumption makes the rate of coalescence constant within subintervals, allowing us to use the exponential density given in (2.16) to compute the likelihood of the genealogy over these short subintervals. In addition to these intervals and subintervals, we allow for the general case that sequence data are sampled serially over time (i.e., the genealogy is heterochronic), such that, altogether, there are four types of time points which divide the genealogy into temporal sections: observation time points $1:T$, time points every dt between these time points, sequence sampling times, and times at which lineages coalesce. The main difference between using a genealogy instead of time series data is that the observed data z_t are now the vector of time subintervals ω_t between two observation time points $t-1$ and t, created by the dt time points, the sequence sampling times, and the coalescent times, rather than time series counts y_t.

To compute the likelihood of the genealogy over a given time interval $p(\omega_t|x_t)$, we can first write it as a joint probability of observing each subinterval time:

$$p(\omega_t|x_t) = \prod_{j=1}^{k} p(\omega_{t_j}|x_{t_j}). \tag{2.19}$$

Here j indexes the subinterval, and k is the number of subintervals in the observation

time interval ending at time t. The likelihood of observing a subinterval time ω_{t_j} is simply given by (2.16) above if the subinterval ends in a coalescent event:

$$p(\omega_{t_j}|x_{t_j}) = \lambda_{t_j} e^{-\lambda_{t_j}\omega_{t_j}}, \tag{2.20}$$

where λ_{t_j} is the instantaneous rate of coalescence at time t_j, which can be computed from the values of the state variables in x_{t_j} using (2.18). If the subinterval does not start at a dt partition time, but instead at a coalescent event or a sampling event, x_{t_j} are the state variables at the closest dt partition time in the future.

The probability of observing subinterval time w_{t_j} if subinterval j does not end in a coalescent event is given by the probability that a coalescent event has not occurred within this time period, which is

$$p(\omega_{t_j}|x_{t_j}) = 1 - \int_{t=0}^{\omega_{t_j}} \lambda_{t_j} e^{-\lambda_{t_j}\omega_{t_j}} dt = e^{-\lambda_{t_j}\omega_{t_j}}, \tag{2.21}$$

as described by Rodrigo and Felsenstein (1999). In the context of particle MCMC, the likelihood of the genealogy over the observation interval given by (2.19) is used to weight each particle at observation time t as described above.

2.2.3 Inference with both time series and a genealogy

Finally, we show how model parameters and past population dynamics can be inferred from both time series data and a genealogy together with particle MCMC. As before, we use the epidemiological model provided in (2.12). The joint likelihood of observing both the time series and the genealogy in the time interval between $t-1$ and t is given by:

$$p(y_t, \omega_t|\theta, x_t). \tag{2.22}$$

Assuming that the genealogy is independent of the time series data, this joint likelihood can be re-written as:

$$p(y_t, \omega_t|\theta, x_t) = p(y_t|\theta, x_t)p(\omega_t|\theta, x_t)). \tag{2.23}$$

Independence can be assumed if the samples in the genealogy are drawn from the infected population independently of which infected hosts are counted in the time series data. This is generally not the case, as the samples present in the genealogy are usually taken from a subset of infected hosts who are counted in the time series data. However, in our case, the fraction of infections counted in the time series data and the fraction present in the genealogy are both chosen at random. Therefore, the joint likelihood of observing both sets of data at time t, given the model and parameters θ, is given by the product of (2.15) and (2.19). In the context of particle MCMC, the unnormalized weight assigned to each particle is then the joint likelihood given in (2.23).

2.3 Results

To illustrate the ability of the particle MCMC algorithm to estimate model parameters θ and the latent state variables $x_{1:T}$ from time series data, we simulated a mock dataset using the SIR process model. Fig. 2.1A shows the simulated dynamics of the latent variable I over time. Fig. 2.1B shows the mock incidence data $y_{1:T}$ that are drawn using simulated c values (i.e. the cumulative incidence) and the distribution given in (2.15) to add normally distributed observation noise. The posterior densities of the process and observation model parameters inferred from the mock time series are shown in Fig. 2.2A-D. As shown, the algorithm provided accurate estimates of the SIR process model parameters, with the true parameter values generally falling well within the 95% Bayesian credible intervals (CI). For the parameters of the observation model, we were able to accurately estimate the reporting rate ρ but found the observation variance τ more difficult to estimate (Fig. 2.2E-F). The series of posterior densities for the latent variable I (i.e. the prevalence of the disease over time) show that the algorithm accurately estimated the dynamics of latent variables (Fig. 2.3A). The wider CI for the prevalence during seasonal peaks in prevalence

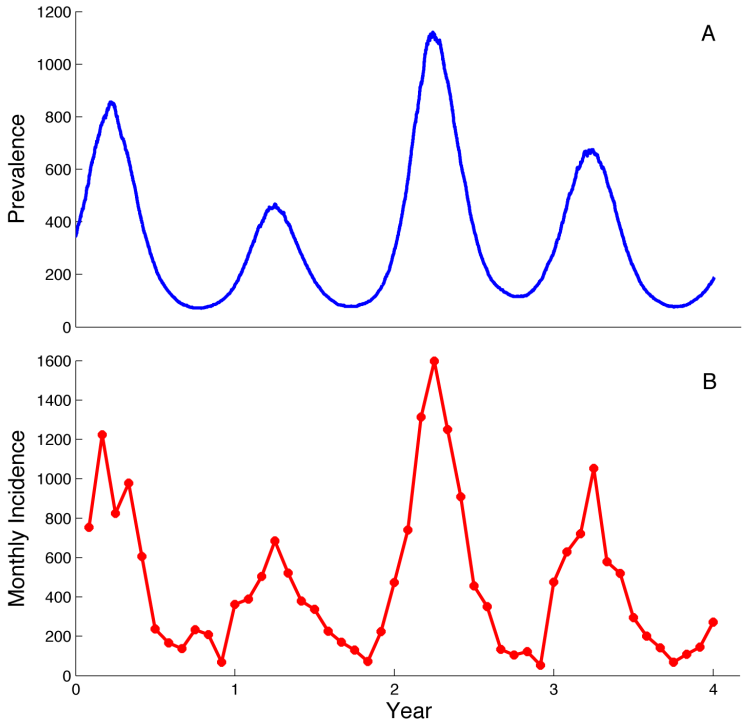

FIGURE 2.1: Simulated infection dynamics and time series used to test the particle MCMC algorithm. (A) Disease dynamics (I) obtained by simulating from the SIR process model over a 4-year period. (B) Corresponding time series of monthly incidence reports simulated from the observation model. Parameters used in the simulation of the process model were: $\gamma = 3$ month^{-1}, $R_0 = 10$, $\alpha = 0.16$, and $F = 0.012$. Other process model parameters that were assumed to be known were: $\mu = 0.0017$ month^{-1}, and $N = 5$ million. Parameters used in the simulation of the time series data were: $\rho = 0.43$, and $\tau = 15$.

relative to the offseason reflects the fact that environmental noise scales with the rate of transmission in our model, which is larger when prevalence is high.

We also tested the ability of the particle MCMC algorithm to infer parameters and past dynamics directly from genealogies. We obtained mock genealogies from our population dynamic simulations by tracking the ancestry of infections in the population and recording times at which infection and recovery events occurred. A subset of infection lineages were then randomly sampled at random times and their ancestry traced backwards through time so that transmission events correspond to

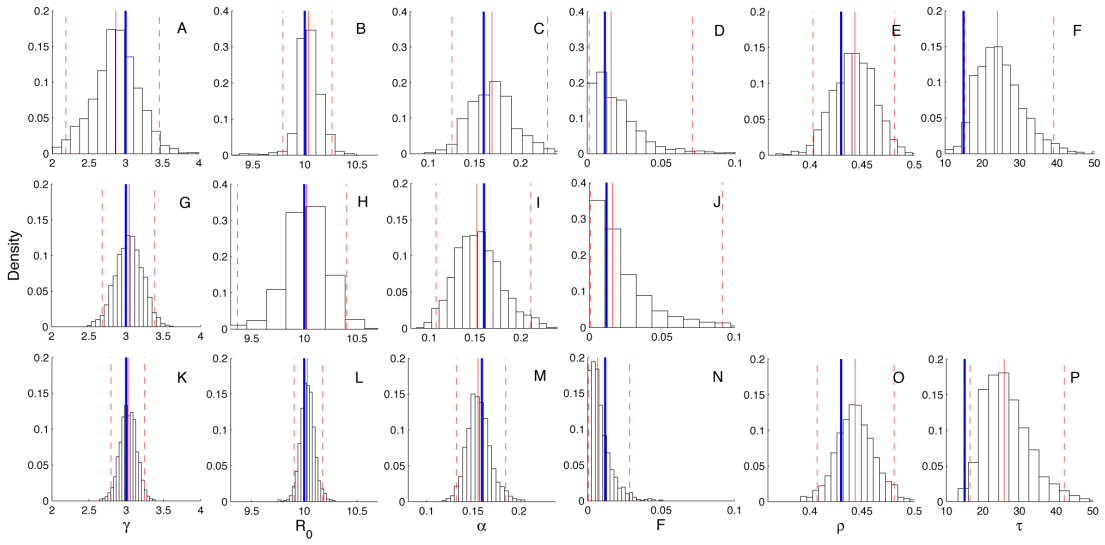

FIGURE 2.2: Frequency histograms representing the marginal posterior densities of the SIR model parameters obtained using the particle MCMC algorithm. Vertical blue lines are placed at the true values of the parameters, solid red lines are the median value of the posterior densities and dashed red lines mark the 95% Bayesian credible intervals. From left to right, the parameters are the recovery rate γ, the basic reproduction number R_0, the strength of seasonality α, the parameter scaling the strength of environmental noise F, the reporting rate ρ, and the observation variance τ. (A-F) Parameters inferred using time series data. (G-J) Parameters inferred using a genealogy. Parameters ρ and τ cannot be inferred using only a genealogy because they are parameters associated with the time series observation model. (K-P) Parameters inferred using both a genealogy and time series.

coalescence events among the sampled lineages. We first checked if the coalescent model could be used to provide accurate and unbiased estimates of epidemiological parameters from genealogies. To check for possible biases, we tested the algorithm using epidemic dynamics with parameter values that lead to an epidemic unfolding and ending within a 12-month period. The shorter length of these simulations allowed us to check the performance of the algorithm using genealogies obtained from simulating the epidemic dynamics 100 times. As can be seen from Fig. 2.4A-B, the epidemiological parameters R_0 and γ could be accurately estimated from the genealogies. However, we found it difficult to estimate the environmental noise term

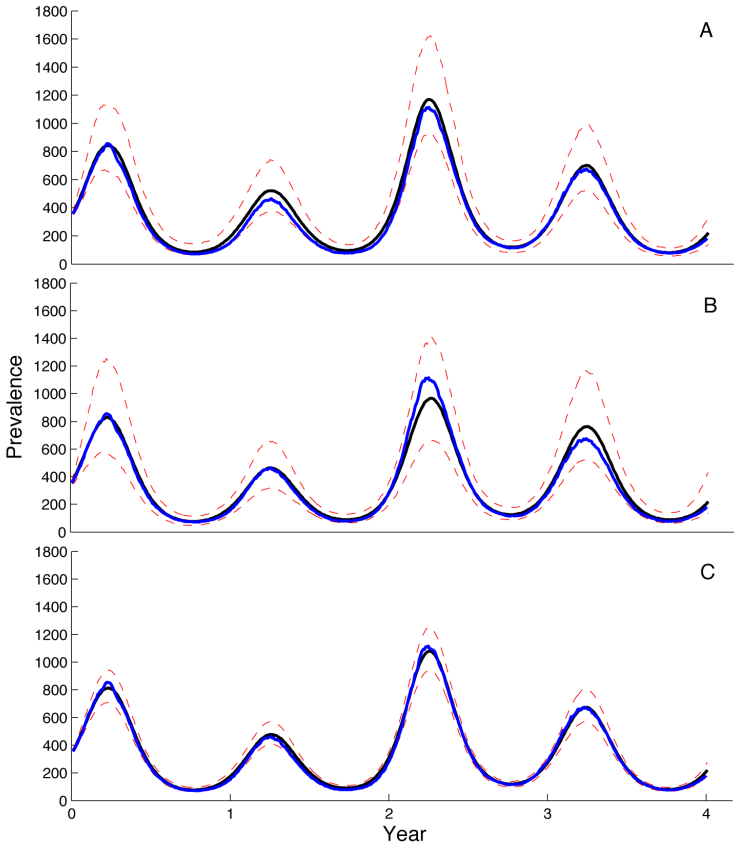

FIGURE 2.3: Series of posterior densities for disease prevalence I over time obtained using particle MCMC. Blue lines represent the exact simulated prevalence, black lines are the median of the posterior density and dashed red lines represent the 95% credible intervals. (A) Prevalence inferred from time series data. (B) Prevalence inferred from a genealogy. (C) Prevalence inferred from both a genealogy and time series.

F from genealogies over such a short time period, so we fixed F at its true value. Fig. 2.4C shows the distribution of the estimated median values of the posterior densities of R_0 and γ in parameter space for all 100 simulations. In spite of the strong negative correlation between these two parameters, the estimates cluster around the true parameter values, with the true values of R_0 and γ falling within the estimated 95% credible intervals 90 and 92 times out of the 100 simulations, respectively.

Because we were able to obtain accurate parameter estimates from genealogies

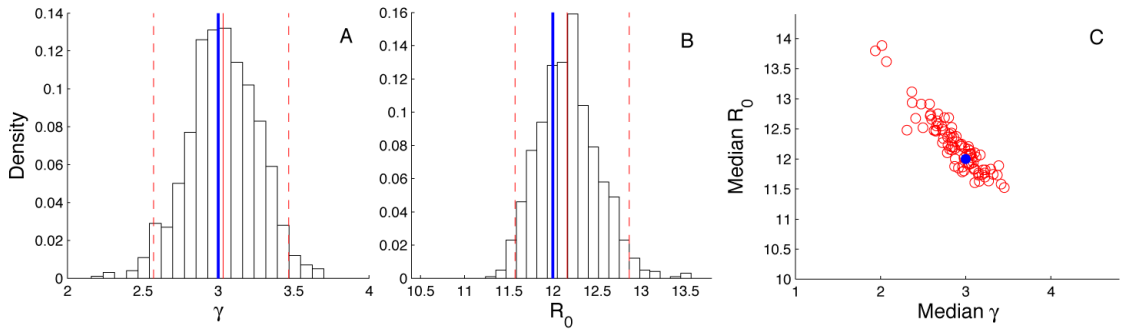

FIGURE 2.4: Posterior densities of the parameters γ and R_0 estimated from 100 independent genealogies obtained from simulated epidemic dynamics. (A-B) Frequency histograms representing the marginal posterior densities of γ and R_0 obtained from a single representative simulation. (C) The distribution of the median values of the posterior densities of γ and R_0 in parameter space for all 100 simulations (open red circles). The solid blue circle marks the true values of the parameters. Note that in our model formulation, γ and R_0 are independent parameters, with the transmission rate computed as $\beta = R_0(\mu + \gamma)$.

under simple epidemic conditions, we next tested the ability of the particle MCMC algorithm to estimate parameters and latent state variables from genealogies under more complex population dynamics. To do this, we generated a mock genealogy containing 200 terminal nodes from the same population dynamic simulation shown in Fig. 2.1. The mock genealogy is shown in Fig. 2.5. The posterior densities of the process model parameters inferred from the mock genealogy show that our method could accurately recover the values of the epidemiological parameters (Fig. 2.2G-J). The series of posterior densities for the latent variable I over time likewise show that our method can accurately estimate past disease dynamics from a genealogy (Fig. 2.3B). This is highly encouraging, as it suggests that both model parameters and past population dynamics can be accurately estimated from a genealogy even in the absence of any time series data as long as the number of sequences sampled over time is sufficiently large.

Although the credible intervals for the process model parameters and past disease

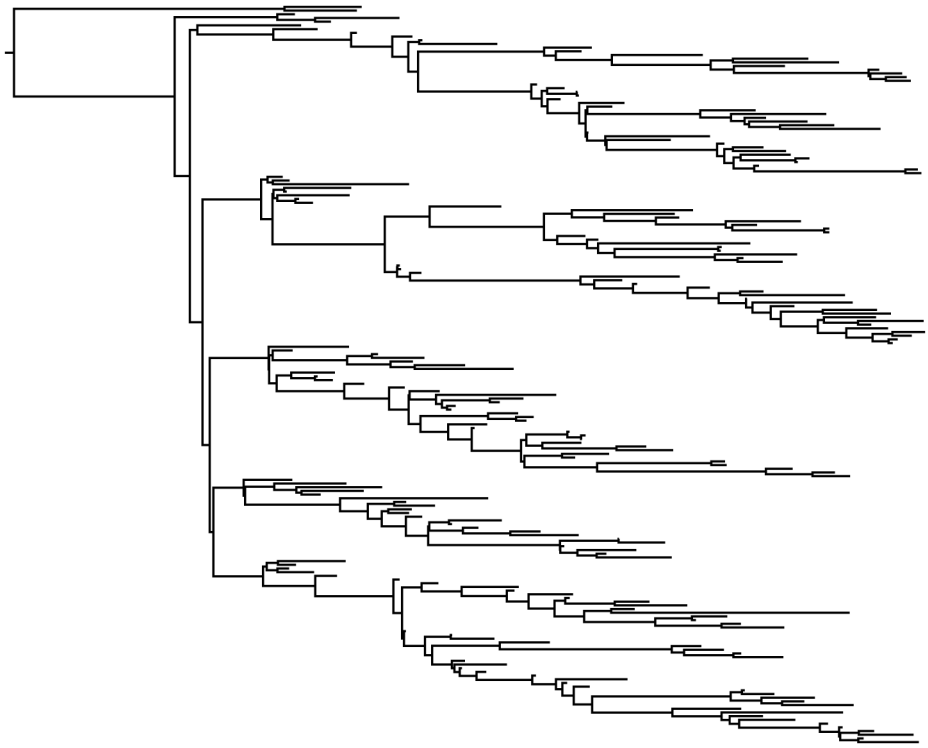

FIGURE 2.5: Genealogy obtained from the simulated disease dynamics shown in Fig. 2.1A. The genealogy contains 200 terminal nodes corresponding to sequence samples being collected sequentially over time with yearly sample sizes of approximately 50 sequences. Sampling events were chosen to occur at random times over the entire interval of the times series.

dynamics are wider when using the genealogy than when using the time series data, the width of the credible intervals likely depends heavily on the sampling effort. We therefore investigated a range of sample sizes to explore how different sample sizes affect the accuracy of and uncertainty associated with our estimates. Summary statistics for the posterior densities of the parameters and past prevalence of the disease are given in Table 2.1. Even with small sample sizes, reasonable estimates were obtained and the loss of accuracy in estimating parameters was most likely due to the difficulty of estimating the environmental noise term F, which is strongly correlated with other parameters, when the sample size was small. If the sample size is initially small, including more samples dramatically improves the accuracy

and reduces the level of uncertainty in parameter estimates. However, going from an intermediate number of samples (100-200) to a large number of samples (400) does not dramatically improve estimates, suggesting only a moderate amount of sequence data is required for accurate inferences to be drawn from genealogies. Similar results were obtained for estimates of the past prevalence of the disease. We quantified the effect of including more sequence data by computing the root mean squared deviation (RMSD) of the inferred median of the posterior densities of disease prevalence from the true prevalence values. Increasing the number of samples initially reduces the RMSD but including more samples provides no further advantage once a sufficient number of samples are included.

Finally, we combined the simulated time series and genealogy to illustrate the ability of particle MCMC to be used with both sources of data. In Fig. 2.2K-P, we show the posterior densities of the parameters when inferred from both the time series and a genealogy. In Fig. 2.3C, we show the series of posterior densities for the latent variable I over time inferred from both the time series data and the genealogy. As shown, including the genealogy along with the time series data considerably reduces the uncertainty in both the estimates of the process model parameters and the past prevalence of the disease.

Table 2.1: Median posterior values and 95% credible intervals for the parameters and past disease dynamics inferred from genealogies with different numbers of samples. Root Mean Squared Deviation (RMSD) for the prevalence was calculated using the deviation of the median of the posterior density from the true value summed over all time points.

Sample size	γ	R_0	α	F	Prevalence RMSD
40	2.36 [1.06-4.18]	3.66 [1.21-10.12]	0.25 [0.09-0.51]	0.56 [0.03-1.78]	244.77
100	2.94 [2.28-3.95]	8.74 [3.97-10.71]	0.19 [0.10-0.32]	0.14 [0.02-0.51]	93.23
200	3.05 [2.68-3.39]	10.02 [9.37-10.41]	0.15 [0.11-0.21]	0.016 [0.001-0.092]	39.55
400	3.00 [2.71-3.29]	10.03 [8.42-10.43]	0.16 [0.12-0.20]	0.026 [0.001-0.141]	73.89
True value	3.00	10.00	0.16	0.012	NA

2.4 Discussion

The framework we have developed extends phylodynamic inference in two major ways. First, stochastic state-space models that consider the biological processes driving population dynamics can be used instead of simple parametric or nonparametric demographic models when inferring past population dynamics. This also allows for key epidemiological parameters to be estimated directly from genealogies. Second, our approach allows for other sources of data such as time series to be considered along with a genealogy when inferring parameters and past population dynamics. Using a particle MCMC algorithm to fit a stochastic SIR model to simulated genealogies and time series data, we found that key epidemiological parameters as well as the past prevalence of the disease could be accurately estimated from genealogies with or without accompanying time series data.

While particle MCMC is computationally expensive because of the need to simulate particle trajectories each MCMC step, we believe it represents a good choice for the purposes of phylodynamic inference. First, particle MCMC allows for efficient MCMC sampling of model parameters and latent variables from their posterior densities even with high-dimensional, nonlinear SSMs. Secondly, particle MCMC is flexible in terms of the form of the SSMs that can be used. Because the particle filtering algorithm used in particle MCMC can be used to approximate the likelihood of the model through simulations, there is no need for an analytical likelihood function. Taken together, this allows for almost any infectious disease model to be used as long as particle trajectories can be simulated from the process model and an observation or coalescent model can be specified (Bretó et al., 2009; Cappe et al., 2007; Doucet et al., 2001). For example, several researchers including us here have used particle filtering methods to fit stochastic, continuous-time dynamic models to time series, even though observations occur only at discrete time points (He et al., 2010; Ionides

et al., 2006; King et al., 2008). Finally, particle MCMC allows for flexibility in terms of the types and structure of the data. As we have shown, fitting dynamic models to different sources of data is straightforward since only the particle weighting scheme needs to be modified. We therefore believe that the computational cost of particle MCMC is outweighed by its flexibility and ease of implementation for most practical purposes in phylodynamics. Still, the efficiency of other statistical methodologies such as approximate Bayesian computation (ABC, see Beaumont (2010)) should be compared against particle MCMC in the future to see if the computational overhead of conducting phylodyamic inference with complex models can be reduced.

The particle MCMC approach described here is also able to incorporate different forms of stochasticity, which is essential for fitting the variation, or over-dispersion, present in real disease data. For simplicity, we only included environmental noise in the transmission process—random variation in the rate at which transmission events occur due to external factors like climatic fluctuations. However, other forms of stochasticity could also be included such as demographic stochasticity—random variation in the timing of demographic events such as the birth and death of individuals. We did not consider demographic stochasticity because it involves event-driven simulation approaches that are much more computationally expensive than the Euler-Maruyama algorithm we used. However, for small populations where demographic stochasticity can play an important dynamical role, other simulation methods could be employed within the particle filtering algorithm. For example, Bretó et al. (2009) recently introduced a simulation method that can include both environmental and demographic stochasticity. While what form of stochasticity is appropriate will be system-dependent, the need for statistical methods that include stochasticity when fitting models to disease data has been demonstrated repeatedly (Rohani et al., 2002; Bretó et al., 2009; King et al., 2008). The particle MCMC approach therefore offers an advantage over other methods for phylodynamic inference that can only be used

to fit deterministic models.

The ability to accurately infer past population dynamics or model parameters from genealogies ultimately depends on how sequences are sampled. Since we were primarily concerned with statistical methodology, we did not extensively explore different sampling strategies and simply considered the case where sequences are sampled randomly over time. However, we did find that only a moderate number of sequences are necessary to obtain reliable parameter estimates. Even when the sampling rate was as low as 10 samples per year, reliable estimates were obtained. Likewise, extremely large sample sizes did not significantly improve estimates, suggesting phylodynamic inference can be conducted without extensive sampling over time. Furthermore, even fewer samples may be necessary if sequences are sampled strategically. For example, in a simulation study, Stack et al. (2010) found that accurate estimates of past population dynamics could be obtained using a variety of sampling protocols and that especially reliable estimates could be obtained if sequences are sampled towards the end of an epidemic rather than at the beginning of an epidemic. Our phylodynamic inference framework should therefore be able to give reliable estimates even if the sampling effort is not uniformly high over time.

We were also interested in when including the information contained within a genealogy alongside of time series data could improve estimation. At the most basic level, considering a genealogy where the coalescence times are known without error provides additional information in that the timing of coalescence events provides information about when transmission events occurred that is not present in temporally aggregated case report data. One could even imagine that knowing the complete genealogy of infections in the population would be preferable to having perfect case report data, since the exact times of infection will still not be known. In practice, we found that considering the genealogy alongside of time series data only significantly improved our estimates if there was observation error in the time series data. For

example, the parameters estimated from the time series data with and without the genealogy in Fig. 2.2 were done with a moderate level of observation error in the mock time series data. However, from our own experience, including the genealogy when there were low levels of observation error in the mock time series data did not significantly improve our estimates (results not shown). We therefore suspect that it will be helpful to include genealogical data only when the observed time series data are relatively uninformative about the true disease dynamics, such as when there is large degree of error in the case report data or when case report data are missing. Genealogies may also aid inference if aspects of the population dynamics such as periodicity or other long-term trends in disease dynamics are obscured by changes in reporting practices.

While we have shown that it is possible to fit complex population dynamic models to simulated genealogies, several challenges remain before this approach can be routinely applied to real data sets. First, while we conditioned our inference on knowing the true genealogy without error, the genealogy will have to be inferred from sequence data in any application of our method. Our uncertainty as to the true topology of the genealogy and the inferred coalescence times will then have to be considered. Fortunately, existing phylogenetic software packages like BEAST allow us to sample from the posterior distribution of trees so as to effectively integrate out phylogenetic uncertainty (Drummond and Rambaut, 2007). Furthermore, programs like BEAST also use an MCMC framework making it possible to estimate population dynamic parameters, the genealogy and the associated molecular evolutionary parameters together in a single MCMC framework by alternately sampling from the appropriate posterior densities.

Another challenge lies in formulating appropriate models for relating population dynamics to the reconstructed genealogy. The coalescent model we used may not be appropriate for all infectious diseases, just as the simple SIR model we used will

not be adequate to describe the population dynamics of all diseases. For one, our coalescent model assumes neutrality with no phenotypic variation in the pathogen population, but real populations will be structured into multiple competing strains with varying antigenicity, pathogenicity and replication rates. Beyond selection, the natural history of a disease and heterogeneities due to population subdivision or contact structure can also have profound effects on genealogies (O'Dea and Wilke, 2010; Wakeley, 2009). Likewise, sequence samples will often not be sampled randomly as assumed under standard coalescent models, leading to potential ascertainment biases if nonrandom sampling is not incorporated into coalescent models. However, the framework for phylodynamic inference presented here is extremely flexible and can be modified to accommodate more realistic population dynamic and coalescent models to account for these complications. For example, we have derived coalescent expressions for models with individual heterogeneity in infectivity and for SEIR models where infected individuals enter an exposed class before becoming infectious (Koelle and Rasmussen, 2012). Finally, when there are discrepancies between the disease dynamics inferred from genealogies and those observed in case report data, the ability to test different population dynamic and coalescent models in a coherent statistical framework will allow us to consider alternative hypotheses for what caused these discrepancies. This in turn should help improve our understanding of the complex ecological and evolutionary processes driving population dynamics—the central goal of phylodynamics (Grenfell et al., 2004; Holmes and Grenfell, 2009).

3

Phylodynamic Inference for Structured Epidemiological Models

3.1 Introduction

Genealogies can provide valuable information about the demographic history of a population because the demography of a population can dramatically shape the structure of a genealogy (Nee et al., 1995; Grenfell et al., 2004). For example, fluctuations in population size will shift the distribution of branching events, or coalescent times, over a genealogy relative to what would be expected for a population with a constant size (Donnelly and Tavaré, 1995). Other aspects of a population's demographic history can also leave behind distinctive genealogical patterns. For example, the structuring of a population into different subpopulations can influence the topology of genealogies, which is often seen as clustering among individuals sampled from the same subpopulation (Lewis et al., 2008). These observations have led to great interest in statistical methods for inferring demographic trends and parameters from genealogies and given rise to the new field of phylodynamic inference (Kuhner et al., 1998; Pybus et al., 2000; Beerli and Felsenstein, 2001; Grenfell et al., 2004; Stadler

et al., 2012).

Most statistical methods for reconstructing the demographic history of a population from genealogies have been motivated by coalescent theory, which provides a probabilistic framework for relating the demographic history of a population to a genealogy of individuals sampled from that population (Kingman, 1982; Wakeley, 2009). Critically, coalescent models provide a way to compute the probability of a given genealogy under a given demographic model. It is therefore possible to estimate parameters of a demographic model, such as population size, from a genealogy using likelihood-based inference methods. Extensions of this basic idea have been used to estimate changes in population size over time, for example by the Bayesian skyline methods available in the BEAST phylogenetic software package (Strimmer and Pybus, 2001; Drummond et al., 2005). Coalescent theory has also been extended to consider different forms of population structure, giving rise to structured coalescent models (Notohara, 1990; Takahata and Slatkin, 1990). Statistical methods that allow fitting of structured coalescent models to genealogies have the ability to estimate parameters relating to population structure, including migration rates between populations (Beerli and Felsenstein, 2001; Kuhner, 2006).

Recent developments in phylodynamics have focused on developing models and statistical methods for more complex demographic scenarios, which have been largely motivated by the application of coalescent methods to pathogens like RNA viruses with rapidly changing population sizes. For example, coalescent models have been developed for populations where birth (i.e. transmission) rates vary over time (Volz et al., 2009; Frost and Volz, 2010). Importantly, the framework of Volz et al. (2009) also considers the coalescent process in populations where transmission rates change over time in a nonlinear manner, as is often the case for epidemiological models like the well-known Susceptible-Infected-Recovered (SIR) model (Anderson and May, 1991). Coalescent models have also been developed for common epidemiological

scenarios with population structure that alters the rate of coalescence in the population (Koelle and Rasmussen, 2012), but these models are limited to populations at equilibrium. Finally, Volz (2012) presented a framework that brings together both complex population dynamics and population structure. This approach has great appeal as it generalizes coalescent models to allow both birth and migration rates to change over time as a function of the underlying population dynamics, which may be nonlinear and far from equilibrium.

Although recent advances with structured coalescent models have enabled the analysis of more complex epidemiological models, the statistical challenge remains of efficiently fitting stochastic population dynamic models to genealogies. These models can be extremely high-dimensional due to a large number of latent state variables for which we have no direct observations. In Rasmussen et al. (2011), a particle filtering approach was used to marginalize out these latent variables by forward simulating population dynamic trajectories from the epidemiological model and then averaging over these trajectories to a compute a marginal likelihood. For unstructured models, adapting particle filtering methods to coalescent-based inference is relatively straightforward as the likelihood of a genealogy is simply a function of the simulated population dynamic trajectories. However, for structured models the likelihood also depends on the internal states of lineages in the genealogy, which may change over time as lineages move between populations (Volz, 2012). The probable state of a lineage can only be calculated retrospectively conditional on the population's demographic history and the state of the lineage at the time of sampling. As we show below, these backward-time dependencies prevent the direct application of forward-time particle filtering methods to structured models.

We therefore present a new statistical approach for fitting stochastic population dynamics models to genealogies using the structured coalescent approach presented in Volz (2012) using a modified particle filtering algorithm. This modified algo-

rithm allows for efficient particle filtering under structured coalescent models where the probability that a lineage is in a certain population may depend on both the past dynamics of the population as well as future sampling of lineages. Using this algorithm, we can fit stochastic, nonlinear epidemiological models with essentially any form of population structure to genealogies as long as the model is Markovian. Because population structure arises naturally in many epidemiological models, we define population structure in a very broad sense and consider any model where the population of infected hosts is structured into different nonequivalent states and therefore lineages in different infected hosts do not necessarily have an equal probability of coalescing. This includes models with spatial structure, multiple stages of infection and models of vector-borne and other multi-host pathogens.

This chapter has the following structure. First, we present the forward-time epidemiological models that we use as examples throughout the paper. Next, we review the framework first developed in Volz (2012) for how coalescent models can be derived for a corresponding forward-time population dynamic model. We then describe how we can fit structured epidemiological models to genealogies given the corresponding structured coalescent model. The statistical method we describe combines MCMC methods with our particle filtering algorithm, and is a variation of the particle MCMC algorithm of Andrieu et al. (2010). Using simulated genealogies, we show that this algorithm can accurately reconstruct population dynamics in structured populations and obtain reliable estimates of epidemiological parameters such as transmission rates. We then apply our approach to the HIV epidemic in Detroit, Michigan in order to estimate stage-specific transmission rates and infer how prevalence and incidence have changed over the course of the epidemic. Finally, we explore under what conditions parameters relating to population structure can be inferred from genealogies and how factors such as sample size affect uncertainty in our estimates.

3.2 Methods

3.2.1 Epidemiological models

We will use epidemiological models to demonstrate how mechanistic population dynamic models can be fit to genealogies. More specifically, we will consider the type of Susceptible-Infected-Recovered (SIR) models widely used to study the transmission dynamics of infectious diseases (Anderson and May, 1991; Keeling and Rohani, 2008). In SIR-type models, the host population is divided into different compartments depending on the host's state (e.g. susceptible or infected). For generality, we let x_t be the vector that holds the number of hosts in each compartment at time t, for example $x_t = \{S_t, I_t, R_t\}$ for the standard SIR model. For stochastic models, the state variables in x_t are treated as random variables. We consider an epidemiological model to be structured if there is more than one class of infected host.

Applications. We use two simple structured epidemiological models throughout the paper as illustrative examples. The first is a SIR model with three stages of infection, which illustrates how our approach can be applied to models where infected hosts progress through different stages of infection. In the Results section, we apply this model to HIV data so we assume that these three stages correspond to the early, chronic and AIDS stages of HIV infection. The deterministic skeleton of the three-stage SIR model can be written as the following system of ordinary differential equations:

$$\frac{dS}{dt} = \mu N - \Lambda(t)S - \mu S \tag{3.1a}$$

$$\frac{dI_E}{dt} = \Lambda(t)S - \gamma_E I_E - \mu I_E \tag{3.1b}$$

$$\frac{dI_C}{dt} = \gamma_E I_E - \gamma_C I_C - \mu I_C \tag{3.1c}$$

$$\frac{dI_A}{dt} = \gamma_C I_C - \gamma_A I_A. \tag{3.1d}$$

Infections progress from one stage to the next according to the rates γ_E and γ_C. We assume there is no recovery and that individuals with AIDS infection die at rate γ_A instead of the normal host mortality rate μ, where generally $\gamma_A > \mu$.

The force of infection $\Lambda(t)$ is given by

$$\Lambda(t) = e^{-\alpha(I_E+I_C+I_A)}\frac{(\beta_E I_E + \beta_C I_C + \beta_A I_A)}{N}, \tag{3.2}$$

where N is the host population size ($N = S + I_E + I_C + I_A$). The exponential term in (3.2) allows incidence to scale nonlinearly with the prevalence of HIV in the population and has been frequently used in HIV models (Williams et al., 2006; Granich et al., 2009; Volz et al., 2012). This nonlinear scaling may reflect heterogeneity in sexual contact rates or behavioral changes as awareness or diagnosis of the disease grows.

The second model we consider is a simple two-population SIR model where transmission can occur both within and between the two populations due to infectious individuals coming into contact with susceptible individuals in either population. While we do not explicitly define the factor that structures the population, population structure could be due to spatial structure or other factors like age that affect the probability of different hosts contacting one another. The deterministic skeleton of this model can be written as follows:

$$\frac{dS_1(t)}{dt} = \mu N_1(t) - \beta_W(t)\frac{S_1(t)}{N_1(t)}I_1(t) - \beta_B(t)\frac{S_1(t)}{N_1(t)}I_2(t) - \mu S_1(t) \tag{3.3a}$$

$$\frac{dI_1(t)}{dt} = \beta_W(t)\frac{S_1(t)}{N_1(t)}I_1(t) + \beta_B(t)\frac{S_1(t)}{N_1(t)}I_2(t) - \nu I_1(t) - \mu I_1(t) \tag{3.3b}$$

$$\frac{dS_2(t)}{dt} = \mu N_2(t) - \beta_W(t)\frac{S_2(t)}{N_2(t)}I_2(t) - \beta_B(t)\frac{S_2(t)}{N_2(t)}I_1(t) - \mu S_2(t) \tag{3.3c}$$

$$\frac{dI_2(t)}{dt} = \beta_W(t)\frac{S_2(t)}{N_2(t)}I_2(t) + \beta_B(t)\frac{S_2(t)}{N_2(t)}I_1(t) - \nu I_2(t) - \mu I_2(t). \tag{3.3d}$$

The parameter μ is the host birth/death rate and ν is the rate at which infected hosts recover. N_1 and N_2 are the host population sizes, respectively. β_W is the within-population transmission rate and β_B is the between-population transmission rate. We write the transmission rates as $\beta_W(t)$ and $\beta_B(t)$ to allow the transmission rate to vary seasonally. Both β_W and β_B are scaled relative to a base transmission rate β such that $\beta_W = \beta\rho$ and $\beta_B = \beta(1-\rho)$, so that the parameter ρ controls the extent of mixing or coupling between the two populations, as in Keeling and Rohani (2002). Under this parameterization, the basic reproductive number $R_0 = \beta/(\mu + \nu)$ and is therefore invariant to changes in ρ so that we can vary the degree of mixing between populations while not significantly altering the overall epidemiological dynamics.

3.2.2 Coalescent models

In this section, we consider formulating structured coalescent models for the type of structured epidemiological models just presented. As shown in Volz (2012), thinking about population dynamic models as simple birth-death processes can be useful when deriving coalescent models that correspond to a given forward-time model. If we randomly sample individuals from a population and trace their ancestry back in time, then coalescent events in the genealogy will correspond to birth events in the population when both the parent and child lineages are ancestral to sampled individuals. While deaths may affect the overall population size, deaths can be ignored along lineages ancestral to sampled individuals because we know that a lineage could not have died out at an earlier time if it persisted to be sampled at some later time. For a structured population, we also must consider individuals transitioning between different subpopulations through migration events that occur independently of birth events, although for the type of models we will consider here a lineage can also transition between populations by being born into a different population than its parent.

The same birth-death-migration framework can be applied to pathogens if we assume that each infected host corresponds to a single individual in the pathogen population. In this case, births in the pathogen population occur at transmission events between hosts. Deaths in the population will correspond to recovery or mortality of infected hosts. If each infected host is represented by a single pathogen lineage, coalescent events in the genealogy will correspond to transmission events if both the infected host and the infector are sampled or give rise to descendent infections that are sampled. For structured epidemiological models, we also must consider a pathogen lineage transitioning among populations, or compartments in SIR-type models, independent of transmission events. For example, in the three-stage model, pathogen lineages can transition between different stages of infection. Here, we will refer to all transitions between states that occur independently of transmission as migration for generality. This allows many epidemiological models with some form of population structure to be thought of as a birth-death-migration process.

To formalize the birth process, we adopt the notation of Volz (2012) and let $F(t)$ be a matrix that specifies the birth rate of new lineages in the population at time t, where $F(t) = F(\theta, x_t)$, meaning that $F(t)$ can be a function of the epidemiological parameters θ and the population state variables x_t. Lineages may be in any one of m states. The rate at which lineages currently in state k give birth to lineages in state l is given by the element f_{kl}. The rate at which migration, or transitions between states independent of birth events, occurs is given by another matrix $G(t) = G(\theta, x_t)$. The rate at which lineages currently in state k migrate to state l is given by the element g_{kl}. We treat birth and migration as distinct processes because, as we will see, they affect the coalescent process in different ways since coalescent events can only occur at birth events but migration events can affect the probability of a particular lineage coalescing with another lineage. The total number of lineages in each state is given by a vector $Y(t)$, such that $y_k(t)$ gives the total number of individuals in the population

in state k at time t. From here in, we drop the time indices and just refer to the matrices F and G or the vector Y, but emphasize that the rates in F and G and the population sizes in Y can be time-dependent.

We illustrate the F and G matrix notation by decomposing the three-stage and two-population SIR models presented above into their component birth and migration processes. For the three-stage model, we have

$$F = \begin{pmatrix} \beta_E \frac{S}{N} I_E & 0 & 0 \\ \beta_C \frac{S}{N} I_C & 0 & 0 \\ \beta_A \frac{S}{N} I_A & 0 & 0 \end{pmatrix}, \tag{3.4}$$

$$G = \begin{pmatrix} 0 & \gamma_E I_E & 0 \\ 0 & 0 & \gamma_C I_C \\ 0 & 0 & 0 \end{pmatrix}. \tag{3.5}$$

In the F matrix, births occur through transmission of the pathogen from any of the three stages of infection to susceptible individuals. Because all new infections begin in the early stage, only the leftmost column of the F matrix has nonzero elements. The nonzero elements in the G matrix correspond to migration between stages through disease progression from early to chronic and from chronic to AIDS.

For the two-population model, we have

$$F = \begin{pmatrix} \beta_W \frac{S_1}{N_1} I_1 & \beta_B \frac{S_2}{N_2} I_1 \\ \beta_B \frac{S_1}{N_1} I_2 & \beta_W \frac{S_2}{N_2} I_2 \end{pmatrix}, \tag{3.6}$$

$$G = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}. \tag{3.7}$$

Because transmission events can move the pathogen within and between the two populations in either direction, all entries in the F matrix are nonzero. The G matrix has all zero entries because there is no migration between populations independent of transmission.

Before moving on, we note that for an infectious pathogen our coalescent models make the implicit assumption that coalescent events in the genealogy correspond to

transmission events between hosts. In essence then, we are ignoring the within-host coalescent process and assuming that all infected hosts are represented by a single lineage. This implies that lineages immediately coalesce once in the same infected host, which may not be true for certain pathogens where multiple lineages can persist within a host for long periods of time. Nevertheless, in general our assumption that each infected host is represented by a single pathogen lineage will be valid as long as super-infection is rare and there is a strong bottleneck in the pathogen population at transmission events so that it is unlikely that more than one lineage is transmitted between hosts.

3.2.3 Coalescent likelihoods

To fit a structured coalescent model to a genealogy, we need to compute the likelihood of the coalescent model given the genealogy. To compute this likelihood, we can partition the genealogy into any number of discrete time intervals. We label the time partitioned genealogy $\mathcal{G}_{1:T}$, where $t = 1$ is the time of the first event in the genealogy and $t = T$ is the final event time going forwards in time (usually the terminal-most sampling event). Time points are chosen to correspond to the times at which events in the genealogy occur such as coalescent and sampling events. We can then further subdivide the genealogy into smaller intervals that correspond to the Δt time steps used to simulate from the epidemiological model so that at any time point t we have the state variables x_t corresponding to that time. With the time partitioned genealogy $\mathcal{G}_{1:T}$, we can compute the likelihood over each interval in the genealogy, $\mathcal{G}_{t-1:t}$, and then take the product over all intervals to compute the total likelihood of the model given $\mathcal{G}_{1:T}$.

Computing the likelihood over a time interval $\mathcal{G}_{t-1:t}$ requires us to first compute the probabilities that the lineages present in the genealogy did or did not coalesce within that time interval. The probability of a coalescent event in turn depends

on the expected rate of coalescence under the model. This expected rate can be computed for a coalescent model with any arbitrary population structure using the formalism summarized above for the rates of birth in F. As shown in Volz (2012), the rate of coalescence λ_{ij} for two lineages i and j is

$$\lambda_{ij} = \sum_{k}^{m} \sum_{l}^{m} \frac{f_{kl}}{y_k y_l} \left(p_{ik} p_{jl} + p_{il} p_{jk} \right) \tag{3.8}$$

where p_{ik} is the probability that lineage i is in state k. How these lineage state probabilities are computed is explained below. We can make intuitive sense of the coalescent rate in (3.8) by noting that f_{kl} is the total rate at which lineages in state k give birth to lineages in state l in the population and that $\frac{1}{y_k y_l}$ is the probability that lineages i and j are the two lineages involved in a particular birth event. However, since we do not know the true states of i and j we must sum over all possible combinations of states for these two lineages.

The total rate of coalescence $\lambda_{\mathcal{A}}$ for all lineages $\mathcal{A}$ present in the genealogy over an interval of time is then

$$\lambda_{\mathcal{A}} = \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{A}, j<i} \lambda_{ij}. \tag{3.9}$$

Given the rates of coalescence, we can then compute the likelihood over a time interval $\mathcal{G}_{t-1:t}$ under the coalescent model. If the time interval does not end in a coalescent event, we have

$$L(\mathcal{G}_{t-1:t}) = e^{-\lambda_{\mathcal{A}} \Delta t}. \tag{3.10}$$

Alternatively, if the interval does end in a coalescent event between two lineages i and j, we have

$$L(\mathcal{G}_{t-1:t}) = \lambda_{ij} e^{-\lambda_{\mathcal{A}} \Delta t}. \tag{3.11}$$

3.2.4 Lineage state probabilities

As alluded to above, computing the coalescent rates requires us to compute the probability of each lineage in the genealogy being in each possible state. At the time of sampling, we may know the state of a lineage from information gathered from the infected host from which the sample was obtained. Alternatively, if we do not know the state of the host at the time of sampling exactly, we can assign prior probabilities to the lineage being in each state under a multinomial distribution. Either way, given the initial state or state probabilities at the time of sampling, we need to be able to compute the probability of the lineage being in each state at any point in the past.

Going backwards in time, the lineages transition between states at the rates given in the F and G matrices, which in turn depends on the population states $x_{1:T}$ and the parameters θ. Given these transition rates, we have a continuous time Markov process on a discrete state space along each branch. We can therefore use master equations to track how the lineage state probabilities change going backwards through time. In other words, we can write down differential equations for how the probability mass assigned to each state flows between states as we move into the past. As shown in Volz (2012), the general form that these master equations take for any lineage i and state k is

$$\frac{d}{dt}p_{ik} = \sum_{l}^{m}\left(p_{il}\frac{g_{kl}}{y_l} - p_{ik}\frac{g_{lk}}{y_k} + p_{il}\frac{f_{kl}}{y_l}\frac{y_k - A_k}{y_k} - p_{ik}\frac{f_{lk}}{y_k}\frac{y_l - A_l}{y_l}\right), \tag{3.12}$$

where $A_k = \sum_{i\in\mathcal{A}} p_{ik}$; that is A_k is the expected number of lineages in state k in the genealogy at a given point in time. The first two terms in (3.12) give the probability mass gained or lost from the lineage transitioning in or out of state k through migration. The second two terms give the probability mass gained or lost from the lineage transitioning between states through a transmission event that was not observed as a coalescent event in the genealogy. In order for a lineage to transition

from state l to state k in this way, there needs be a coalescent event between the lineage in state l and another lineage in state k that is not among the A_k sampled lineages in the genealogy so that it is not observed in the tree. The probability that the lineage in state k is not among the sampled lineages is $\frac{(y_k - A_k)}{y_k}$. This probability is then multiplied by the total rate at which lineages transition from state l to state k going backwards in time, $\frac{f_{kl}}{y_l}$, to get the total rate at which probability mass is gained by state k.

We also have to take into consideration how the lineage state probabilities get updated after a coalescent event. Given that lineages i and j coalesce, the parent lineage h may be either lineage i or j because we cannot observe from the tree which of the two lineages was the donor. To compute the probability that the parent lineage h was in state k when in transmitted, we therefore have to take into consideration all of the different ways h could have transmitted either lineage i or j. Conditioning on the current lineage state probabilities for lineages i and j, we therefore have

$$p_{hk} = \frac{1}{\lambda_{ij}} \sum_{l}^{m} \frac{f_{kl}}{y_k y_l} \left(p_{ik} p_{jl} + p_{il} p_{jk} \right). \tag{3.13}$$

Given these updates, we have everything needed to compute the lineage state probabilities over an entire genealogy. For convenience, we introduce the notation $\mathcal{P}_t$ to denote the lineage state probabilities for all lineages in the genealogy at time t and $\mathcal{P}_{1:T}$ to denote the complete mapping of lineage state probabilities onto the genealogy over the entire time partitioned genealogy $\mathcal{G}_{1:T}$.

3.2.5 Statistical inference

The goal of phylodynamic inference for the type of models presented above will generally be to infer the parameters of interest from the genealogy along with the latent population state variables, such as the number of infected or susceptible hosts over time. In a Bayesian context then, we would like to infer the joint posterior

density of the model parameters θ and the latent state variables $x_{1:T}$. Up to a normalizing constant, this posterior density is given by

$$p(\theta, x_{1:T}|\mathcal{G}_{1:T}) \propto p(\mathcal{G}_{1:T}|\theta, x_{1:T})p(x_{1:T}|\theta)p(\theta). \tag{3.14}$$

From (3.14), we see that this joint density can be factored into three parts: the coalescent likelihood $p(\mathcal{G}_{1:T}|\theta, x_{1:T})$ which we outlined how to compute above; the prior density on the population state variables $p(x_{1:T}|\theta)$ as defined by the epidemiological process model; and the prior density on the parameters $p(\theta)$. Although we may be able to compute each component individually and thereby the posterior probability of a given set of parameters θ and population states $x_{1:T}$, the posterior density is not analytically tractable in general and we must resort to sampling from the posterior using MCMC methods.

However, it may be difficult or impossible to sample from complex, high-dimensional densities such as $p(\theta, x_{1:T}|\mathcal{G}_{1:T})$ using standard MCMC methods. We could, for example, use a Gibbs sampler to iteratively sample from the conditional posterior densities of θ and any component of $x_{1:T}$, but this strategy can be extremely inefficient owing to strong correlations among the parameters and the state variables, leading to slow MCMC mixing (Andrieu and Roberts, 2009). In Rasmussen et al. (2011), a particle MCMC approach known as the particle marginal Metropolis-Hastings (PMMH) algorithm was therefore used to sample from the joint posterior density of θ and $x_{1:T}$. The main motivation behind using the PMMH algorithm is that we can jointly update θ and $x_{1:T}$ together (Andrieu et al., 2010). Each MCMC iteration, we first propose new parameter values θ^* and then run a particle filtering algorithm to get a numerical approximation of the posterior density of the latent state variables $p(x_{1:T}|\mathcal{G}_{1:T}, \theta^*)$, which we refer to as $\hat{p}(x_{1:T}|\mathcal{G}_{1:T}, \theta^*)$. Particle filtering, also known as sequential Monte Carlo, provides a computational means of approximating high dimensional densities by providing samples (i.e the particles) distributed according to the desired

density, and are often used in the context of nonlinear and non-Gaussian state space models (Doucet et al., 2001; Cappe et al., 2007; Doucet and Johansen, 2009). How particle filters can be used to fit epidemiological models to genealogies is reviewed in Appendix A.

After running the particle filtering step in the PMMH algorithm, we can then sample a particle from $\hat{p}(x_{1:T}|\mathcal{G}_{1:T}, \theta^*)$ to get a proposal $x^*_{1:T}$ for the latent state variables that is adapted to the parameters in θ^*. We can also use the particle filter to compute the marginal likelihood of θ^* by marginalizing out the state variables. Because we jointly accept θ^* and $x^*_{1:T}$ based on the marginal likelihood, we do not have to independently update $x_{1:T}$, leading to a much more efficient MCMC sampler. Despite marginalizing out the latent state variables, the remarkable feature of the PMMH algorithm is it provides an exact (i.e. unbiased) approximation to the density of interest, $p(\theta, x_{1:T}|\mathcal{G}_{1:T})$. The PMMH algorithm is summarized in pseudo-code below.

Algorithm 1: The PMMH sampler targeting $p(\theta, x_{1:T}|\mathcal{G}_{1:T})$

At each MCMC iteration, with current parameter values θ:

1. Sample θ^* from a proposal density $q(\theta^*|\theta)$.

2. Run particle filter to sample $x^*_{1:T}$ from $\hat{p}(x_{1:T}|\mathcal{G}_{1:T}, \theta^*)$ and obtain the marginal likelihood estimate $\hat{p}(\mathcal{G}_{1:T}|\theta^*)$.

3. Accept θ^* and $x^*_{1:T}$ with probability

$$\min\left(\frac{\hat{p}(\mathcal{G}_{1:T}|\theta^*)p(\theta^*)}{\hat{p}(\mathcal{G}_{1:T}|\theta)p(\theta)}\frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}, 1\right). \tag{3.15}$$

While the PMMH algorithm described above works for unstructured epidemiological models where all infected hosts are assumed to be in the same population,

we encounter an additional problem for structured epidemiological models. In this case, the inference task at hand becomes more difficult because we need to take into account the unknown lineage states. This is done by conditioning on the lineage state probabilities $\mathcal{P}_{1:T}$ when computing the coalescent likelihood. We can make this dependence on the lineage state probabilities clear by rewriting the likelihood as $p(\mathcal{G}_{1:T}|\theta, x_{1:T}, \{\mathcal{P}_{1:T}\})$. We use the notation $\{\mathcal{P}_{1:T}\}$ to indicate that while the lineage state probabilities are required to compute the coalescent likelihood, we are not treating the lineage states as random variables but rather as probabilities that are completely determined by the master equations shown in (3.12) given θ and a population state trajectory $x_{1:T}$.

Recall that the probability of a lineage being in a certain state in the past depends conditionally on the state of the lineage at the time of sampling. This creates a backwards-time dependence structure that cannot easily be accommodated by the forwards in time particle filtering methods used in the PMMH approach. This is because the computational efficiency of the particle filter largely relies on the ability to resample—replacing particles with low weights with particles with high weights. In order to resample, we need to be able to compute the particle weights at any time t, which in turn requires the ability to compute the likelihood $p(\mathcal{G}_{t-1:t}|\theta, x_t, \{\mathcal{P}_t\})$ over any time interval. Computing this likelihood therefore entails being able to compute the lineage state probabilities $\mathcal{P}_t$, which will depend on the future states of the system $x_{t+1:T}$ for any structured model. The backward-time dependency of the lineage state probabilities therefore prohibits resampling and thus compromises the efficiency of the particle filter. We therefore use a modified particle filtering scheme that allows us to resample by computing expected lineage state probabilities before running the particle filter and then applying a correction step to counteract any bias introduced by using the expected rather than the true lineage state probabilities while filtering.

In more detail, the algorithm proceeds as follows. We initially simulate a deterministic trajectory from the epidemiological model for the state variables in $x_{1:T}$, which we refer to as $\bar{x}_{1:T}$. We can then compute the expected lineage state probabilities $\bar{\mathcal{P}}_{1:T}$ going backwards in time conditional on $\bar{x}_{1:T}$. We then run the particle filter forward in time to approximate the density $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\bar{\mathcal{P}}_{1:T}\})$. Although $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\bar{\mathcal{P}}_{1:T}\})$ is not ultimately the target density we are interested in, it serves a useful intermediate purpose. Once we have sampled particles representing trajectories from $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\bar{\mathcal{P}}_{1:T}\})$, we can re-weight these particles and use an additional round of importance sampling to get particles representing samples from $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\mathcal{P}_{1:T}\})$, our density of interest.

To do this, for each particle j sampled using the particle filter we compute the true lineage state probabilities $\mathcal{P}^j_{1:T}$ conditional on the actual population state trajectory of the particle $x^j_{1:T}$. We can therefore compute the corrected weights $v^j_T = p(x^j_{1:T}|\mathcal{G}_{1:T}, \theta, \{\mathcal{P}^j_{1:T}\})$ using the true lineage state probabilities. We also store the expected weights $u^j_T = p(x^j_{1:T}|\mathcal{G}_{1:T}, \theta, \{\bar{\mathcal{P}}_{1:T}\})$ computed using the expected lineage state probabilities. With both the expected weights u^j_T and corrected weights v^j_T we can assign final importance weights $w^j_T = \frac{v^j_T}{u^j_T}$ and resample particles again according to the final weights in w_T. This final round of importance sampling corrects for any bias we may have introduced by resampling particles using the expected lineage state probabilities while filtering and thereby gives us particles approximately distributed according to the target density $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\mathcal{P}_{1:T}\})$.

Algorithm 2: The particle filter/importance sampler targeting $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\mathcal{P}_{1:T}\})$

1. Run deterministic simulation to obtain $\bar{x}_{1:T}$ and compute the expected lineage state probabilities $\bar{\mathcal{P}}_{1:T}$ conditional on $\bar{x}_{1:T}$.

2. Initialize the particle filter at time $t = 1$ with N particles.

   (a) Set x_1^j to initial values for all particles.

   (b) Assign normalized weights, $U_1^j = \frac{1}{N}$.

3. Run filter from $t = 2$ to $t = T$.

   (a) Propagate particles forward by simulating from the process model $p(x_t^j | x_{t-1}^j, \theta)$.

   (b) Set $x_{1:t}^j = (x_{1:t-1}^j, x_t^j)$ for all particles.

   (c) Compute unnormalized weights conditional on $\bar{\mathcal{P}}_{1:T}$,

$$u_t^j = (u_{t-1}^j) p(\mathcal{G}_{t-1:t} | \theta, x_t^j, \{\bar{\mathcal{P}}_{t-1:t}\}). \tag{3.16}$$

   (d) Normalize weights, so that $U_t^j = \frac{u_t^j}{\sum_{j=1}^N u_t^j}$.

   (e) If resampling at t, resample according to U_t^j.

4. At time T, resample particles again according to U_T^j to get particles distributed according to $p(x_{1:T} | \mathcal{G}_{1:T}, \theta, \{\bar{\mathcal{P}}_{1:T}\})$.

5. Compute corrected lineage state probabilities $\mathcal{P}_{1:T}^j$ for each particle conditional on $x_{1:T}^j$.

6. Compute corrected weights $v_T^j = p(x_{1:T}^j | \mathcal{G}_{1:T}, \theta, \{\mathcal{P}_{1:T}^j\})$.

7. Compute final importance weights $w_T^j = \frac{v_T^j}{u_T^j}$ and normalize to get W_T^j.

8. Sample $x_{1:T}^*$ from $\hat{p}(x_{1:T} | \mathcal{G}_{1:T}, \theta, \{\mathcal{P}_{1:T}\})$ according W_T^j.

9. Compute marginal likelihood estimate

$$\hat{p}(\mathcal{G}_{1:T} | \theta) = \prod_{t=1}^{T} \sum_{j=1}^{N} W_T^j v_t^j. \tag{3.17}$$

The particle filter/importance sampler therefore provides us with a proposal for the population state variables $x^*_{1:T}$ and an estimate of the marginal likelihood. We can therefore plug the particle filter/importance sampler into step 2 of Algorithm 1 to obtain a PMMH algorithm for sampling from the joint posterior density of θ and $x_{1:T}$ under structured epidemiological models. Moreover, because we marginalize over the population state variables $x_{1:T}$ using the particle filter and then marginalize over the lineage states by summing over all possible lineage states when computing the likelihood, we can efficiently sample from the posterior density using the PMMH algorithm without having to design proposal updates for the population states or lineage states.

Before moving on, we make a few notes about the potential limitations and efficiency of the particle filter/importance sampler. As a basic requirement of importance sampling, the support of the importance density $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\bar{\mathcal{P}}_{1:T}\})$ must span the support of the target density $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\mathcal{P}_{1:T}\}))$, so that wherever $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\mathcal{P}_{1:T}\}) > 0$ so must $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\bar{\mathcal{P}}_{1:T}\}) > 0$. However, in order for the particle filter/importance sampler to be efficient, the density $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\mathcal{P}_{1:T}\})$ should also be similar to $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \bar{\mathcal{P}}_{1:T}\})$. Of course, this might not always be the case. If the stochastic particle trajectories can diverge largely from $\bar{x}_{1:T}$ or the lineage state probabilities are highly correlated with $x_{1:T}$ (such that small changes in the population states lead to large jumps in the lineage state probabilities), then these two densities may be quite different, causing the importance sampler to become very inefficient and requiring us to sample many particle trajectories in order to obtain a reasonable particle approximation to $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\mathcal{P}_{1:T}\})$. In such cases, it may be unwise to resample particles according to their expected weights u_T during the particle filtering stage because these weights will not be predictive of the corrected weights v_T, meaning we may be discarding particles with high posterior probability under the desired density $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\mathcal{P}_{1:T}\})$. In practice, this can

easily be checked by making sure that there is a strong positive correlation between the expected and corrected weights. We found this to be true for all cases considered here and found that resampling according to the expected weights during the filtering stage measurably improved the performance of the algorithm by reducing the variance in the marginal likelihood estimates, which tends to improve MCMC mixing overall.

3.2.6 Simulations

We simulated mock genealogies under each model to test the performance of the PMMH algorithm before applying the method to real data. Mock genealogies were obtained by first forward simulating from the population dynamic model while tracking all infected hosts in the population and the parent-offspring relationships at transmission events. From the forward simulations, we could then trace the lineages of infected individuals backwards through time to obtain the true genealogy for a fraction of sampled lineages. All population dynamic simulations were performed using the tau-leaping algorithm so that the epidemiological dynamics included demographic noise (Gillespie, 2007).

The three-stage model was parameterized to reflect the natural history of HIV because we planned to apply our method to real HIV genealogies (see Table 3.1). We set the disease progression and AIDS death rate to values that give an average time between infection and death of about 10 years, consistent with observed patterns. The incidence scaling parameter α was set to zero so that in the simulations there was a linear scaling between incidence and prevalence. The epidemic simulations were seeded with one early-stage infection at time zero and run for 37 years to reflect the timespan of the HIV epidemic in the U.S. To obtain mock genealogies from the complete infection trees, we sampled 200 individuals in the last six years of the epidemic to reflect the fact that most HIV sequences have been sampled in the

Table 3.1: Fixed parameters in the epidemiological models.

Three-stage model		Two-pop model	
Initial pop size	$N = 10,000$	Initial pop sizes	$N_1 = N_2 = 2$ mil
Birth/death rate	$\mu = \frac{1}{40.28}$ yr^{-1}	Birth/death rate	$\mu = \frac{1}{70}$ yr^{-1}
Progression rate	$\gamma_E = 1$ yr^{-1}	Recovery rate	$\nu = \frac{1}{7}$ yr^{-1}
Progression rate	$\gamma_C = \frac{1}{6.31}$ yr^{-1}	Seasonal amplitude	$\alpha = 0.08$
AIDS death rate	$\gamma_A = \frac{1}{2.55}$ yr^{-1}	Seasonal phase	$\delta_1 = 0.0$ yrs
			$\delta_2 = 0.5$ yrs

recent past. For all parameters, we chose to use uniform priors over a wide range of biologically plausible values so that the choice of prior would have minimal influence on our estimates.

For the two-population model, we added seasonality to the model by seasonally forcing the base transmission rate $\beta(t)$ using a sinusoidal forcing function where

$$\beta(t) = \beta\left(1 + \alpha\cos\left(\frac{(t+\delta)}{2\pi}\right)\right). \tag{3.18}$$

The strength of seasonality α was the same in both populations but we allowed δ to differ between the two populations to get asynchronous dynamics between populations. The values of all fixed parameters in the model are also shown in Table 3.1. For the genealogies, 120 infected hosts were randomly sampled over time with sampling effort proportional to disease prevalence in each population. For the two-population model, we fixed the initial conditions for the number of susceptible and infected hosts in each population.

For the simulation experiments, we wished to compare estimates obtained by fitting stochastic models using the PMMH algorithm against estimates obtained by fitting deterministic models. To fit deterministic models, we used a Metropolis Hastings sampler where, whenever new parameters were proposed, we computed the likelihood of the genealogy under the new parameters by conditioning on a determin-

istic trajectory of the state variables $x_{1:T}$ simulated from the model using the new parameters.

3.2.7 HIV data

We applied our method to a set of HIV-1 partial *pol* sequences collected from men who have sex with men (MSM) in the metropolitan area of Detroit, Michigan. The dataset contained 437 HIV-1 subtype B sequences which were originally collected for drug resistance testing between 2004 and 2011. More information about this dataset can be found in Volz et al. (2012). Data were anonymized by staff at the Michigan Department of Community Health before being provided to investigators. Because this research falls under the original mandate for HIV surveillance and was de-identified, it was classified as human subjects research but was exempt from further IRB review.

We reconstructed time-scaled genealogies from the HIV sequences in BEAST using a relaxed molecular clock (Drummond and Rambaut, 2007). All sequences identified as likely recombinants were removed from the alignment prior to the analysis. Tips in the genealogy corresponding to sampled infected individuals were assigned prior probabilities of being in each infection stage based on the time since infection estimated from CD4 cell counts and genetic diversity within the host (Volz et al., 2013a).

From the HIV genealogies, we estimated the transmission rates β_E, β_C and β_A as well as the incidence scaling parameter α. All other parameters were fixed at the values given in Table 3.1. Rather than estimate initial conditions, the time of the initial introduction of HIV into Detroit was estimated, at which point the epidemic was seeded with one early-stage infection in a completely susceptible population. All priors on the parameters were uniform. For the time of initial introduction the prior was truncated at 1973 as a lower bound and the root time of each tree as an

upper bound. To ensure our phylodynamic estimates of HIV incidence were reasonable, we compared our estimates against incidence back-calculated from Michigan Department of Community Health surveillance data using the method of Yan et al. (2011).

3.2.8 Implementation

For all results shown in this paper, the PMMH algorithm was run for at least 100,000 iterations or until the MCMC fully converged. For the Metropolis-Hastings step, we chose a multivariate normal proposal density for $q(\theta|\theta^*)$, which can take into account the correlations among different parameters by optimizing the covariance parameters that specify the density.

For the particle filter, we found that using a small number of particles ($N = 10$) was sufficient. Running the particle filter with a small number of particles tends to increase the error, or variance, in the marginal likelihood estimates. However, this error will not affect inference as long as the marginal likelihood estimates are not systematically biased because the error in the estimates will get averaged out in the encompassing MCMC algorithm. Nevertheless, with too few particles we run the risk of the MCMC getting stuck at erroneously high values of the likelihood. Our choice of $N = 10$ was therefore a compromise between minimizing the error in the marginal likelihood estimates and the time taken to run the particle filter. Resampling within the particle filter was done by multinomial sampling with replacement. Resampling times were chosen to minimize the variance in the marginal likelihood estimates and were usually placed around coalescent events, as most of the variation in particle weights arises at coalescent times.

The PMMH algorithm was implemented in the software package PHYLter and Java source code is freely available at `http://code.google.com/p/phylter/`. Running the PMMH algorithm for 100,000 iterations using the simulated HIV genealogies

took approximately 10 hours (0.36s per iteration) on a 3.4 GHz Intel i7 processor without any parallelization across cores. The most computationally intensive component of the algorithm is computing the lineage state probabilities, which involves numerically solving the master equations for each lineage in the genealogy and has a time complexity of $\mathcal{O}(m^2)$, where m is the number of possible lineage states. On the other hand, run times scale linearly with the number of particles and lineages in the genealogy. Thus, the efficiency of the algorithm is mainly limited by the number of states in the model.

3.3 Results

3.3.1 Testing the algorithm

Before applying the PMMH algorithm to genealogies reconstructed from real data, we ran extensive simulations to ensure that we could accurately recover epidemiological parameters and population dynamics from mock genealogies. We simulated 100 stochastic realizations of an epidemic from the three-stage model, keeping track of the underlying infection tree so that we could obtain the true genealogy for a fraction of sampled lineages. From the simulated epidemic dynamics, we can see that demographic stochasticity generates considerable variation in when the epidemic begins and peaks (Fig. 3.1). Even with this variability, we accurately inferred stage-specific prevalence and transmission rates from the mock genealogies using the PMMH algorithm (Fig. 3.2). The 95% credible intervals generally contained the true prevalence for all three stages of infection (Fig. 3.2A). We were also able to estimate the stage-specific transmission rates associated with each stage of infection (Fig. 3.2B-D), even though there were strong correlations among the different transmission rates as seen in the pairwise joint posterior densities (Fig. 3.2E-G). Overall, out of all 100 simulations, the 95% credible intervals contained all three transmission rates 94 times, while the posterior coverage was greater than 95% for each parameter individually.

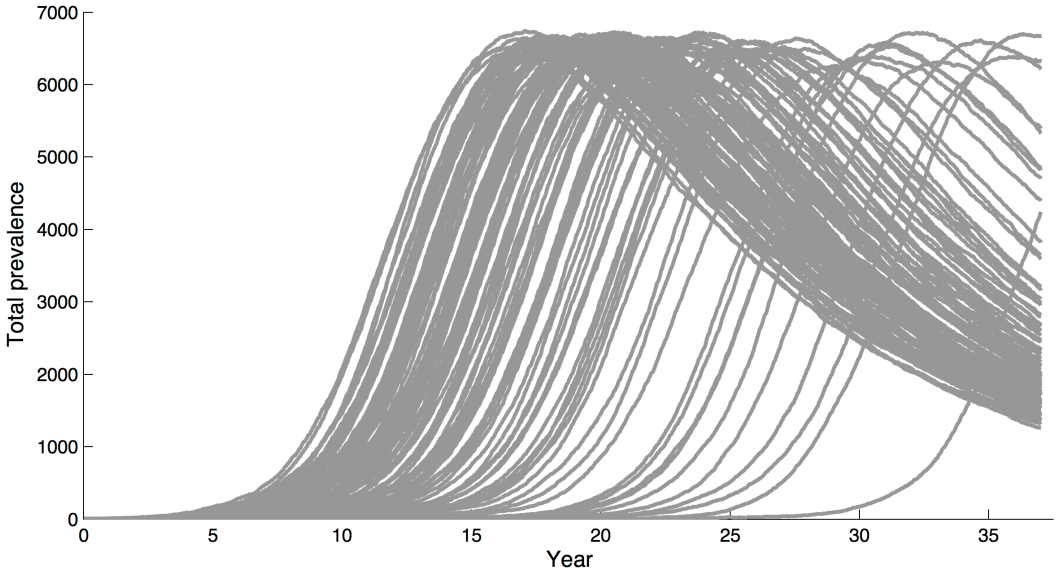

FIGURE 3.1: Simulated epidemic dynamics for 100 stochastic realizations of the three-stage SIR model. Total prevalence includes all three stages of infection.

In contrast, when we fit deterministic models to the same set of genealogies, the credible intervals contained the true parameters only 79% of the time. The PMMH algorithm therefore appears to give reliable estimates of parameters and epidemiological dynamics and outperforms deterministic methods when stochasticity plays a role in the epidemic dynamics.

3.3.2 HIV in Detroit

Given that we were able to reliably estimate transmission parameters and prevalence in our simulation study, we next applied the method to HIV genealogies reconstructed from sequences collected in Detroit, Michigan. A critical question in HIV epidemiology is to what extent transmission during the early stages of infection contributes to overall HIV incidence. Transmission during early infection may influence the effectiveness of interventions based on antiretroviral treatment in limiting the epidemic (Cohen et al., 2012; Kretzschmar et al., 2013). If most new cases of HIV result from recently infected individuals, then prevention strategies that rely on treating diag-

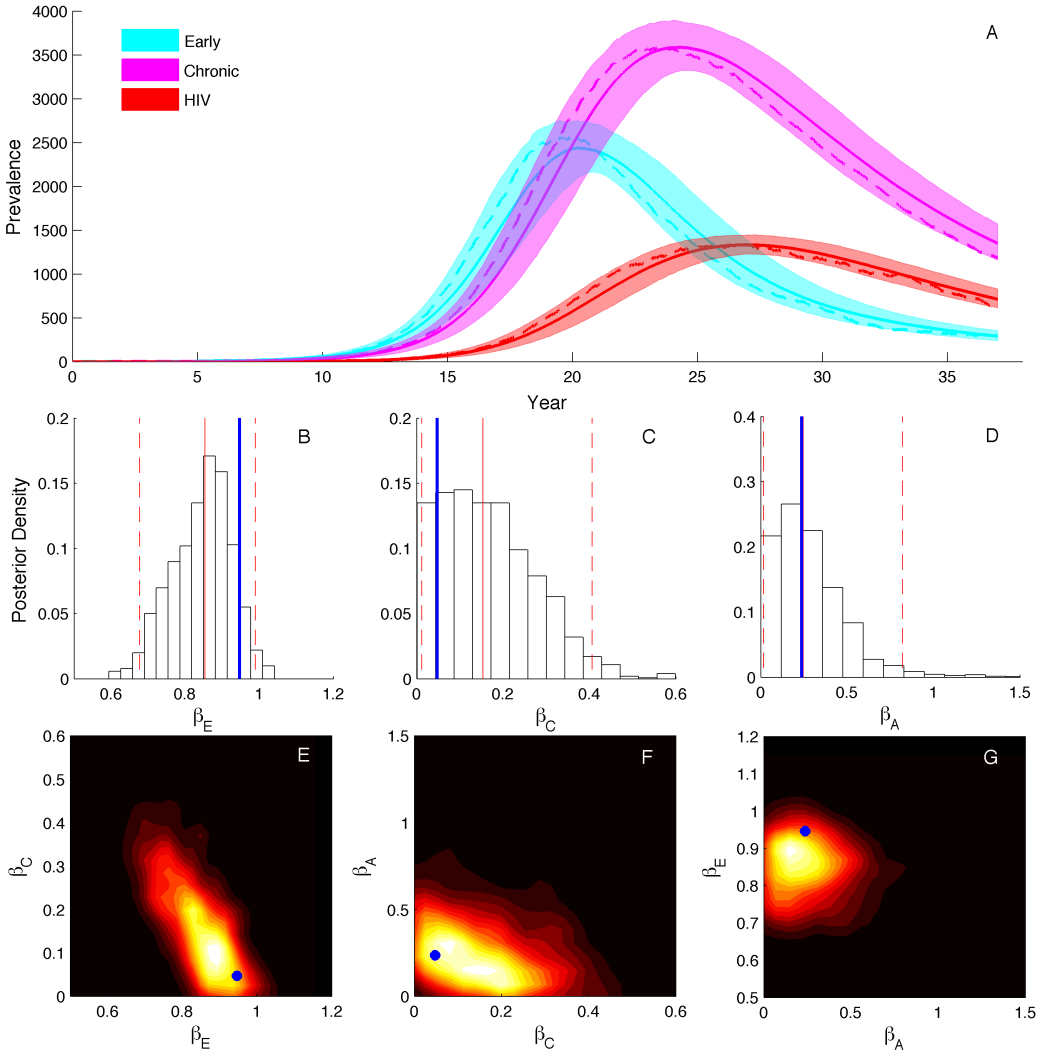

FIGURE 3.2: Prevalence and transmission rates estimated from a representative genealogy simulated under the three-stage SIR model. (A) Stage-specific prevalence estimates with the 95% credible intervals shaded and the posterior medians shown as solid lines. Dashed lines show the true prevalence. (B-D) The marginal posterior densities of the stage-specific transmission rates. (E-G) The corresponding pairwise joint densities of the transmission rates, which were constructed from the MCMC samples using nonparametric kernel density estimation.

nosed individuals, who are likely in later stages of infection, will directly prevent few transmissions. Thus, the transmission rate from early HIV infections (EHI) is a key parameter of great interest, although difficult to measure directly from traditional surveillance data. Phylogenetic studies of HIV have used the high degree of clustering and short branch times within these clusters to argue for a high EHI transmission rate (Lewis et al., 2008; Brown et al., 2011). However, clustering alone cannot be taken as definitive evidence for high EHI transmission as similar patterns can arise simply from epidemic transmission dynamics (Volz et al., 2012). In this section, we demonstrate that our inference framework can be used to estimate the EHI transmission rate and the number of new HIV infections attributable to EHI from HIV genealogies using models that explicitly consider HIV's transmission dynamics, as well as the stochastic nature of the epidemic dynamics.

Time-scaled genealogies were reconstructed using BEAST from HIV-1 partial *pol* sequences isolated from men who have sex with men (MSM) in the metropolitan area of Detroit. A representative genealogy randomly sampled from the BEAST posterior is shown in Fig. 3.3. We then fit our three-stage SIR model to 10 genealogies sampled from the BEAST posterior to take into account uncertainty in the genealogy. From these genealogies, we estimated the transmission rate for each stage, including the EHI transmission rate, along with the stage-specific dynamics of prevalence and the incidence (i.e number of new cases) attributable to each stage over the course of the epidemic.

Parameters estimated from the representative HIV genealogy are shown in Fig. 3.4 and estimates from all 10 genealogies are given in Table 3.2. We estimated that transmission rates are higher during the early and AIDS stages than during the chronic stage, as expected from previous studies (Pilcher et al., 2004; Hollingsworth et al., 2008; Powers et al., 2011). The transmission rate from EHI is about 20 times higher than during the chronic stage and about five times higher than during the AIDS

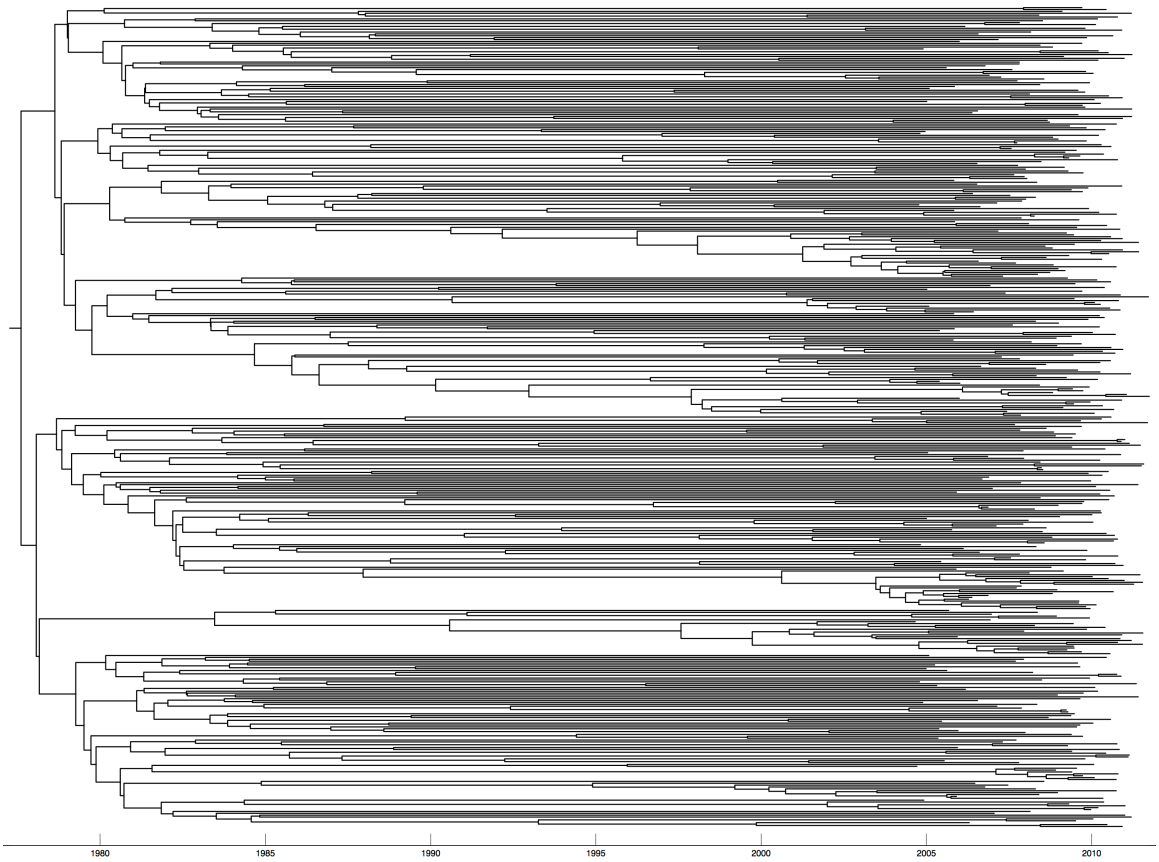

FIGURE 3.3: Representative time-scaled HIV genealogy from Detroit, Michigan.

stage (Fig. 3.4A-C). We also found evidence for a nonlinear dependence of incidence on prevalence, quantified through the incidence scaling parameter α. Although estimated values of α are small, the posterior density is clearly centered away from zero, indicating that incidence scales nonlinearly with prevalence (Fig. 3.4D). Overall, parameter estimates were largely consistent across genealogies, although there was considerable variation in the time of initial introduction of HIV into Detroit estimated from different trees. This is likely attributable to the large amount of variation in the root times inferred for different trees, as we inferred earlier times of introduction from trees with earlier root times.

Stage-specific HIV prevalence inferred from the genealogies shows a predictable

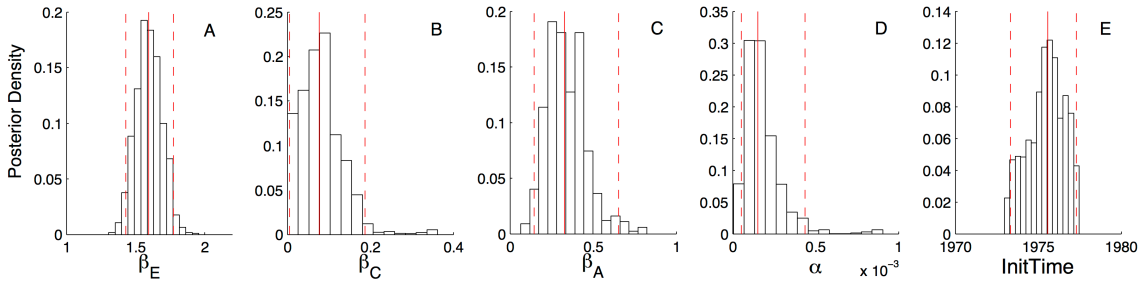

FIGURE 3.4: Posterior densities of parameters inferred from one HIV genealogy. Solid red lines mark the median values and dashed lines indicate the 95% credible intervals. The estimated parameters are the early stage transmission rate β_E, the chronic stage transmission rate β_C, the AIDS stage transmission rate β_A, the incidence scaler α and the initial introduction time of HIV into Detroit

transition from most infections being in the early stage at the beginning of the epidemic to most infections being in the chronic or AIDS stages later in the epidemic (Fig. 3.5A). This is expected given the longer duration of the chronic and AIDS stages. In general, our phylodynamic estimates of the epidemic dynamics closely track HIV incidence imputed from surveillance data from the beginning of the epidemic through the peak (Fig. 3.5B). While our phylodynamic estimates do not capture the fluctuations in incidence that occur after 1990, there was nothing in our model that would allow us to reproduce this pattern, which likely results from complex changes in HIV treatment and behavioral changes (Volz et al., 2013a). Although there was also considerable variability in the population dynamics inferred from different genealogies, this variation occurs primarily during the early stages of the epidemic (Fig. 3.5C). Again, this appears to be associated with uncertainty in the root times of trees; dynamics inferred from trees with earlier root times show an earlier rise and peak in incidence. After the epidemic peaks, the incidence estimated from different trees seems to converge on similar values.

Estimates of incidence attributable to each stage show that EHI contributed to most new infections at the beginning of epidemic when EHI prevalence was high

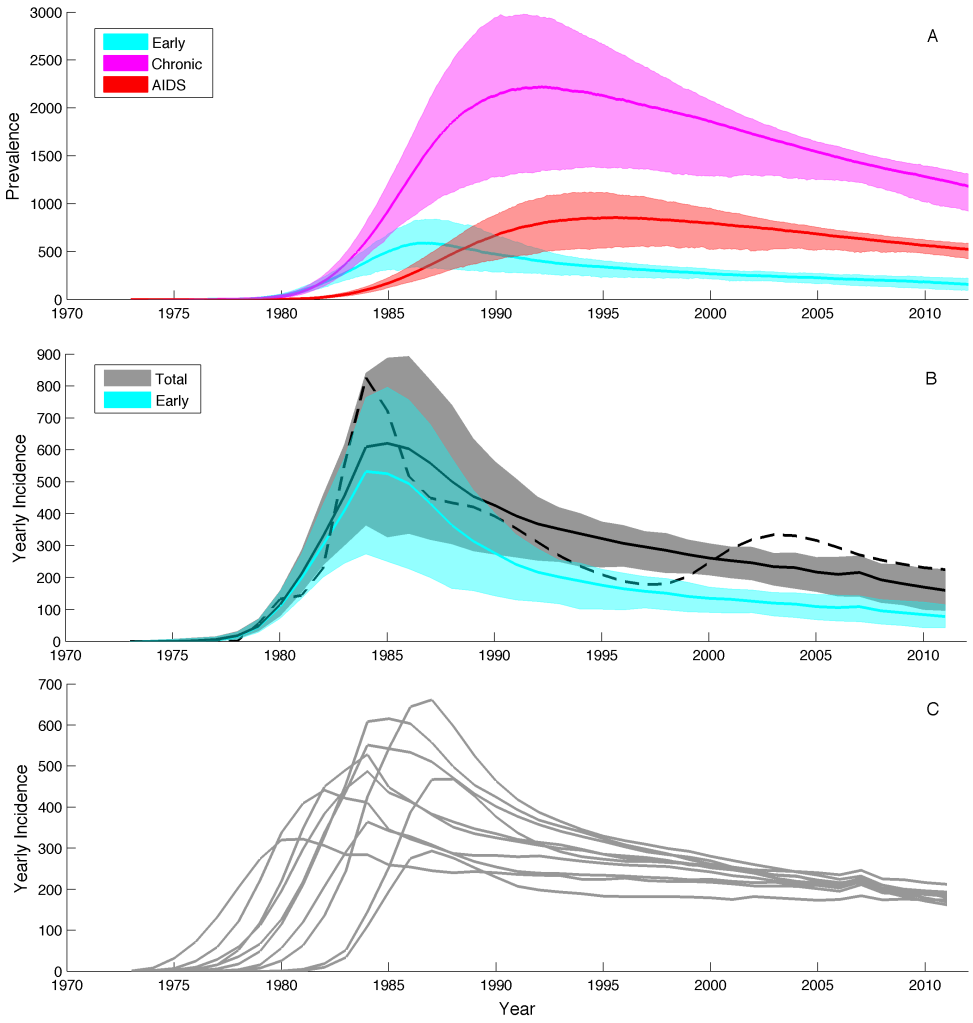

FIGURE 3.5: Population dynamics inferred from the Detroit HIV genealogies. (A) Stage-specific prevalence estimates from one genealogy with shaded regions showing the 95% credible intervals and lines the median of the posterior densities. (B) Estimated total yearly incidence and the estimated incidence attributable to early stage infections. The dashed black line shows the incidence back-calculated from Michigan Department of Community Health surveillance data. (C) Total incidence estimated from 10 randomly sampled HIV genealogies.

(Fig. 3.5B). After the epidemic peak, infections arising from EHI remains high proportional to EHI prevalence, consistent with the higher transmission rate we estimated for EHI. In the late 2000's, we estimated that between 40 to 50% of all new infections arise from EHI, indicating that early stage infections still play a major role in driving HIV transmission. These large estimates for number of new infections arising from EHI are consistent with the phylodynamic estimates of Volz et al. (2013a), who fit a more complex but deterministic epidemiological model to the same set of HIV sequences.

Table 3.2: Median posterior and 95% credible intervals for parameters estimated from 10 HIV genealogies.

Tree	β_E	β_C	β_A	$\alpha \times 10^{-4}$	Initial Time
1	1.59 (1.43 - 1.77)	0.077 (0.006 - 0.187)	0.328 (0.146 - 0.653)	1.52 (0.51 - 4.35)	1975 (1973 - 1977)
2	1.69 (1.51 - 1.9)	0.079 (0.007 - 0.184)	0.318 (0.126 - 0.59)	1.72 (0.62 - 4.02)	1976 (1973 - 1979)
3	1.69 (1.39 - 1.99)	0.184 (0.02 - 0.539)	0.593 (0.151 - 1.26)	5.78 (1.43 - 11.8)	1976 (1973 - 1978)
4	1.51 (1.32 - 1.73)	0.079 (0.007 - 0.191)	0.316 (0.114 - 0.648)	1.55 (0.48 - 4.44)	1974 (1973 - 1976)
5	1.6 (1.29 - 1.92)	0.146 (0.016 - 0.395)	0.786 (0.235 - 1.64)	5.99 (1.91 - 11.1)	1973 (1973 - 1974)
6	1.69 (1.43 - 1.99)	0.114 (0.008 - 0.28)	0.543 (0.193 - 1.0)	3.75 (1.23 - 8.0)	1974 (1973 - 1975)
7	1.67 (1.45 - 1.91)	0.1 (0.012 - 0.224)	0.431 (0.155 - 0.823)	2.93 (1.21 - 5.57)	1974 (1973 - 1977)
8	1.87 (1.62 - 2.14)	0.095 (0.006 - 0.239)	0.521 (0.214 - 0.964)	4.06 (1.22 - 8.3)	1979 (1974 - 1981)
9	1.9 (1.51 - 2.31)	0.295 (0.025 - 0.636)	0.823 (0.247 - 1.74)	10.2 (2.76 - 17.5)	1980 (1974 - 1982)
10	1.52 (1.32 - 1.73)	0.08 (0.009 - 0.248)	0.433 (0.184 -1.29)	2.5 (0.73 - 10.5)	1974 (1973 - 1976)

3.3.3 Inferring population structure

While our results for the three-stage model suggest that the PMMH algorithm works effectively and can be used to estimate key epidemiological parameters like HIV transmission rates, we were also interested in how much information genealogies contain about the structure of populations in general. To explore this question, we used the two-population model presented in (3.3), for which we can tune the strength of population structure by altering the mixing rate ρ between populations. Mock genealogies were simulated under three values of ρ: low (0.01), medium (0.05) and high(0.2). At $\rho = 0.01$, for example, about one in every one hundred transmission events occurs between populations. For all three values of ρ, we were able to accurately infer the epidemiological parameters of interest and the population dynamics from the simulated genealogies (Fig. 3.6). While we can easily estimate β under all three demographic scenarios, the posterior densities become skewed towards increasingly high values of ρ as mixing increases between the populations (Fig. 3.6A-C). This indicates that it may be very difficult to obtain precise estimates of ρ or other parameters pertaining to population structure when populations are only weakly structured.

We can visually explore how much information a genealogy contains about population structure and pathogen movement by comparing the true lineage states to the computed lineage state probabilities. In Fig. 3.7A-C, the true state of each lineage over time is mapped onto the genealogies. For ease of viewing, we only display a representative subtree of each genealogy. As expected, under low mixing lineages change states very slowly leading to a high degree of clustering among lineages sampled from the same population, whereas under high mixing lineages move rapidly between states and there is little clustering. We can then compare the true lineage states with the state probabilities computed under the median posterior values of the

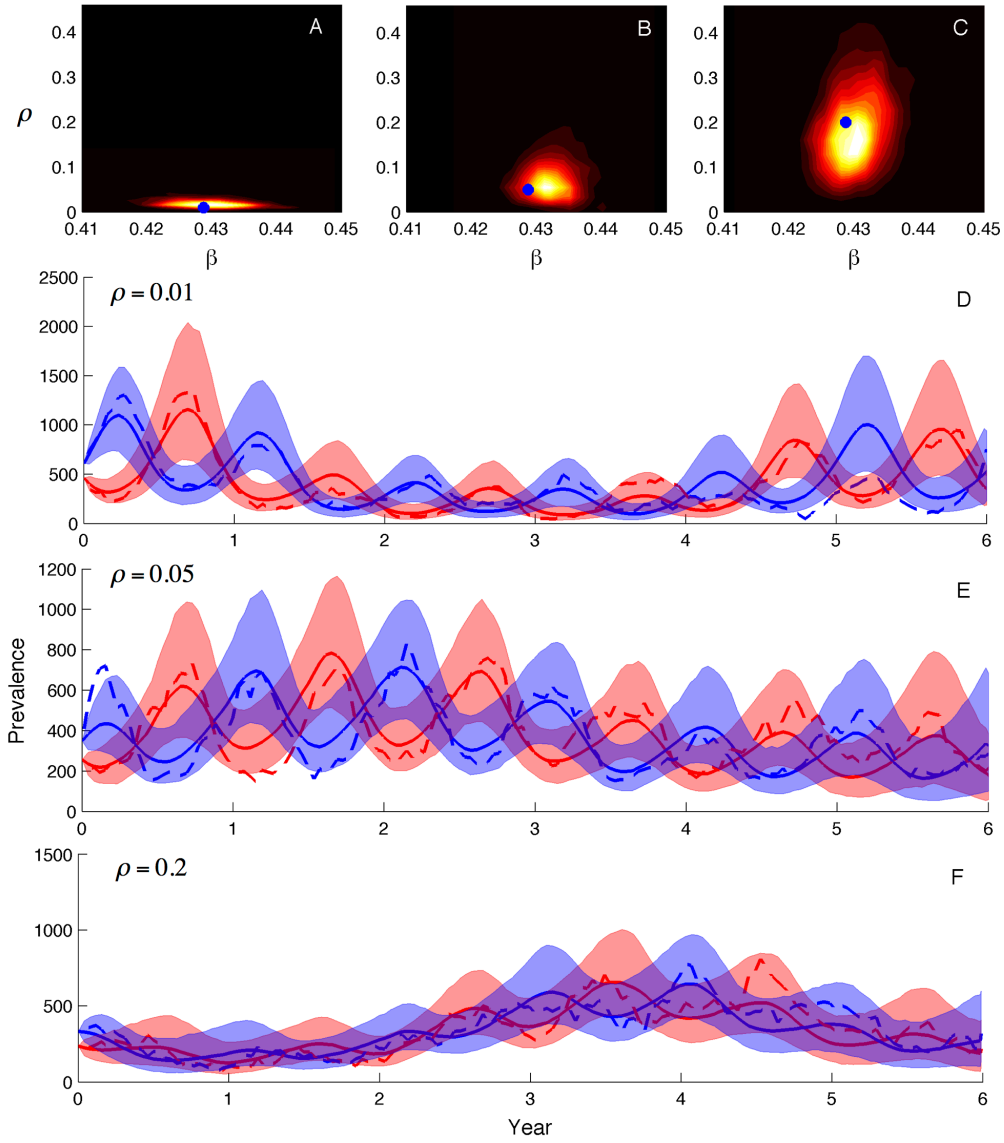

FIGURE 3.6: Parameter and prevalence estimates for the two-population model. Mixing rates between the two populations were varied from low ($\rho = 0.01$), medium ($\rho = 0.05$) to high ($\rho = 0.2$). (A-C) Joint posterior densities for the transmission rate β and the mixing parameter ρ. (D-F) Prevalence estimates for the two populations with the 95% credible intervals shaded and the posterior medians shown as solid lines. Dashed lines show the true prevalence. Initial conditions for the number of susceptible and infected individuals in each population were fixed at their true values for these simulations.

estimated parameters (Fig. 3.7D-F). When ρ is low, the state of the lineages at the time of sampling is highly informative about the state of the lineage going into the past. However, when we increase ρ to 0.05, the state of the sampled lineages is less informative about the past states and we can see that the lineage state probabilities fluctuate seasonally according to the asynchronous dynamics between populations. When ρ is high, the lineages move between states so rapidly that there is high uncertainty in the lineage states over the entire tree. This loss of information regarding the lineage states is readily observed by considering how the entropy, or uncertainty, in the lineage states changes going backwards in time (Fig. 3.7G-H).

Visualizing the flow of information along the lineages in the trees shows how uncertainty in parameters like ρ depends on how rapidly information about the lineage states decays. When ρ is low, lineages remain in the same state long enough that once a coalescent event is reached, information about the probable state of the lineages is still present. In this case, the probable states of the coalescing lineages provides additional information about the transmission event with respect to whether the transmission event occurred within or between populations. By combining information from coalescent events across the entire tree, we can then estimate the rates at which transmission occurs within and between populations. However, if all information about the past lineage states is lost before lineages coalesce, the observed coalescent events will no longer be informative about whether transmission occurred within or between populations and therefore parameters like ρ will be difficult to precisely estimate.

The preceding observations about uncertainty in lineage states suggest that it may be possible to estimate ρ more precisely if we increase the number of sampled lineages. Increasing the sampling fraction will also increase the coalescent rate among lineages, thereby increasing the probability of lineages coalescing before all information about their probable state is lost. To test this idea, we simulated genealogies under the

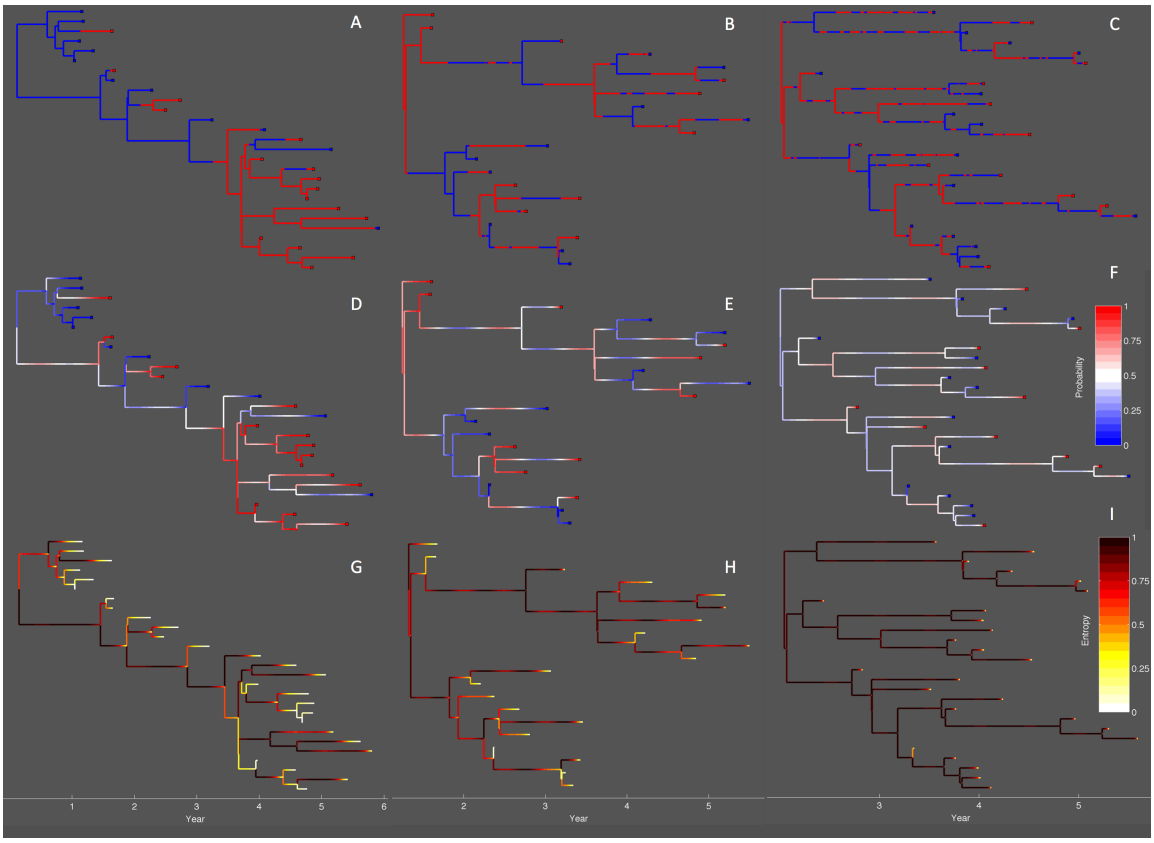

FIGURE 3.7: Genealogies simulated under different mixing rates. Mixing rates between the red and blue population were varied from low ($\rho = 0.01$), medium ($\rho = 0.05$) to high ($\rho = 0.2$). (A-C) The true lineage states mapped onto the genealogy. (D-F) Lineage state probabilities given with respect to the probability that the lineage is in the red state. (G-I) Entropy in the lineage states, which shows how much uncertainty there is in the lineage states. For each lineage i, the entropy $H_i = -\sum_k^m p_{ik} \log_2 \left(\frac{1}{p_{ik}}\right)$.

same three values of ρ but varied the sample size. With a sample size of 100, the same as used above, we see that the likelihood is peaked around the true value of ρ when mixing is low but the likelihood profile is fairly flat when mixing is high (Fig. 3.8A-C). Increasing the sample size to 500 resulted in more curved likelihood profiles but the likelihood remains relatively flat with high mixing (Fig. 3.8D-F). Doubling the sample size again to 1,000, the likelihood profiles show significant curvature for all values of ρ (Fig. 3.8G-I). This suggests that while the sample size does play a significant role in

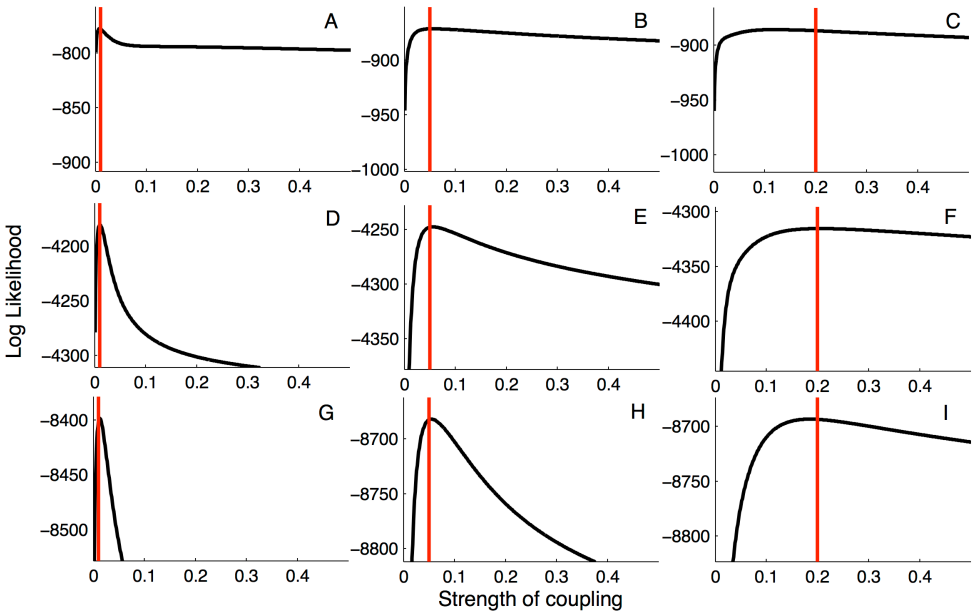

FIGURE 3.8: Likelihood profiles for the strength of coupling ρ obtained from genealogies simulated under different values of ρ. Red lines correspond to the true value of ρ. The likelihoods were computed from genealogies with 100 samples in (A-C), 500 samples in (D-F) and 1000 samples in (G-I). These sample sizes correspond, respectively, to approximately 0.2%, 1.0% and 2% of all infected individuals being sampled.

determining whether parameters like ρ can be precisely estimated from genealogies, extremely large sample sizes may be required to estimate parameters pertaining to population structure when the population is only weakly structured.

3.4 Discussion

The approach outlined above allows for structured, stochastic epidemiological and other population dynamic models to be fit to genealogies in order to jointly infer past population dynamics and model parameters. We believe this to be an important step forward in the field of phylodynamics because many populations are structured in ways that could bias estimates of demographic parameters when using coalescent-based methods if population structure is not properly taken into account. Further-

more, unlike earlier methods for fitting structured coalescent models to genealogies (Beerli and Felsenstein, 2001; Kuhner, 2006), our framework can accommodate non-equilibrium and nonlinear population dynamics and allows birth and migration rates to vary over time. We can also include stochasticity in our models when fitting them to data obtained from real populations, which may behave very differently than what would be expected under deterministic models. We can therefore fit the type of mechanistic population dynamic models typically used by epidemiologists and ecologists, which often include population structure, to genealogies.

As we have shown, fitting stochastic population dynamic models to genealogies through a structured coalescent model poses some challenges to statistical inference not normally dealt with in the statistical literature on fitting generic state-space models to observational data. Under our structured coalescent models, the probability of a genealogy depends conditionally on both the population state variables as well as the states of individual lineages over time. However, going backwards in time, the probability that a lineage is in a certain state can strongly depend on the state that the lineage was sampled in at some future point in time. Particle filtering methods, which are widely used to fit state space models to other sources of data, can perform very poorly under these circumstances because the state of the system, in this case the lineage states, can depend strongly on the future states of the system. One strategy we initially tried was therefore to use a Gibbs sampling approach to iteratively sample from the conditional posterior densities of the population state and lineage state variables in independent steps to avoid the problem of having both forward and backward time dependencies in the model. Unfortunately, we found that such a Gibbs sampling strategy can be very inefficient and suffer from extremely poor MCMC mixing when there are strong correlations among the parameters and the lineage states. For example, in our two-population model, the mixing parameter ρ controls how rapidly lineages move between states and is thus highly correlated

with the lineage states. If we update ρ conditional on our current lineage states, the proposed value of ρ will need to be very close to the current value in order for the proposal to have high enough probability to be accepted conditional on the current lineage states. We therefore explore a potentially very large parameter space taking only small steps at a time.

Given these issues, we decided to use a modified version of the PMMH algorithm originally proposed by Andrieu et al. (2010). In this approach, we simply propose new parameter values each MCMC iteration and then run the particle filter to numerically integrate over the population state variables. To make the particle filtering algorithm as efficient as possible within each MCMC step, we allow for resampling by first weighting the particles according to the expected lineage state probabilities. Once we have run the particle filter forwards in time, we can then compute the true lineage state probabilities backwards in time and apply an additional round of importance sampling to correct for any bias introduced by using the expected lineage state probabilities. With the true lineage state probabilities of each particle, we can compute the coalescent likelihood of the genealogy while summing over all possible lineage states. We can therefore integrate over both the unobserved population state variables and the lineage state variables when computing the marginal likelihood of the parameter proposal. We thus have an efficient MCMC algorithm for sampling from the posterior density of the parameters without having to design independent proposals for the population states or the lineage states. The PMMH sampler therefore has a major practical advantage over other MCMC approaches that can be easily quantified. For the models considered in this paper, the PMMH algorithm typically converged in less than 100,000 iterations whereas for the Gibbs sampler we could run millions of MCMC iterations and still not converge. The efficiency of this approach will hopefully make it possible to also consider phylogenetic uncertainty in the future by sampling genealogies in addition to epidemiological parameters in the

MCMC algorithm.

Whether or not the type of coalescent models considered here are appropriate for a particular pathogen is another important issue. The coalescent models assume that each infected host corresponds to a single pathogen lineage. If this were indeed always the case then coalescent events in the genealogy would always correspond to transmission events in the population. In reality, coalescent events will not occur instantaneously at transmission events but at some time before the actual transmission event because there will be a waiting time between when a lineage is transmitted and when it coalesces with another sampled lineage in the host. How closely the actual transmission event corresponds in time with the coalescent event will likely depend on the within-host dynamics of the pathogen (Ypma et al., 2013). For chronic viral infections like HIV where multiple lineages can persist within a given host for months or years, this may result in a large discrepancy in the timing of transmission and coalescent events. Nevertheless, a simulation study using a realistic distribution of within-host coalescent times for HIV found that the difference in timing between coalescent and transmission events was not sufficient to bias estimates of epidemiological parameters (Volz et al., 2013a). This may be due to the fact that a large fraction of HIV transmissions are due to recently infected individuals, in which case the within-host coalescent event cannot have occurred very long before the actual transmission event. A more principled approach to pursue in the future may be to impute the actual times of transmission conditional on the time of the coalescent events using information about within-host population dynamics. For example, additional information about pathogen population sizes over the course of a typical infection could provide an informative prior on waiting times between transmission events and coalescent events within hosts.

Another possible violation of the coalescent model occurs if sampled individuals have descendants that are themselves sampled, which can occur when samples are

collected serially over time. The coalescent model implicitly assumes that when a new lineage is sampled, that lineage is sampled from a different host than any other lineage already in the genealogy. However, if a lineage is sampled from a host that has other sampled descendant lineages in the genealogy, then this results in a coalescent event in the tree that does not correspond to a transmission event in the population. A similar problem would arise if we unwittingly sampled more than one lineage from a single infected host. However this is likely to occur only if sampling is dense relative to prevalence over time. For example, if sampling is dense at the beginning and the end of an epidemic, then with a high probability hosts sampled at the beginning of the epidemic will likely have sampled descendants at the end of the epidemic. We acknowledge that the coalescent models used in this paper cannot adequately handle these types of situations, although for the HIV analysis it is unlikely that this is a serious problem seeing as all sequences were sampled in the recent past when prevalence was high. In cases where this is likely to be a serious problem, it may be worth developing metapopulation coalescent models, such as those introduced by Dearlove and Wilson (2013), that allow hosts to be infected by more than a single lineage.

As our application to HIV showed, the PMMH algorithm allowed us to infer key epidemiological parameters like stage-specific transmission rates directly from genealogies. However, in the case of HIV, individuals stay in the same stage of infection for long periods of time relative to the timescale of the epidemic. The stage of infection of sampled individuals is therefore highly informative about the state of the lineage going into the past. Our experience with HIV may therefore not be representative of our general ability to infer parameters pertaining to pathogen transmission or movement in structured populations. In fact, our simple two-population SIR model revealed certain conditions under which it may be inherently difficult to estimate parameters relating to population structure. When lineages move between

states rapidly due to transmission or migration any particular lineage is likely to have changed states multiple times before a coalescent event is reached, leading to high uncertainty about the state of lineage over the majority of the genealogy. This is somewhat analogous to the problem of site saturation in phylogenetic inference, where multiple transitions at a particular site along branches can render that site phylogenetically uninformative (Yang, 1998). In the case of rapid transition rates among population states, observing the state of lineages at the time of sampling offers little or no information about the structure of the population because all information about the state of the lineage is quickly lost. Under these circumstances, it will be difficult to precisely estimate migration rates or other parameters relating to population structure from genealogies as we saw from the likelihood profiles of the mixing parameter in the two-population model, although it may be possible with many samples or a large sample fraction. This echoes earlier work on inference with structured coalescent models, where researchers have found it difficult to estimate migration rates from genealogies even without the complication of complex population dynamics (Beerli and Felsenstein, 1999, 2001).

Although it may not always be possible to precisely estimate parameters relating to population structure from genealogies, we can imagine several cases in which the ability to fit mechanistic epidemiological models to genealogies that include population structure may be extremely useful. For example, our methods could be used to fit spatially structured models to genealogies of samples collected in different locations and could potentially complement recently developed phylogeographic methods that consider spatial structure but do not generally take into account local population dynamics at any particular location (Lemey et al., 2009; Pybus et al., 2012). For instance, incorporating both spatial and temporal dynamics could be important when the structure of a population is not static but changes over time due to changes in migration rates, which themselves may vary due to non-stationary population dy-

namics across locations. Our approach can also be applied in cases where sampling effort is distributed unevenly among populations so that the assumption of random sampling in unstructured coalescent models has obviously been violated. In this case, structured coalescent models can be used to control for non-random sampling as long as sampling is random within the subpopulations defined in the coalescent model. Finally, our methods can be applied to multi-host or vectored pathogens where lineages can move among different host or vector species. As shown in Rasmussen et al. (2014b) for the case of dengue, including the dynamics of both the host and vector populations in coalescent models may be necessary in order for population dynamics inferred from genealogies of vector-borne pathogens to be accurate.

We end by noting that the methods presented here can be used to fit epidemiological models to genealogies as well as other sources of data simultaneously. For example, we previously showed how unstructured epidemiological models can be fit to a genealogy and a time series of case reports simultaneously and it would be straightforward to extend the methods presented here to include time series or other observational data (Rasmussen et al., 2011). This could be especially helpful when certain parameters or aspects of the dynamics are difficult to infer from one data source but for which an alternative data source could be highly informative. For example, case report data may be aggregated over different subpopulations obscuring some of the heterogeneity present in the population but could be revealed by also considering information present in a genealogy. Consolidating data sources in this way will likely play an important role in epidemiological modeling in the future, especially as molecular sequence data become increasingly available and phylodynamic methods become integrated into modern epidemiology.

4

Reconciling Phylodynamics with Epidemiology: The Case of Dengue Virus in Southern Vietnam

4.1 Introduction

The field of phylodynamics is concerned with how various ecological and evolutionary processes act or interact to shape genealogies and patterns of genetic diversity (Grenfell et al., 2004; Volz et al., 2013a). A major focus of phylodynamics has also been on what can be considered the inverse problem—given a genealogy, can the processes that generated the genealogy be inferred? With respect to this question, most effort has been focused on inferring the demographic history of populations from genealogies using coalescent-based methods such as the popular Bayesian Skyline approach (Strimmer and Pybus, 2001; Drummond et al., 2005). These methods have become especially popular among epidemiologists studying the population dynamics of infectious diseases, particularly rapidly evolving RNA viruses like influenza, dengue, hepatitis C and HIV (Pybus et al., 2001; Rambaut et al., 2008; Gray et al., 2009; Bennett et al., 2010).

Infectious diseases also present an opportunity to test phylodynamic methods

in situations where epidemiological data like time series of case reports are available alongside sequence data, allowing phylodynamic reconstructions of population dynamics to be compared against patterns observed through hospital- or community-reported incidence. Reassuringly, in many cases phylodynamic estimates have been in line with observed disease dynamics. A very striking example of such congruence was provided by Rambaut et al. (2008), who reconstructed seasonal influenza A dynamics consistent with the strongly annual fluctuations observed in surveillance data. Phylodynamic methods have also been used to successfully reconstruct the early, exponential growth phase of emerging epidemics (Pybus et al., 2001; Lemey et al., 2003; Dearlove and Wilson, 2013). Yet, in other cases, phylodynamic estimates have differed widely from observed or expected disease dynamics. This has often been the case with pathogens undergoing complex seasonal or multi-annual dynamics (Amore et al., 2010; Bennett et al., 2010; Siebenga et al., 2010; Lin et al., 2013). While the inability to capture fluctuations in population size at fine temporal resolution can partially be attributed to insufficiently dense sampling, cases have even been found where dynamics inferred from genealogies are out of phase with case report data (Bennett et al., 2010).

Discrepancies between phylodynamic estimates and observed dynamics highlight some of the technical issues that need to be addressed if phylodynamic methods are to become a reliable tool in epidemiology and other fields. One major concern is whether the coalescent models often used in phylodynamic inference are appropriate for populations undergoing complex population dynamics, as is often the case for infectious diseases. This is important for inference because it is the coalescent model that provides the probabilistic framework necessary to compute the likelihood of a particular demographic model given a genealogy. Coalescent models commonly used in traditional population genetics assume that the coalescent rate is inversely proportional to the effective population size N_e. For infectious diseases, changing

transmission rates can also affect coalescent rates (Volz et al., 2009; Frost and Volz, 2010). Therefore, the dynamics of N_e inferred from genealogies using standard coalescent models need to be interpreted carefully for pathogens as they may not reflect the true underlying disease dynamics. Additional ecological complexities can also seriously bias estimates if not properly taken into account. For example, different forms of population structure can bias estimates obtained using coalescent models that do not take into account the possibility of different lineages being in different populations (Carrington et al., 2005; Pybus and Rambaut, 2009; Heller et al., 2013). These issues make it difficult to assess whether inferences drawn from phylodynamic analyses are reliable or are, at least in part, artifacts of the coalescent models used for inference.

To explore some of these issues, we used dengue virus as a case study in phylodynamic inference. Dengue is a mosquito-borne flavivirus and has been the subject of several previous phylodynamic studies, which have had various degrees of success reconstructing dengue's complex epidemiological dynamics (Schreiber et al., 2009; Bennett et al., 2010; McElroy et al., 2011; Raghwani et al., 2011). Here, we limit our attention to dengue serotype 1 (DENV-1) in southern Vietnam, for which a large number of sequence samples and reliable hospitalization data are both available.

We were also interested in DENV-1 because, as shown below, we were unable to reconstruct the highly seasonal incidence patterns observed in hospitalization data using Bayesian Skyline methods. While there are many plausible explanations for this discrepancy, we explored three factors particularly relevant to dengue. These were: **(1)** Dengue's seasonality and nonlinear transmission dynamics, which lead to rapid fluctuations in dengue incidence; **(2)** Vector-borne transmission and the population dynamics of mosquitoes; and **(3)** Spatial structure in the host population arising from the spatial heterogeneity of southern Vietnam. While all three of these factors play a crucial role in dengue's ecological dynamics, it is less clear how each

factor acts to shape viral genealogies and therefore affects inferences drawn using coalescent-based methods.

To understand how each of these factors affect phylodynamic estimates drawn from the DENV-1 genealogy, we used a mechanistic modeling framework that allowed us to formulate each of the three proposed factors as a simple compartmental epidemiological model: a seasonal susceptible-infected-recovered (SIR) model, a vector-borne SIR model, and a spatially structured SIR model. We then derived coalescent models corresponding to each of the epidemiological models using the framework presented in Volz et al. (2009) and Volz (2012). With these coalescent models, we were able to directly fit each of the epidemiological models to the DENV-1 genealogy and explore how each factor affects the coalescent process. By comparing the relative fit of each model to the genealogy, we were able to gain insight into which factors are most important in shaping the DENV-1 genealogy. Moreover, the best fitting epidemiological models did much better than standard coalescent models in reconstructing population dynamics consistent with the dengue hospitalization data, showing that incorporating mechanistic modeling approaches into phylodynamic inference can greatly improve estimates of historical population dynamics.

4.2 Materials and methods

4.2.1 Epidemiological data

Dengue hospital admission data was compiled from the Hospital for Tropical Diseases and Children's Hospitals 1 and 2 in Ho Chi Minh City, as described in Anders et al. (2011). Here, we report the absolute number of dengue hospital admissions occurring each month. RT-PCR data on relative serotype frequencies is from Vu et al. (2010).

4.2.2 Sequence data and tree reconstruction

Whole genome viral sequences were obtained through the Broad Institute's Genome Resources in Dengue (GRID) website: `www.broadinstitute.org/annotation/viral/Dengue/Home.html`. For our analysis, 237 sequences were randomly subsampled from the larger set of 757 sequences used in the analysis of Vu et al. (2010). This larger set of sequences contained many samples collected during the same dengue season. We therefore subsampled sequences in years where large numbers of samples were sequenced so that approximately 40 sequences were included from each year between 2003 and 2008. Including more sequences did not appear to have any substantial effect on the population dynamics inferred from the genealogy. For each sequence we provide the Broad Institute's ID, the GenBank accession number, the date of isolation and whether or not the sample was isolated from an individual identified as living in HCMC in Rasmussen et al. (2014b).

The DENV-1 genealogy was inferred using the Bayesian MCMC methods available in BEAST version 1.6.1 (Drummond and Rambaut, 2007). Phylogenetic inference was performed using a General Time Reversible substitution model with gamma rate heterogeneity across sites and a strict molecular clock across lineages. Coalescent times inferred under the strict molecular clock were very close to those inferred under a relaxed clock. A Bayesian Skyline prior was chosen as the tree prior with 20 different population size intervals (Drummond et al., 2005). Including more population size intervals did not substantially change the Bayesian Skyline Plots.

4.2.3 Phylodynamic inference

For each of the three mechanistic models considered, we were interested in estimating the posterior density of parameters θ and latent state variables $x_{1:T}$ given the fixed DENV-1 genealogy $\mathcal{G}$. The variables in $x_{1:T}$ track the state of the population, such as the number of susceptible and infected individuals in the population. We

can compute the trajectory of all state variables in $x_{1:T}$ given a particular set of parameters θ by forward simulating the population dynamics from the deterministic ordinary differential equations (ODEs) that define the epidemiological model. For efficiency, forward simulations were performed using the Euler method of numerical integration with a sufficiently small integration time step.

Given a particular parameter set θ and population state trajectory $x_{1:T}$, we need to be able to compute the likelihood of the coalescent model given the genealogy in order to compute the posterior probability of θ and $x_{1:T}$. Methods for computing this likelihood for generic state space models were described in Rasmussen et al. (2011). For all of the coalescent models we consider here, the likelihood can be computed using an exponential probability distribution with rate parameter λ, the expected coalescent rate, which we derive for each model below.

A Metropolis-Hastings algorithm was used to sample from the posterior density of θ and $x_{1:T}$. Each iteration, new parameters were proposed and either accepted or rejected based on the posterior probability of the parameters and the state trajectory simulated under the model. Uniform priors were placed on all parameters. The algorithm was tested on multiple genealogies simulated under each model before being applied to the DENV-1 genealogy. The algorithm was implemented in the program EpiTreeFit and Java source code is available from the project website: `http://code.google.com/p/epitreefit/`.

Bayes factors were used to compare the fit of different models to the DENV-1 genealogy. Bayes factors give the ratio of posterior to prior odds favoring one model over another and thus serve as a summary of the evidence provided by the data in favor of a given model (Kass and Raftery, 1995). To compute Bayes factors from the MCMC samples, we used the standard harmonic mean estimator, which takes the harmonic mean of the posterior probabilities of the MCMC samples. While the harmonic mean estimator is known to be unstable when MCMC methods are used

to integrate over a very high-dimensional or complex parameter space (Lartillot and Philippe, 2006), we found that Bayes factors computed from different MCMC runs were quite stable, with variances less than one.

4.2.4 Epidemiological and coalescent models

Below we describe the three epidemiological models we fit to the DENV-1 genealogy and show how the corresponding coalescent model for each of these model can be derived using the coalescent framework of Volz et al. (2009) and Volz (2012).

Seasonal SIR model

The first model we consider is a simple, unstructured SIR model with direct transmission between humans for a single dengue serotype. By considering DENV-1 dynamics in the absence of the other DENV serotypes, we are assuming that susceptibility to and infectivity with DENV-1 is not, or is only weakly, impacted by the other DENV serotypes over this time period. The model is given by the following system of ODEs:

$$\frac{dS}{dt} = \mu N - \beta(t)\frac{S}{N}I - \mu S \tag{4.1a}$$

$$\frac{dI}{dt} = \beta(t)\frac{S}{N}I - (\mu + \nu)I \tag{4.1b}$$

$$\frac{dR}{dt} = \nu I - \mu R, \tag{4.1c}$$

where μ is the human birth and death rate, ν is the recovery rate in humans, and $\beta(t)$ is the seasonally-varying transmission rate. This transmission rate is given by:

$$\beta(t) = \bar{\beta}\left(1 + \alpha\cos\left(\frac{t+\delta}{2\pi}\right)\right), \tag{4.2}$$

where $\bar{\beta}$ is the average transmission rate over the entire year, α is the seasonal amplitude parameter, and δ controls the seasonal phase. R_0 in this model is given by $\frac{\beta}{\mu+\nu}$.

To reduce the number of parameters in the model that need to be estimated directly from the genealogy, we fixed several parameters available from other demographic or clinical data. We fixed the human birth/death rate μ at $\frac{1}{60}$ per year, reflecting the current birth rate in Vietnam, and the human population size at 10 million to reflect the population of HCMC, which was officially 7.5 million in 2007 but likely much larger (General Statistics Office of Vietnam, 2008). The recovery rate ν was set at $\frac{1}{7}$ per day, consistent with observed durations of viremia between 2 and 12 days (Gubler et al., 1981; Tricou et al., 2011). The free parameters in the model that we estimated were the average transmission rate $\bar{\beta}$, the seasonal amplitude α, the seasonal phase δ, and the initial conditions for the number of susceptible and infected individuals in the population.

As shown in Volz (2012), the pairwise rate of coalescence λ for an unstructured SIR model depends on the transmission rate, as well as the number of infected individuals and the fraction of the population susceptible to infection:

$$\lambda = \frac{2\beta(t)\frac{S}{N}}{I}. \tag{4.3}$$

Vector-borne model

Our vector-borne transmission model for an unstructured human population is given by the following ODEs:

$$\frac{dS_h}{dt} = \mu_h N_h - \beta_{vh}\frac{S_h}{N_h}I_v - \mu_h S_h \tag{4.4a}$$

$$\frac{dI_h}{dt} = \beta_{vh}\frac{S_h}{N_h}I_v - (\mu_h + \nu_h)I_h \tag{4.4b}$$

$$\frac{dS_v}{dt} = B_v(t) - \beta_{hv}S_v\frac{I_h}{N_h} - \mu_v S_v \tag{4.4c}$$

$$\frac{dI_v}{dt} = \beta_{hv}S_v\frac{I_h}{N_h} - \mu_v I_v. \tag{4.4d}$$

The subscripts h and v denote variables and parameters for humans and vectors, respectively. B_v is the vector birth rate, which we assume varies seasonally. The force of infection to both humans and mosquitoes is frequency-dependent with respect to humans. The transmission rates β_{vh} and β_{hv} are proportional to the per capita biting rate of a mosquito times a factor that determines the probability of a bite being infectious. For this model $R_0 = \frac{\beta_{hv}\beta_{vh}N_m}{\mu_v(\mu_h+\nu_h)N_h}$, as shown in Keeling and Rohani (2008).

We varied the size of the mosquito population by sinusoidally forcing the vector birth rate B_v:

$$B_v(t) = \bar{B_v}\left(1 + \alpha\cos\left(\frac{t+\delta}{2\pi}\right)\right), \tag{4.5}$$

where $\bar{B_v}$ is the seasonal average of B_v. We set $\bar{B_v} = \mu_v N_v$, so that the average seasonal mosquito population size does not change over time. However, because we do not know the size of the mosquito population N_v, we redefine $\bar{B_v}$ as equal to $\mu M N_h$, where M is a free parameter in the model that represents the ratio of the mosquito population size to the human population size.

When fitting the vector-borne model, we fixed the human birth and death rate μ_h, population size N_h and recovery rate ν_h at the same values as in the directly transmitted model. We also fixed the vector death rate μ_v at $\frac{1}{7}$ per day, which was chosen to represent the average of the daily mortality rates reported in the literature for *Aedes aegypti* adult females (Sheppard et al., 1969; McDonald, 1977; Harrington et al., 2001). We initially allowed the transmission rates β_{vh} and β_{hv} to differ depending on the directionality of transmission but model comparisons showed that a model with assymetric transmission rates did not fit the genealogy significantly better than a model with symmetric rates (Bayes factor < 3.0). We therefore only estimated a single transmission rate, β. The other estimated parameters were seasonal amplitude α, the seasonal phase δ, the ratio of mosquitoes to humans M, and the initial

conditions for the number of susceptible and infected humans in the population.

Given the forward-time dynamics, we need to derive the rate of coalescence under the vector-borne model. However, the population is now structured because viral lineages can either be in an infected human or an infected mosquito. We therefore use the structured coalescent framework of Volz (2012), who showed that for a generic structured population where lineages can be in any of m different states, the rate of coalescence λ_{ij} for two lineages i and j is

$$\lambda_{ij} = \sum_{k}^{m} \sum_{l}^{m} \frac{f_{kl}}{Y_k Y_l} \left(p_{ik} p_{jl} + p_{il} p_{jk} \right), \tag{4.6}$$

where p_{ik} is the probability that lineage i is in state k and p_{jl} is the probability that lineage j is in state l. f_{kl} is the rate at which lineages are transmitted from state k to state l and Y_k and Y_l are the total number of infected individuals in states k and l, respectively.

Adapting (4.6) to the vector-borne model, the rate of coalescence becomes

$$\lambda_{ij} = \frac{\beta_{vh} \frac{S_h}{N_h} I_v + \beta_{hv} S_v \frac{I_h}{N_h}}{I_v I_h} \left(p_{iv} p_{jh} + p_{ih} p_{jv} \right). \tag{4.7}$$

From (4.7) we can see that we need to compute the probabilities that lineages are in either an infected vector or human. We discuss how these lineage state probabilities can be computed in Appendix B along with our mathematical analysis of this vector-borne coalescent model.

Spatially structured model

Our spatially structured model partitions the total population into two sub-populations, which we refer to as the focal and global populations. For our analysis of DENV-1, the focal population corresponds to HCMC and the global population to the non-

HCMC population. The model is given by the following ODEs:

$$\frac{dS_f}{dt} = \mu N_f - \beta_{ff}(t)\frac{S_f}{N_f}I_f - \beta_{gf}(t)\frac{S_f}{N_f}I_g - \mu S_f \tag{4.8a}$$

$$\frac{dI_f}{dt} = \beta_{ff}(t)\frac{S_f}{N_f}I_f + \beta_{gf}(t)\frac{S_f}{N_f}I_g - (\mu+\nu)I_f \tag{4.8b}$$

$$\frac{dS_g}{dt} = \mu N_g - \beta_{gg}(t)\frac{S_g}{N_g}I_g - \beta_{fg}(t)\frac{S_g}{N_g}I_f - \mu S_g \tag{4.8c}$$

$$\frac{dI_g}{dt} = \beta_{gg}(t)\frac{S_g}{N_g}I_g + \beta_{fg}(t)\frac{S_g}{N_g}I_f - (\mu+\nu)I_g \tag{4.8d}$$

We assume that the human birth/death rate μ and recovery rate ν is the same in both populations, fixed at the values used for the previous two models. The global population size N_g was set at 25 million to reflect the population size of the southern-most 20 provinces excluding HCMC (General Statistics Office of Vietnam, 2008).

Transmission between the two populations occurs when an infected individual from one population contacts a susceptible individual in the other population. Bayes factor comparisons revealed that a model with separate transmission rates β_{gf} and β_{fg} did not fit the DENV-1 genealogy significantly better than a model with a single between-population transmission rate β_b (Bayes factor < 3.0). We therefore set $\beta_b = \beta_{gf} = \beta_{fg}$. However, the transmission rates within the focal population β_{ff} and global population β_{gg} are allowed to differ.

We first fit a model with seasonality in the focal population using the same sinusoidal forcing function as in (2) and assuming no seasonality in the non-HCMC population. For this model, we estimated the transmission rates β_{ff}, β_{gg}, β_b as well as the seasonality parameters α and δ for the focal population. We also estimated the initial number of susceptible and infected individuals in the focal population but, to reduce the number of parameters being fit, we set the initial conditions for the global population to their expected values at endemic equilibrium. We also fit a second model that allowed for seasonality in both populations. In this case,

we estimated the seasonality parameters α and δ for both populations, as well as the initial conditions in the non-HCMC population. While Bayes factors indicated that the more complex model with seasonality in both populations did not fit the genealogy significantly better, we retain this parameterization because it allowed us to detect the differing seasonal phase between populations.

We can again use the coalescent rate given in (4.6) to derive the coalescent rate for our spatially structured model. For two lineages i and j the pairwise rate of coalescence is:

$$\lambda_{ij} = \sum_{k}^{m} \sum_{l}^{m} \frac{\beta_{kl} \frac{S_l}{N_l} I_k}{I_k I_l} \left(p_{ik} p_{jl} + p_{il} p_{jk} \right). \tag{4.9}$$

In this case, there are only two populations and the subscripts k and l refer either to the focal or the global population. Given our epidemiological model, the lineage state probabilities p_{ik} can be computed backwards in time using equation 42 in Volz (2012) given the state of each lineage at the time of sampling.

4.3 Results

Dengue is hyperendemic in southern Vietnam with all four serotypes commonly circulating together. Previous epidemiological studies have shown that incidence is consistently high in the region with an annual attack rate in children estimated to be around 10% (Thai et al., 2005, 2011). Case reports collected at hospitals in Ho Chi Minh City (HCMC) between 2003 and 2008 indicate that transmission can occur year-round, although incidence is highly seasonal with a strong annual periodicity (Fig. 4.1A).

The hospitalization data shown in Fig. 4.1A include all four serotypes but may not be representative of any particular serotype. We therefore used viral isolates serotyped using RT-PCR to determine the fraction of isolates belonging to each of the four dengue serotypes over time. As shown in Fig. 4.1A, the proportion of

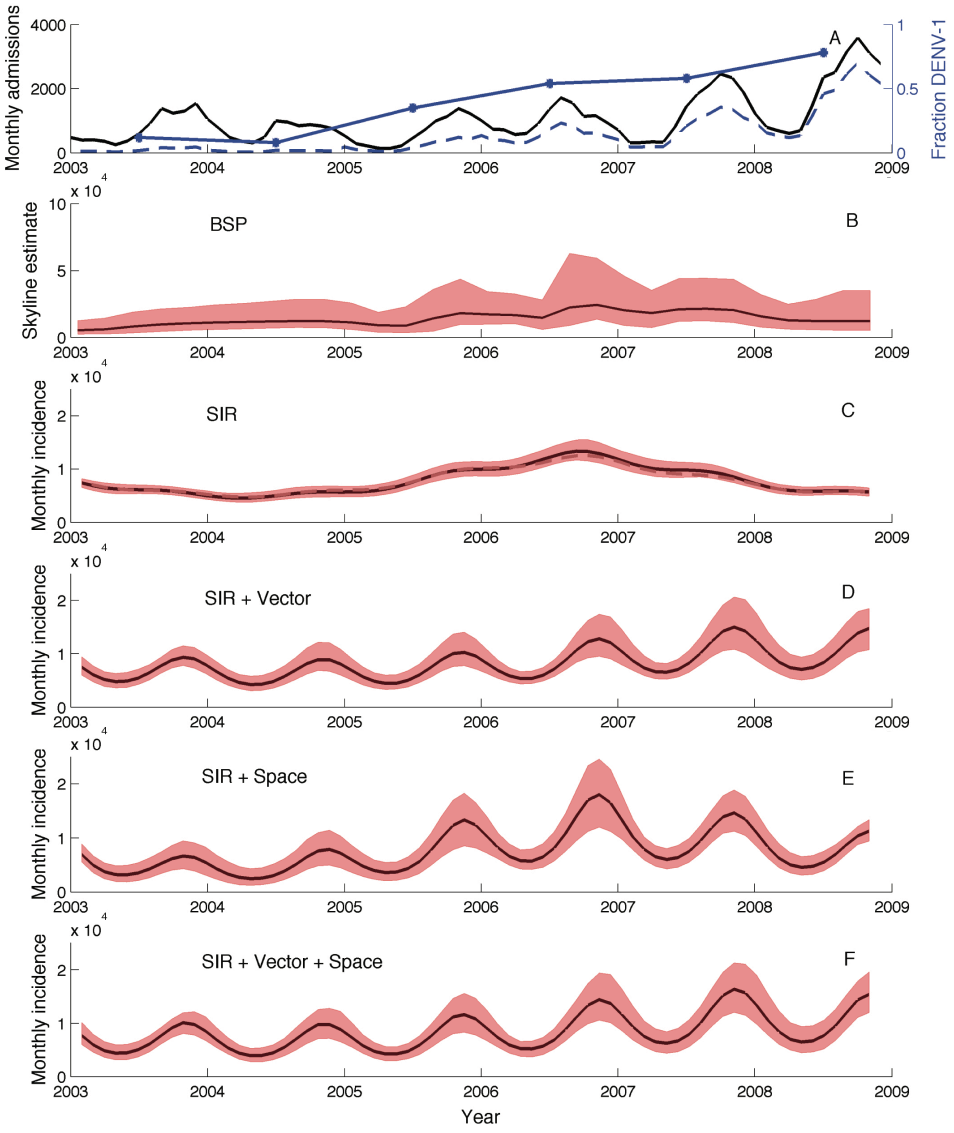

FIGURE 4.1: Population dynamics of dengue in southern Vietnam. (A) Absolute number of dengue hospital admissions each month in HCMC (black), yearly relative abundance of DENV-1 among RT-PCR positive cases (blue), and the extrapolated number of DENV-1 hospitalizations (dashed blue). (B) Bayesian Skyline Plot inferred from the DENV-1 sequences. Black lines show the median posterior estimates and shaded red regions give the 95% credible intervals. (C) Incidence inferred under the seasonal SIR model from the DENV-1 genealogy. Incidence estimates are reported as the absolute number of cases occurring each month. The dashed grey line shows the median estimate obtained from the HCMC-specific genealogy. (D) Incidence inferred under the vector-borne model. (E) Incidence inferred under the spatially structured model in HCMC. (F) Incidence inferred under the combined model with vectors and spatial structure.

DENV-1 isolates dramatically increased from around 2004 onwards. This trend is consistent with regional level data that indicate DENV-1 replaced DENV-2 as the dominant serotype in southern Vietnam while the relative abundances of DENV-3 and DENV-4 remained low over this period of time (Vu et al., 2010). Because of the predominance of DENV-1 over the time period studied, we focused on this serotype in our phylodynamic analysis, using the fraction of DENV-1 viral isolates to estimate monthly DENV-1 incidence from the hospitalization data (Fig. 4.1A). While the hospitalization data are likely representative of DENV-1 dynamics, the total incidence of DENV-1 is likely much higher because only a small fraction of dengue cases result in hospitalization.

To determine if we could reconstruct the dynamics observed in the dengue hospitalization data from sequence data, we inferred the genealogy of 237 DENV-1 whole genome sequence samples collected between 2003 and 2008 from dengue patients living throughout southern Vietnam. The maximum clade credibility (MCC) genealogy for these samples is shown in Fig. 4.2. Fig. 4.1B shows the population dynamics inferred, along with the genealogy, using BEAST in the form of a Bayesian Skyline Plot (BSP). While we do recover the increase in DENV-1 that occurred starting around 2004, other aspects of the dynamics observed in the hospitalization data are absent in the BSP. Most noticeably, the small fluctuations of DENV-1 inferred from the genealogy do not seem consistent with the large seasonal fluctuations in the hospitalization data (Fig. 4.1A-B). While in theory this could be due to inadequate sampling, exploratory simulations using sequence data simulated under dengue-like dynamics showed that the large seasonal fluctuations should be recoverable given the current sample size (simulations not shown). Aside from the discrepancy in seasonal dynamics, the BSP also shows DENV-1 incidence peaking in 2006 and then declining whereas the hospitalization data shows the peak in seasonal incidence increasing each year from 2004 to 2008.

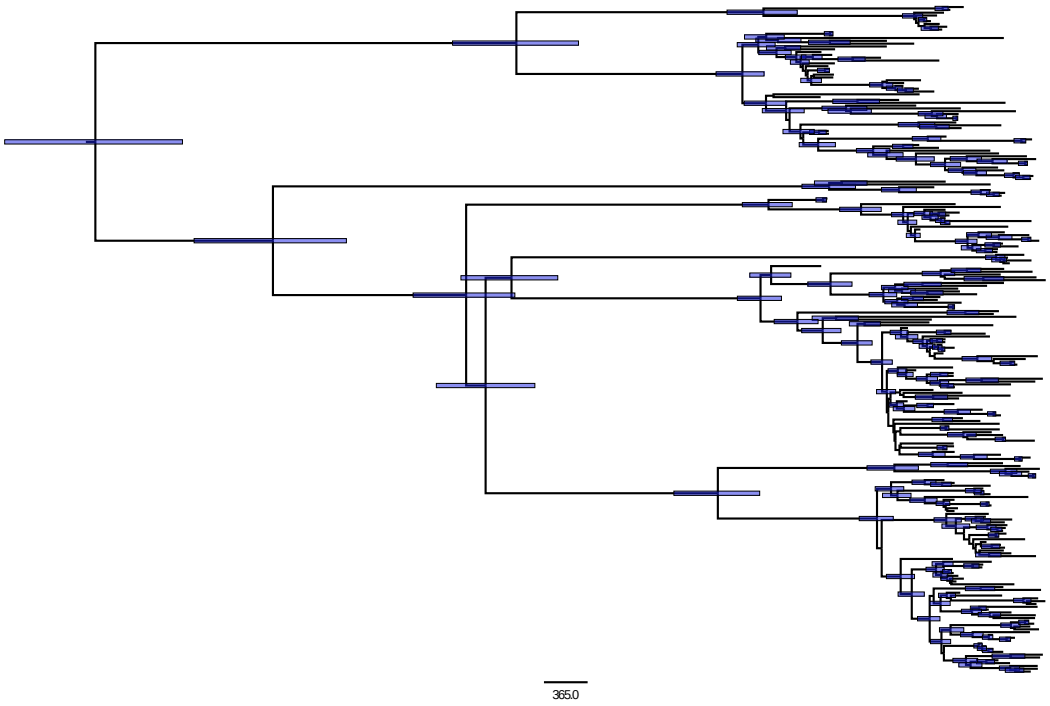

FIGURE 4.2: Maximum clade credibility tree for DENV-1. The 95% credible intervals on the coalescent times are shown as blue bars. The scale bar shows time in days.

Because the sequence samples were collected from patients living within a large geographic region, we also tried to reconstruct DENV-1 dynamics only within HCMC, reasoning that it may be easier to reconstruct seasonal dynamics on a more limited spatial scale than all of southern Vietnam. To do so, we performed a second Bayesian Skyline analysis with a genealogy from which all non-HCMC samples were removed. However, the Bayesian Skyline reconstruction of dynamics within HCMC also failed to recover the large seasonal fluctuations in DENV-1 incidence (Fig. 4.3).

4.3.1 Seasonality

Given the large discrepancy in seasonal dengue dynamics between the BSP and the hospitalization data, we first considered whether an epidemiological model that explicitly considered seasonality and nonlinear transmission dynamics might outper-

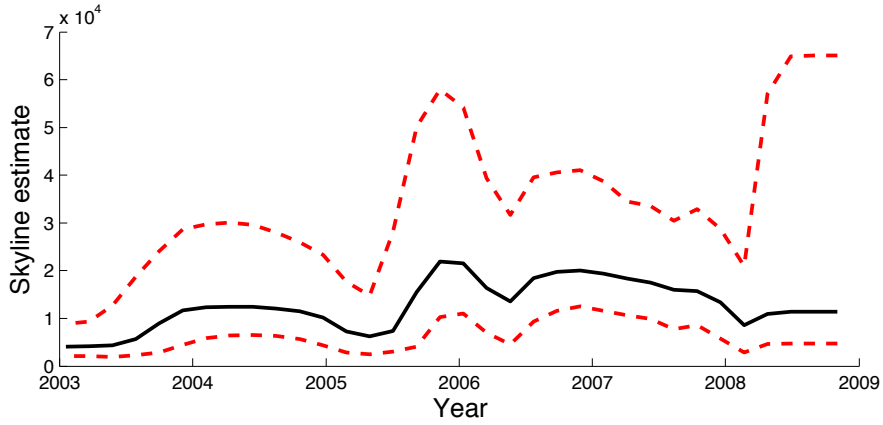

FIGURE 4.3: Bayesian Skyline Plot inferred from the HCMC DENV-1 sequences with all non-HCMC sequences removed. Black lines show the median posterior estimates and dashed red lines give the 95% credible intervals.

form the BSP. We therefore fit a SIR model with seasonal forcing to the DENV-1 genealogy using a coalescent model derived from the SIR model.

The population dynamics inferred from the DENV-1 MCC genealogy under the seasonal SIR model were qualitatively very similar to the dynamics in the BSP, with the seasonal fluctuations in incidence still an order of magnitude lower than those observed in the hospitalization data (Fig. 4.1C). Coinciding with the small fluctuations in incidence, epidemiological parameters estimated directly from the genealogy also indicated a very low seasonal amplitude (quantifying the strength of seasonality) and a difficulty in identifying the seasonal phase (Fig. 4.4A-B). Incidence estimated from the genealogy is also much higher than the number of hospital admissions, which we expected based on the fact that most dengue cases are not severe enough to require hospitalization. The basic reproduction number R_0 was estimated to be slightly higher than three (Fig. 4.4C), consistent with the range of serotype-specific R_0 values previously reported for dengue in southeast Asia (Ferguson et al., 1999; Thai et al., 2005). As in the BSP, the inferred dynamics show DENV-1 incidence peaking in 2006 and then steadily declining, at odds with the continued growth in

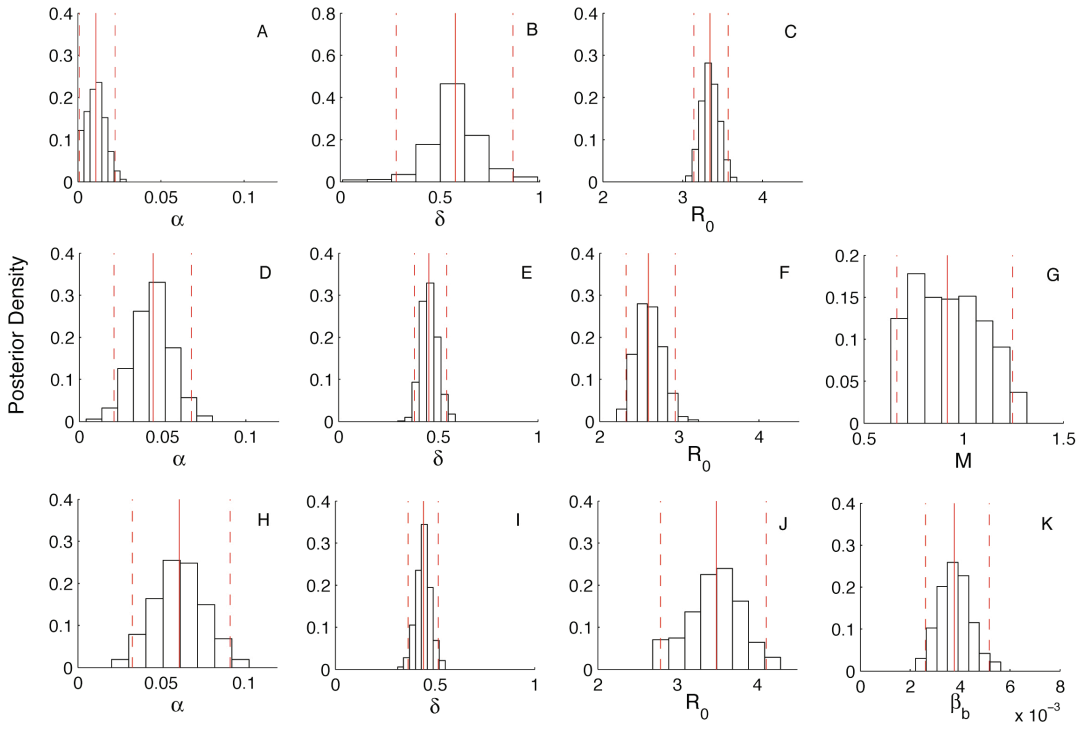

FIGURE 4.4: Posterior densities of the parameters inferred from the DENV-1 genealogy. Solid red lines indicate the median and dashed red lines indicate the 95% credible intervals of the posterior densities. The parameter α is the seasonal amplitude, δ is seasonal phase parameter, R_0 is the basic reproduction number, M is the ratio of mosquito to human population sizes in the vector-borne model, and β_b is the transmission rate between populations in the spatially structured model. (A-C) Estimates for the seasonal SIR model. (D-G) Estimates for the vector-borne model. (H-K) Estimates for the spatially structured model.

peak incidence each season observed in the hospitalization data. Similar dynamics were inferred from the genealogy containing only samples from HCMC (Fig. 4.1C).

To explore how uncertainty in the genealogy, especially with respect to the coalescent times, might affect our estimates, we additionally fit the seasonal SIR model to ten random trees sampled from the BEAST posterior tree distribution. Reconstructed dynamics did not significantly differ between trees, suggesting estimates were largely robust to phylogenetic uncertainty (Fig. 4.5A). The seasonal SIR model therefore appears unable to reconstruct dynamics consistent with the hospitalization

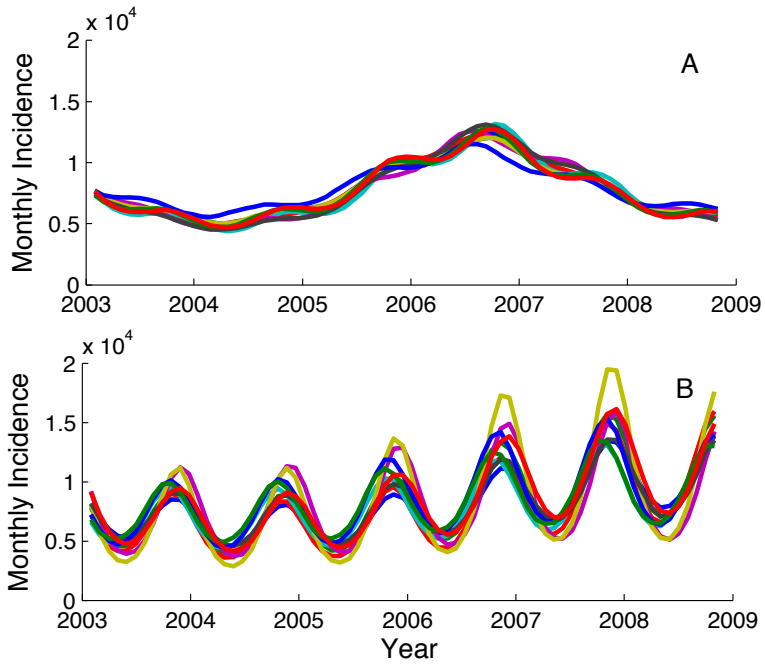

FIGURE 4.5: Reconstructed incidence of DENV-1 inferred from ten genealogies randomly sampled from the posterior distribution of trees. Only the median estimate from each tree is shown. (A) Estimates under the unstructured seasonal SIR model. (B) Estimates under the combined model with both spatial structure and vector-borne transmission.

data regardless of the geographic distribution of samples or the particular genealogy used for inference.

4.3.2 Vector dynamics

Dengue is a vector-borne virus spread by *Aedes* mosquitoes and the seasonality in dengue transmission presumably arises from fluctuations in mosquito population densities. Yet the seasonal SIR model fit above does not explicitly consider vector-borne transmission or mosquito population dynamics. To see if ignoring the vector population in the coalescent model could be distorting population dynamic inferences drawn from the genealogy, we fit a mechanistic vector-borne SIR model with seasonality in mosquito birth rates to the DENV-1 genealogy.

Table 4.1: Comparison of the models fit to the DENV-1 genealogy. Median posterior probabilities and Bayes factors are on the log scale.

Model	Median posterior	Bayes factor
Seasonal SIR	-2342.4	-
SIR + Vector	-2271.1	71.0
SIR + Space	-2253.9	88.3
SIR + Vector + Space	-2247.9	93.7

The population dynamics inferred from the DENV-1 genealogy under the vector-borne model show much larger seasonal fluctuations in incidence and correspond to the hospitalization data much better than those inferred under the directly transmitted model (Fig. 4.1D). These pronounced seasonal fluctuations arise from an estimated amplitude parameter that is much higher under the vector-borne model than the directly transmitted SIR model (Fig. 4.4D). We were also able to reconstruct the sustained growth in peak DENV-1 incidence each season through 2008, which we were unable to capture using the BSP or the directly transmitted model. Overall, a model comparison using Bayes factors showed that the vector-borne model provided a much better fit to the DENV-1 genealogy, with the posterior odds highly favoring the vector-borne model over the directly transmitted model (Table 4.1).

We were also able to obtain much more precise estimates of the seasonal phase parameter using the vector-borne model (Fig. 4.4E). The estimated phase coincides with a peak in mosquito population densities occurring in May or June, the same time at which *Aedes aegypti* densities peak in independent data from the Pasteur Institute in HCMC (Coudeville and Garnett, 2012). R_0 under the vector-borne model was estimated to be slightly lower than three (Fig. 4.4F), again consistent with the range of R_0 estimates in the literature (Ferguson et al., 1999; Thai et al., 2005). We were also able to obtain an estimate of the seasonal average of M, the ratio of mosquito to human population sizes, at a value close to one (Fig. 4.4G). For

comparison, estimates from other areas of the world have reported the number of *A. aegypti* per person to range from 0.2 to over 60.0, although most reported values fall below one (Focks and Chadee, 1997; Morrison et al., 2004; Koenraadt et al., 2008; Jeffery et al., 2009).

To gain intuition about why the vector-borne model was able to capture the population dynamics of DENV-1 better than the directly transmitted model, we studied the coalescent process for a vector-borne pathogen in Appendix B. Our mathematical analysis revealed that a vector-borne pathogen will in general have a lower rate of coalescence than a directly transmitted pathogen, although how much lower depends on the ratio of mosquito to human population sizes M. As M increases, so does the number of infected mosquitoes. A larger number of infected mosquitoes decreases the coalescent rate in a way similar to how larger population sizes decrease the coalescent rate in standard population genetics models. Thus, the larger M is, the lower the coalescent rate for a vector-borne disease will be relative to directly transmitted pathogen, although the relationship between M and the coalescent rate is nonlinear (Appendix B).

The seasonal fluctuations in the coalescent rate also become increasingly damped for the vector-borne model relative to the direct transmission model as M increases (Fig. 4.6). At high values of M, the coalescent rate is low year-round because the number of infected mosquitoes remains large year-round. The damped fluctuations in the coalescent rate will result in coalescent events being more uniformly distributed throughout the year in the genealogy, which will be interpreted as small fluctuations in human incidence under a coalescent model for a directly transmitted pathogen. It is therefore possible for a vector-borne pathogen to induce large seasonal fluctuations in human incidence, but to infer low-amplitude oscillations in human incidence under a coalescent model that ignores the vector population. Interestingly, our estimate of M around one falls in a part of parameter space in which this would likely oc-

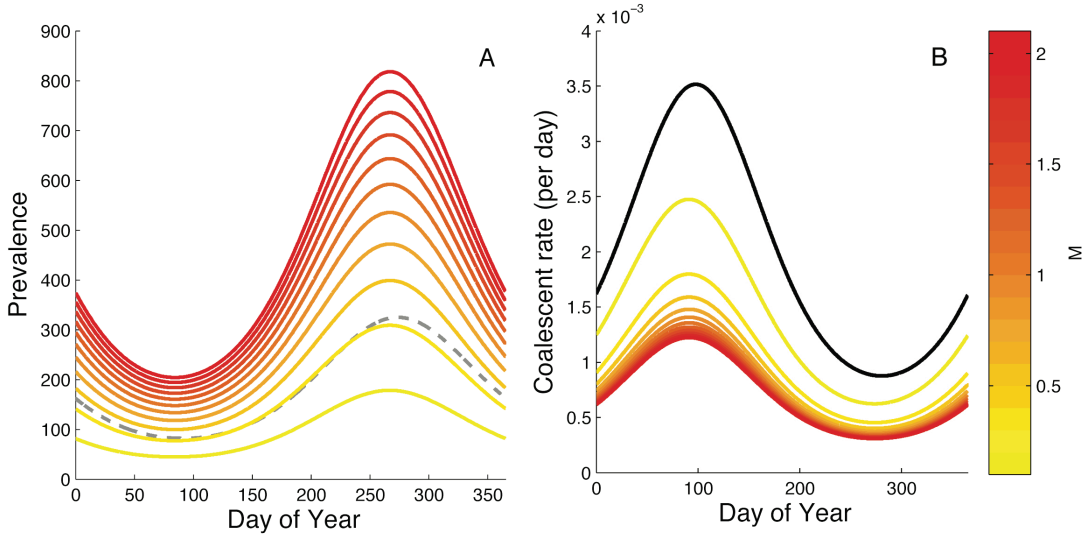

FIGURE 4.6: Comparison of seasonal coalescent rates and mosquito population sizes for the direct transmission model and the vector-borne model. (A) Simulated seasonal prevalence of the disease in humans and mosquitoes. The different colored lines show disease prevalence in mosquitoes assuming different values of M. Prevalence in humans (dashed-gray) is also seasonal, and constrained to be the same for both the direct and vector-borne model by keeping R_0 constant. (B) Seasonal coalescent rates for both models. The black line shows the seasonal coalescent rate for the direct transmission model and the different colored lines are the coalescent rates for the vector-borne model.

cur. These results therefore explain why the vector-borne model was able to better reconstruct the highly seasonal patterns of human DENV-1 incidence.

We also note that, on average, incidence inferred under the vector-borne model was about 10% lower than under the direct transmission model. This makes sense given the lower rate of coalescence for a vector-borne pathogen. Under both coalescent models, there is a certain number of infected humans that maximizes the likelihood of observing a given coalescent event. However, for the vector-borne model, this number of infected humans needs to be lower in order to increase the coalescent rate to compensate for the effect of the vector. In order to have fewer infected humans, the basic reproduction number R_0 is estimated to be lower under the vector-borne model (Fig. 4.4F). This in turn likely explains why we were able to capture

the continued rise in peak DENV-1 incidence each season through 2008 under the vector-borne model while incidence peaked too early under the direct model. The higher R_0 estimated under the direct transmission model causes the susceptible human population to be rapidly depleted and therefore incidence to decline after 2006. In comparison, the lower R_0 estimated under the vector-borne model allows for a more gradual depletion of susceptible humans and therefore a sustained, gradual rise in DENV-1 incidence each season.

4.3.3 Spatial structure

There is considerable spatial heterogeneity in dengue transmission dynamics across southern Vietnam, which includes large urban centers like HCMC as well as the less densely populated provinces to the north and west and the rural Mekong Delta region to the south. In our third model, we therefore considered how spatial structure may affect inferences drawn from the DENV-1 genealogy. As a starting point, we considered a spatially structured model with two populations: a HCMC and a non-HCMC population. While this simple model cannot account for all of the spatial heterogeneity in the region, including a non-HCMC population may allow us to more accurately infer dynamics within HCMC by controlling for the movement of lineages in and out of the city. To fit this structured model, we used the coalescent framework developed in Volz (2012) to compute the probability of each lineage being in either the HCMC or non-HCMC population conditional on the location of the external lineages at the time of sampling. Under this model, the coalescent rate between different lineages can differ depending on each lineage's probability of being in each population. For example, two lineages with high probabilities of being in HCMC will have a higher expected coalescent rate than two lineages with a high probability of being in different populations.

Incidence patterns inferred from the genealogy using the spatially structured

model show large seasonal fluctuations in incidence consistent with the hospitalization data (Fig. 4.1E). The seasonality parameters and R_0 for the structured model are shown in Fig. 4.4H-K. However, the dynamics inferred under the spatially structured model show the highest seasonal peak in incidence occurring in 2006, with subsequent years having lower peak incidence. We therefore also fit a combined model with both vector-borne transmission and spatial structure. Incidence patterns inferred under the combined model show both large seasonal fluctuations and the continued growth in peak incidence each season from 2004 to 2008, consistent with the hospitalization data (Fig. 4.1F). Bayes factor comparisons also showed that while both the vector-borne and spatially structured model fit the genealogy significantly better than the unstructured model, the combined model fits better than either of the two models individually (Table 4.1). The population dynamics reconstructed under the best-fitting combined model also appear robust to phylogenetic uncertainty (Fig. 4.5B).

The spatially structured model we fit above only assumed seasonality in the HCMC population. However, both hospitalization and notifiable disease data (Cuong et al., 2013) indicate that all of Vietnam's southern provinces experience strong seasonal fluctuations in incidence. These data further indicate that seasonal outbreaks begin and peak one to three months earlier in the provinces than in HCMC (Fig. 4.7A-B). We therefore fit a second model that included seasonality in both the HCMC and non-HCMC populations and allowed the amplitude and phase of seasonality to vary between the two populations. While this more complex model did not fit the genealogy significantly better (Bayes factor < 1.0), we were able to reconstruct the differences in seasonality between the HCMC and non-HCMC populations observed in hospitalization data (Fig. 4.7C). The reconstructed incidence clearly shows that the dengue season begins in the provinces about one to three months earlier than in HCMC. Thus, including spatial structure in the coalescent model not only allowed us to improve our estimates of the population dynamics in HCMC, but to detect

spatiotemporal differences in dengue transmission across the region.

Previous phylogeographic analyses of dengue in southern Vietnam have found evidence for frequent movement of lineages in and out of HCMC (Rabaa et al., 2010; Raghwani et al., 2011). Consistent with these findings, we estimated a relatively high between-population transmission rate (Fig. 4.4K). Using this rate along with the other estimated parameters and population dynamics, we computed the probability of each lineage being in HCMC over time (Fig. 4.8). The mapping of lineage state probabilities onto the tree indicated that many different lineages have been imported and exported in and out of HCMC; it is likely that some lineages have even moved in and out of HCMC multiple times since DENV-1 reemerged as the dominant serotype in the early 2000s.

To better understand how the movement of lineages in a spatially structured population shapes the genealogy, we simulated dengue-like dynamics under the spatial SIR model with parameters close to what we inferred from the DENV-1 genealogy (Fig. 4.9A). The expected coalescent rates for two hypothetical lineages sampled in HCMC are shown in Fig. 4.9B. If we ignore the spatial structure of the population and assume that the two lineages remain in HCMC over time, the rate at which these lineages coalesce fluctuates between high and low as the prevalence cycles between low and high, giving the strong signal of seasonality we expect to see in the timing of coalescent events.

In a spatially structured population however, our two hypothetical lineages may not remain in the same population indefinitely going into the past because of movement of lineages between populations. This results in a decline in the probability that our two hypothetical lineages remain in the same population as we recede into the past. As this happens, the coalescent rate decreases and the seasonal fluctuations in the coalescent rate also dampen going back in time (Fig. 4.9B). In the very recent past, both lineages retain a high probability of being in HCMC and so the

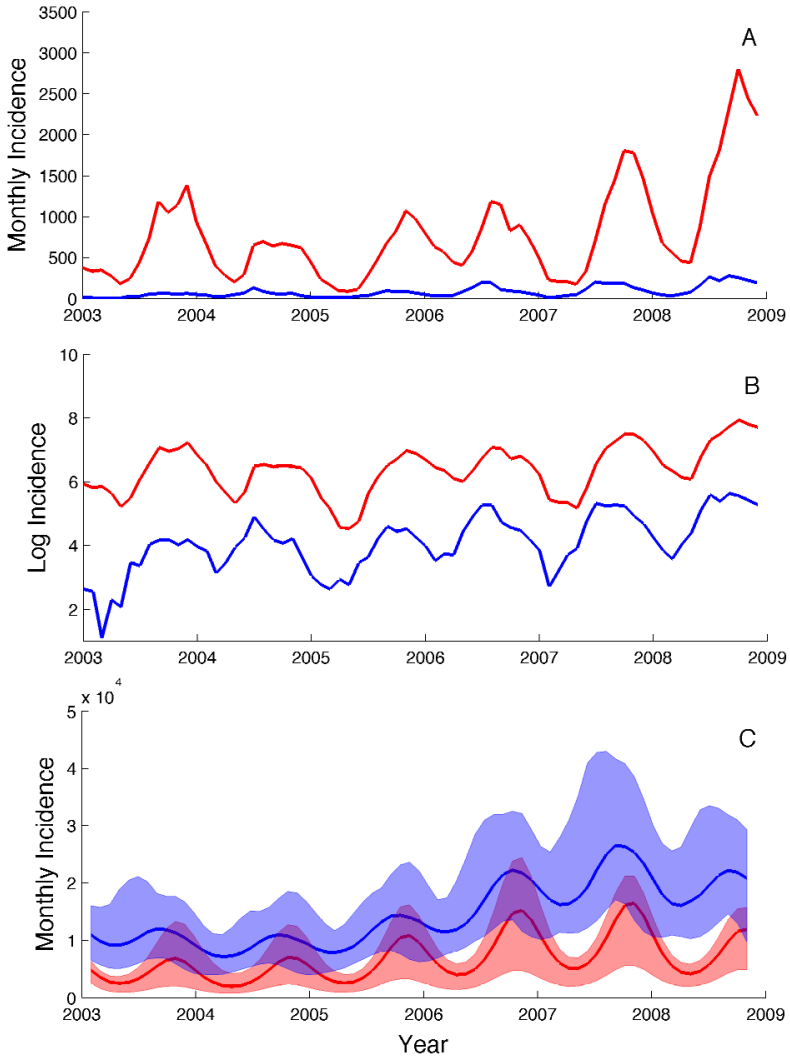

FIGURE 4.7: Population dynamics of dengue in HCMC (red) and the non-HCMC provinces (blue). (A) Monthly dengue hospital admissions in HCMC by location of patients' primary place of residence. The small number of cases from the provinces likely reflects the low probability of dengue patients in the provinces being hospitalized in HCMC. (B) Same as in A but on a log scale to emphasize the difference in seasonal phase between HCMC and the provinces. (C) Incidence inferred under the spatially structured model for the HCMC and non-HCMC populations. Shaded regions give the 95% credible intervals.

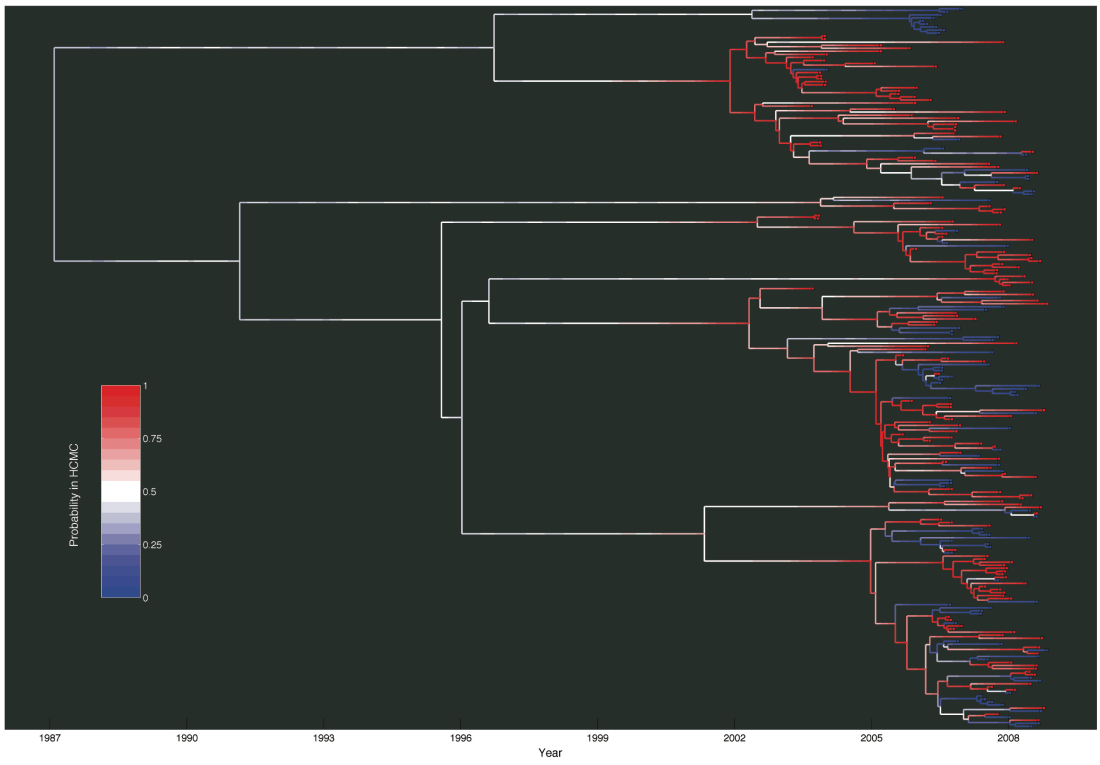

FIGURE 4.8: DENV-1 genealogy showing the probability that each lineage is in HCMC. Lineage state probabilities were computed under the spatially structured model using the median posterior values of all parameters. The colored boxes at the tips indicate the population from which the lineage was sampled. Red indicates HCMC and blue indicates the non-HCMC population.

coalescent rate reflects the highly seasonal coalescent process in HCMC. However, in the more distant past, the coalescent rate remains low year-round because of the higher probability of the lineages being in different populations. Thus, spatial structure destroys the strong signal of seasonality we expect in the timing of coalescent events in an unstructured population. This likely explains why were able to infer strong seasonality using the structured coalescent approach but were unable to do so simply by removing samples from outside of HCMC from the genealogy. The rapid movement of lineages in and out of HCMC means that many of the lineages sampled in HCMC were not in HCMC in the recent past. Only by taking into account the

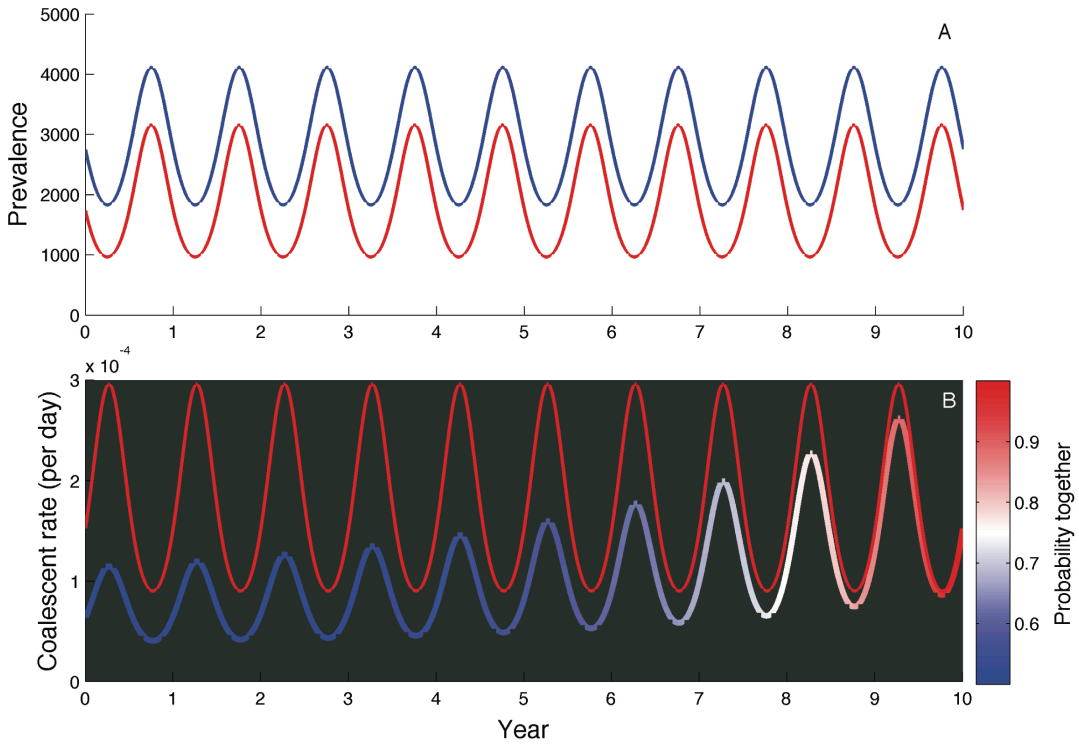

FIGURE 4.9: (A) Simulated seasonal dynamics for a structured population with a focal (red) and global (blue) population representing the HCMC and non-HCMC populations, respectively. (B) Expected coalescent rates for two lineages both sampled in HCMC at the end of year ten. The solid red line shows the strong seasonal fluctuations in coalescent rates in HCMC under an unstructured model. The colored line is the coalescent rate under the spatially structured model where the color shows how the probability of the lineages being together in the same population changes over time.

probable location of lineages through time can we detect the signal of seasonality in the timing of coalescent events within a given population.

4.4 Discussion

Our phylodynamic analysis of DENV-1 shows that, while it is possible to reconstruct complex population dynamics from genealogies, additional ecological factors may need to be included in coalescent models in order for demographic inferences to

be accurate. For DENV-1, we were unable to detect the large seasonal fluctuations in dengue incidence using the popular Bayesian Skyline method or even a coalescent model derived from a SIR epidemiological model that allowed for seasonality. However, using models that included either vector population dynamics or spatial structure in the host population, we were able to successfully reconstruct DENV-1 dynamics. The substantially better fit of these two more complex models indicates that vector dynamics and spatial heterogeneity likely play a large role in shaping the genealogy of dengue.

More generally, our results add to the mounting body of evidence that both population dynamics and structure can strongly impact the shape of viral genealogies (Frost and Volz, 2010; Bahl et al., 2011; Pybus et al., 2012; Duke-Sylvester et al., 2013; Robinson et al., 2013; Stadler and Bonhoeffer, 2013). When conducting phylodynamic analyses, this dependence of phylogeny on ecology can be both good and bad. On the upside, the strong influence of ecological factors means that genealogies contain valuable information about the dynamics and structure of populations that may be absent in other sources of data. For example, we were able to infer the transmission rate of DENV-1 between HCMC and the non-HCMC population, about which hospitalization records contain no information. On the downside, our results for DENV-1 suggest that we may need to include ecological factors in coalescent models that may not be of primary interest to us or we know little about when inferring population dynamics from genealogies.

While it is difficult to know *a priori* what ecological factors need to be included in a coalescent model for a particular pathogen, different factors can be tested by formulating them in terms of a mechanistic model that can be fit to genealogies through the appropriate coalescent model. These models do not need to be very complex—our vector-borne model and spatial model were both very simple. The appropriate coalescent model can then be derived from the forward-time model using

the coalescent framework of Volz (2012). Different models can then be compared using model selection, as we did using Bayes factors. The advantage of this approach is that mechanistic insights into the factors shaping genealogies can be found, increasing our general knowledge about which factors are most important in shaping the genealogies of different pathogens.

For dengue, it is interesting to consider why vector-borne transmission and spatial structure had such a large impact on our estimates. While mosquitoes play an integral role in dengue's ecological dynamics, it is not clear from standard coalescent theory why the vector population needs to be considered. The pairwise coalescent rate under most standard population genetics models depends only on the population size. If we assume that the population size for an infectious pathogen is equivalent to the number of infected hosts, we might think that the coalescent rate should show large fluctuations as the number of infected humans rises and falls. However, the coalescent model derived from the vector-borne SIR model tells us that it is not only prevalence in the human population that is important, but that mosquito population densities are also important. As the mosquito population increases so too does the number of infected mosquitoes, resulting in a lower probability that a given lineage in a human will coalesce with a given lineage in a mosquito. If mosquito population densities are high year-round, the coalescent rate will remain low year-round even if there are large fluctuations in human infections. Thus, unless the coalescent model includes the vector, estimates of the strength of seasonality will be biased.

Given the large amount of spatial heterogeneity in dengue dynamics in southern Vietnam and the widespread movement of people in the region [Rabaa et al. 2010, Raghwani et al. 2011], it does not seem too surprising that including spatial structure in the coalescent model improved our ability to reconstruct population dynamics in a particular population like HCMC. In highly structured populations, lineages in different isolated communities may have little probability of coalescing

with one another, especially if transmission between those communities is rare. In this case, the distribution of coalescent events over the genealogy will depend more on the spatial structure of the population than on the dynamics within any particular community. This is one of the reasons why phylodynamic estimates of population sizes in spatially structured populations are usually taken to be a measure of the relative genetic diversity of the population, which may not reflect the true population dynamics (Carrington et al., 2005; Pybus and Rambaut, 2009). However, in many cases we may not be interested in patterns of relative genetic diversity but actually want to reconstruct the population dynamics within a particular focal population. As we showed for HCMC, it is possible to reconstruct the dynamics in a focal population by taking into account the movement of lineages and their probable locations through time in the coalescent model. Remarkably, including spatial structure in the coalescent model even allowed us to detect the short lag in time between the beginning of the dengue season in the provinces and the beginning of the seasonal outbreak in HCMC.

Our experience with DENV-1 may also shed some light on why phylodynamic estimates of seasonal population dynamics have been successful for some pathogen populations but not others. Annual seasonal dynamics have been inferred from viral sequence data before, most notably for influenza A in temperate regions (Rambaut et al., 2008). However, in the case of influenza, there is no year-round transmission in temperate regions and the viral population is seeded by imported viruses each year (Nelson et al., 2007; Bedford et al., 2010). Therefore, looking back in time, all lineages sampled during a given season that descended from one of the imported lineages will coalesce at the beginning of the season. Because of this, influenza genealogies contain a strong signature of exponential growth each year—seasonality is so strong that it masks the effects of global population structure on the genealogy. However, if prevalence varies seasonally but the pathogen can still persist in the focal

population year-around, accounting for population structure might be necessary. In less seasonal populations, some lineages may remain in the focal population for many seasons going into the past while other lineages may have left the focal population, obscuring the local population dynamics in the genealogy. This may account for why previous phylodynamic studies of populations with seasonal dynamics but year-round persistence were unable to reconstruct accurate seasonal fluctuations in prevalence from genealogies (Amore et al., 2010; Bennett et al., 2010). In such cases, it would be interesting to see if our strategy of subdividing the population into a global and a focal population in the coalescent model would improve estimates of seasonality.

Adding ecological realism to our coalescent models greatly improved our ability to accurately infer DENV-1 dynamics, but do our estimates of dengue incidence accurately reflect the true number of dengue infections? If they do, it would be of great significance to dengue epidemiology, as determining overall disease burden remains challenging because clinical cases represent only a small fraction of all cases. However, we are somewhat skeptical that our estimates accurately reflect the true incidence of dengue because there are several ecological factors that we did not consider in our models that could bias our estimates. For one, our models assume that there is no heterogeneity in transmission rates, whereas in reality there is likely a large amount of variation in the rate at which different mosquitoes bite and the rate at which different humans are bitten (Scott and Morrison, 2010). Variation in transmission rates will increase coalescent rates, akin to how reproductive variance reduces effective population size and increases coalescent rates in standard population genetics models (Griffiths and Tavare, 1994; Pybus et al., 2001; Charlesworth, 2009; Koelle and Rasmussen, 2012). Transmission heterogeneity would therefore cause us to underestimate the true number of infections from the genealogy. One in theory could use the ratio of the observed number of infections to the estimated effective population size to infer the extent of transmission heterogeneity, as was

done by Magiorkinis et al. (2013), but again for dengue we have no way of knowing the true number of infections. We also did not consider fine-scale spatial structure within HCMC and the provinces. In contrast to the effect of transmission variance, unaccounted for spatial heterogeneity would decrease the rate of coalescence, similar to an increase in the effective population size in standard population genetics models (Wright, 1943; Beerli and Felsenstein, 2001; Laporte and Charlesworth, 2002). Thus, if there is strong local spatial structure, our estimates of incidence will be biased upwards. It is possible that local population structure counteracts the affects of variable transmission rates so that these two sources of potential bias cancel each other out, but the relative magnitude of each is unknown. We therefore urge caution in interpreting our estimates as representative of the true number of dengue cases.

There are certainly many other ecological factors that could distort inferences from genealogies that we did not consider for DENV-1. Notably, we did not consider interactions between DENV-1 and the remaining three dengue serotypes, nor interactions between different DENV-1 genotypes. However, there was only a single dominant DENV-1 genotype circulating in the population and single-serotype SIR models were sufficient to capture the rise in DENV-1 incidence that occurred in Vietnam during the period we considered. However, we cannot rule out selection acting within this genotype. Theoretical work has shown that both purifying and directional selection increases coalescent rates deeper in the genealogy, similar to a decrease in the past effective population size (O'Fallon et al., 2010; Walczak et al., 2012; Neher and Hallatschek, 2013). Selection could therefore result in a spurious inference of population growth, but we believe it is far more likely that the rise in DENV-1 inferred from the genealogy reflects the actual rise in DENV-1 observed in the hospitalization data. In the future, however, it would be interesting to look at multi-strain models that could encompass the competitive and facilitative interactions between different dengue genotypes and serotypes, as long as sufficient data

are available.

We end by noting that the methods we used to estimate population dynamics and parameters from the DENV-1 genealogy could be greatly improved upon in the future. One of the shortcomings of the methods used here is that phylogenetic uncertainty in the genealogy is not fully taken into account. For DENV-1, the availability of whole genome sequence data meant that there was relatively little uncertainty in the genealogy and we showed our phylodynamic estimates were robust to this level of uncertainty. But in other cases where sequence data are less informative about the genealogy, phylogenetic uncertainty will need to be considered. Another shortcoming is that we only fit deterministic epidemiological models, although we now have methods for fitting stochastic models to genealogies (see Chapters 2 and 3). However, for dengue in southern Vietnam, stochasticity can reasonably be ignored because the large number of infections and the strong seasonal dynamics ensure that dynamics are unlikely to widely differ from what is expected under deterministic models. Yet, in other cases, stochasticity can play an important dynamical role, like at the beginning of epidemics when prevalence is low. As we have shown, fitting mechanistic models to genealogies can improve our understanding of the forces shaping genealogies and improve phylodynamic estimates; extending current methods to include phylogenetic uncertainty and stochasticity will help to further improve the robustness of phylodynamic inference.

5

Conclusion

While the term "phylodynamics" was coined only 10 years ago, the field of phylodynamics has developed at a very rapid pace since then (Grenfell et al., 2004; Volz et al., 2013b). The first phylodynamic studies of infectious pathogens only considered how genealogies were generated under rather simple ecological and evolutionary conditions. The models currently being fit to genealogies are far more complex and contain far greater ecological realism than was possible only a few years ago. The increasing complexity of these models has also brought about the need for increasingly sophisticated statistical methods. Rather than try to develop statistical methods for any particular model, in my dissertation I have tried to develop a general framework for the types of models typically used in mathematical epidemiology—stochastic models that can accommodate nonlinear population dynamics as well as different forms of population structure. As the application of these methods to pathogens like dengue and HIV has illustrated, the greater complexity of these epidemiological models is often necessary in order to accurately estimate epidemiological parameters and reliably reconstruct complex disease dynamics. With these advances in modeling and phylodynamic methods, it has also become possible to extract information from

genealogies not easily obtainable from other sources of epidemiological data, such as the stage-specific transmission rates estimated for HIV in Chapter 3. Yet, while new advances in the field are occurring all the time, some fundamental challenges in phylodynamics remain. I would therefore like to conclude by sharing my own perspective on some of these recent advances and some of the challenges that still need to be addressed if phylodynamics is to become fully integrated into modern epidemiology.

Given that phylodynamics is such a young, rapidly evolving field, it seems inevitable that new approaches and improved methods will continue to arise. One particularly active area of research is on birth-death models, which can be derived from branching processes to describe the genealogies of pathogens, as well as statistical methods for fitting these models to genealogies (Leventhal et al., 2014; Stadler, 2010; Stadler et al., 2012). Indeed, these birth-death approaches do have some advantages over the backward-time, coalescent-based approaches my colleagues and I have taken. For one, birth-death models are forward-time models that naturally take into account demographic stochasticity due to finite population sizes (Kendall, 1948). It is therefore not necessary to first forward simulate a stochastic trajectory and then condition on this trajectory when calculating the likelihood of a genealogy, as our coalescent-based particle filtering methods require. Moreover, under certain simple epidemiological models, it is possible to analytically compute the likelihood of a genealogy under a birth-death model (Stadler et al., 2012), removing the need for computationally expensive methods like particle filtering all together.

However, the birth-death framework has its own limitations. Computing the likelihood of a genealogy requires knowledge about how lineages are sampled over time or that the sampling fraction can be estimated by other means (Stadler, 2010). This is not necessary with coalescent-based approaches, as the coalescent provides the likelihood of a genealogy conditional on the sample. Analytical methods for

birth-death models are also currently limited to fairly simple epidemiological models, although these methods can be extended to include basic forms of population structure (Stadler and Bonhoeffer, 2013). However, for more complex epidemiological models with nonlinear, stochastic dynamics and multiple different forms of population structure, it does not seem possible to derive analytical likelihood expressions for birth-death models, as even the forward-time dynamics of these more complex models cannot be solved analytically. It therefore seems inevitable to resort to simulation-based methods and I suspect that the coalescent methods I have developed will remain useful for their generality and the diversity of models that they can be applied to. Moreover, for simple epidemiological models where the simulation-based methods can be directly compared against exact birth-death methods, the two methods give remarkably consistent estimates of population dynamics and model parameters (personal communication, Veronika Boskova and Tanja Stadler), suggesting simulation-based methods are generally reliable and a perfectly adequate choice when more exact methods are not available.

Beyond these methodological issues, a more general question in phylodynamics concerns how much we can realistically hope to learn from genealogies about epidemiology. Now that phylodynamic inference methods have been applied to a variety of pathogens with different epidemiological dynamics, it appears safe to conclude that pathogen genealogies are highly informative about past population dynamics. While in some cases inadequate sampling or weak phylogenetic signal in sequence data may have prevented phylodynamic methods from detecting fluctuations in prevalence (de Silva et al., 2012; Siebenga et al., 2010), the simulation studies I have conducted demonstrate that it is generally possible to reliably reconstruct population dynamics from genealogies using sampling protocols and sizes roughly equivalent to what is currently available for most well-studied human pathogens. However, certain demographic features appear harder to infer from genealogies than others. Most

notably, population structure and parameters like migration rates are difficult to estimate from genealogies unless lineages transition between populations relatively slowly compared to the timescale on which coalescent events occur in the genealogy, as shown in Chapter 3. Nevertheless, this is only a problem when population structure is relatively weak and lineages move rapidly between populations. In more strongly structured populations, genealogies can be informative about population structure, especially relative to traditional sources of epidemiological data like time series where case counts may be aggregated across different subpopulations, obscuring the underlying population structure.

It therefore appears that we can learn quite a lot about the epidemiology of pathogens from studying their phylodynamics, but this raises yet another question: how much confidence should we place in phylodynamic estimates of epidemiological parameters and dynamics? This question seems most pertinent when, unlike for the dengue and HIV examples I explored, there is no reference in terms of independent observational data to compare phylodynamic estimates against. While using more ecologically realistic coalescent models can help improve phylodynamic estimates, it is difficult to know what complexities should be included in a coalescent model for a particular pathogen *a priori*. Although formal model selection criteria can help compare models of differing complexity, even estimates obtained under the best fitting models can be arbitrarily bad in reality. At the same time, it is impractical and probably unnecessary to include every ecological complexity that we might consider in our models. I would therefore argue that what we need is a more broad conceptual understanding of what ecological and demographic factors have the most impact on pathogen genealogies, and therefore the most potential to bias estimates if not considered. In particular, I think phylodynamics would greatly benefit from a better understanding of how different types of heterogeneity in both host and pathogen populations shape genealogies and the consequences this has for inference.

With respect to heterogeneity in host populations, we already have some understanding of how large-scale population structure shapes genealogies. As the dengue example in Chapter 4 illustrates, geographic isolation due to spatial structure can skew the distribution of coalescent events over a genealogy in a way that masks signals of population growth and decline. Fortunately though, even relatively simple structured coalescent models like the two-population model used for dengue can control for the effect of spatial structure and provide much more accurate demographic estimates. However, binning lineages into discrete geographic regions is a relatively coarse way of accounting for spatial structure that may not take advantage of all available information about sampling locations. Incorporating spatial structure in such a way that allows lineages to diffuse along a continuous landscape, as has recently been done in phylogeography (Lemey et al., 2010; Pybus et al., 2012), may provide better spatiotemporal resolution when inferring disease dynamics from genealogies.

At smaller scales, variability at the level of individual hosts may play an important role in shaping pathogen genealogies, but this type of variation is difficult to capture with compartmental epidemiological models that simply aggregate hosts into a small number of discrete classes. Real populations exhibit tremendous amounts of variation in contact rates and other epidemiological traits. This is especially true for sexually transmitted diseases, where the number of sexual contacts between different hosts can vary by orders of magnitude. For these pathogens, network models that explicitly consider contacts between individual hosts may be more appropriate than compartmental models. But while the population dynamics of pathogens spreading on epidemiological networks have been well-studied in recent years (Bansal et al., 2007; Keeling and Eames, 2005), how network structure shapes genealogies and phylodynamic estimates has only just begun to be explored. Early simulation studies have shown that network topology can have a strong impact on pathogen genealogies

and potentially bias coalescent-based inferences if not properly taken into account (Leventhal et al., 2012; Robinson et al., 2013). A theoretical framework that incorporates network structure and individual-level heterogeneity into phylodynamics would therefore be very useful, especially if developed in a way that would allow for inference of network properties from genealogies. While it may not be possible to reconstruct detailed networks from sparsely sampled populations, it may still be possible to estimate statistical properties of networks such as the degree distribution of contacts in the host population from genealogies, which are typically not readily identifiable from other common sources of epidemiological data.

Heterogeneity within pathogen populations may also play a considerable role in shaping genealogies, especially if there is phenotypic variation in traits relevant to fitness that selection can act on. In this case, pathogen evolution cannot be considered as a purely neutral process. But while there has recently been work done on how non-neutral evolutionary dynamics shape pathogen genealogies using phylodynamic simulations (Koelle et al., 2006), non-neutral evolutionary processes have yet to be incorporated into phylodynamic inference methods, reflecting the inherent difficulty of the problem. One relatively simple way of including non-neutral evolution in phylodynamics is through multi-strain epidemiological models that allow different strains to interact and compete for resources, usually susceptible hosts. If different strains form monophyletic clades within a larger genealogy, one can divide strains into independent genealogies and fit multi-strain models simultaneously to these trees. While the genealogy of each strain is considered independently, the strains can still interact through competition for shared resources and thus influence one another's ecological dynamics. I experimented with this approach while working on my dissertation and it does appear to be feasible when only a small number of strains are present, but beyond two or three strains multi-strains models become too high-dimensional to be of practical use. This approach is also somewhat unsatisfactory in that it fails

to take into account genuine evolutionary novelty arising from *de novo* mutations entering the population, and thus not applicable to pathogens rapidly adapting to continual changes in their environment. Extending phylodynamic models to incorporate evolutionary change beyond standing variation remains an open problem, but also a formidable one as it is not clear how to incorporate newly emerging strains in standard models that treat pathogen populations as closed systems.

Incorporating these types of host and pathogen heterogeneity into phylodynamic inference will allow for greater ecological realism, and should make phylodynamic estimates more reliable and less prone to biases. Hopefully, it will also open the way for researchers to learn about new aspects of the epidemiology of pathogens that have not previously been possible from more traditional sources of data. But as I have argued here, it will also be important to understand at a conceptual level how different ecological and evolutionary factors shape genealogies and the relative importance of each factor. Doing so should not only improve phylodynamic inference, but also lead to an improved understanding of the ecological and evolutionary processes that drive disease dynamics.

Appendix A

Particle Filtering with a Genealogy

In Rasmussen et al. (2011), it was shown how particle filters could also be applied to genealogies instead of standard observational data by using a coalescent model to relate the genealogy to the unobserved state variables. To briefly review the algorithm, the particle filter is run forward in time from time $t = 1$ to time $t = T$, sequentially updating the particle states x_t^j and assigning importance weights w_t^j for each particle j at each time step. Particle states are updated at each time step by simulating from a proposal density $q(x_t^j|\bullet)$. Particle weights are then updated to reflect the posterior probability of each particle trajectory $x_{1:t}$ up to time t given the data observed up to time t, which can either be some generic observation data $z_{1:t}$ as described in Chapter 2 or, as considered here, the genealogy up to time t, $\mathcal{G}_{1:t}$. Therefore, at any time t, the weighted system of particles gives an importance sampling approximation to the density $p(x_{1:t}|\mathcal{G}_{1:t}, \theta)$. Once we reach time $t = T$, we sample a state trajectory $x_{1:T}^*$ by randomly selecting a particle according to the final normalized particle weights W_T to obtain a random sample from $\hat{p}(x_{1:T}|\mathcal{G}_{1:T}, \theta)$. We can also use the weights assigned to the particles to approximate the marginal

likelihood of the parameters $p(\mathcal{G}_{1:T}|\theta)$.

Algorithm A1: The particle filter targeting $p(x_{1:T}|\mathcal{G}_{1:T}, \theta)$

1. Initialize the particle filter at time $t = 1$ with N particles.

   (a) Set x_1^j to initial values for all particles.

   (b) Assign normalized weights, $W_1^j = \frac{1}{N}$.

2. Run filter from $t = 2$ to $t = T$.

   (a) Propagate particles forward by drawing from the proposal density $q(x_t^j|\bullet)$.

   (b) Set $x_{1:t}^j = (x_{1:t-1}^j, x_t^j)$ for all particles.

   (c) Compute unnormalized weights,

$$w_t^j = \frac{(w_{t-1}^j)p(\mathcal{G}_{t-1:t}|\theta, x_t^j)p(x_t^j|x_{t-1}^j, \theta)}{q(x_t^j|\bullet)}. \tag{A.1}$$

   (d) Normalize weights, so that $W_t^j = \frac{w_t^j}{\sum_{j=1}^N w_t^j}$.

   (e) If resampling at t, choose parent particle indexes a_t^j according to their weights, such that $p(a_t^j = k) = W_t^k$. Set $x_t^j = x_t^k$ and set $w_t^j = 1$. Otherwise, set $a_t^j = j$.

3. Sample $x_{1:T}^*$ from $\hat{p}(x_{1:T}|\mathcal{G}_{1:T}, \theta)$ by tracing the ancestry of one particle back through time.

   (a) Sample a single particle index k such that $p(k) = W_T^k$ and set $b_T^k = k$.

   (b) For $t = T - 1$ to $t = 1$, set $b_t^k = a_t^{b_{t+1}^k}$.

   (c) Set $x_{1:T}^* = x_{1:T}^{b_{1:T}^k}$.

4. Compute marginal likelihood estimate

$$\hat{p}(\mathcal{G}_{1:T}|\theta) = \prod_{t=1}^{T} \frac{1}{N} \sum_{j=1}^{N} w_t^j. \tag{A.2}$$

Note that we have left the exact form of the proposal density $q(x_t^j|\bullet)$ unspecified in lack of an ideal proposal density. Nevertheless, we can update the particle states by simulating directly from the epidemiological process model $p(x_t|x_{t-1}, \theta)$ (Ionides et al., 2006; Cappe et al., 2007). In this case, the weighting function simplifies to

$$w_t^j = (w_{t-1}^j) p(\mathcal{G}_{t-1:t}|\theta, x_t^j). \tag{A.3}$$

This has the fortuitous result that the term $p(x_t^j|x_{t-1}^j, \theta)$ does not appear in the weighting function so that we do not need to compute these transition densities explicitly, which is often not possible for continuous-time, nonlinear epidemiological models.

The particle filtering algorithm also allows for resampling to occur at the end of each time step, which is often necessary to ensure the practical feasibility of the algorithm. Resampling removes unpromising particle trajectories before we reach time T by replacing particles with low weights, and therefore very likely low posterior probabilities, with particles with high weights. However, it is often unnecessary and computationally wasteful to resample after each time step, especially if most particles have high unnormalized weights or there is little variance in weights across the particle population (Doucet and Johansen, 2009). For this reason, we allow for adaptive resampling by making sampling after each step of the algorithm optional and generally resample as infrequently as possible. However, if we do resample, it requires us to track the ancestry of each particle in the population so that we can sample a single particle state trajectory at time T. We do this by recording the parent index a_t^j of each particle in the population at each time step. At time T, we

choose a single particle index k and can trace that particle's ancestry back through time by setting $b_t^k = a_t^{b_{t+1}^k}$ for all times $t < T$. Thus $b_{1:T}^k$ gives the ancestral lineage of particle k in that b_t^k gives the index of the ancestor of particle k at time t. The state trajectory associated with particle k is then $x_{1:T}^{b_{1:T}^k}$.

Appendix B

A Coalescent Model for a Vector-Borne Pathogen

In Chapter 4, a vector-borne coalescent model is used to understand how the coalescent process for a pathogen transmitted by a mosquito vector differs from the coalescent process of a directly transmitted pathogen. For dengue in southern Vietnam, it was shown that including the vector population in the coalescent model can have a substantial effect on the population dynamics inferred from the genealogy of a vector-borne disease. Here, I show how the vector-borne coalescent model was derived from the more general structured coalescent framework of Volz (2012). I also show that under equilibrium epidemiological conditions, the rate of coalescence for a vector-borne pathogen will in general be slower than for a directly transmitted pathogen, but how much slower depends on the size of the vector population relative to the host population size.

For a directly transmitted pathogen, all pathogen lineages can be assumed to be in a single infected host population. However, for a vector-borne pathogen like dengue virus, viral lineages can be in either an infected human or an infected mosquito and thus the population is structured. We therefore use the structured coalescent

framework of Volz (2012), who showed that for a generic structured population where lineages can be in any of m different states, the rate of coalescence λ_{ij} for two lineages i and j is

$$\lambda_{ij} = \sum_{k}^{m} \sum_{l}^{m} \frac{f_{kl}}{Y_k Y_l} \left(p_{ik} p_{jl} + p_{il} p_{jk} \right), \tag{B.1}$$

where p_{ik} is the probability that lineage i is in state k and p_{jl} is the probability that lineage j is in state l. f_{kl} is the rate at which lineages are transmitted from state k to state l and Y_k and Y_l are the total number of infected individuals in states k and l, respectively.

Adapting (B.1) to the vector-borne SIR model presented in (4.4), the pairwise rate of coalescence under the vector-borne model is:

$$\lambda_{ij} = \frac{\beta_{vh} \frac{S_h}{N_h} I_v + \beta_{hv} S_v \frac{I_h}{N_h}}{I_v I_h} \left(p_{iv} p_{jh} + p_{ih} p_{jv} \right), \tag{B.2}$$

where p_{iv}, for example, gives the probability that lineage i is in a vector.

In general, we can compute the probability that a given lineage is in a certain state if we know the initial state of the lineage at the time of sampling and the rates at which lineages move between states. How the lineage state probabilities change as we move backwards in time can be tracked using master equations (Volz, 2012). For our vector-borne model, assuming that the number of infected humans and vectors is large relative to the number of lineages in the genealogy, the master equations for the vector and human states are:

$$\frac{dp_{iv}}{ds} = p_{ih} \frac{\beta_{vh} \frac{S_h}{N_h} I_v}{I_h} - p_{iv} \frac{\beta_{hv} S_v \frac{I_h}{N_h}}{I_v} \tag{B.3a}$$

$$\frac{dp_{ih}}{ds} = p_{iv} \frac{\beta_{hv} S_v \frac{I_h}{N_h}}{I_v} - p_{ih} \frac{\beta_{vh} \frac{S_h}{N_h} I_v}{I_h}. \tag{B.3b}$$

From these master equations, we can see that the rate at which probability mass flows between states depends on the rates at which lineages move between states

through transmission events.

Here, we assume that the lineage state probabilities are at equilibrium with respect to the overall epidemiological dynamics. This is a reasonable assumption as long as the lineages move between states much faster than the overall epidemiological dynamics change. With this assumption, we can then solve for the equilibrium probabilities p_{iv}^* and p_{ih}^* using equation (B.3). To do so, we set $\frac{dp_{iv}}{ds} = 0$ and substitute in $1 - p_{iv}$ for p_{ih}. Solving, p_{iv}^* becomes:

$$p_{iv}^* = \frac{\frac{\beta_{vh}\frac{S_h}{N_h}I_v}{I_h}}{(\frac{\beta_{vh}\frac{S_h}{N_h}I_v}{I_h} + \frac{\beta_{hv}S_v\frac{I_h}{N_h}}{I_v})}, \tag{B.4}$$

and

$$p_{ih}^* = 1 - p_{iv}^* = \frac{\frac{\beta_{hv}S_v\frac{I_h}{N_h}}{I_v}}{(\frac{\beta_{vh}\frac{S_h}{N_h}I_v}{I_h} + \frac{\beta_{hv}S_v\frac{I_h}{N_h}}{I_v})}. \tag{B.5}$$

Plugging these equilibrium lineage state probabilities into (B.2), the pairwise rate of coalescence becomes:

$$\lambda_{ij} = \frac{\beta_{vh}\frac{S_h}{N_h}I_v + \beta_{hv}S_v\frac{I_h}{N_h}}{I_vI_h}\left(\frac{2\frac{\beta_{vh}\frac{S_h}{N_h}I_v}{I_h}\frac{\beta_{hv}S_v\frac{I_h}{N_h}}{I_v}}{\left(\frac{\beta_{vh}\frac{S_h}{N_h}I_v}{I_h} + \frac{\beta_{hv}S_v\frac{I_h}{N_h}}{I_v}\right)^2}\right). \tag{B.6}$$

We can make sense of this coalescent rate by decomposing it into two parts. One part, the term in parentheses on the right hand side, gives the overall probability of the two lineages being in different states: one in an infected vector and one in an infected human. Intuitively, this term enters into the coalescent rate because coalescent events can only occur at transmission events, which requires the two lineages to be in opposite states; one in an infected vector and the other in an infected human. Because of this requirement, we can see that the rate of coalescence will generally be

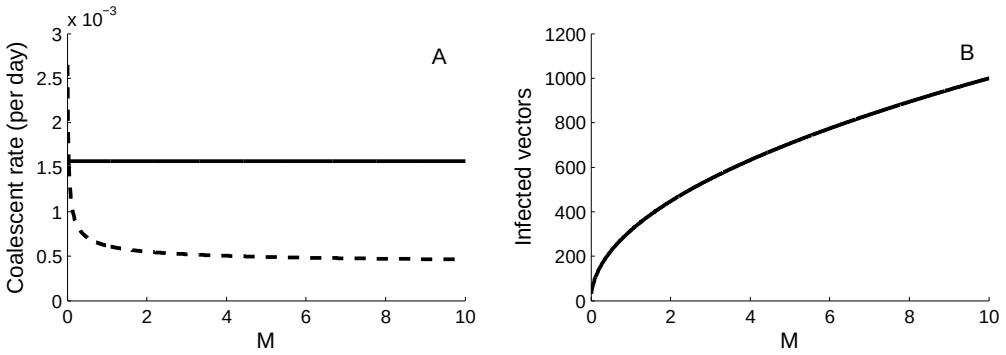

FIGURE B.1: (A) Comparison of coalescent rates at equilibrium for the direct transmission model (solid) and for the vector-borne model (dashed) over a range of M, the ratio of the vector population size to the human population size. For the direct transmission model, the coalescent rate does not depend on M. (B) The number of infected mosquitoes at equilibrium under different values of M. The number of infected humans remains constant regardless of M because we are holding R_0 constant.

lower for a vector-borne pathogen than a directly transmitted pathogen. The probability that two lineages are in opposite states reaches a maximum when $p_{iv} = p_{ih} = \frac{1}{2}$, which means that the highest attainable probability of the lineages being in opposite states is also $\frac{1}{2}$. All else being equal then, the rate of coalescence for a vector-borne pathogen will be at most half that of a directly transmitted pathogen.

The other part of the coalescent rate, the leading term on the right hand side of (B.6), gives the rate at which two lineages coalesce conditional on one lineage being in a vector and the other in a human. We can see that the coalescent rate inversely depends on the product of the number of infected vectors and humans, as this gives the probability that of all the lineages circulating in the population, the pair of lineages that we are considering are the two lineages that coalesce at a given transmission event. Because the term $I_v I_h$ tends to dominate the overall rate of coalescence, the number of infected humans and vectors plays a very important role in determining the overall coalescent rate. The number of infected humans and vectors in turn depends on a key parameter M, which we define as the ratio of the

vector population size N_v to the human population size N_h.

To understand how the vector population size N_v affects the rate of coalescence, we can hold R_0 constant, so that the number of infected humans remains the same at equilibrium, but vary the ratio of vector to human population sizes M (note that we decrease the transmission rates β_{vh} and β_{hv} as we increase M to keep R_0 constant). At equilibrium, the rate of coalescence for the vector-borne pathogen drops off asymptotically with increasing M relative to a directly transmitted pathogen (Fig. B.1A). This is because the number of infected vectors also increases with M (Fig. B.1B), resulting in a larger product $I_v I_h$ in the denominator of (B.6), and consequently a lower coalescent rate.

Bibliography

Amore, G., Bertolotti, L., Hamer, G. L., Kitron, U. D., Walker, E. D., Ruiz, M. O., Brawn, J. D., and Goldberg, T. L. (2010), "Multi-year evolutionary dynamics of West Nile virus in suburban Chicago, USA, 2005–2007," *Philos Trans R Soc Lond B Biol Sci*, 365, 1871–1878.

Anders, K. L., Nguyet, N. M., Chau, N. V. V., Hung, N. T., Thuy, T. T., Farrar, J., Wills, B., Hien, T. T., Simmons, C. P., et al. (2011), "Epidemiological factors associated with dengue shock syndrome and mortality in hospitalized dengue patients in Ho Chi Minh City, Vietnam," *Am J Trop Med Hyg*, 84, 127–134.

Anderson, R. M. and May, R. M. (1991), *Infectious Diseases of Humans: Dynamics and Control*, Oxford University Press, Oxford, U.K.

Anderson, R. M., May, R. M., et al. (1979), "Population biology of infectious diseases: Part I," *Nature*, 280, 361–367.

Andreasen, V., Lin, J., and Levin, S. A. (1997), "The dynamics of cocirculating influenza strains conferring partial cross-immunity," *J Math Biol*, 35, 825–842.

Andrieu, C. and Roberts, G. O. (2009), "The pseudo-marginal approach for efficient Monte Carlo computations," *The Annals of Statistics*, 37, 697–725.

Andrieu, C., Doucet, A., and Holenstein, R. (2010), "Particle Markov chain Monte Carlo methods." *Journal of the Royal Statistical Society: Series B*, 72, 269–342.

Bahl, J., Nelson, M. I., Chan, K. H., Chen, R., Vijaykrishna, D., Halpin, R. A., Stockwell, T. B., Lin, X., Wentworth, D. E., Ghedin, E., et al. (2011), "Temporally structured metapopulation dynamics and persistence of influenza A H3N2 virus in humans," *Proc Natl Acad Sci USA*, 108, 19359–19364.

Bansal, S., Grenfell, B. T., and Meyers, L. A. (2007), "When individual behaviour matters: homogeneous and network models in epidemiology," *J R Soc Interface*, 4, 879–891.

Beaumont, M. A. (2003), "Estimation of population growth or decline in genetically monitored populations," *Genetics*, 164, 1139–1160.

Beaumont, M. A. (2010), "Approximate Bayesian computation in evolution and ecology," *Annual Review of Ecology, Evolution, and Systematics*, 41, 379–406.

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

Bedford, T., Cobey, S., Beerli, P., and Pascual, M. (2010), "Global migration dynamics underlie evolution and persistence of human influenza A (H3N2)," *PLoS Pathog*, 6, e1000918.

Beerli, P. and Felsenstein, J. (1999), "Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach," *Genetics*, 152, 763–73.

Beerli, P. and Felsenstein, J. (2001), "Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach," *Proc Natl Acad Sci USA*, 98, 4563–8.

Bennett, S., Drummond, A., Kapan, D., Suchard, M., Munoz-Jordan, J., Pybus, O., Holmes, E., and Gubler, D. (2010), "Epidemic dynamics revealed in dengue evolution," *Mol Biol Evol*, 27, 811–818.

Biek, R., Henderson, J. C., Waller, L. A., Rupprecht, C. E., and Real, L. A. (2007), "A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus," *Proc Natl Acad Sci USA*, 104, 7993–7998.

Bolker, B. and Grenfell, B. (1995), "Space, persistence and dynamics of measles epidemics," *Philos Trans R Soc Lond B Biol Sci*, 348, 309–320.

Bretó, C., He, D., Ionides, E., and King, A. (2009), "Time series analysis via mechanistic models," *The Annals of Applied Statistics*, 3, 319–348.

Brown, A. J. L., Lycett, S. J., Weinert, L., Hughes, G. J., Fearnhill, E., and Dunn, D. T. (2011), "Transmission network parameters estimated from HIV sequences for a nationwide epidemic," *Journal of Infectious Diseases*, 204, 1463–1469.

Cappe, O., Godsill, S., and Moulines, E. (2007), "An overview of existing methods and recent advances in sequential Monte Carlo." *Proc Inst Electr Elect*, 95, 899–924.

Carrington, C., Foster, J., Pybus, O., Bennett, S., and Holmes, E. (2005), "Invasion and maintenance of dengue virus type 2 and type 4 in the Americas," *J Virol*, 79, 14680–14687.

Cauchemez, S. and Ferguson, N. M. (2008), "Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in London," *J R Soc Interface*, 5, 885–897.

Charlesworth, B. (2009), "Effective population size and patterns of molecular evolution and variation," *Nature Reviews Genetics*, 10, 195–205.

Chopin, N. (2004), "Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference," *The Annals of Statistics*, 32, 2385–2411.

Cohen, M. S., Dye, C., Fraser, C., Miller, W. C., Powers, K. A., and Williams, B. G. (2012), "HIV treatment as prevention: debate and commentary—will early infection compromise treatment-as-prevention strategies?" *PLoS medicine*, 9, e1001232.

Coudeville, L. and Garnett, G. P. (2012), "Transmission dynamics of the four dengue serotypes in southern Vietnam and the potential impact of vaccination," *PLoS One*, 7, e51244.

Coulson, T., Rohani, P., and Pascual, M. (2004), "Skeletons, noise and population growth: the end of an old debate?" *Trends Ecol Evol*, 19, 359–364.

Cuong, H., Vu, T., Cazelles, B., Boni, M., Thai, K., Rabaa, M., Quang, L., Simmons, C., Huu, T., and Anders, K. (2013), "Spatiotemporal dynamics of dengue epidemics, southern Vietnam," *Emerg Infect Dis*, 19, 945–953.

de Silva, E., Ferguson, N. M., and Fraser, C. (2012), "Inferring pandemic growth rates from sequence data," *J R Soc Interface*, 9, 1797–1808.

Dearlove, B. and Wilson, D. J. (2013), "Coalescent inference for infectious disease: meta-analysis of hepatitis C," *Philos Trans R Soc Lond B Biol Sci*, 368.

Donnelly, P. and Tavaré, S. (1995), "Coalescents and genealogical structure under neutrality," *Annu Rev Genet*, 29, 401–21.

Doucet, A. and Johansen, A. (2009), "A tutorial on particle filtering and smoothing: Fifteen years later," *Handbook of Nonlinear Filtering*, pp. 656–704.

Doucet, A., De Freitas, N., Gordon, N., et al. (2001), *Sequential Monte Carlo methods in practice*, Springer New York.

Drummond, A. J. and Rambaut, A. (2007), "BEAST: Bayesian evolutionary analysis by sampling trees," *BMC Evol Biol*, 7, 214.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., and Solomon, W. (2002), "Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data," *Genetics*, 161, 1307–1320.

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005), "Bayesian coalescent inference of past population dynamics from molecular sequences," *Mol Biol Evol*, 22, 1185–92.

Duke-Sylvester, S. M., Biek, R., and Real, L. A. (2013), "Molecular evolutionary signatures reveal the role of host ecological dynamics in viral disease emergence and spread," *Philos Trans R Soc Lond B Biol Sci*, 368.

Earn, D. J., Rohani, P., Bolker, B. M., and Grenfell, B. T. (2000), "A simple model for complex dynamical transitions in epidemics," *Science*, 287, 667–670.

Ferguson, N. M., Donnelly, C. A., and Anderson, R. M. (1999), "Transmission dynamics and epidemiology of dengue: insights from age–stratified sero–prevalence surveys," *Philos Trans R Soc Lond B Biol Sci*, 354, 757–768.

Finkenstädt, B. F. and Grenfell, B. T. (2000), "Time series modelling of childhood diseases: a dynamical systems approach," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49, 187–205.

Focks, D. A. and Chadee, D. D. (1997), "Pupal survey: an epidemiologically significant surveillance method for Aedes aegypti: an example using data from Trinidad." *Am J Trop Med Hyg*, 56, 159–167.

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, 1557–1561.

Frost, S. and Volz, E. (2010), "Viral phylodynamics and the search for an 'effective number of infections'," *Philos Trans R Soc Lond B Biol Sci*, 365, 1879–90.

General Statistics Office of Vietnam (2008), "Average population by province," .

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996), *Markov chain Monte Carlo in practice*, CRC Press.

Gillespie, D. T. (2007), "Stochastic simulation of chemical kinetics," *Annu Rev Phys Chem*, 58, 35–55.

Gordon, N. J., Salmond, D. J., and Smith, A. F. (1993), "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *IEE Proceedings F (Radar and Signal Processing)*, 140, 107–113.

Granich, R. M., Gilks, C. F., Dye, C., De Cock, K. M., and Williams, B. G. (2009), "Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: a mathematical model," *The Lancet*, 373, 48–57.

Gray, R. R., Tatem, A. J., Lamers, S., Hou, W., Laeyendecker, O., Serwadda, D., Sewankambo, N., Gray, R. H., Wawer, M., Quinn, T. C., et al. (2009), "Spatial phylodynamics of HIV-1 epidemic emergence in east Africa," *AIDS (London, England)*, 23, F9.

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L. N., Daly, J. M., Mumford, J. A., and Holmes, E. C. (2004), "Unifying the epidemiological and evolutionary dynamics of pathogens," *Science*, 303, 327–32.

Griffiths, R. C. and Tavare, S. (1994), "Sampling theory for neutral alleles in a varying environment," *Philos Trans R Soc Lond B Biol Sci*, 344, 403–410.

Gubler, D., Suharyono, W., Tan, R., Abidin, M., and Sie, A. (1981), "Viraemia in patients with naturally acquired dengue infection," *Bulletin of the World Health Organization*, 59, 623.

Harrington, L. C., Buonaccorsi, J. P., Edman, J. D., Costero, A., Kittayapong, P., Clark, G. G., and Scott, T. W. (2001), "Analysis of survival of young and old Aedes aegypti (Diptera: Culicidae) from Puerto Rico and Thailand," *J Med Entomol*, 38, 537–547.

He, D., Ionides, E. L., and King, A. A. (2010), "Plug-and-play inference for disease dynamics: measles in large and small populations as a case study," *J R Soc Interface*, 7, 271–283.

Heller, R., Chikhi, L., and Siegismund, H. R. (2013), "The confounding effect of population structure on Bayesian skyline plot inferences of demographic history," *PloS One*, 8, e62992.

Hollingsworth, T. D., Anderson, R. M., and Fraser, C. (2008), "HIV-1 transmission, by stage of infection," *Journal of Infectious Diseases*, 198, 687–693.

Holmes, E. and Grenfell, B. (2009), "Discovering the phylodynamics of RNA viruses," *PLoS Comput Biol*, 5, e1000505.

Holmes, E. C., Nee, S., Rambaut, A., Garnett, G. P., and Harvey, P. H. (1995), "Revealing the history of infectious disease epidemics through phylogenetic trees," *Philos Trans R Soc Lond B Biol Sci*, 349, 33–40.

Hudson, R. R. et al. (1990), "Gene genealogies and the coalescent process," *Oxford Surveys in Evolutionary Biology*, 7, 44.

Ionides, E. L., Bretó, C., and King, A. A. (2006), "Inference for nonlinear dynamical systems," *Proc Natl Acad Sci USA*, 103, 18438–43.

Jeffery, J. A., Yen, N. T., Nam, V. S., Hoffmann, A. A., Kay, B. H., Ryan, P. A., et al. (2009), "Characterizing the Aedes aegypti population in a Vietnamese village in preparation for a Wolbachia-based mosquito control strategy to eliminate dengue," *PLoS Negl Trop Dis*, 3, e552.

Kass, R. E. and Raftery, A. E. (1995), "Bayes factors," *J Am Stat Assoc*, 90, 773–795.

Keeling, M. and Rohani, P. (2002), "Estimating spatial coupling in epidemiological systems: a mechanistic approach," *Ecology Letters*, 5, 20–29.

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

Keeling, M. J. and Eames, K. T. (2005), "Networks and epidemic models," *J R Soc Interface*, 2, 295–307.

Kendall, D. G. (1948), "On the generalized birth-and-death process," *The Annals of Mathematical Statistics*, pp. 1–15.

King, A. A., Ionides, E. L., Pascual, M., and Bouma, M. J. (2008), "Inapparent infections and cholera dynamics," *Nature*, 454, 877–880.

Kingman, J. (1982), "On the Genealogy of Large Populations," *Journal of Applied Probability*, 19, 27–43.

Koelle, K. and Rasmussen, D. (2012), "Rates of coalescence for common epidemiological models at equilibrium," *J R Soc Interface*, 9, 997–1007.

Koelle, K., Cobey, S., Grenfell, B., and Pascual, M. (2006), "Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans," *Science*, 314, 1898–1903.

Koenraadt, C. J., Aldstadt, J., Kijchalao, U., Sithiprasasna, R., Getis, A., Jones, J. W., and Scott, T. W. (2008), "Spatial and temporal patterns in pupal and adult production of the dengue vector Aedes aegypti in Kamphaeng Phet, Thailand," *Am J Trop Med Hyg*, 79, 230–238.

Kretzschmar, M. E., van der Loeff, M. F. S., Birrell, P. J., De Angelis, D., and Coutinho, R. A. (2013), "Prospects of elimination of HIV with test-and-treat strategy," *Proc Natl Acad Sci USA*, 110, 15538–15543.

Kuhner, M. (2006), "LAMARC 2.0: Maximum likelihood and Bayesian estimation of population parameters," *Bioinformatics*, 22, 768–70.

Kuhner, M., Yamato, J., and Felsenstein, J. (1998), "Maximum likelihood estimation of population growth rates based on the coalescent," *Genetics*, 149, 429–34.

Laporte, V. and Charlesworth, B. (2002), "Effective population size and population subdivision in demographically structured populations," *Genetics*, 162, 501–519.

Lartillot, N. and Philippe, H. (2006), "Computing Bayes factors using thermodynamic integration," *Systematic Biol*, 55, 195–207.

Lemey, P., Pybus, O. G., Wang, B., Saksena, N. K., Salemi, M., and Vandamme, A.-M. (2003), "Tracing the origin and history of the HIV-2 epidemic," *Proc Natl Acad Sci USA*, 100, 6588–6592.

Lemey, P., Pybus, O. G., Rambaut, A., Drummond, A. J., Robertson, D. L., Roques, P., Worobey, M., and Vandamme, A.-M. (2004), "The molecular population genetics of HIV-1 group O," *Genetics*, 167, 1059–1068.

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

Lemey, P., Rambaut, A., Welch, J. J., and Suchard, M. A. (2010), "Phylogeography takes a relaxed random walk in continuous space and time," *Mol Biol Evol*, 27, 1877–1885.

Leventhal, G. E., Kouyos, R., Stadler, T., Von Wyl, V., Yerly, S., Böni, J., Cellerai, C., Klimkait, T., Günthard, H. F., and Bonhoeffer, S. (2012), "Inferring epidemic contact structure from phylogenetic trees," *PLoS Comput Biol*, 8, e1002413.

Leventhal, G. E., Günthard, H. F., Bonhoeffer, S., and Stadler, T. (2014), "Using an Epidemiological Model for Phylogenetic Inference Reveals Density Dependence in HIV Transmission," *Mol Biol Evol*, 31, 6–17.

Lewis, F., Hughes, G. J., Rambaut, A., Pozniak, A., and Leigh Brown, A. J. (2008), "Episodic sexual transmission of HIV revealed by molecular phylodynamics," *PLoS Medicine*, 5.

Lin, J.-H., Chiu, S.-C., Lin, Y.-C., Cheng, J.-C., Wu, H.-S., Salemi, M., and Liu, H.-F. (2013), "Exploring the Molecular Epidemiology and Evolutionary Dynamics of Influenza A Virus in Taiwan," *PLoS One*, 8, e61957.

Magiorkinis, G., Sypsa, V., Magiorkinis, E., Paraskevis, D., Katsoulidou, A., Belshaw, R., Fraser, C., Pybus, O. G., and Hatzakis, A. (2013), "Integrating Phylodynamics and Epidemiology to Estimate Transmission Diversity in Viral Epidemics," *PLoS Comput Biol*, 9, e1002876.

Maslow, J. N., Mulligan, M. E., and Arbeit, R. D. (1993), "Molecular epidemiology: application of contemporary techniques to the typing of microorganisms," *Clinical Infectious Diseases*, 17, 153–162.

McDonald, P. (1977), "Population characteristics of domestic Aedes aegypti (Diptera: Culicidae) in villages on the Kenya Coast I. Adult survivorship and population size," *J Med Entomol*, 14, 42–48.

McElroy, K., Santiago, G., Lennon, N., Birren, B., Henn, M., and Muñoz-Jordán, J. (2011), "Endurance, refuge, and reemergence of dengue virus type 2, Puerto Rico, 1986–2007," *Emerg Infect Dis*, 17, 64.

Minin, V. N., Bloomquist, E. W., and Suchard, M. A. (2008), "Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics," *Mol Biol Evol*, 25, 1459–1471.

Morens, D. M., Folkers, G. K., and Fauci, A. S. (2004), "The challenge of emerging and re-emerging infectious diseases," *Nature*, 430, 242–249.

Morrison, A. C., Gray, K., Getis, A., Astete, H., Sihuincha, M., Focks, D., Watts, D., Stancil, J. D., Olson, J. G., Blair, P., et al. (2004), "Temporal and geographic patterns of Aedes aegypti (Diptera: Culicidae) production in Iquitos, Peru," *J Med Entomol*, 41, 1123–1142.

Nee, S., Holmes, E. C., Rambaut, A., and Harvey, P. H. (1995), "Inferring population history from molecular phylogenies," *Philos Trans R Soc Lond B Biol Sci*, 349, 25–31.

Neher, R. A. and Hallatschek, O. (2013), "Genealogies of rapidly adapting populations," *Proc Natl Acad Sci USA*, 110, 437–442.

Nelson, M. I., Simonsen, L., Viboud, C., Miller, M. A., and Holmes, E. C. (2007), "Phylogenetic analysis reveals the global migration of seasonal influenza A viruses," *PLoS Pathog*, 3, e131.

Notohara, M. (1990), "The coalescent and the genealogical process in geographically structured population," *J Math Biol*, 29, 59–75.

O'Dea, E. B. and Wilke, C. O. (2010), "Contact heterogeneity and phylodynamics: how contact networks shape parasite evolutionary trees," *Interdisciplinary Perspectives on Infectious Diseases*, 2011.

O'Fallon, B. D., Seger, J., and Adler, F. R. (2010), "A continuous-state coalescent and the impact of weak selection on the structure of gene genealogies," *Mol Biol Evol*, 27, 1162–1172.

O'Neill, P. D. (2002), "A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods," *Math Biosci*, 180, 103–114.

O'Neill, P. D. (2010), "Introduction and snapshot review: Relating infectious disease transmission models to data," *Statistics in Medicine*, 29, 2069–2077.

O'Neill, P. D. and Roberts, G. O. (1999), "Bayesian inference for partially observed stochastic epidemics," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162, 121–129.

Pascual, M., Rodó, X., Ellner, S. P., Colwell, R., and Bouma, M. J. (2000), "Cholera dynamics and El Nino-southern oscillation," *Science*, 289, 1766–1769.

Pilcher, C. D., Tien, H. C., Eron, J. J., Vernazza, P. L., Leu, S.-Y., Stewart, P. W., Goh, L.-E., and Cohen, M. S. (2004), "Brief but efficient: acute HIV infection and the sexual transmission of HIV," *Journal of Infectious Diseases*, 189, 1785–1792.

Powers, K. A., Ghani, A. C., Miller, W. C., Hoffman, I. F., Pettifor, A. E., Kamanga, G., Martinson, F. E., and Cohen, M. S. (2011), "The role of acute and early HIV infection in the spread of HIV and implications for transmission prevention strategies in Lilongwe, Malawi: a modelling study," *The Lancet*, 378, 256–268.

Pybus, O., Rambaut, A., and Harvey, P. (2000), "An integrated framework for the inference of viral population history from reconstructed genealogies," *Genetics*, 155, 1429–37.

Pybus, O. G. and Rambaut, A. (2009), "Evolutionary analysis of the dynamics of viral infectious disease," *Nature Reviews Genetics*, 10, 540–550.

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

Pybus, O. G., Suchard, M. A., Lemey, P., Bernardin, F. J., Rambaut, A., Crawford, F. W., Gray, R. R., Arinaminpathy, N., Stramer, S. L., Busch, M. P., and Delwart, E. L. (2012), "Unifying the spatial epidemiology and molecular evolution of emerging epidemics," *Proc Natl Acad Sci USA*, 109, 15066–71.

Rabaa, M., Ty Hang, V., Wills, B., Farrar, J., Simmons, C., and Holmes, E. (2010), "Phylogeography of recently emerged DENV-2 in southern Viet Nam," *PLoS Negl Trop Dis*, 4, e766.

Raghwani, J., Rambaut, A., Holmes, E., Hang, V., Hien, T., Farrar, J., Wills, B., Lennon, N., Birren, B., Henn, M., and Simmons, C. (2011), "Endemic dengue associated with the co-circulation of multiple viral lineages and localized density-dependent transmission," *PLoS Pathog*, 7, e1002064.

Rambaut, A., Pybus, O. G., Nelson, M. I., Viboud, C., Taubenberger, J. K., and Holmes, E. C. (2008), "The genomic and epidemiological dynamics of human influenza A virus," *Nature*, 453, 615–619.

Rasmussen, D., Ratmann, O., and Koelle, K. (2011), "Inference for nonlinear epidemiological models using genealogies and time series," *PLoS Comput Biol*, 7.

Rasmussen, D., Volz, E., and Koelle, K. (2014a), "Phylodynamic inference for structured epidemiological models," *PLoS Comput Biol*, 10(4).

Rasmussen, D., Boni, M., and Koelle, K. (2014b), "Reconciling epidemiology with phylogeny: The case of dengue virus in southern Vietnam," *Mol Biol Evol*, 31, 258–271.

Robinson, K., Fyson, N., Cohen, T., Fraser, C., and Colijn, C. (2013), "How the Dynamics and Structure of Sexual Contact Networks Shape Pathogen Phylogenies," *PLoS Comput Biol*, 9, e1003105.

Rodrigo, A. G. and Felsenstein, J. (1999), *Coalescent approaches to HIV population genetics*, pp. 233–272, Johns Hopkins Univ. Press, Baltimore, MD.

Rohani, P. and King, A. A. (2010), "Never mind the length, feel the quality: the impact of long-term epidemiological data sets on theory, application and policy," *Trends Ecol Evol*, 25, 611–618.

Rohani, P., Earn, D. J., Finkenstädt, B., and Grenfell, B. T. (1998), "Population dynamic interference among childhood diseases," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265, 2033–2041.

Rohani, P., Keeling, M. J., and Grenfell, B. T. (2002), "The interplay between determinism and stochasticity in childhood diseases," *The American Naturalist*, 159, 469–481.

Schreiber, M., Holmes, E., Ong, S., Soh, H., Liu, W., Tanner, L., Aw, P., Tan, H., Ng, L., Leo, Y., et al. (2009), "Genomic epidemiology of a dengue virus epidemic in urban Singapore," *J Virol*, 83, 4163–4173.

Scott, T. W. and Morrison, A. C. (2010), "Vector dynamics and transmission of dengue virus: implications for dengue surveillance and prevention strategies," in *Dengue Virus*, ed. A. L. Rothman, vol. 338 of *Current Topics in Microbiology and Immunology*, pp. 115–128, Springer.

Sheppard, P., Macdonald, W., Tonn, R., and Grab, B. (1969), "The dynamics of an adult population of Aedes aegypti in relation to dengue haemorrhagic fever in Bangkok," *Journal of Animal Ecology*, pp. 661–702.

Siebenga, J. J., Lemey, P., Pond, S. L. K., Rambaut, A., Vennema, H., and Koopmans, M. (2010), "Phylodynamic reconstruction reveals norovirus GII. 4 epidemic expansions and their molecular determinants," *PLoS Pathog*, 6, e1000884.

Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007), "Sequential monte carlo without likelihoods," *Proceedings of the National Academy of Sciences*, 104, 1760–1765.

Slatkin, M. and Hudson, R. R. (1991), "Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations." *Genetics*, 129, 555–562.

Stack, J. C., Welch, J. D., Ferrari, M. J., Shapiro, B. U., and Grenfell, B. T. (2010), "Protocols for sampling viral sequences to study epidemic dynamics," *J R Soc Interface*, 7, 1119–1127.

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

Stadler, T. and Bonhoeffer, S. (2013), "Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods," *Philos Trans R Soc Lond B Biol Sci*, 368.

Stadler, T., Kouyos, R., von Wyl, V., Yerly, S., Böni, J., Bürgisser, P., Klimkait, T., Joos, B., Rieder, P., Xie, D., et al. (2012), "Estimating the basic reproductive number from viral sequence data," *Mol Biol Evol*, 29, 347–357.

Strimmer, K. and Pybus, O. G. (2001), "Exploring the demographic history of DNA sequences using the generalized skyline plot," *Mol Biol Evol*, 18, 2298–305.

Takahata, N. and Slatkin, M. (1990), "Genealogy of neutral genes in two partially isolated populations," *Theor Popul Biol*, 38, 331–50.

Tavare, S., Balding, D. J., Griffiths, R., and Donnelly, P. (1997), "Inferring coalescence times from DNA sequence data," *Genetics*, 145, 505–518.

Thai, K., Binh, T., Giao, P., Phuong, H., Hung, L., Van Nam, N., Nga, T., Groen, J., Nagelkerke, N., and de Vries, P. (2005), "Seroprevalence of dengue antibodies, annual incidence and risk factors among children in southern Vietnam," *Trop Med Int Health*, 10, 379–86.

Thai, K., Nishiura, H., Hoang, P., Tran, N., Phan, G., Le, H., Tran, B., Nguyen, N., and de Vries, P. (2011), "Age-specificity of clinical dengue during primary and secondary infections," *PLoS Negl Trop Dis*, 5, e1180.

Tricou, V., Minh, N. N., Farrar, J., Tran, H. T., and Simmons, C. P. (2011), "Kinetics of viremia and NS1 antigenemia are shaped by immune status and virus serotype in adults with dengue," *PLoS Negl Trop Dis*, 5, e1309.

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

Volz, E. M., Kosakovsky Pond, S. L., Ward, M. J., Leigh Brown, A. J., and Frost, S. D. W. (2009), "Phylodynamics of infectious disease epidemics," *Genetics*, 183, 1421–30.

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

Volz, E. M., Ionides, E., Romero-Severson, E. O., Brandt, M.-G., Mokotoff, E., and Koopman, J. S. (2013a), "HIV-1 Transmission during Early Infection in Men Who Have Sex with Men: A Phylodynamic Analysis," *PLoS Medicine*, 10, e1001568.

Volz, E. M., Koelle, K., and Bedford, T. (2013b), "Viral phylodynamics," *PLoS Comput Biol*, 9, e1002947.

Vu, T., Holmes, E., Duong, V., Nguyen, T., Tran, T., Quail, M., Churcher, C., Parkhill, J., Cardosa, J., Farrar, J., Wills, B., Lennon, N., Birren, B., Buchy, P., Henn, M., and Simmons, C. (2010), "Emergence of the Asian 1 genotype of dengue virus serotype 2 in Viet Nam: in vivo fitness advantage and lineage replacement in South-East Asia," *PLoS Negl Trop Dis*, 4, e757.

Wakeley, J. (2009), *Coalescent Theory: An Introduction*, Roberts and Company Publishers, Greenwood Village, Colo.

Walczak, A. M., Nicolaisen, L. E., Plotkin, J. B., and Desai, M. M. (2012), "The structure of genealogies in the presence of purifying selection: A fitness-class coalescent," *Genetics*, 190, 753–779.

Williams, B. G., Lloyd-Smith, J. O., Gouws, E., Hankins, C., Getz, W. M., Hargrove, J., De Zoysa, I., Dye, C., and Auvert, B. (2006), "The potential impact of male circumcision on HIV in sub-Saharan Africa," *PLoS Medicine*, 3, e262.

Woolhouse, M. E. (2002), "Population biology of emerging and re-emerging pathogens," *Trends in Microbiology*, 10, s3–s7.

Wright, S. (1943), "Isolation by distance," *Genetics*, 28, 114.

Yan, P., Zhang, F., and Wand, H. (2011), "Using HIV diagnostic data to estimate HIV incidence: method and simulation," *Statistical Communications in Infectious Diseases*, 3.

Yang, Z. (1998), "On the best evolutionary rate for phylogenetic analysis," *Syst Biol*, 47, 125–33.

Ypma, R. J., van Ballegooijen, W. M., and Wallinga, J. (2013), "Relating Phylogenetic Trees to Transmission Trees of Infectious Disease Outbreaks," *Genetics*, 195, 1055–1062.

Zanotto, P. d., Gould, E. A., Gao, G. F., Harvey, P. H., and Holmes, E. C. (1996), "Population dynamics of flaviviruses revealed by molecular phylogenies," *Proc Natl Acad Sci USA*, 93, 548–553.

Biography

David Alan Rasmussen was born March 19, 1985 in Cleveland, Ohio. He graduated from Midpark High School in 2003. He then studied at Reed College, where he received his B.A. in Biology in 2007. After college, he attended graduate school at Duke University, where he was a National Science Foundation graduate research fellow and received his Ph.D. in Biology in 2014. He plans to continue working on the ecology and evolution of infectious diseases as a postdoctoral fellow at ETH Zurich in Switzerland.