Inference of Infectious Disease Dynamics from Genetic Data via Sequential Monte Carlo

by

R. Alexander Smith

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2018

Doctoral Committee:

Professor Edward L. Ionides, Co-Chair
Professor Aaron A. King, Co-Chair
Professor Stephen A. Smith
Professor Evan Snitkin
Professor Jianzhi George Zhang

"We shall not cease from exploration,
and the end of all our exploring
will be to arrive where we started
and know the place for the first time."

- T.S. Eliot

Richard Alexander Smith
alxsmth@umich.edu
ORCID iD:0000-0001-7197-517X

ACKNOWLEDGEMENTS

I would like to thank my advisors, Aaron King and Ed Ionides. This work would not have been possible without their teaching, guidance, and support. I would also like to thank the members of my dissertation committee: Jianzhi George Zhang, Stephen Smith, and Evan Snitkin. Their advice and insights have greatly improved this work. I would like to thank my parents for their support and encouragement over many years. Finally, I would like to thank Felicia Magpantay, whose careful thinking as well as her love and support have helped me at many points along the way.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii

LIST OF FIGURES vi

LIST OF TABLES viii

LIST OF APPENDICES ix

ABSTRACT x

CHAPTER

I. Introduction 1

II. An argument for the importance of joint inference of phylogeny and transmission 7

2.1 Abstract 7
2.2 Introduction 7
2.3 Methods 8
2.3.1 A seasonally-forced SIR model 8
2.3.2 Simulation 10
2.3.3 Inference 13
2.4 Results 15
2.5 Discussion 16
2.5.1 The consequences of two-stage inference 16
2.5.2 The influence of sampling regime on inference 17
2.5.3 The Rasmussen Criterion 17
2.5.4 The role of relaxed clock models 18
2.5.5 Joint inference 19
2.6 Future Directions 19

III. A flexible method for joint phylodynamic inference 23

3.1 Abstract . . . 23
3.2 Introduction . . . 24
3.3 New Approaches . . . 25
3.3.1 The latent process . . . 26
3.3.2 The observable process . . . 27
3.3.3 Relaxed molecular clocks . . . 28
3.3.4 The plug and play property . . . 28
3.4 A model of HIV transmission . . . 29
3.4.1 The latent and observable processes . . . 29
3.5 Results . . . 31
3.5.1 A study on simulated data . . . 31
3.5.2 Analysis of an HIV subepidemic in Detroit, MI . . . 39
3.6 Discussion . . . 47
3.7 Materials and methods . . . 50
3.7.1 Overview of sequential Monte Carlo estimation of the likelihood . . . 50
3.7.2 Maximization of the likelihood via iterated filtering . . . 54
3.7.3 Computational Structure . . . 54
3.7.4 A model of HIV transmission: computation of the measurement model . . . 56
3.7.5 Data analysis methods: the sequence data . . . 57
3.7.6 Data analysis methods: selecting a subepidemic . . . 57
3.8 Supplementary Material . . . 62
3.9 Acknowledgments . . . 62

IV. Inferring transmission in a simulated hospital outbreak . . . 63

4.1 Abstract . . . 63
4.2 Introduction . . . 63
4.3 Methods . . . 65
4.3.1 A model of transmission on a hospital with two wards . . . 65
4.3.2 Simulation . . . 67
4.3.3 Inference . . . 69
4.4 Results . . . 70
4.5 Discussion . . . 72
4.6 Future Directions . . . 72
4.6.1 Fitting a genetically-defined cluster . . . 72
4.6.2 A more complicated model of VRE transmission . . . 73
4.6.3 Testing the feasibility of inference . . . 74

V. Conclusion . . . 78

5.1 Other ideas . . . 80
5.2 Final remarks . . . 81

APPENDICES 83
A.1 The GenPOMP model: linking infectious disease dynamics with genetic data 84
A.2 A GenSMC algorithm for filtering the GenPOMP model 91
A.2.1 The implementation of GenSMC in the `genPomp` program 98
A.2.2 Extending GenSMC to infer unknown parameters: The GenIF algorithm 98
A.2.3 Scalability of GenSMC 99
A.3 A theoretical derivation of the GenSMC algorithm 100
A.3.1 A basic SMC algorithm 100
A.3.2 A targeted SMC approach with a partial plug-and-play property 103
A.3.3 SMC with hierarchical sampling 105
A.3.4 Just-in-time evaluation of some state variable components 106
A.3.5 Moving from discrete time to continuous time 107
A.4 Details of the HIV model used in the main text 108
A.4.1 Initial values for the HIV model 109
A.4.2 Algorithmic parameters used for the numerical results . . 112
B.1 Background 114
B.1.1 The untargeted proposal 115
B.1.2 A targeted proposal 116
B.1.3 The correction term 118
B.2 A simple test of the targeted proposal 118

BIBLIOGRAPHY 121

LIST OF FIGURES

Figure

2.1	A schematic of an SIR model	10
2.2	Components of an SIR epidemic, simulated and estimated	20
2.3	Likelihood profiles for β_L and β_H derived from using the true phylogeny and the estimated phylogeny.	21
2.4	A comparison of the quality of inference across different scenarios	22
3.1	The latent transmission and phylogeny processes	33
3.2	A model of HIV transmission	34
3.3	A simulated epidemic, its underlying phylogeny, and its observed sequences	37
3.4	Grid-based estimates of likelihood surfaces and likelihood profiles from fitting to simulated data.	38
3.5	The distribution of age at diagnosis through time for black MSM in Detroit, MI.	42
3.6	Estimated likelihood profiles from fits to data from the black, MSM cohort.	46
3.7	A schematic of the particle filter for a GenPOMP model	53
3.8	A schematic of a hierarchical sampling scheme for the particle filter	59
3.9	A schematic of a Monte Carlo approach to estimate the conditional probability of a sequence under a relaxed clock	60
3.10	A schematic of quantities used in calculation of the conditional density of a diagnosis and the conditional probability of a genetic sequence.	61
4.1	A schematic of a model of transmission on a hospital with two wards	66
4.2	A simulated epidemic on two wards showing the latent states of individuals and their culture test results.	69
4.3	Grid-based likelihood profiles over β_w and β_b.	71
4.4	Patient traces and their relation to pairwise genetic distances between their pathogen sequences	76
4.5	A model of VRE transmission on a hospital with 12 wards	77
5.1	A state-space model that includes both a susceptible pool and the infected population.	79
A.1	A directed acyclical graph representation of dependencies among GenPOMP model components.	91
A.2	A diagram of just-in-time construction of state variable to compute the probability of a new sequence.	94

A.3 Results from an experiment exploring how the standard deviation of the log likelihood estimate scales with both the number of sequences fit and the number of particles used. 99
A.4 Trajectories of counts of each class of infected individuals over four years prior to $t_0 = 2004$, used to determine initial conditions in the data analysis. 111
B.1 A comparison of the distribution of likelihood estimates computed using the untargeted proposal and using the targeted proposal. 120

LIST OF TABLES

Table

2.1	Parameters of the transmission model used in simulation of datasets.	11
2.2	Parameters of the genetic model used in simulation of datasets.	12
3.1	Parameters of the transmission model used in simulation of datasets.	35
3.2	Parameters of the genetic model used in simulation of datasets.	36
3.3	Parameters fixed in the data analysis.	43
3.4	Parameters fit in the data analysis. This table is continued in Table 3.5.	44
3.5	Parameters fit in the data analysis (continued).	45
4.1	Parameters values used in simulating the outbreak on two wards.	68
A.1	Algorithmic parameters used in the simulation study and the data analysis.	113
B.1	Parameters of the transmission model used in the simple test of the targeted proposal approach.	119

LIST OF APPENDICES

Appendix

A. Supplementary Material for GenPomp: A flexible method for joint phylodynamic inference 84

B. A targeted proposal to combat particle depletion due to perfect measurement 114

ABSTRACT

When an epidemic moves through a population of hosts, the process of transmission may leave a signature in the genetic sequences of the pathogen. Patterns in pathogen sequences may therefore be a rich source of information on disease dynamics. Genetic sequences may replace or supplement other epidemiological observations. Furthermore, sequences may contain information not present in other datatypes, opening the possibility of inferences inaccessible by other means. The field of phylodynamic inference aims to reconstruct disease dynamics from pathogen genetic sequences.

Although a wide variety of phylodynamic inference methods have been proposed, most methods for fitting mechanistic models of disease operate in two disjoint steps, first estimating the phylogeny of the pathogen and then fitting models of disease dynamics to properties of the estimated phylogeny. Logical inconsistency in demographic assumptions underlying the two stages of inference may create bias in resulting parameter estimates. Joint inference of disease dynamics and phylogeny ensures consistent assumptions, but few methods for joint inference are currently available. The central work of this thesis is a new method for joint inference of disease dynamics and phylogeny from pathogen genetic sequences. This likelihood-based method, which we call `genPomp`, allows for fitting mechanistic models of arbitrary complexity to genetic sequences.

The organization of this thesis is as follows. In Chapter I, we present background on the field of phylodynamic inference. In Chapter II, we use simulation to study a two-stage inference approach proposed by Rasmussen et al. (2011). We find that errors in phylogenetic reconstruction may drive bias in two-stage phylodynamic inference. This result underscores the need for methodology for joint inference of the transmission model and the pathogen phylogeny. In Chapter III, we propose a flexible method for joint inference and demonstrate the feasibility of this method through simulation and a study on stage-specific infectiousness

of HIV in Detroit, MI. This method is comprised of a class of algorithms that use sequential Monte Carlo to estimate and maximize likelihoods. In Appendix A we show theoretical support for our algorithms. In Chapter IV, we demonstrate the flexibility of our approach by developing a model of transmission of Vancomycin-resistant enterococcus in a hospital setting. To allow for fitting this model to patient-level data we developed a targeted proposal, detailed in Appendix B. We present exploratory analysis of a hospital outbreak at NIH that motivates the form of the model, and carry out a study on simulated data. Although some assumptions of the simulated example are unrealistic, these initial results will inform future efforts at fitting real data. In Chapter V, we summarize the progress represented in this thesis and consider possibilities for future work.

CHAPTER I

Introduction

Grenfell et al. (2004) coined the term phylodynamics, defining it to be the study of how immunological, epidemiological and evolutionary processes combine to shape pathogen phylogenies. This paper called for a synthesis of multiple fields to better understand processes operating across different scales, from within-host infections to between-host interactions to population-level dynamics. Thinking broadly across the range of patterns in observed RNA virus phylogenies, Grenfell et al. (2004) presented what they termed "a phylodynamic framework". This framework was in essence a categorization of commonly observed phylogenetic shapes coupled with well-reasoned hypotheses about the complex processes that formed them. In turn, they speculated about the type of inferences one could make about process from the shape of phylogenies. In order to move beyond a heuristic link between mechanism and observed phylogeny shape, this paper concluded with a call for research effort in three areas:

1. A focus on determining the "immunological implications of genetic change in the virus".

2. A better understanding of "the quantitative interaction between the strength of the immune response, the kinetics of viral adaptation, and the timing of transmission".

3. Work to "establish how epidemic and metapopulation disease dynamics modulate selective forces [...] to drive long-term phylogenetic patterns".

The field of phylodynamics therefore encompasses a broad research agenda, and much of what Grenfell et al. (2004) highlighted as important operates at the level of within-host processes. At the other end of the spectrum lies the related goal of understanding how population-level processes, such as disease dynamics and patterns of disease transmission, operate to shape pathogen phylogenies.

For patterns of transmission to affect the shape of phylogenies, the rate of pathogen evolution and the process of transmission must occur on similar timescales. Pathogen populations that meet this criteria have been called measurably evolving populations (Drummond et al., 2003). For example, RNA viruses, despite having relatively short genomes, evolve at a rapid enough rate that their phylogenies may contain a signature of disease dynamics (Pybus and Rambaut, 2009). In recent years, as the quality of sequencing technology has improved, the definition of what constitutes a measurably evolving population has expanded to include populations of bacterial pathogens (Biek et al., 2015). While the mutation rate of bacteria is lower than that of RNA viruses, the length of whole genome sequences from bacteria allow for detecting informative mutations over the timeframe of an outbreak.

Resting on the key the assumption of studying a measurably evolving population, phylodynamic inference aims to infer epidemiological process from patterns in pathogen genetic sequences. Phylodynamic inference methods include a broad range of goals and techniques; methods differ not only in mechanics, but also in objective. For example, some methods aim to estimate the effective population size of infected individuals through time. In some cases, a simple scaling of effective population size is an adequate proxy for census population size. Other methods aim to estimate transmission trees (e.g. Didelot et al. (2014)). More recently, methods have been developed to fit mechanistic models of transmission to pathogen genetic sequences. To achieve any of these aims, methods require a means for relating population dynamics to phylogeny. Below we discuss methods that fall into three major categories: coalescent-based approaches, birth-death process methods, and feature matching approaches.

Proposed by Kingman (1982a,b), the coalescent is the basis for many inference approaches. The coalescent is a backward-in-time approximation for the distribution of trees that arise from simple forward-in-time models, such as the Moran model (Moran, 1958) and the Fisher-Wright model (Wright, 1931; Fisher et al., 1999). Extensions to the coalescent allow for modeling more complicated population dynamics. Coalescent-based phylodynamic inference approaches connect models of infectious disease to phylogenies by explicitly specifying how demographic quantities relate to the distribution of coalescence times of the phylogeny. Coalescent-based methods for estimation of effective population size through time are perhaps the most widely-used phylodynamic inference approaches. Methods for inference of effective population size are often integrated into the software BEAST (Drummond and Rambaut, 2007). In many cases, BEAST allows for joint estimation of phylogeny and effective population size. Early methods for estimation of effective population size used simple parametric models, such as constant population size or exponential growth (Kuhner

et al., 1998; Drummond et al., 2002). While these parametric models are easy to work with, they lack the flexibility one may desire in modeling more complicated disease dynamics that are commonly observed, such as seasonality in population size. The Bayesian skyline plot was one of the first nonparametric approaches for estimating a flexible trajectory of effective population size through time from a phylogeny (Pybus et al., 2000). This technique generates a piecewise constant estimate of the effective population size through time, a trajectory which may resemble a city skyline. The basic ideas underlying this method have been extended and modified to yield other related approaches. The generalized skyline plot (Strimmer and Pybus, 2001) smooths the estimate of the effective population size through time by aggregating adjacent coalecent intervals. The skyride (Minin et al., 2008), and a yet more flexible version of the skyride, the skygrid (Gill et al., 2013) , both use a Gaussian Markov random field to smooth the estimate of effective population size. Most recently, the skygrowth method (Volz and Didelot, 2018) defines the prior for epidemic history in terms of the growth rate of the effective population size. This prior may be more realistic for commonly observed disease dynamics than options available in earlier methods.

Birth-death models provide an alternative basis for estimation of effective population size from phylogeny. Using a birth-death model may be advantageous when the assumptions of the coalescent are problematic. For example, the coalescent approximation assumes a small sample from a large population. If the proportion of sampled infections is large then it may be more appropriate to use a birth-death model. Stadler (2010) derived the density for a tree with sampled extant and extinct ancestors. This work allowed for using a birth-death prior for phylogenies in Bayesian inference approaches. In particular, Stadler et al. (2013) proposed the birth-death Bayesian skyline plot as an alternative to coalescent-based approaches. Recently, this method was extended to allow for modeling multitype birth-death processes. This extension allows for modeling structured populations, such as transmission between metapopulations (Kühnert et al., 2016).

Methods for estimation of effective population size have the advantage of fitting naturally into a Bayesian framework that allows for joint inference of dynamics and phylogeny. A disadvantage is that effective population size is an abstract quantity that relates to census population size in a straightforward manner only under idealized models of reproduction. Appropriate scaling of effective population size to extract census population size is therefore not always possible. Furthermore, census population size may not be of primary interest. In many cases, researchers may instead be interested in understanding underlying mechanisms that drive observed disease dynamics. Fitting mechanistic models of transmission is an alternative approach to inference that allows for asking a broader range of questions. Mechanistic models can be carefully formulated to represent a wide range of

scientific hypotheses.

The coalescent can be used as a basis for fitting mechanistic models of transmission to phylogenies. For example, Volz et al. (2009) derived a coalescent-based approximation for relating an SIR model to the coalescent times of a phylogeny. This approximation served as the basis for a particle MCMC approach presented in Rasmussen et al. (2011). Furthermore, modifications to this approximation have allowed for fitting more complex models (Rasmussen et al., 2014).

Birth-death models can also form the basis of fitting mechanistic models of transmission. For example, the birth-death skyline model can be parametrized so as to approximate a stochastic SIR model (Kühnert et al., 2014). This approach allows for joint inference of epidemiological parameters of interest, such as the basic reproductive number, and phylogeny. In simulation, Kühnert et al. (2014) demonstrate that this approximation of a stochastic SIR model is sufficient to accurately estimate parameters of interest. However, this approach allows for fitting an approximation of a specific mechanistic model to sequence data. Modifying the parameterization of the birth-death skyline model is not a general strategy for fitting models of arbitrary complexity to sequences of pathogens. Therefore this avenue of methodological development is unlikely to serve as a foundation for fitting mechanistic models that represent the diverse hypothesis one may desire to consider.

Feature matching approaches provide a general basis for fitting mechanistic models of disease. These methods bypass computing a likelihood and instead fit models of disease transmission to phylogenies by using simulation to match summary statistics. A number of feature matching approaches have been proposed for phylodynamic inference. Ratmann et al. (2012) proposed an Approximate Bayesian computation (ABC) method, which was then applied to study the dynamics of influenza. Poon (2015) developed an ABC method based on a kernel function from computational linguistics, and showed its utility in a study on HIV. Giardina et al. (2017) used an approach that combines ABC and sequential Monte Carlo to infer contact network structure from phylogeny. This methodology is based on a more general ABC-SMC approach proposed by Toni et al. (2009). Feature matching methods have the advantage of being simulation-based, and therefore in principle allow for fitting mechanistic models of arbitrary complexity. However, they have the disadvantage of having no systematic criteria to define which summary statistics are optimal. Likelihood-based methods, when feasible, allow for more efficient use of information in the data.

Most phylodynamic inference methods that allow for fitting mechanistic models operate by first estimating a pathogen phylogeny and then relating a dynamic model of disease to coalescent times or summary statistics on the phylogeny. We refer to these approaches as two-stage inference methods. These methods are a natural first step in refining the

phylodynamic framework outlined by Grenfell et al. (2004), and have involved much careful work in crafting connections between models of disease dynamics and pathogen phylogenies. Although these two-stage methods may differ in how they relate transmission models to phylogenies, all may suffer from a common weakness. In particular, two-stage inference methods do not guarantee agreement between the demographic assumptions underlying phylogenetic reconstruction and those of the transmission model. Disagreement in these demographic assumptions, for example fitting an SIR model to a phylogeny estimated under the assumption of a constant population size, could drive bias in parameter estimates. To explore this hypothesis, in Chapter II we study a two-stage inference approach proposed by Rasmussen et al. (2011). Through a simulation study, we show that errors in phylogenetic reconstruction may bias estimates of transmission rate derived from this particular two-stage phylodynamic inference approach. The results of this chapter, if broadly applicable, emphasize the need for methodology for joint inference of the transmission model and the pathogen phylogeny.

There are few currently available methods for joint inference using mechanistic models, and those that do exist are limited in various ways. The birth-death SIR model of Kühnert et al. (2014), discussed above, allows for joint inference only for an approximation of specific mechanistic model. Recently, Rasmussen et al. (2017) proposed Bayesian method for joint inference on contact network structure and phylogeny. This method is limited to fitting pairwise coalescent models, which may be of use only under certain scenarios. Lau et al. (2015) proposed a more general Bayesian method for joint inference. All of these approaches suffer from the drawback that tuning of the MCMC sampler may be required for efficient inference.

In Chapter III, we describe the central work of this thesis: `genPomp`, a method for likelihood-based joint inference of mechanistic models of transmission and phylogeny. This method is comprised of a class of algorithms that use sequential Monte Carlo to estimate and maximize likelihoods of models. In Appendix A we present theoretical support for our approach. We also describe a broad class of individual-based partially observed Markov process models for which our methods are applicable. The flexibility of this class of models allows for tailoring mechanistic models to represent specific scientific hypotheses. We demonstrate the feasibility of `genPomp` through a study on simulated data and show its potential for inference in a study on stage-specific infectiousness of HIV in a subepidemic in Detroit, MI.

A recent review on phylodynamics highlighted the increasing importance of data integration as a path to stronger inference (Baele et al., 2017). The idea is simple: instead of using only pathogen genetic sequences, it may be critical to incorporate additional sources

of information to maximize potential insights from phylodynamic analyses. However, integrating multiple data types into one analysis is not necessarily an easy task for most phylodynamic methodologies. The approach of Rasmussen et al. (2011) is a notable example of one of the first methods to allow for using both pathogen genetic data and case count data to inform a model of disease transmission. The ability to easily incorporate multiple data types in a single model fitting procedure is also one of the strengths of the `genPomp` approach.

In Chapter IV, we develop a model of transmission of Vancomycin-resistant enterococcus in a hospital setting. We perform a study on simulated data, using a model with only two wards, to explore the feasibility of our approach to inference. This study on simulated data makes a number of assumptions that depart from the more difficult scenario we will encounter in fitting the real data. Nevertheless, results from the simulated example represent a nontrivial step forward. Fitting detailed individual-level data required development of a targeted proposal to allow for computing likelihoods. We describe the targeted proposal in Appendix B. Furthermore, the study on simulated data showed that incorporating multiple datatypes will likely be essential for leveraging information in the genetic sequences when fitting data from real outbreaks. We present exploratory analysis of a hospital outbreak at NIH that indicates the data may contain information on transmission history. Scaling up to fitting this data from this outbreak, or other outbreaks, may require further simulation studies designed to be more similar to the form of real data in scale and structure.

As our ability to sequence pathogens continues to improve, the potential of phylodynamic inference to revise our understanding of infectious disease systems will increase. However, a critical component to success will be the synthesis of multiple fields of study in a meaningful way. The central work of this thesis aims to unite mechanistic models of transmission and neutral models of pathogen evolution in a consistent and useful fashion. In doing so, we lay the foundation for asking and answering questions not accessible via other approaches. In Chapter V, we summarize the progress represented in this thesis and consider possible future directions.

CHAPTER II

An argument for the importance of joint inference of phylogeny and transmission

2.1 Abstract

Many phylodynamic inference methods operate by fitting models of disease transmission to features of phylogenies, such as coalescence times or summary statistics. These methods proceed in two stages: (1) estimate the pathogen phylogeny, and (2) fit the transmission model to features of the phylogeny. In practice, except for very simple models, these methods allow the demographic assumptions underlying phylogenetic reconstruction to disagree with dynamics of the fitted transmission model. In this paper, we use simulation to show that errors in phylogenetic reconstruction may drive bias in parameter estimates from two-stage phylodynamic inference. The fragility of inference when working with an estimated phylogeny underscores the need for methods that jointly infer transmission dynamics and pathogen phylogeny.

2.2 Introduction

When transmission dynamics and pathogen evolution occur on a similar timescale, the process of transmission may play a central role in shaping the pathogen phylogeny. Under such conditions, the pathogen phylogeny may have similar characteristics to the transmission tree. Working from this assumption, many phylodynamic inference methods aim to reconstruct infectious disease dynamics by fitting models of transmission to features of estimated phylogenies, such as coalescence times or summary statistics (Volz et al., 2009; Rasmussen et al., 2011, 2014; Ratmann et al., 2012; Volz and Frost, 2014; Boskova et al., 2014; Leventhal et al., 2014; Poon, 2015; Giardina et al., 2017). These methods operate in two stages: (1) estimate the pathogen phylogeny from time-stamped genetic sequences and (2) fit a model of transmission to properties of the estimated phylogeny. Although

these methods differ in the details of how they relate transmission models to features of pathogen phylogenies, all may suffer from a common weakness. Because these methods operate in two uncoupled stages, with a separate model for each stage, inconsistency in model assumptions may compromise inference. In particular, except for very simple models (e.g., the birth-death SIR model of Kühnert et al. (2014)), in many currently available methods the demographic assumptions of the model used to estimate the phylogeny may be incompatible with those of the fitted transmission model. Inconsistency in demographic assumptions underlying these two stages of inference may generate bias of unknown magnitude in parameter estimates.

In this paper, we use simulation to explore scenarios under which a specific two-stage inference method may be suitable as well as scenarios under which it may be problematic. We use a stochastic, individual-based, seasonally-forced SIR model to simulate transmission trees and pathogen sequences under three different sampling regimes. With this toy example, we compare the performance of two-stage inference when provided the true phylogeny of the pathogen versus its performance when using an estimated phylogeny. Furthermore, we explore the influence of different sampling regimes on inference.

2.3 Methods

2.3.1 A seasonally-forced SIR model

In this section we describe a stochastic, individual-based SIR model, seasonally forced with a square wave. We construct the model as a particular instance of the class of partially observed Markov process models described in Chapter III, Section 3.3. The latent state of the system at time t is a Markov process with three components: $X(t) = \big(\mathcal{T}(t),\, \mathcal{P}(t),\, \mathcal{U}(t)\big)$. Here, $\mathcal{T}(t)$ is the transmission forest, $\mathcal{P}(t)$ is the pathogen phylogeny, and $\mathcal{U}(t)$ is itself a Markov process describing the infection status of each individual in the population at time t. There are three possible infection statuses: S, susceptible; I, infected; or R, removed (see Fig. 2.1). The infection status of individual i at time t is given by a random process $\{X_i(t)\}$. The probabilities of a change in infection status for each individual over an interval

of duration δ are given by

$$\begin{aligned}
\mathbb{P}\big[\mathrm{S} \to \mathrm{I}\big] &= \delta\beta(t)I_t + o(\delta), && (2.1)\\
\mathbb{P}\big[\mathrm{I} \to \mathrm{R}\big] &= \delta\gamma + o(\delta), && (2.2)\\
\mathbb{P}\big[\mathrm{S} \to \dagger\big] &= \delta\mu + o(\delta), && (2.3)\\
\mathbb{P}\big[\mathrm{I} \to \dagger\big] &= \delta\mu + o(\delta), && (2.4)\\
\mathbb{P}\big[\mathrm{R} \to \dagger\big] &= \delta\mu + o(\delta), && (2.5)\\
\mathbb{P}\big[\odot \to \mathrm{S}\big] &= \delta\mu[S_t + I_t + R_t] + o(\delta), && (2.6)
\end{aligned}$$

where

$$\beta(t) = \begin{cases} \beta_L, & \text{if } t - \lfloor t \rfloor > 0.5 \\ \beta_H, & \text{otherwise,} \end{cases} \tag{2.7}$$

γ is the recovery rate, S_t is the number of susceptible individuals at time t, I_t is the number of infected individuals at time t, R_t is the number of recovered individuals at time t, $\odot$ indicates an individual is outside the study population, and $\dagger$ indicates an individual is dead or removed from the population.

We assume that the topology of the pathogen phylogeny, $\mathcal{P}(t)$, maps onto that of the transmission tree. That is, each node or edge in $\mathcal{P}(t)$ has a corresponding node or edge in $\mathcal{T}(t)$. However, we allow for the edge lengths of $\mathcal{P}(t)$ to differ from those of $\mathcal{T}(t)$ to model heterogeneity in the rate of molecular evolution. We assume that each edge of $\mathcal{P}(t)$ has length conditionally Gamma distributed with expectation equal, and variance proportional, to the length of the corresponding edge of $\mathcal{T}(t)$. That is, if L is the length of an edge of $\mathcal{P}(t)$ corresponding to an edge of length D in $\mathcal{T}(t)$, we let $L|D$ be Gamma distributed with $\mathbb{E}\left[L|D = d\right] = d$ and $\mathrm{Var}[L|D = d] = \sigma\, d$. Once the structure of $\mathcal{P}(t)$ is specified, then the time-reversible model of molecular evolution applied across this structure yields the joint distribution of the sequences at each tip of $\mathcal{P}(t)$. For this study, we used the TN93 model of molecular evolution (Tamura and Nei, 1993), which is fully specified by the following rate matrix:

$$Q = \begin{bmatrix} * & \beta\pi_T & \beta\pi_C & \alpha_R\pi_G \\ \beta\pi_A & * & \alpha_Y\pi_C & \beta\pi_G \\ \beta\pi_A & \alpha_Y\pi_T & * & \beta\pi_G \\ \alpha_R\pi_A & \beta\pi_T & \beta\pi_C & * \end{bmatrix}$$

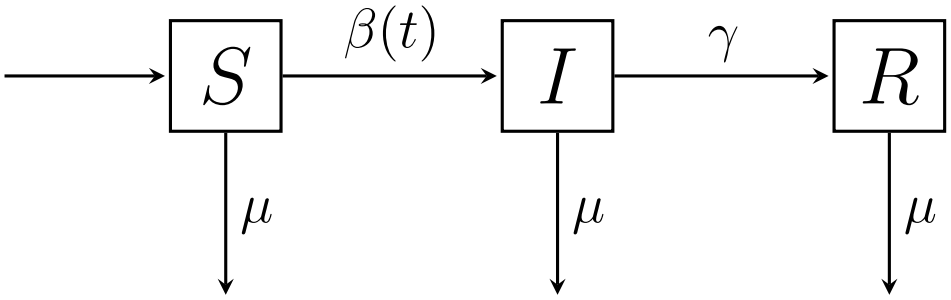

Figure 2.1: A schematic of the possible infection statuses for individuals. The per capita birth rate of susceptible individuals is the same as the per capita death rate, μ, such that on average the total population remains constant. The transmission rate, $\beta(t)$, is time-dependent; high in the first half of each year and low in the second. Infected individuals recover at rate γ.

2.3.2 Simulation

Using the model described above, we simulated epidemics with annual peaks over a three year period. We implemented this model in the software `genPomp` and simulated from the model using an exact method (Gillespie, 1977). Simulated data consist of time-stamped pathogen genetic sequences as well as the pathogen phylogeny that underlies these sequences.

We chose to simulate using three different sampling regimes: (1) uniform sampling, in which samples were spaced evenly across the period of simulation; (2) proportional sampling, in which samples were taken proportional to the number of infected individuals at a given time; and (3) late sampling, in which samples were taken uniformly over a short period of time during the declining phase of the epidemic each year. We chose these sampling regimes in part because they have been studied before (Stack et al., 2010; Hall et al., 2016). Also, these sampling regimes correspond to three scenarios one may encounter in practice. Proportional sampling corresponds to an observation process that is directly tied to the dynamics of the disease. For example, in a highly seasonal infectious disease the number of cases varies significantly over the course of the year. A convenience sample from such a process may yield more sequences from the peaks of epidemics than from the troughs. Late sampling corresponds to a surveillance system that misses the initial onset of an epidemic. Finally, uniform sampling represents a possible strategy one could employ when conducting a retrospective study of an epidemic with a limited budget for

sequencing. This sampling regime allows for capturing some measure of the pathogen evolutionary process over the entire course of the epidemic.

We simulated using parameters that generate annual dynamics with strong bottlenecks. For each sampling regime we simulated 100 epidemics and for all sampling regimes we allow each individual to be sampled only once. The parameters used in simulation are given in Table 2.1 and Table 2.2.

Table 2.1: Parameters of the transmission model used in simulation of datasets.

Parameter	**Interpretation**	**Value**
S_{t_0}	Number of susceptible individuals at t_0	1485
I_{t_0}	Number of infected individuals at t_0	15
R_{t_0}	Number of recovered individuals at t_0	23000
β_L	Low transmission rate (during first half of the year)	312 yr^{-1}
β_H	High transmission rate (during second half of the year)	624 yr^{-1}
γ	Recovery rate of infected individuals	13 yr^{-1}
μ	Death rate (and birth rate)	0.1 yr^{-1}
t_0	Start time of simulation	0 yr
t_{end}	End time of simulation	3 yr
ζ	Sampling rate (only for proportional sampling regime)	0.2 yr^{-1}

Table 2.2: Parameters of the genetic model used in simulation of datasets.

Parameter	**Interpretation**	**Value**
β	Rate of transversions	1 yr^{-1}
α_Y	Rate of transitions between purines	0.5 yr^{-1}
α_R	Rate of transitions between pyrimidines	0.5 yr^{-1}
π_A	Equilibrium frequency of adenine	0.25
π_G	Equilibrium frequency of guanine	0.25
π_C	Equilibrium frequency of cytosine	0.25
π_T	Equilibrium frequency of thymine	0.25
σ_{site}	Relaxation of the molecular clock with respect to sites	0
σ	Relaxation of the molecular clock with respect to branches	0.5 yr
δ_{fixed}	The initial component of the sequence stem	0.001 yr
δ_{prop}	Proportion of time since infection to add to the sequence stem	0
n_L	Number of loci in simulated genetic sequences	100
t_p	Time of the polytomy	0 yr

2.3.3 Inference

To perform two-stage phylodynamic inference, we used a particle filtering approach developed by Rasmussen et al. (2011). This approach is based on coalescent theory, and relates the counts of S, I and R individuals of an SIR process (such as the one described in Sect. 2.3.1) to the coalescent times of the pathogen phylogeny. Following the work of Volz et al. (2009), this approach assumes that the hazard of an infection event being observed as a coalescence in the pathogen phylogeny at time t is:

$$\lambda(t) = \frac{\binom{i_t}{2}}{\binom{I_t}{2}} \beta(t) \frac{S_t}{N_t} I_t, \tag{2.8}$$

where i_t is the number of lineages in the pathogen phylogeny at time t, S_t is the number of susceptible individuals, I_t is the number of infected individuals, N_t is the total number of individuals in the population, and $\beta(t)$ is the transmission rate at time t. In the analyses in this paper, we modify this equation slightly:

$$\lambda(t) = \frac{\binom{i_t+1}{2}}{\binom{I_t+1}{2}} \beta(t) \frac{S_t}{N_t} I_t. \tag{2.9}$$

We made this modification in part for numerical tractability when the count of infected individuals becomes small. In our formulation, the hazard is defined even when $I_t = 1$. We also prefer this formulation as it fits with a forward-in-time, mechanistic interpretation of the equation. If, instead of thinking of $\lambda(t)$ as a coalescence rate, we think of it as a branching rate then we can interpret $\lambda(t)$ as the hazard of incrementing the number of infected lineages by one and observing that branching event in the phylogeny. That it is not clear which is the correct choosing statement hints at a fundamental inconsistency in this approach.

Over any time interval, we can compute the cumulative hazard of an infection event being observed as a coalescence in the pathogen phylogeny:

$$\Lambda_j = \int_{t_{j-1}}^{t_j} \lambda(t) dt \tag{2.10}$$

The hazard and the cumulative hazard are the two quantities used to construct the measurement model used in the particle filter.

This particle filter takes as input the count of lineages in the pathogen phylogeny through time. We can represent this data reduction of the pathogen phylogeny, which we call the *lineage count trajectory*, as an ordered set of $N + 1$ times, $\{t_0, t_1, ..., t_N\}$, and

$N+1$ corresponding lineage counts, $\{i_{t_0}, i_{t_1}, ..., i_{t_N}\}$. At time t_0, the initial number of lineages in the pathogen phylogeny is i_{t_0}. At time t_1, the count of lineages changes to i_{t_1}, and so on. At times of sampling, the lineage count decrements by one; at times of coalescence, the lineage count increments by one.

The particle filter begins at t_0 and sequentially processes the subsequent times, generating a score for each time interval, $\omega_j = t_j - t_{j-1}$, $j \in \{1, 2, ..., n\}$, to ultimately build up an aggregate score for all of the data. As described above, there are two types of intervals: an interval either ends in a sampling event or in a coalescent event. Using the hazard, cumulative hazard, and known trajectories of state variables in each particle, we can compute a density for each interval of time. The more compatible the simulations are with the observed lineage count trajectory of the pathogen phylogeny, the higher the density. The steps of this particle filter are detailed in Algorithm 1.

Algorithm 1: The particle filter of Rasmussen et al. (2011) applied to the model of Sect. 2.3.1

input: $\{S_{t_0}, I_{t_0}, R_{t_0}\}$, initial state values;
$\{\beta_L, \beta_H, \gamma, \mu\}$, parameter values;
$\{t_0, t_1, ..., t_n\}$, event times;
$\{i_{t_0}, i_{t_1}, ..., i_{t_N}\}$, lineage counts;
J, the number of particles.

Initialize each particle with initial state values: $\{S_{t_0}, I_{t_0}, R_{t_0}\}$

for n in $1 : N$

  for j in $1 : J$

    Simulate states of particle j forward in time from t_{n-1} to t_n

    Compute particle weight, w_j, the log of the density of the interval:

      If $i_{t_n} > i_{t_n - 1}$: $w_j = \ln(\lambda(t_j)) - \Lambda_j$ (a coalescent event at t_n)

      Else: $w_j = -\Lambda_j$ (a sampling event at t_n)

    Resample particles proportional to their weights

  end for

  Compute interval score, $W_n = \ln \frac{1}{J} \sum_{j=1}^{J} e^{w_j}$

end for

output: Rasmussen criterion: $\mathcal{R} = \sum_{n=1}^{N} W_n$

We implemented this particle filter using the R package pomp (King et al., 2016b). See archived codes for details of this implementation. The particle filter calculation described in Rasmussen et al. (2011) is not a likelihood, at least not for the mechanistic model for which we aim to calculate a likelihood. We expand on why this calculation is not a likelihood and speculate on the potential implications of this approximation in the discussion. For the purpose of generating point estimates and confidence intervals, however, we follow the current standard practice and treat this calculation as if it were a likelihood. To indicate

that it is not a likelihood, we refer to this calculation as the Rasmussen criterion, $\mathcal{R}$.

We derived point estimates and confidence intervals for β_L and β_H from estimated likelihood profiles. To estimate likelihood profiles, we fixed all other parameters at their known true values and then estimated a two-dimensional likelihood surface over a grid of values for β_L and β_H. From this surface we then extracted likelihood profiles for each parameter. For each simulation, we estimated profiles in two ways: (1) using the true pathogen phylogeny, and (2) using an estimated pathogen phylogeny. We used BEAST (Drummond and Rambaut, 2007) with a Bayesian skyline and an independent lognormal relaxed clock to estimate the phylogeny. By comparing results from the estimated versus the true phylogeny we were able to assess how much the process of phylogenetic reconstruction affected the conclusions of two-stage inference.

2.4 Results

We find that the process of estimating the pathogen phylogeny degrades or erases the signal of seasonality contained in the true phylogeny. Figures 2.2 and 2.3 show a representative example of this result from a single simulation in which samples were distributed uniformly across the span of the epidemic. In Fig. 2.2 we see that although the estimated and true phylogenies share some general structural characteristics, the lineage count trajectories of the two trees differ markedly. In particular, the lineage of trajectory of the estimated phylogeny has a dampened signal of seasonality relative to that of the true phylogeny. What we see by eye in the plot of lineage count trajectories is confirmed in the likelihood profiles for β_L and β_H (Fig. 2.3). When using the lineage count trajectory of the true phylogeny, parameter estimates fall near the truth and 95% confidence intervals encompass the truth. On the other hand, when using the lineage count trajectory from the estimated phylogeny the point estimate for β_H is low, with a confidence interval that does not contain the true value.

Examining point estimates and confidence intervals across many simulated examples reveals that downward bias in the point estimate for β_H is a common problem when using an estimated phylogeny, but the severity of bias depends on the sampling regime (Fig. 2.4). Of the three sampling regimes considered in this study, the distributions of point estimates and confidence intervals overlap to the greatest degree for the late sampling regime. For both uniform and proportional sampling regimes there is clear bias in the estimates for β_H when using the estimated phylogeny; the distributions of point estimates are both centered below the true value of β_H. However, the signal of seasonality in these two cases is not completely obscured as one can still see some separation in the distributions of point

estimates for β_H and β_L.

When provided the true pathogen phylogeny, two-stage inference may perform quite well, regardless of the sampling regime. We can think of this scenario as two-stage inference in the case when the first stage is executed without error. As shown in Fig. 2.4, under all sampling regimes we find that the distributions of point estimates for both β_L and β_H are centered about their true values, and their corresponding 95% confidence intervals appear to have approximately the correct level of coverage.

In Fig. 2.4 we excluded all cases in which the estimated phylogeny implied initial conditions that were incompatible with the true initial conditions. This occurs when the estimated phylogeny has a deep root relative to the true phylogeny such that the lineage count of the estimated phylogeny at t_0 exceeds the known number of infected individuals at t_0. Under this scenario the estimated lineage trajectory is entirely incompatible with the assumption of the model that each infected individual carries only one pathogen lineage. Excluding these examples means that Fig. 2.4 shows an overly favorable picture of the performance of this two-stage inference approach when using the estimated phylogeny.

2.5 Discussion

2.5.1 The consequences of two-stage inference

This study shows that error in phylogenetic reconstruction may drive bias in parameter estimation via two-stage phylodynamic inference. Although inference using the true phylogeny yields a clear signal of seasonality, when performing inference with an estimated phylogeny the signal of seasonality is degraded or lost almost entirely. These results are compatible with those of Rasmussen et al. (2011), which examined the quality of inference only when using the true phylogeny. Rasmussen et al. (2011) found that their particle filtering approach, used in particle MCMC, was able to successfully estimate parameters of a seasonal SIR model when given the true phylogeny. Our simulation study adds an important perspective by exploring the performance of this approach when using an estimated phylogeny. The fragility of inference when using an estimated phylogeny underscores the need for methods that jointly infer the transmission model and the pathogen phylogeny.

This study presents only a handful of scenarios; there may be others in which estimation of the phylogeny is sufficiently accurate so as to merit the use of two-stage inference. For example, in some cases, inference of divergence times is likely to be quite accurate when estimated using BEAST on heterchronous genetic sequences. This may indeed be the case when the molecular model of evolution is correctly specified.

2.5.2 The influence of sampling regime on inference

When working with the estimated phylogeny we found that the sampling regime played an important role in how well two-stage inference was able to recover the true parameters. Uniform and proportional sampling both performed better than late sampling. These two sampling regimes share the property that they have some degree of sampling throughout the full course of the epidemic. Late sampling, on the other hand, captures sequences only from period when the annual epidemic is declining.

This result does not agree with those of Stack et al. (2010), which found that sampling during the declining phase of the epidemic was the best strategy to capture information on disease dynamics with strong seasonality. Stack et al. (2010) compared several different sampling regimes, some not included in our study, but did include both late sampling and uniform sampling. In many respects the simulation study of the Stack et al. (2010) study is similar this study; a key difference is that Stack et al. (2010) used a strict molecular clock to simulate the sequences. This could be the main reason that our results differ. A strict clock would preserve a great deal more information than a relaxed clock. Stack et al. (2010) argue that late sampling allows for capturing ephemeral lineages as well as persistent lineages that together yield information on both growth and decline. A relaxed clock could obscure the distinction between these two types of lineages. When estimating the phylogeny from sequences generated by a relaxed clock it may be best to have samples spaced throughout the epidemic Distributed samples would serve as anchor points that facilitate interpolation as opposed to the clustered nature of late sampling, which would leave more to extrapolation. This is speculation as to why our results differ from those of Stack et al. (2010). One could explore whether there is support for this idea by carrying out a simulation study in which sequences were simulated under a strict clock.

2.5.3 The Rasmussen Criterion

The computation generated by Algorithm 1 is not a likelihood for the model for which we aim to compute a likelihood. One way to see this is to compare the process that generates the data with the model represented in Algorithm 1. The process that generates the data involves multiple steps: first, an epidemic unfolds forward in time and a subset of infected individuals are sampled for sequencing; second, the time-stamped pathogen sequences are used to estimate a phylogeny; last, we extract from the phylogeny a set of coalescence times. The process to generate the data involves both a forward-in-time realization of a stochastic system and a backward-in-time reconstruction of past events. Importantly, the data cannot be observed until the forward-in-time process of simulation and sampling is complete; only

when all samples have been taken can one construct the set of all coalescence times.

In the particle filter described in Algorithm 1, we aim to compute a likelihood for data generated from this process only through a set of forward-in-time operations carried out in a sequential fashion. By using a particle filter to estimate the likelihood, our implicit assumption is that we are fitting a POMP model. By definition, a POMP model includes a simulator of the latent state and a measurement model. The simulator of the latent state used in Algorithm 1 is correct; our claim is that the measurement model used is not the appropriate counterpart to the process that generates the data.

One could use the measurement model of Algorithm 1, together with the simulator of the latent state, to simulate forward in time a set of coalescence times (without constructing trees of any sort, merely by using the hazard as measurement model for a generative process). The distribution of these coalescence times would not match the distribution of those generated by the forward-backward process described above. For example, a set coalescence times generated purely forward in time could produce realizations in which the number of infected individuals is less than the number of lineages in the phylogeny. In the forward-backward generative process, this outcome is not possible.

Although this argument shows that the Rasmussen criterion is at best an approximation, it may be quite a good approximation under certain circumstances. The empirical results from our simulation study suggest that this approximation holds up well enough to be useful for inference when the true phylogeny is provided. The limits of this approximation, however, are yet to be fully explored and there may be cases when it breaks down.

2.5.4 The role of relaxed clock models

The relaxed clock model that we used to estimate the phylogeny using BEAST is not the same as the relaxed clock model that we used to simulate. We simulated using a relaxed clock with gamma white noise on the rate and fit using a lognormal relaxed clock. This difference could matter quite a lot. A potential criticism of our results is that they may be largely driven by model misspecification. Were we able to give BEAST the correct relaxed clock model, we might have seen much better performance for inference using the estimated phylogeny. A rejoinder to this criticism is that in practice we may be more likely to encounter the scenario we present. In general, one never knows how well one has estimated the phylogeny. Furthermore, the lognormal relaxed clock is widely used in BEAST analyses, yet this model fails to meet basic properties that we expect to see in real evolutionary systems. For example, evolutionary branch lengths under an independent lognormal relaxed clock model are not additive. The relaxed clock with gamma noise does have properties that we would expect from a physical process of evolution, such as additivity

of branch lengths, and is a reasonable model to simulate the process of evolution.

2.5.5 Joint inference

Joint inference of transmission and phylogeny holds the promise to remedy the problem of bias that may arise in two-stage inference. Until recently, joint inference has only been possible for simple models of transmission. In the last few years, researchers have begun to develop joint inference methodology for fitting more complex models that correspond more directly with questions of scientific interest. A handful of methods are now available. Lau et al. (2015) developed a Bayesian method for joint inference. Rasmussen et al. (2017) proposed a method for joint inference on network structure and phylogeny. In Chapter III we propose a plug-and-play method for likelihood-based joint inference that allows for fitting stochastic mechanistic models of arbitrary complexity.

2.6 Future Directions

1. A worthwhile comparison will be to run a simulation study using a strict molecular clock as opposed to a relaxed clock. It will be especially interesting to see if the results for the late sampling regime change relative to the other two sampling regimes.

2. We will add results from estimating β_L and β_H using `genPomp`. Replication on the scale needed to add a column to Fig. 2.4 is not feasible, but we could show results for a subset of the simulations.

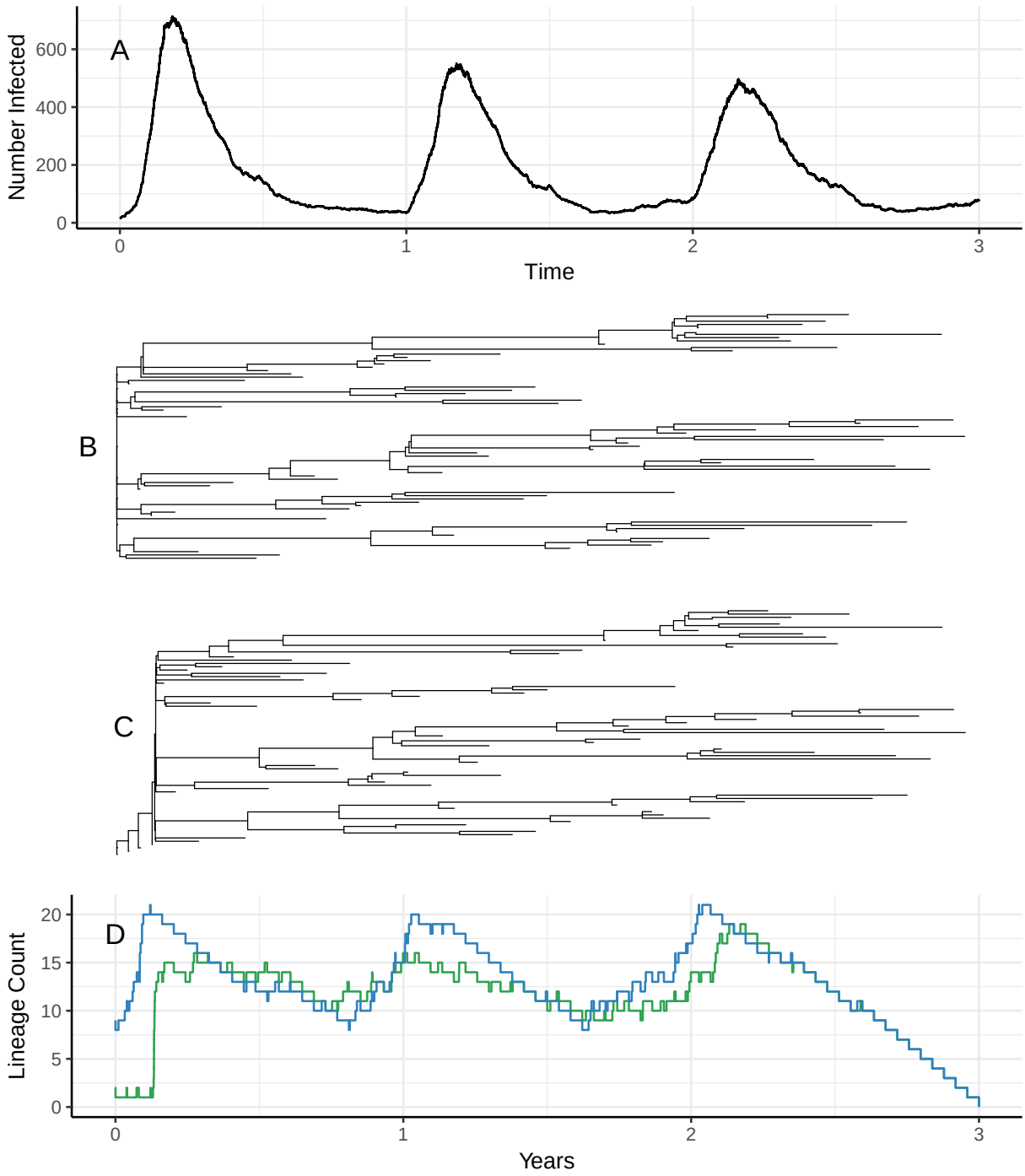

Figure 2.2: Components of an SIR epidemic, simulated and estimated. Panel A shows a simulated trajectory of infected individuals. Panel B shows the true phylogeny that connects a set of uniformly sampled sequences from the epidemic above. Panel C shows an estimated phylogeny for the sampled sequences. Panel D shows lineage count trajectories of true and estimated phylogenies, in blue and green respectively. The lineage count trajectory of the true tree appears to have a stronger signal of seasonality than the lineage count trajectory of the estimated tree; peaks and troughs of the blue curve are more extreme than those of the green curve.

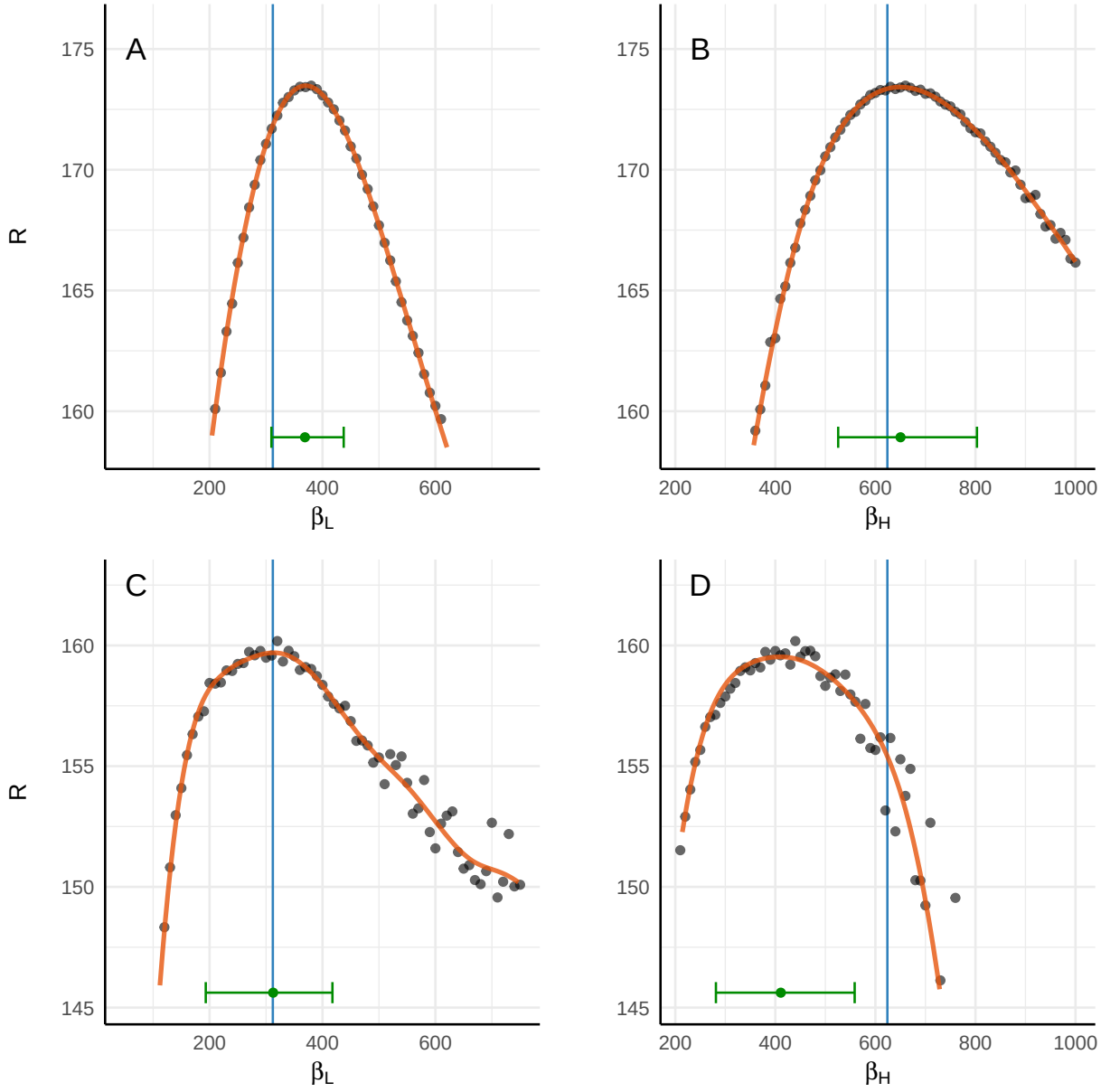

Figure 2.3: Likelihood profiles for β_L and β_H derived from using the true phylogeny (panels A and B) and the estimated phylogeny (panels C and D). Black dots show likelihood estimates, red lines show smoothed likelihood profiles, and blue vertical lines show the true values of β_L and β_H. Point estimates and confidence intervals are shown in green at the bottom of each panel. This plot shows evidence for downward bias in the estimate for β_H. This result is in line with the dampened lineage count trajectory of the estimated phylogeny, relative to the lineage count trajectory of the true phylogeny, shown in Fig. 2.2.

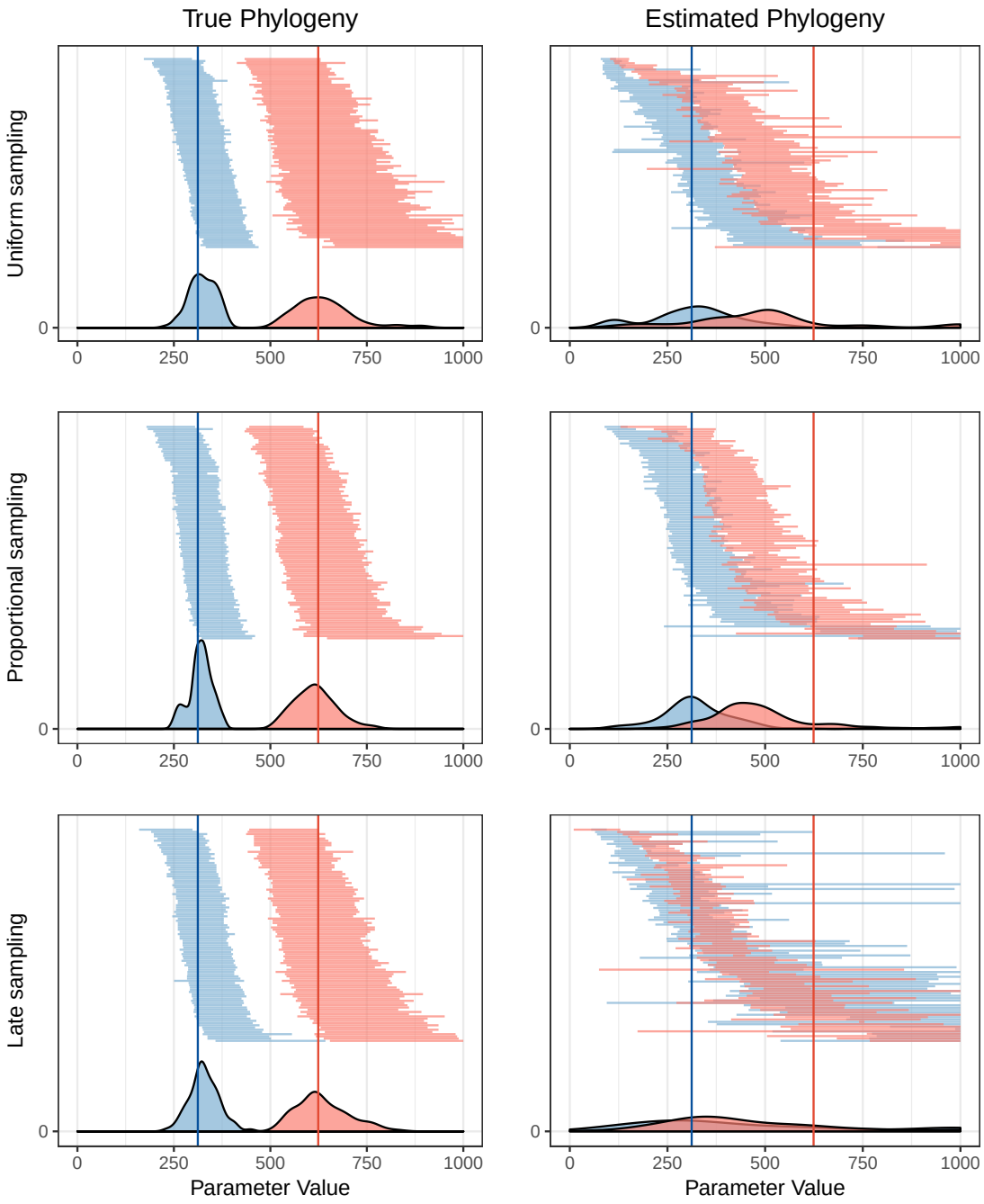

Figure 2.4: Distributions of point estimates and 95% confidence intervals for β_L (in blue) and β_H (in red) when using true phylogenies (left column) and estimated phylogenies (right column) under three different sampling regimes (top row, uniform sampling; middle row, proportional sampling; bottom row, late sampling). Distributions of point estimates are shown as empirical densities; horizontal bars above the densities show the distributions of confidence intervals (sorted by β_L point estimates). Dark vertical lines in blue and red show the true values of β_L and β_H, respectively. The degree of overlap between the distributions of estimates for β_L and for β_H demonstrates how much inference breaks down in the case of different sampling schemes when using the estimated phylogeny.

CHAPTER III

A flexible method for joint phylodynamic inference

This chapter of my dissertation was published in the *Molecular Biology and Evolution* in 2017 (DOI: https://doi.org/10.1093/molbev/msx124). The authors are myself, E. L. Ionides, and A. A. King. The supplementary material is included in this thesis as Appendix A. I drafted the main paper and ELI drafted the supplement. All three of us edited both the main paper and the supplement. I developed the software `genPomp` and ran the analyses, with guidance from both coauthors.

3.1 Abstract

Genetic sequences from pathogens can provide information about infectious disease dynamics that may supplement or replace information from other epidemiological observations. Most currently available methods first estimate phylogenetic trees from sequence data, then estimate a transmission model conditional on these phylogenies. Outside limited classes of models, existing methods are unable to enforce logical consistency between the model of transmission and that underlying the phylogenetic reconstruction. Such conflicts in assumptions can lead to bias in the resulting inferences. Here, we develop a general, statistically efficient, plug-and-play method to jointly estimate both disease transmission and phylogeny using genetic data and, if desired, other epidemiological observations. This method explicitly connects the model of transmission and the model of phylogeny so as to avoid the aforementioned inconsistency. We demonstrate the feasibility of our approach through simulation and apply it to estimate stage-specific infectiousness in a subepidemic of HIV in Detroit, Michigan. In a supplement, we prove that our approach is a valid sequential Monte Carlo algorithm. While we focus on how these methods may be applied to population-level models of infectious disease, their scope is more general. These methods may be applied in other biological systems where one seeks to infer population dynamics from genetic sequences, and they may also find application for evolutionary models with

phenotypic rather than genotypic data.

Keywords: phylodynamics, iterated filtering, sequential Monte Carlo, maximum likelihood, virus evolution, human immunodeficiency virus

3.2 Introduction

Phylodynamic methods extract information from pathogen genetic sequences and epidemiological data to infer the determinants of infectious disease transmission (Grenfell et al., 2004). For successful phylodynamic inference, mechanisms of transmission must leave their signature in genetic sequences. This occurs when pathogen transmission and evolution occur on similar timescales (Drummond et al., 2003). By explicitly relating models of disease dynamics to their predictions with respect to pathogen sequences, it is possible to estimate aspects of the mechanisms of transmission (Rasmussen et al., 2011; Stadler et al., 2013; Volz et al., 2013; Frost et al., 2015; Poon, 2015; Karcher et al., 2016). Most existing phylodynamic inference methods proceed in three stages. First, one estimates the pathogen phylogeny using sequence data. Next, one fits models of disease dynamics to properties of the pathogen phylogeny, such as coalescent times or summary statistics on the tree. Finally, one assesses the robustness of the results to variation in the estimated phylogeny to account for phylogenetic uncertainty. Frequently, such methods harbor logical inconsistencies between the assumptions of the model used to estimate the phylogeny and those of the model of disease dynamics. In particular, it may happen that population dynamics, as estimated by the transmission model, are inconsistent with those assumed when estimating the phylogeny. In the absence of consistent methods, it may be difficult to assess the loss of accuracy due to the use of inconsistent methods.

Researchers developing Markov chain Monte Carlo (MCMC) approaches to phylodynamic inference have recognized the need to develop fully consistent approaches. In particular, Lau et al. (2015) have proposed a Bayesian method for joint inference. This work builds off phylodynamic inference that uses MCMC to fit deterministic population models (Bouckaert et al., 2014). However, to achieve efficiency, it is typically necessary to tailor an MCMC sampler to the specific model being fit (Vaughan et al., 2014). The required investment makes it costly to entertain competing models and to base inference directly on the models of greatest scientific interest. In practice, phylodynamic inference for infectious diseases has therefore tended to focus on the three-stage methods described above.

In this paper, we develop methodology for jointly inferring both phylogeny and transmission, as well as estimating unknown model parameters. Our central contribution is an algorithm which we call GenSMC, an abbreviation of *sequential Monte Carlo with genetic*

sequence data. Sequential Monte Carlo (SMC), also known as the particle filter, provides a widely used basis for inference on complex dynamic systems (Kantas et al., 2015) with several appealing properties. Because basic SMC methods rely only on forward-in-time simulation of stochastic processes, it can accommodate a wide variety of models: essentially any model that can be simulated is formally admissible. Thus, the algorithm enjoys a variant of the plug-and-play property (Bretó et al., 2009; He et al., 2010). An SMC computation results in an evaluation of the likelihood, which is a well-understood and powerful basis for both frequentist and Bayesian inference. Finally, again because SMC requires only forward-in-time computation, it is straightforward to construct a model of genetic sequence evolution upon the basis of a transmission model, thus avoiding all conflict between these models.

SMC techniques have previously been used for inferring phylogenies (Bouchard-Côté et al., 2012) and for phylodynamic inference conditional on a phylogeny (Rasmussen et al., 2011). However, using SMC to solve the joint inference problem through forward-in-time simulation of tree-valued processes is a high-dimensional, computationally challenging problem. We found that several innovations were necessary to realize a SMC approach that is computationally feasible on models and datasets of scientific interest. The key innovations that provided a path to feasibility were: just-in-time construction of state variables, hierarchical sampling, algorithm parallelization, restriction to a class of physical molecular clocks, and maximization of the likelihood using the iterated filtering algorithm of Ionides et al. (2015).

In the following, we first give an overview of the class of models for which our SMC algorithms are applicable. A formal specification is given in the supplement, and the source code for our implementation is also available. Next, we present a study on a simulated dataset as evidence of the algorithm's feasibility. Finally, we use our methods to estimate determinants of the epidemic of human immunodeficiency virus (HIV) among the population of young, black, men who have sex with men (MSM) in Detroit, Michigan from 2004 to 2011. This analysis uses time-of-diagnosis and consensus protease sequences to estimate the rates of infection attributable to sources inside and outside the focal population.

3.3 New Approaches

The key novelty in our approach to phylodynamics is in formulating a flexible class of phylodynamic models and a class of sequential Monte Carlo algorithms in such a way that the latter can be efficiently applied to the former. We refer to our phylodynamic model class as GenPOMP models, in recognition of the fact that they are partially observed

Markov processes (POMPs). As such, a GenPOMP model consists of an unobserved Markov process—called the latent process—and an observable process. In the following sections, we specify the structure of each of these components. An additional, more formal, description of the GenPOMP model is given in the supplement (Section S1). Our GenSMC algorithm for GenPOMP models is introduced in the Materials and Methods section. GenSMC is presented at greater length in the supplement (Section S2) and also provided with a mathematical justification (Section S3). Our extension of GenSMC to parameter estimation, via iterated filtering, is called the GenIF algorithm and is discussed briefly in the Material and Methods section and at greater length in Section S2.2. For computational implementation of the GenPOMP framework and the GenSMC and GenIF algorithms, we wrote the open-source `genPomp` program discussed further in Section S2.1.

For concreteness, we focus here on an infectious disease scenario, wherein the model describes transmission of infections among hosts and the sequences come from pathogens in those infections. In this context, measurements on infected individuals are called diagnoses. In the concluding discussion section, we briefly consider other contexts within which the models and methods we have developed may prove useful.

3.3.1 The latent process

We adopt the convention of denoting random variables using uppercase symbols; we denote specific values assumed by random variables using the corresponding lowercase symbol. We use an asterisk to denote the data, which are treated as a specific realization of random variables in the model.

The latent Markov process, $\{X(t),\ t \in \mathbb{T}\}$, defined over a time interval $\mathbb{T} = [t_0,\ t_{\text{end}}]$, explicitly models the population dynamics and also includes any other processes needed to describe the evolution of the pathogen. Specifically, we suppose that we can write $X(t) = \big(\mathcal{T}(t),\ \mathcal{P}(t),\ \mathcal{U}(t)\big)$, where $\mathcal{T}(t)$ is the *transmission forest*, $\mathcal{P}(t)$ is the *pathogen phylogeny* equipped with a relaxed molecular clock, and $\mathcal{U}(t)$ represents the state of the pathogen and host populations. For example, $\mathcal{U}(t)$ may categorize each individual in the host population into classes representing different stages of infection. We suppose that $\{\mathcal{U}(t), t \in \mathbb{T}\}$ is itself a Markov process.

The transmission forest represents the history of transmission among hosts. We assume that hosts cannot be multiply infected; this implies that $\mathcal{T}(t)$ is a forest, i.e., a collection of trees. Nodes in $\mathcal{T}(t)$ are time-stamped and of several types. Internal nodes represent transmission events. Terminal nodes are of three types: (a) *active nodes* represent infections active at time t; (b) *observed nodes* correspond to diagnosis events, possibly associated with genetic sequences; (c) *dead nodes* correspond to death or emigration events. Root nodes

at time t_0 correspond to infections present in the initial population; root nodes at times $t > t_0$ correspond to immigration events. Since all nodes are time-stamped, edges of $\mathcal{T}(t)$ have lengths measured in units of calendar time.

The pathogen phylogeny $\mathcal{P}(t)$ represents the history of divergences of pathogen lineages. Internal nodes of $\mathcal{P}(t)$ represent branch-points of pathogen lineages, which, we assume, coincide with transmission events. The terminal nodes of $\mathcal{P}(t)$ are in 1-1 correspondence with the terminal nodes of $\mathcal{T}(t)$. The distinction between $\mathcal{P}(t)$ and $\mathcal{T}(t)$ allows for random variation in the rate of molecular evolution, i.e., relaxed molecular clocks (see below). Specifically, the edge lengths of $\mathcal{T}(t)$ measure calendar time between events, whereas edge lengths in $\mathcal{P}(t)$ can have additional random variation describing non-constant rates of evolution.

The transmission forest $\mathcal{T}(t)$ can grow in only five distinct ways: (1) active nodes can split in two, when a transmission event occurs, (2) active nodes can become dead nodes, upon emigration, recovery, or death of the corresponding host, (3) immigration events can give rise to new active nodes, each with its own distinct root, (4) sampling events cause active nodes to spawn diagnosis nodes, and (5) active nodes for which none of the above occur simply grow older. Likewise, the pathogen phylogeny $\mathcal{P}(t)$ grows along with $\mathcal{T}(t)$ (Fig. 3.1). The Markov process $\{\mathcal{U}(t)\}$ can contain additional information about the system at time t, e.g., states of individual hosts. $\{\mathcal{U}(t)\}$ can affect, but must not be affected by, the $\{\mathcal{T}(t)\}$ and $\{\mathcal{P}(t)\}$ processes. That is, given any sequence of times $t_1 < \cdots < t_k < t$, $\{\mathcal{U}(t)\}$ is independent of $\big\{\big(\mathcal{T}(t_j), \mathcal{P}(t_j)\big),\ j = 1, \ldots, k\big\}$ conditional on $\{\mathcal{U}(t_j),\ t_1 < \cdots < t_k < t\}$. The dependence relationships among $\mathcal{T}$, $\mathcal{P}$, $\mathcal{U}$, and the data are diagrammed in Fig. A.1.

We assume subsequently that $\mathcal{P}(t)$ and $\mathcal{T}(t)$ agree topologically, but we note that this assumption is not essential. In particular, the sequential Monte Carlo algorithms we apply could be straightforwardly extended to allow the topology and timing of genetic lineage divergences to deviate from those of transmission events and to allow multiple pathogen lineages within each host. Such extensions might be useful, for example, in accounting for within-host pathogen diversity.

3.3.2 The observable process

We now describe the model explicitly linking the latent process to the data. Let $\mathbb{Y}$ be the set of all finite collections of dated genetic sequences, with an element of $\mathbb{Y}$ being a collection $\{(g_k, t_k),\ k = 1, \ldots, n\}$ where g_k is a sequence and t_k is the associated diagnosis time. We allow g_k to be an empty sequence, in the event that the corresponding diagnosis had no associated sequence. The observable process is a $\mathbb{Y}$-valued process, $\{Y(t),\ t \in \mathbb{T}\}$, where $Y(t)$ consists of all sequences that have accumulated up to time t. Thus, $Y(t)$ is

expanding, i.e., $Y(t) \subset Y(t')$ whenever $t \leq t'$, and if $Y(t) = \{(G_k, T_k),\ k = 1, \ldots, N\}$, then $T_k \leq t$ for all k. The data are modeled as a realization of the observable process, $Y(t_{\text{end}}) = y^*$.

Suppose each diagnosis has has an equal and independent chance to give rise to a pathogen sequence, and each diagnosis event in $Y(t)$ corresponds to a unique diagnosis node in $\mathcal{T}(t)$. Suppose also that some time-reversible molecular substitution model is defined to describe sequence evolution on the pathogen phylogeny $\mathcal{P}(t)$. These modeling assumptions implicitly define a conditional distribution for $Y(t)$ given $X(t)$.

3.3.3 Relaxed molecular clocks

A strict molecular clock assumes that the rate of evolution is constant through time and across lineages. Relaxation of this assumption has been shown to improve the fit of phylogenetic models to observed genetic sequences in many cases (Drummond et al., 2006) and for HIV in particular (Posada and Crandall, 2001). A relaxed molecular clock models the rate of evolution as random. In our approach, this corresponds to constructing each edge length of $\mathcal{P}(t)$ as a stochastic process on the corresponding edge of $\mathcal{T}(t)$. Various forms of such processes have been assumed in the literature (Lepage et al., 2007; Ho and Duchêne, 2014), but not all of these are compatible with a mechanistic approach. In particular, a mechanistic molecular clock must be defined at all times and must have non-negative increments. Many relaxed clocks commonly employed in the literature do not enjoy the latter property: in effect, such clocks allow evolutionary time to run backward. The class of suitable random processes includes the class of nondecreasing Lévy processes, i.e., continuous-time processes with independent, stationary, non-negative increments.

3.3.4 The plug and play property

The formulation of the latent and observable processes as above is flexible enough to embrace a wide range of individual-based models. In particular, models that describe actual or hypothetical mechanisms of transmission and disease progression are readily formulated in this framework. Moreover, with this formulation, it becomes clear that the models described are partially observed Markov processes (Bretó et al., 2009). This fact makes sequential Monte Carlo methods for likelihood-based inference available for use in the present context. The supplementary material makes the formal connections between this class of models and sequential Monte Carlo methodology.

It is worth noting that models formulated as above are compatible with inference techniques that only require simulation from the model, not closed-form expressions for tran-

sition probabilities. Such algorithms are said to have the *plug-and-play* property (Bretó et al., 2009; He et al., 2010). The particle filter and iterated filtering, which we describe in the Methods section, are two algorithms that have this property. Because these algorithms only require the ability to simulate from the model, they allow for consideration of a wide class of models. Greater freedom in choice of the form of the model allows one to pose scientific questions closed to non-plug-and-play approaches. In the following, we demonstrate this potential in a study of HIV transmission dynamics.

3.4 A model of HIV transmission

Our study focuses on the expanding HIV epidemic among young, black, MSM within the Detroit metropolitan area. Specifically, we ask two questions: (1) How much transmission originates inside the study population relative to that originating outside? (2) Within the study population, how does transmission vary with respect to stage of disease (e.g., early, chronic, AIDS) and diagnosis status? To address these questions we construct a basic model of HIV transmission, similar to that of Volz et al. (2013). We describe our model as a special case of the general class of models described above. This model contains assumptions that can be altered and examined within our methodological framework. In the following, we describe both the form of the model and how we relate it to two data types: diagnosis times and genetic sequences.

3.4.1 The latent and observable processes

The latent state of the system at time t, $\big(\mathcal{T}(t),\ \mathcal{P}(t),\ \mathcal{U}(t)\big)$, is of the form described above. To specify it completely, it remains to describe the Markov process $\{\mathcal{U}(t)\}$ and the transitions of $\{\mathcal{T}(t)\}$ and $\{\mathcal{P}(t)\}$. $\mathcal{U}(t)$ contains information about all infected individuals in the population. Following Volz et al. (2013), we do not explicitly track uninfected individuals and thus disallow depletion of the susceptible pool. There are reasons to suspect that this assumption may be problematic (Kenah et al., 2016), but its adoption here facilitates comparison of our results with those of Volz et al. (2013). Specifically, $\mathcal{U}(t) = \big\{\big(\tau_i, B_i(t)\big) : i \text{ infected at time } t\big\}$, where τ_i is the time at which individual i was infected and $B_i(t) \in \mathbb{C}$, the class of individual i at time t, where $\mathbb{C} = \{I_0, I_1, I_2, J_0, J_1, J_2\}$. $B_i(t) = I_k$ indicates that individual i has an infection at stage $k \in \{0, 1, 2\}$ but has not yet been diagnosed; $B_i(t) = J_k$ indicates that individual i has been diagnosed and has an infection at stage k. We think of $k = 0$ as indicating the early stage of infection; $k = 1$, the chronic stage; $k = 2$, AIDS. Individuals move between classes according to Fig. 3.2. New infections can occur, as can deaths, emigrations, and diagnosis events. Transmission

events, immigration events, deaths, and diagnoses all result in events of the corresponding type being recorded in the structure of $\mathcal{T}(t)$.

New infections arise from two distinct sources: immigration and transmission within the population. Immigrations occur at a constant rate, ψ. Each currently infected individual inside the population seeds new infections at rate ε_c, where $c \in \mathbb{C}$ indicates infection class. Thus, we allow transmissibility to vary between different infection classes, but assume homogeneous transmissibility within each class. It follows that the incidence of new infections is $h(t)+\psi$, where $h(t) = \varepsilon_{I_0} N_{I_0}(t)+\varepsilon_{I_1} N_{I_1}(t)+\varepsilon_{I_2} N_{I_2}(t)+\varepsilon_{J_0} N_{J_0}(t)+\varepsilon_{J_1} N_{J_1}(t)+\varepsilon_{J_2} N_{J_2}(t)$, and $N_c(t)$ is the number of individuals in class c at time t. Defining all nonzero transition rates between states is sufficient to specify a Markov process; a full set of model equations for $\{\mathcal{U}(t)\}$ is presented in the supplement (Section S4).

The inclusion of individual time-of-infection, τ_i, within $\{\mathcal{U}(t)\}$ allows us to model within-host pathogen evolution. In particular, when an individual is diagnosed at time t, a diagnosis node is added to $\mathcal{T}(t)$, together with a diagnosis edge, the length of which is linearly related to how long the diagnosed individual has been infected (Fig. 3.1). This edge may account for sequencing error; it can also describe the emergence of new pathogen strains within a host having reduced between-host transmission potential (Lythgoe and Fraser, 2012).

We assume for simplicity that the topology of $\mathcal{P}(t)$ matches that of $\mathcal{T}(t)$. Thus, we explicitly disallow the possibility of incomplete lineage sorting, though, as mentioned before, this choice is not forced by the algorithm. We assume a relaxed molecular clock: the edge lengths of $\mathcal{P}(t)$ are random. Specifically, each edge of $\mathcal{P}(t)$ has length conditionally Gamma distributed with expectation equal, and variance proportional, to the corresponding edge of $\mathcal{T}(t)$. That is, if L is the length of an edge of $\mathcal{P}(t)$ corresponding to an edge of length D in $\mathcal{T}(t)$, we posit that $L|D$ is Gamma distributed with $\mathbb{E}[L|D=d] = d$ and $\mathrm{Var}[L|D=d] = \sigma\, d$. The parameter σ scales the noise on the rate of evolution. This relaxation, identical to the white noise model of Lepage et al. (2007), is a Lévy process with non-negative increments, as we require. Having specified $\mathcal{P}(t)$, the joint distribution of observed sequences is determined by the choice of the time-reversible molecular substitution model. Here, we used the TN93 model of molecular evolution (Tamura and Nei, 1993). This model distinguishes between the rate of transitions between purines, the rate of transitions between pyrimidines, and the rate of transversions. It is fully specified by the following

rate matrix (see also Table 3.2):

$$Q = \begin{bmatrix} * & \beta\pi_T & \beta\pi_C & \alpha_R\pi_G \\ \beta\pi_A & * & \alpha_Y\pi_C & \beta\pi_G \\ \beta\pi_A & \alpha_Y\pi_T & * & \beta\pi_G \\ \alpha_R\pi_A & \beta\pi_T & \beta\pi_C & * \end{bmatrix}$$

3.5 Results

We present results from both a study on simulated data and an analysis of actual data. The primary goal of the simulation study is to show how our methods can be used to extract information about transmission dynamics from pathogen genetic sequence data within the framework of likelihood-based inference. This study was carried out with 30 sequences of length 100 bases. The goals of the data analysis are to demonstrate the numerical feasibility of our implementation as well as illustrate the role of likelihood-based inference as part of the cycle of data-informed model development for a phylodynamic model. The data analysis was carried out using 100 protease consensus sequences of length 297 bases. Due to the intensive nature of the computations, further developments will be required to handle considerably larger datasets. Some empirical results concerning how our GenSMC implementation scales with number of sequences are given in the supplement (Section S2.3). We discuss applicability to the range of current phylodynamic challenges in the discussion section.

3.5.1 A study on simulated data

Using the individual-based, stochastic model of HIV described above (Fig. 3.2), we set parameters governing the rate of evolution at relatively high values to generate a high proportion of variable sites. As computation scales with the number of variable sites, the computational effort in this simulation study could be comparable to fitting real sequences of greater length. Parameters values and their interpretations are specified in Tables 3.1 and 3.2. Algorithmic parameters are specified in Section S4.2. Each simulated epidemic consisted of a transmission forest and a set of pathogen genetic sequences. We randomly selected 5 epidemics to fit. Each dataset consists of two types of data: times of diagnoses and pathogen genetic sequences. A representative simulated transmission forest and its associated pathogen genetic sequences are shown in Fig. 3.3.

For each of the selected epidemics we ask two questions. First, when all other parameters are known, is it possible to infer ε_{I_0} and ε_{I_1} using only diagnosis times? Second, how does inference change when we supplement the diagnosis data with pathogen genetic sequences? To perform this comparison we estimated two likelihood surfaces for each epidemic: one using only the diagnosis likelihood, and one using both the diagnosis likelihood and the genetic likelihood. We estimated each surface by using the particle filter to compute a grid of likelihood estimates with respect to the two parameters of interest: ε_{I_0}, the infectiousness of early-stage undiagnosed individuals, and ε_{I_1}, the infectiousness of chronic-stage undiagnosed individuals. Equilibrium base frequencies were set to the empirical values in the simulated data. All other parameters were fixed at the known values used for simulation. We extracted grid-based likelihood profiles for each parameter by taking maxima over the columns or rows of the grid. For each parameter we therefore obtained two profiles: one using only the diagnosis likelihood and one using the joint likelihood. The difference in curvature between these profiles tells how much the genetic data improves, or weakens, inference on the parameters.

When only the diagnosis data are used, we find a tradeoff between ε_{I_0} and ε_{I_1} (Fig. 3.4). The diagnoses provide information on upper bounds for each infectiousness parameter, but otherwise only inform their sum. In other words, when estimated using only the diagnosis times, ε_{I_0} and ε_{I_1} are nonidentifiable. Supplementing the data on diagnoses with pathogen genetic sequences resolves this uncertainty (Fig. 3.4). Note that including the genetic data increases noise in the likelihood estimate. This is expected, as computing the likelihood estimate for the genetic sequences requires a numerical approximation to an integral over tree space. Nevertheless, the genetic data increase the curvature of the likelihood surface. From Fig. 3.4, we see that this additional curvature leads to more precise identification of the parameters despite the increased Monte Carlo noise. In principle, Monte Carlo variation can be reduced to negligibility by increased computational effort. This may not be practical when computational expense is high, as it is here. Therefore, it is necessary to bear in mind the tradeoff between the benefits of the information accessed for inference versus the computational burden of extracting this information.

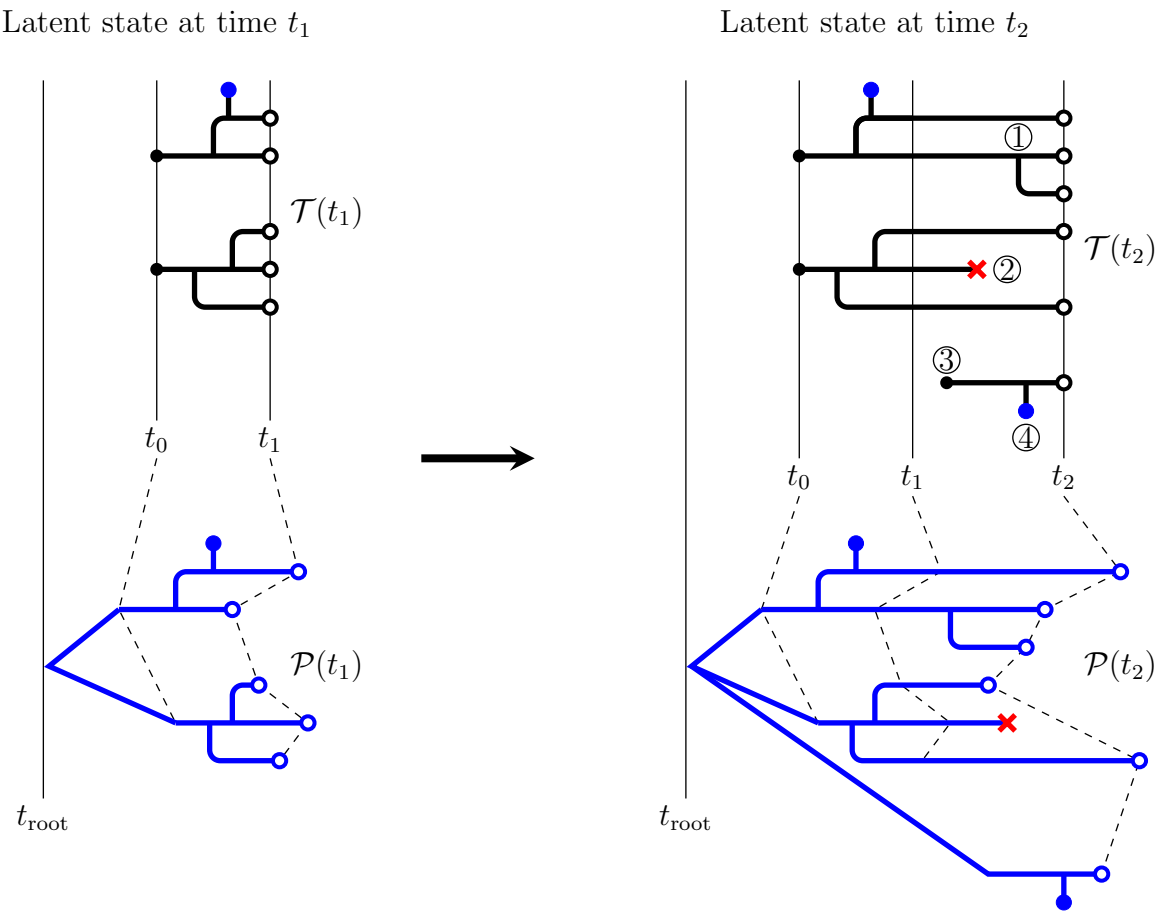

Figure 3.1: A schematic showing the nature and evolution of the latent transmission and phylogeny processes. The transmission forest, $\mathcal{T}(t)$, is shown in black; the pathogen phylogeny, $\mathcal{P}(t)$, in blue. On the left, we see the latent state at time t_1; it evolves by time t_2 to the state shown on the right. At time t_1, $\mathcal{T}(t_1)$ consists of two disconnected trees, representing the transmission histories of five active infections ($\circ$). These infections derive from two infections present at t_0 (black dots). The branching pattern of the pathogen phylogeny mirrors that of $\mathcal{T}(t)$ over the interval $[t_0, t_1]$. This diagram assumes that pathogen lineages branch exactly at transmission events; alternative models could allow for differences in the branching pattern between $\mathcal{T}(t)$ and $\mathcal{P}(t)$. This diagram displays a model with a relaxed molecular clock; randomness in the rate of evolution along lineages is depicted via random edge lengths in $\mathcal{P}(t)$. Over the time interval $[t_1, t_2]$, changes of each of the five permissible types are shown. At ①, an active node splits in two when a transmission event occurs. At ②, an active node becomes a dead node ($\times$) when an infected host emigrates, recovers, or dies. At ③, an immigration event gives rise to a new active node with its own root. At ④, a sequence node ($\bullet$) is spawned when a sample is taken. Finally, active nodes for which none of the above occur simply persist. The Markovian property insists that the latent state at time t_2 be an extension of the latent state at time t_1. In other words, changes to the latent state over the interval $[t_1, t_2]$ must not retroactively modify elements of the latent state prior to time t_1.

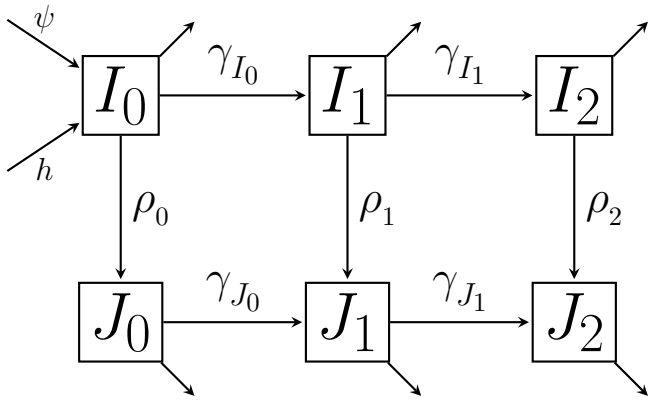

Figure 3.2: A flow diagram showing the possible classes for infected individuals. The columns represent stage of disease: with subscripts 0, 1 and 2 representing early, chronic, and AIDS stages respectively. The rows represent diagnosis status, with the top row representing undiagnosed individuals, I_k, and the bottom row representing diagnosed individuals, J_k, where $k \in \{0, 1, 2\}$. ρ_k are per capita rates of diagnosis and γ_c are rates of disease progression. Arrows out of classes that do not flow into other classes represent the combined flow out of the infected population due to death and emigration.

Table 3.1: Parameters of the transmission model used in simulation of datasets.

Parameter	**Interpretation**	**Value**
ε_{I_1}	Infectiousness of undiagnosed chronic stage individuals	0.25 yr^{-1}
ε_{I_2}	Infectiousness of undiagnosed AIDS individuals	0 yr^{-1}
ε_{J_0}	Infectiousness of diagnosed acute stage individuals	0.125 yr^{-1}
ε_{J_1}	Infectiousness of diagnosed chronic stage individuals	0.025 yr^{-1}
ε_{J_2}	Infectiousness of diagnosed AIDS individuals	0 yr^{-1}
μ_{I_0}	Death rate + Aging rate of undiagnosed acute stage individuals	1/3 yr^{-1}
μ_{I_1}	Death rate + Aging rate of undiagnosed chronic stage individuals	1/3 yr^{-1}
μ_{I_2}	Death rate + Aging rate of undiagnosed AIDS individuals	5/6 yr^{-1}
μ_{J_0}	Death rate + Aging rate of diagnosed acute stage individuals	1/3 yr^{-1}
μ_{J_1}	Death rate + Aging rate of diagnosed chronic stage individuals	1/3yr^{-1}
μ_{J_2}	Death rate + Aging rate of diagnosed AIDS individuals	2/3 yr^{-1}
γ_{I_0}	Progression rate from undiagnosed acute to undiagnosed chronic	1 yr^{-1}
γ_{I_1}	Progression rate from undiagnosed chronic to undiagnosed AIDS	1/6.3 yr^{-1}
γ_{J_0}	Progression rate from diagnosed acute to diagnosed chronic	1 yr^{-1}
γ_{J_1}	Progression rate from diagnosed chronic to diagnosed AIDS	1/6.3 yr^{-1}
ρ_0	Diagnosis rate of acute stage individuals	0.5 yr^{-1}
ρ_1	Diagnosis rate of chronic stage individuals	0.225 yr^{-1}
ρ_2	Diagnosis rate of AIDS individuals	50 yr^{-1}
ψ	Immigration rate of infected individuals	0 yr^{-1}
ϕ	Emigration rate of infected individuals	0 yr^{-1}
t_{root}	Root (polytomy) time	0 yr
t_0	Time to begin simulation of the transmission model	2 yr
t_{end}	Time to end simulation of the transmission model	10 yr
n_{loci}	Length of the sequences to simulate	100 base pairs
p_G	Probability of a sequence given diagnosis	0.48
$N_{I_0}(t_0)$	Number of undiagnosed early-stage individuals at t_0	11
$N_{I_1}(t_0)$	Number of undiagnosed chronic-stage individuals at t_0	15
$N_{I_2}(t_0)$	Number of undiagnosed AIDS individuals at t_0	0
$N_{J_0}(t_0)$	Number of diagnosed early-stage individuals at t_0	4
$N_{J_1}(t_0)$	Number of diagnosed chronic- stage individuals at t_0	8
$N_{J_2}(t_0)$	Number of diagnosed AIDS individuals at t_0	6

Table 3.2: Parameters of the genetic model used in simulation of datasets.

Parameter	**Interpretation**	**Value**
β	Rate of transversions	0.013 yr^{-1}
α_Y	Rate of transitions between purines	0.03 yr^{-1}
α_R	Rate of transitions between pyrimidines	0.1 yr^{-1}
π_A	Equilibrium frequency of adenine	0.37
π_G	Equilibrium frequency of guanine	0.23
π_C	Equilibrium frequency of cytosine	0.18
π_T	Equilibrium frequency of thymine	0.22
σ_{site}	Relaxation of the molecular clock with respect to sites	0
σ	Relaxation of the molecular clock with respect to edges	0.1 yr
δ_{fixed}	The initial component of the sequence stem	0.001 yr
δ_{prop}	Proportion of time since infection to add to the sequence stem	0.05

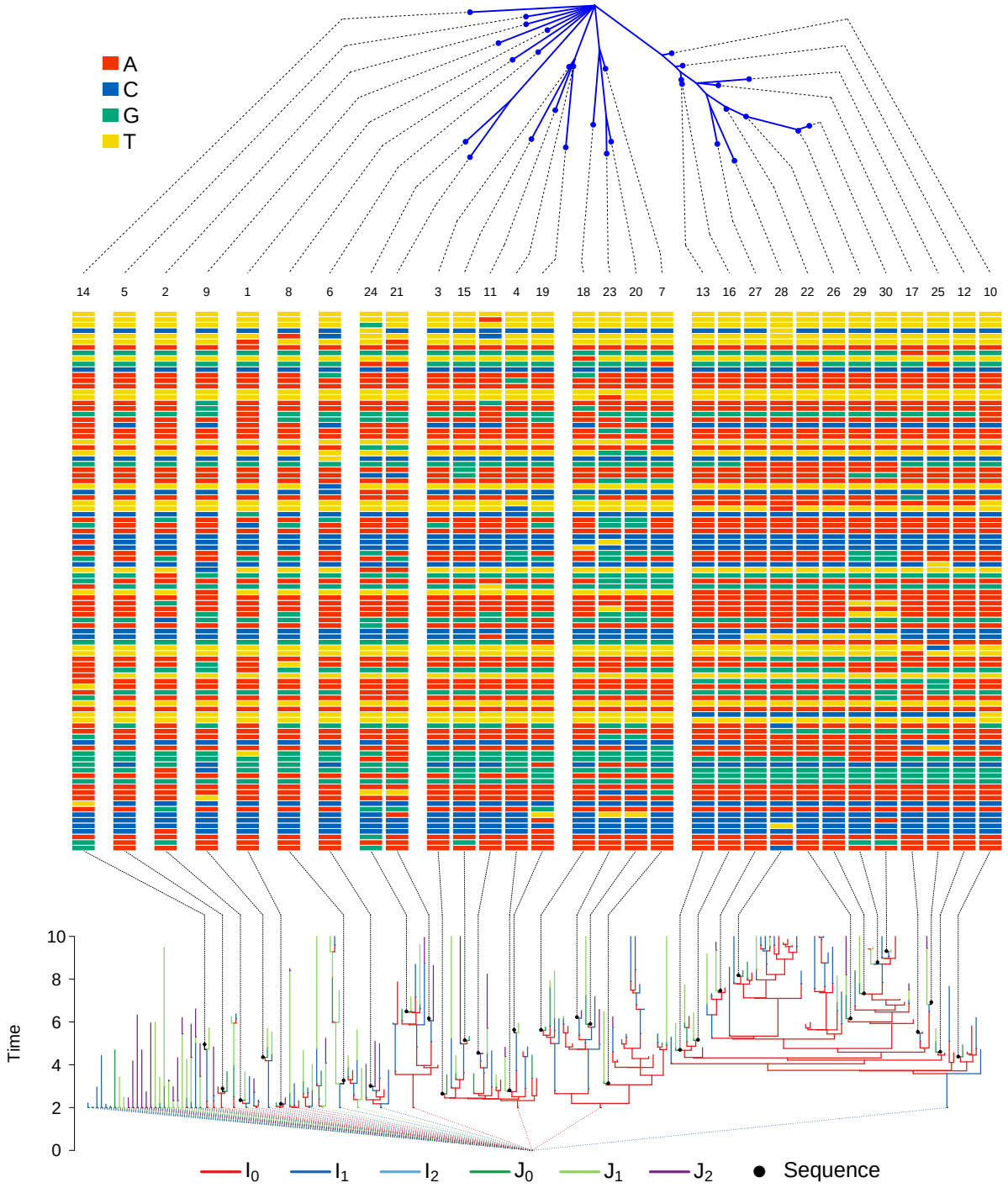

Figure 3.3: A simulated transmission forest (bottom), its associated pathogen genetic sequences (middle), and the phylogeny of the sequences (top). The class of the infected individual in the transmission forest is indicated by its color. Black dots represent genetic sequences. Black dashed lines connect sequence locations on the transmission tree, or the phylogeny, to visualizations of the sequences in the middle panel. Colored dashed lines from the roots of transmission trees connect at the polytomy at $t_{\text{root}} = 0$. Numbers at the top of the sequences indicate the rank of the sequence, with rank 1 being the first observed.

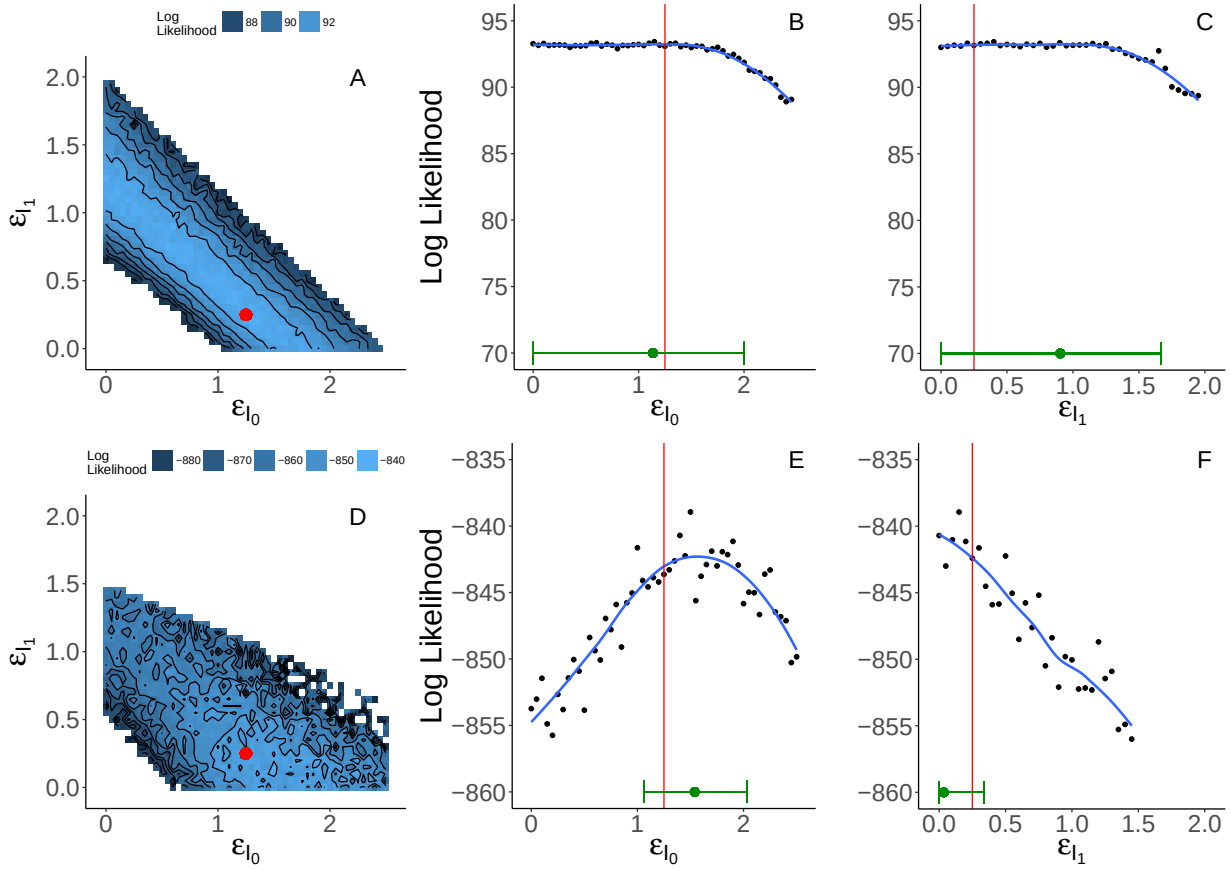

Figure 3.4: Grid-based estimates of likelihood surfaces and likelihood profiles from fitting to simulated data. The top row shows the surface (A) and profiles (B and C) estimated using only the diagnosis likelihood. The bottom row shows the surface (D) and profiles (E and F) estimated using both the diagnosis and the genetic likelihood. Red dots and red lines indicate true values of ε_{I_0} and ε_{I_1} used in simulation. Point estimates and 95% confidence intervals are shown in green just above the horizontal axis of the likelihood profile plots. Confidence intervals for E and F account for both statistical uncertainty and Monte Carlo noise (Ionides et al., 2016) using a square root transformation appropriate for non-negative parameters. Augmenting the diagnosis data with genetic data yields smaller confidence intervals for ε_{I_0} and ε_{I_1}, and resolves the nonidentifiability of these parameters when estimated using only the diagnoses. Note that scales of the likelihood surfaces shown in A and D are not the same; E and F have the same scale as B and C but with a vertical shift.

3.5.2 Analysis of an HIV subepidemic in Detroit, MI

In this data analysis, we explored whether our full-information approach could estimate key transmission parameters using HIV protease consensus sequences and diagnosis times. We focused our analysis on a subepidemic in the young, black, MSM community. The cohort of individuals that we chose to study is shown in Fig. 3.5. See the Materials and Methods section for details on how we selected the subepidemic and cleaned the sequence data.

As in the study on simulated data, we were interested in what the genetic data yield beyond what we can see using the diagnoses alone. Therefore, we again estimated likelihood profiles in two ways: using only the diagnosis data and using both the diagnosis data and the genetic sequences. We estimated likelihood profiles for three parameters of interest: ε_{I_0}, ε_{J_0}, and ψ. In contrast to the simulation study, in this analysis we were faced with a parameter space of much higher dimension. To reduce the dimension of the problem we fixed some parameters: rates of disease progression, rates of diagnosis, and the rate of emigration. Parameters that were fixed and fit are shown in Tables 3.3 and 3.4, respectively. Algorithmic parameters are specified in Section S4.2. For each likelihood profile we first used iterated filtering (Ionides et al., 2015) to maximize the likelihood for a sequence of values that spanned the reasonable range of the parameter. Second, we used the particle filter to estimate likelihoods for each parameter set obtained from iterated filtering. We repeated this process of maximization followed by evaluation until the profile stabilized. All initial-value parameters were fixed, with the exception of t_{root}. Initial counts for individuals in each class were fixed. See the supplement for details on how we arrived at these counts.

When only the diagnosis data are used, we find that the model prefers to explain all infections as originating outside the cohort, with the maximum likelihood estimate (MLE) for $\psi \approx 120$ infections per year (Fig. 3.6). Under this explanation for the data, little or no transmission occurs inside the cohort: this covariate-defined subgroup acts as a sentinel of the broader epidemic. Equivalently, this result would imply that the covariates we used to select these cases do not define a meaningful subepidemic.

On the other hand, when the genetic data are folded in, the estimate of ψ is greatly revised: the MLE for ψ becomes ≈ 6 infections per year. On its face, this is evidence for a low rate of transmission into the cohort and, therefore, evidence that the cohort subepidemic is much more self-contained. Although this may in part be true, the lower estimate of ψ is also potentially driven by assumptions of the genetic model. Supposing, as it does, that all immigrant lineages coalesce at a single, global polytomy, the model insists that sequences from immigrant infections derive from a broad genetic pool. The breadth

of this pool—the average genetic distance between an imported infection and any other observed sequence—is determined by the depth of the polytomy, an estimated parameter. Nevertheless, the low estimate of ψ implies that few infections derive from this broader pool. The model's disallowance of a more structured immigrant pool makes it difficult to say more, however. In particular, the low value of ψ is not inconsistent with the existence of chains of transmission originating within the cohort, leaving it, and returning. Such chains would produce sequence clustering despite the openness of the cohort to transmission. Future work, incorporating genetic and diagnosis information from the broader epidemic will be needed to better quantify the latter effect.

Joint likelihood profiles over ε_{I_0} and ε_{J_0} show support for transmission from both of the early-stage groups, with evidence for higher infectiousness in the early-stage diagnosed class than in the early-stage undiagnosed class. However, it is epidemiologically implausible that diagnosis increases transmission: this is a paradox. Since the paradox did not arise in the simulation study, it cannot be due to a coding error in the implementation of the model or the statistical methodology. Assuming no errors in the data, therefore, it must derive from some inappropriate feature of the model. We propose two possible explanations for how the model and data combine to yield this result.

One possibility is that temporal clusters of genetically related diagnoses favor high infectiousness for the early-stage diagnosed. For example, this could be an artifact of unmodeled clusters in HIV testing. We searched the data for such clusters, but found no conclusive evidence for their presence.

A second possibility is understood by noting that, under the model, any significant amount of transmission from the undiagnosed classes leads necessarily to an exponentially growing accumulation of diagnoses, in conflict with the data. When the genetic data were left out, the model accounted for the observed, roughly linear, ramp-up in diagnoses using immigration, hence the relatively high estimated ψ. Incorporating the genetic data eliminates this option, forcing the model to explain the epidemic's sub-exponential growth as a consequence of diagnosis itself.

To illustrate the second possibility, we estimated likelihood profiles using only the diagnosis likelihood, fixing the immigration rate, ψ, at zero. These profiles show that, when forced to explain the diagnoses without any imported infection, the model prefers to do so by making the early-stage diagnosed class most infectious (Fig. 3.6). This suggests that the model lacks flexibility to explain the pattern in the diagnoses without immigration; this constraint likely limits efficient use of information in the genetic sequences. To remedy this problem, one could modify the model by explicitly introducing a small and ephemeral population of susceptible hosts.

In this methodological paper, we display but one iteration of the scientific method and it is clear that our motivating scientific questions remain incompletely answered. Our principal goal, however, is to illustrate how the methodology facilitates the formulation and testing of scientific hypotheses. For example, the results above suggest a number of straightforward model modifications: the plug-and-play property of the methodology makes it nearly as straightforward to evaluate the evidence for these new hypotheses just as we have done for the old. Moreover, we have shown how probing the data with a mechanistic model can lead to clear identification of flaws in model structure, along with indications for improvements.

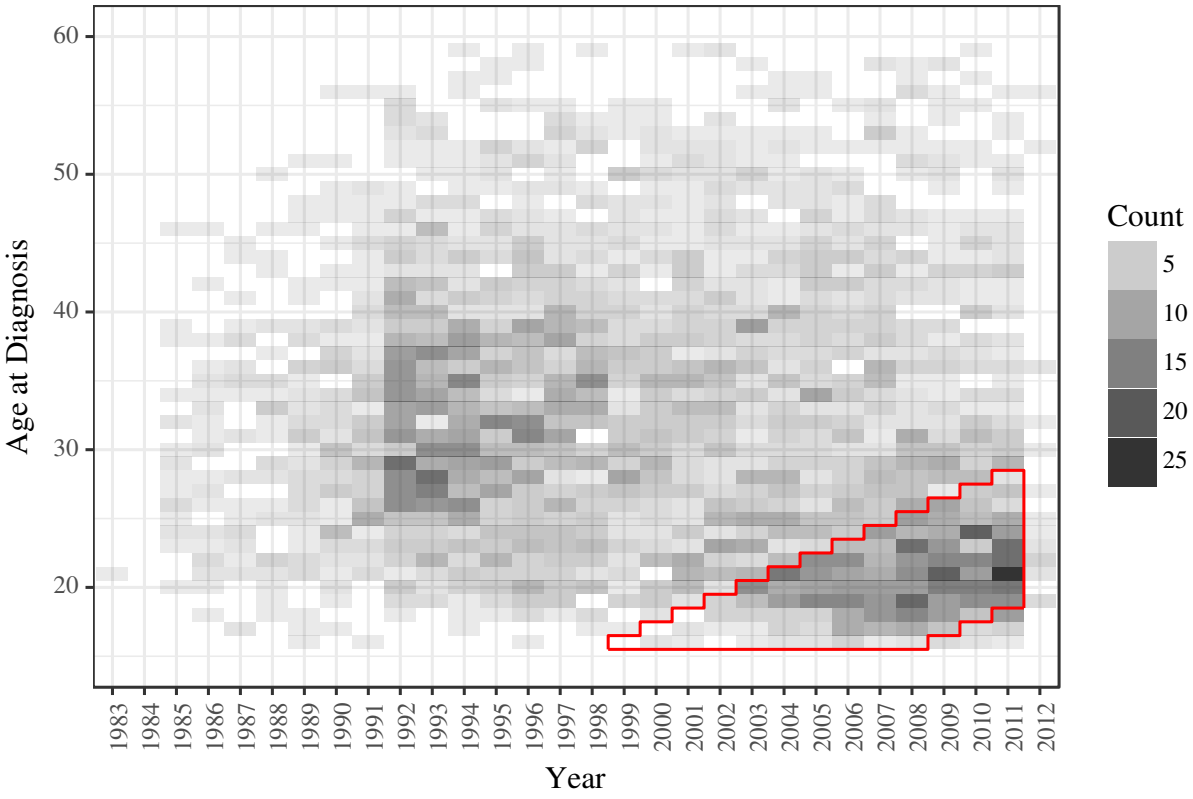

Figure 3.5: The distribution of age at diagnosis through time for black MSM in Detroit, MI. The cohort that we selected for analysis is outlined in red. We excluded the data from 2012 to limit effects from delays in updating the MDCH database. 29 individuals that were diagnosed at ages greater than or equal to 60 years are not shown on this plot.

Table 3.3: Parameters fixed in the data analysis.

Parameter	**Interpretation**	**Value**
μ_{I_0}	Death rate of undiagnosed acute stage individuals	1/70 yr^{-1}
μ_{I_1}	Death rate of undiagnosed chronic stage individuals	1/70 yr^{-1}
μ_{I_2}	Death rate of undiagnosed AIDS individuals	1/2 yr^{-1}
μ_{J_0}	Death rate of diagnosed acute stage individuals	1/70 yr^{-1}
μ_{J_1}	Death rate of diagnosed chronic stage individuals	1/70 yr^{-1}
μ_{J_2}	Death rate of diagnosed AIDS individuals	1/70 yr^{-1}
γ_{I_0}	Progression rate from undiagnosed acute to undiagnosed chronic	1 yr^{-1}
γ_{I_1}	Progression rate from undiagnosed chronic to undiagnosed AIDS	1/6.3 yr^{-1}
γ_{J_0}	Progression rate from diagnosed acute to diagnosed chronic	1 yr^{-1}
γ_{J_1}	Progression rate from diagnosed chronic to diagnosed AIDS	1/6.3 yr^{-1}
ρ_0	Diagnosis rate of acute stage individuals	0.225 yr^{-1}
ρ_1	Diagnosis rate of chronic stage individuals	0.225 yr^{-1}
ρ_2	Diagnosis rate of AIDS individuals	50 yr^{-1}
ϕ	Emigration rate of infected individuals	0 yr^{-1}
$N_{I_0}(t_0)$	Number of undiagnosed early-stage individuals at t_0	20
$N_{I_1}(t_0)$	Number of undiagnosed chronic-stage individuals at t_0	36
$N_{I_2}(t_0)$	Number of undiagnosed AIDS individuals at t_0	0
$N_{J_0}(t_0)$	Number of diagnosed early-stage individuals at t_0	4
$N_{J_1}(t_0)$	Number of diagnosed chronic- stage individuals at t_0	22
$N_{J_2}(t_0)$	Number of diagnosed AIDS individuals at t_0	16
σ_{site}	Relaxation of molecular clock with respect to sites	0 yr
t_0	Time to start filtering	1 Jan 2004

Table 3.4: Parameters fit in the data analysis. We present confidences intervals for parameters for which we computed likelihood profiles. For all other parameters, we present only the point estimate. This table is continued in Table 3.5.

Parameter	**Interpretation**	**Diagnosis data**	**Diagnosis data and genetic sequences**	**Diagnosis data, with ψ fixed at 0**
ψ	Immigration rate of infected individuals	120 (104, 134) yr^{-1}	5.82 (2.55, 11.2) yr^{-1}	0 yr^{-1}
ε_{I_0}	Infectiousness of undiagnosed acute stage individuals	0 (0, 0.413)	0.257 (0.0399, 0.623)	0 (0, 0.192)
ε_{I_1}	Infectiousness of undiagnosed chronic stage individuals	0.0042	0.00048	0.0056
ε_{I_2}	Infectiousness of undiagnosed AIDS individuals	0	0	0
ε_{J_0}	Infectiousness of diagnosed acute stage individuals	0.0675 (0, 1.17)	3.36 (3.13, 4.2)	7.34 (5.78, 9.25)
ε_{J_1}	Infectiousness of diagnosed chronic stage individuals	0.0089	0.17	0.032
ε_{J_2}	Infectiousness of diagnosed AIDS individuals	0	0	0

Table 3.5: Parameters fit in the data analysis (continued).

Parameter	**Interpretation**	**Diagnosis data**	**Diagnosis data and genetic sequences**	**Diagnosis data, with ψ fixed at 0**
β	Rate of transversions	-	0.0042 yr^{-1}	-
α_Y	Rate of transitions between purines	-	0.047 yr^{-1}	-
α_R	Rate of transitions between pyrimidines	-	0.043 yr^{-1}	-
π_A	Equilibrium frequency of adenine	-	0.37	-
π_G	Equilibrium frequency of guanine	-	0.24	-
π_C	Equilibrium frequency of cytosine	-	0.18	-
π_T	Equilibrium frequency of thymine	-	0.21	-
σ	Relaxation of molecular clock with respect to edges	-	2 yr	-
δ_{prop}	Proportion of time since infection to use for diagnosis edge	-	0.064	-
δ_{fixed}	Amount of calendar time to add on to diagnosis edge	-	0.00049 yr	-
t_{root}	Time of the polytomy that joins all genetic lineages	-	27 Aug 2000	-

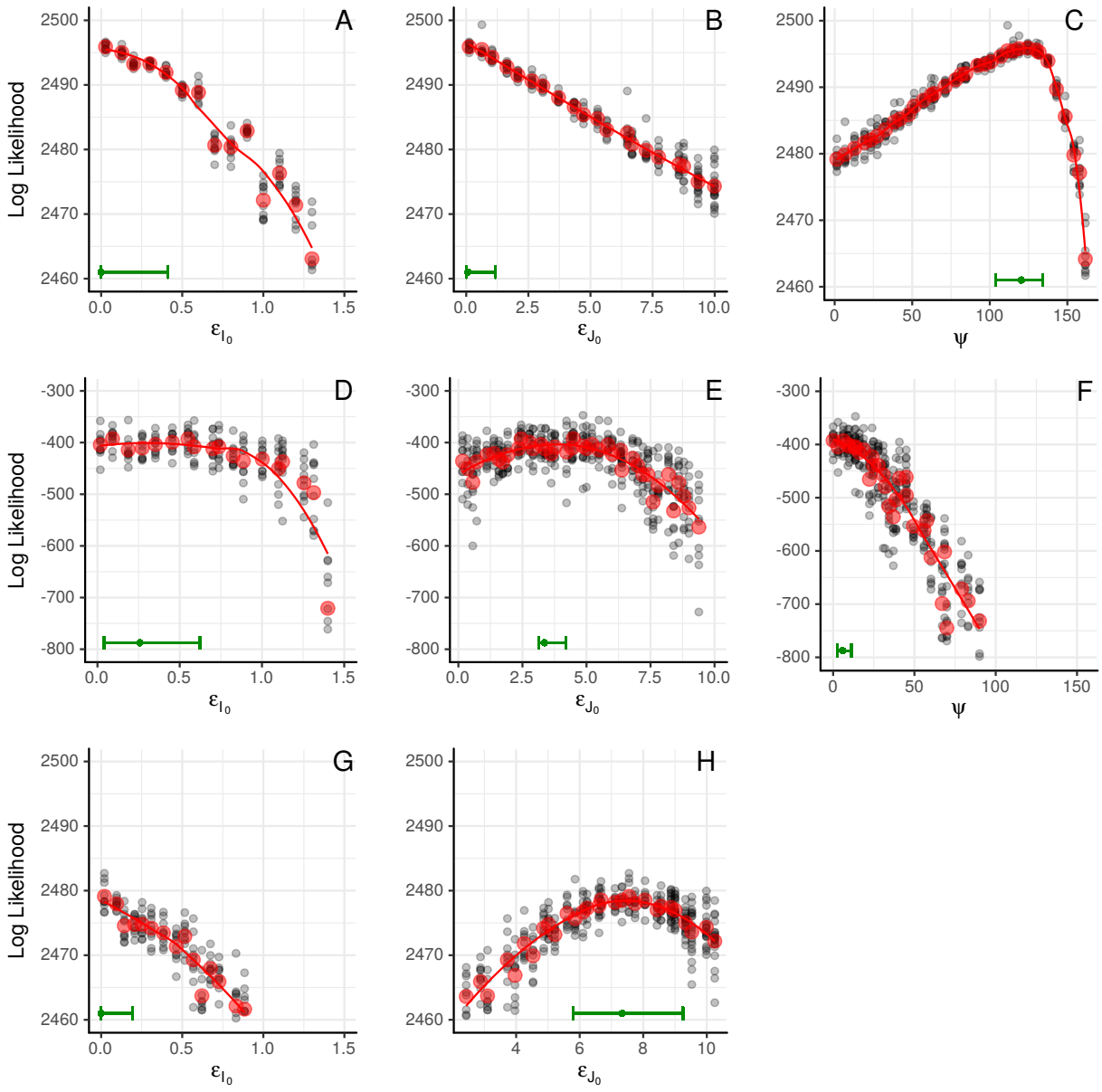

Figure 3.6: Estimated likelihood profiles from fits to data from the black, MSM cohort. A-C show likelihood profiles computed using only the diagnosis likelihood. D-F show likelihood profiles computed using both the diagnosis likelihood and the genetic likelihood. G and H show likelihood profiles computed using only the diagnosis likelihood when ψ is fixed at zero. Black dots represent particle filter likelihood evaluations of parameter sets obtained using iterated filtering. Red dots represent mean log likelihoods of the multiple likelihood evaluations (black dots) at each point in the profile. Red lines are loess fits to the red dots. Green bars along the lower margin of each panel encompass 95% confidence intervals for each parameter. Confidence intervals account for both statistical uncertainty and Monte Carlo noise (Ionides et al., 2016). The smoothed profile was calculated on the square root scale, appropriate for non-negative parameters, with a green dot indicating the maximum.

3.6 Discussion

We demonstrated, via a simulation study, that our algorithms provide access to the likelihood surface of a population dynamic model fit to genetic sequence data. This opens the door to likelihood-based phylodynamic inference. As this study shows, incorporating information from genetic data has the potential to improve on inference that we obtain using diagnosis data alone.

In our analysis of an HIV subepidemic in Detroit, MI, we showed that our methods can be used to ask questions of current public health interest by fitting practical models to data of nontrivial size. This study illustrates how the ability to confront the model with different data types, alone or in combination, can be essential to understanding how the model interacts with the data, to uncovering shortcomings of the model, and to pointing the way toward improved model formulations. The ability of our methods to incorporate different data types made it possible to assess each source of information's contribution to the overall inference. In turn, the ability to easily restructure the model, guaranteed by the plug-and-play property, will allow us to push forward model development.

The scope of our methodology goes beyond the examples presented: the algorithms described here are applicable across a wide range of host-pathogen systems and may find application in realms beyond genetics. From an abstract perspective, these algorithms provide the ability to relate demographic processes with a growing tree-like structure to the evolution of discrete characters that are carried and passed along the branches of that tree. So long as this evolution occurs on a similar timescale to that of the demographic process, and measurements of the discrete process are heterochronous, the methods presented here apply.

In this paper, we demonstrated the methods using relatively short consensus sequences derived from Sanger sequencing. While our methods may be well suited to analysis of data from fast-evolving RNA viruses, they may also apply in studies of pathogens that evolve more slowly. Advances in sequencing are increasing the range of problems for which phylodynamic inference is applicable (Biek et al., 2015). The ability to apply phylodynamic inference to bacterial and protozoan genomes opens the door to many epidemiological applications. One area that may be particularly interesting to explore using our methods is hospital outbreaks of drug resistant bacteria. Hospital records on location and duration of stay may provide fine-scale information on populations of susceptible and infected individuals. Accurate measures of these demographic quantities may allow for efficient use of information held in genetic data. Furthermore, the relatively small size of outbreaks in hospitals means that stochasticity may play a large role in their dynamics, and our methods

are designed to explicitly account for the role of different sources of stochasticity.

We conclude by placing our new methodology in the context of the eight current challenges identified by Frost et al. (2015) for inferring disease dynamics from pathogen sequences. We will make some relevant comments on each challenge, in order.

1. **Accounting for sequence sampling patterns**. Our methodology explicitly models sequence sampling. The chance of an individual being diagnosed, or subsequently having their pathogen sequenced, is permitted to depend on the state of the individual. This state could contain geographic information, or whatever other aspect of the sampling procedure one desires to investigate. Sampling issues revolve around how the dynamics and the measurement process affect the relatedness of sequences, and are more naturally handled in a framework that deals jointly with estimation of the population dynamics and the phylogeny. Thus, our main innovation of joint estimation is directly relevant to this challenge.

2. **Using more realistic evolutionary models to improve phylodynamic inferences**. In this paper, we have used simple evolutionary models that have been widely used for previous phylodynamic inference investigations. Our methodology does not particularly facilitate the use of more complex evolutionary models, since the large number of trees under consideration puts a premium on rapid likelihood computation. However, our methodology is primarily targeted at drawing inference on the population dynamics rather than the micro-evolutionary processes. For this purpose, it may be sufficient to employ an evolutionary model which captures the statistical relationship between genetic distance and temporal distance on the transmission tree, together with an appropriate estimate of the uncertainty in this relationship. Better evolutionary models would be able to extract information more efficiently from the data, but from our perspective this challenge may not be a primary concern.

3. **The role of stochastic effects in phylodynamics**. Our methodology explicitly allows for stochastic effects in the population dynamics and sequence collection.

4. **Relating the structure of the host population to pathogen genetic variation**. Our framework explicitly models this joint relationship. Further scientific investigations, fitting models using methods accounting properly for the joint relationship, will lead to progress in understanding which aspects of dynamics (such as super-spreading) might be especially important to include when carrying out phylodynamic inference.

5. **Incorporating recombination and reassortment**. In principle, our methodology is flexible enough to include co-infection and its evolutionary consequences. Due to computational considerations, it will be important to capture parsimoniously the key aspects of these processes.

6. **Including phenotypic as well as genotypic information**. Our framework naturally combines genotypic information with other information sources. For example, in our data analysis we complemented genetic sequence data with diagnosis times for unsequenced patients.

7. **Capturing pathogen evolution at both within-host and between-host scales**. The diagnosis edges on our phylogenetic tree allow for differences between observed and transmissible strains, and therefore give a representation of within-host diversity or measurement noise. Other approaches to within-host pathogen diversity are possible within our general framework. For example, one could include within-host branching of the phylogenetic tree. More complete investigation of within-host pathogen dynamics will require additional modeling. Due to the larger models and datasets involved, applying our methodology to such investigations will require further methodological work on scaling.

8. **Scaling analytical approaches to keep up with advances in sequencing**. In this manuscript, our goal was to develop generally applicable and statistically efficient methodology. Our methodology is structured with computational efficiency in mind, subject to that goal. Our approach combines various algorithms that have favorable computational properties: peeling, particle filtering with hierarchical resampling and just-in-time variable construction, and iterated filtering. There is scope for computational enhancement by adapting the methodology to high performance architectures. In particular, parallel particle filtering is an active research topic (Paige et al., 2014) that is directly applicable to our methodology. There are also possibilities for improving scaling by imposing suitable situation-specific approximations; for example, it might be appropriate to reduce the computational burden by supposing that some deep branches in the phylogeny are known.

In summary, our new methodology has potential for making progress on many of the challenges identified by Frost et al. (2015). Beyond that, the methodology offers a full-information, plug-and-play approach to phylodynamic inference that gives the scientist flexibility in selecting appropriate models for the research question and dataset at hand. Although technical challenges remain, especially in scaling these methods to large data,

these algorithms hold the potential to ask and answer questions not accessible by alternative approaches.

3.7 Materials and methods

3.7.1 Overview of sequential Monte Carlo estimation of the likelihood

Sequential Monte Carlo (SMC) is a family of stochastic algorithms originally designed to estimate imperfectly observed states of a system via a collection of dynamically interacting simulations (Arulampalam et al., 2002). Each such simulation is called a *particle*; SMC is often referred to as the *particle filter*. The simplest SMC algorithm sequentially estimates the latent state at the time of each observation by iteratively repeating three steps: (1) for each particle, simulate the latent process forward in time to the next data point, (2) for each particle, compute the conditional probability density of the observation given the proposed latent state, and (3) resample the particles with replacement with probabilities proportional to their conditional probabilities. While inference of unobserved states is one use of the particle filter, we are primarily interested in using the filter for likelihood estimation. The average of the conditional likelihoods across particles is an estimator of the conditional likelihood of each observation, and the product of these conditional likelihoods is an unbiased estimator of the full likelihood of the data (Theorem 7.4.2 on page 239 in Del Moral, 2004).

The basic particle filter described above requires only the ability to simulate realizations of the latent state and to evaluate the density of an observation given the latent state. As explained above, in the present case, the latent state contains both the full transmission forest and the phylogeny of the pathogen lineages. At minimum, the observations consist of a time-ordered set of pathogen genetic sequences. Although in principle these methods could be applied to homochronous sequences, we primarily envision using them to fit models to heterochronous sequences. Additional datatypes can be incorporated into the likelihood evaluation if desired so long as there is a means to relate these data to the latent state.

We implemented the particle filter such that the algorithmic code is independent of the code that specifies the model. This structure allows for realizing the advantanges of the plug-and-play paradigm by facilitating quick comparisons between models of different forms. Pseudocode for the algorithm is provided in the supplement.

In our framework, the user specifies the model by writing three functions:

1. **A simulator for the initial state of the latent process.** This function initializes $\mathcal{T}(t_0)$ and $\mathcal{U}(t_0)$. For example, in a model with only one class of infected individuals,

this function would initialize $\mathcal{T}(t_0)$ by specifying the number of infected individuals at t_0. Additional information about the states of those individuals may be contained in $\mathcal{U}(t_0)$. Each of these individuals then becomes a root of a tree in the transmission forest. Each root of the transmission forest has its own genetic lineage; these comprise $\mathcal{P}(t_0)$. In our implementation, the initializer does not construct $\mathcal{P}(t_0)$; the structure of $\mathcal{P}(t)$ is built as needed (see below in the section 'Just-in-time construction of state variables').

2. **A forward simulator for the latent state.** This function simulates $\mathcal{T}(t)$ and $\mathcal{U}(t)$ forward in time from one observation to the next. This function also places the next observation on $\mathcal{T}(t)$, assigning the sequence to an individual by augmenting $\mathcal{T}(t)$ with a diagnosis edge and a sequence node. Note that this function does not simulate evolution of genetic sequences. Rather, the algorithm proposes ancestral relationships between genetic sequences via the simulated transmission forest. While formally, the pathogen phylogeny $\mathcal{P}(t)$ is part of the latent state, for computational efficiency we choose not to simulate its structure in full. The function in (3) builds the necessary components of $\mathcal{P}(t)$ given the simulated transmission forest and placement of sequences on the forest.

3. **An evaluator for the conditional probability of observing a sequence.** This function returns the conditional probability of observing a sequence given the latent state and all previously observed sequences. In particular, this function conditions on the structure of the subtree of $\mathcal{P}(t)$ that connects the observed sequences. The simplest choice for this function is to (1) make the strong assumption that $\mathcal{P}(t)$ maps directly onto $\mathcal{T}(t)$, and therefore build the phylogeny based strictly on the topology of $\mathcal{T}(t)$ and (2) evaluate the conditional likelihood of the genetic sequence using the peeling algorithm (Felsenstein, 1981). These two choices are equivalent to assuming a strict molecular clock. However, one may choose more complicated functions, such as mappings that allow for discrepency between $\mathcal{T}(t)$ and $\mathcal{P}(t)$ or a relaxed molecular clock, to better match the mechanistic processes that generate real data. The branching pattern of the transmission forest and of the phylogeny may differ for a number of reasons (Romero-Severson et al., 2014), so there may be strong arguments for allowing for discrepency between these trees.

Algorithm 2: GenSMC [Corresponding step numbers for the complete description in Section S2 are in brackets]

input: simulator for the initial state; a dynamic model; diagnosis times; genetic sequence data; number of particles; number of nested particles; number of relaxed clock samples.

initialize filter particles [step 1]

for each diagnosis time **do** [step 2]

simulate particles through to next diagnosis time [steps 3, 5]

propose multiple candidate individuals for the next diagnosis [steps 6, 7]

propose multiple relaxed clock edge lengths for each candidate assignment [steps 8-11]

compute particle weights: the probability density of the diagnosis and sequence [steps 4, 12, 13]

resample according to particle weights [steps 14-21]

compute conditional log likelihood [step 22]

end for output: log likelihood estimate; latent states estimates.

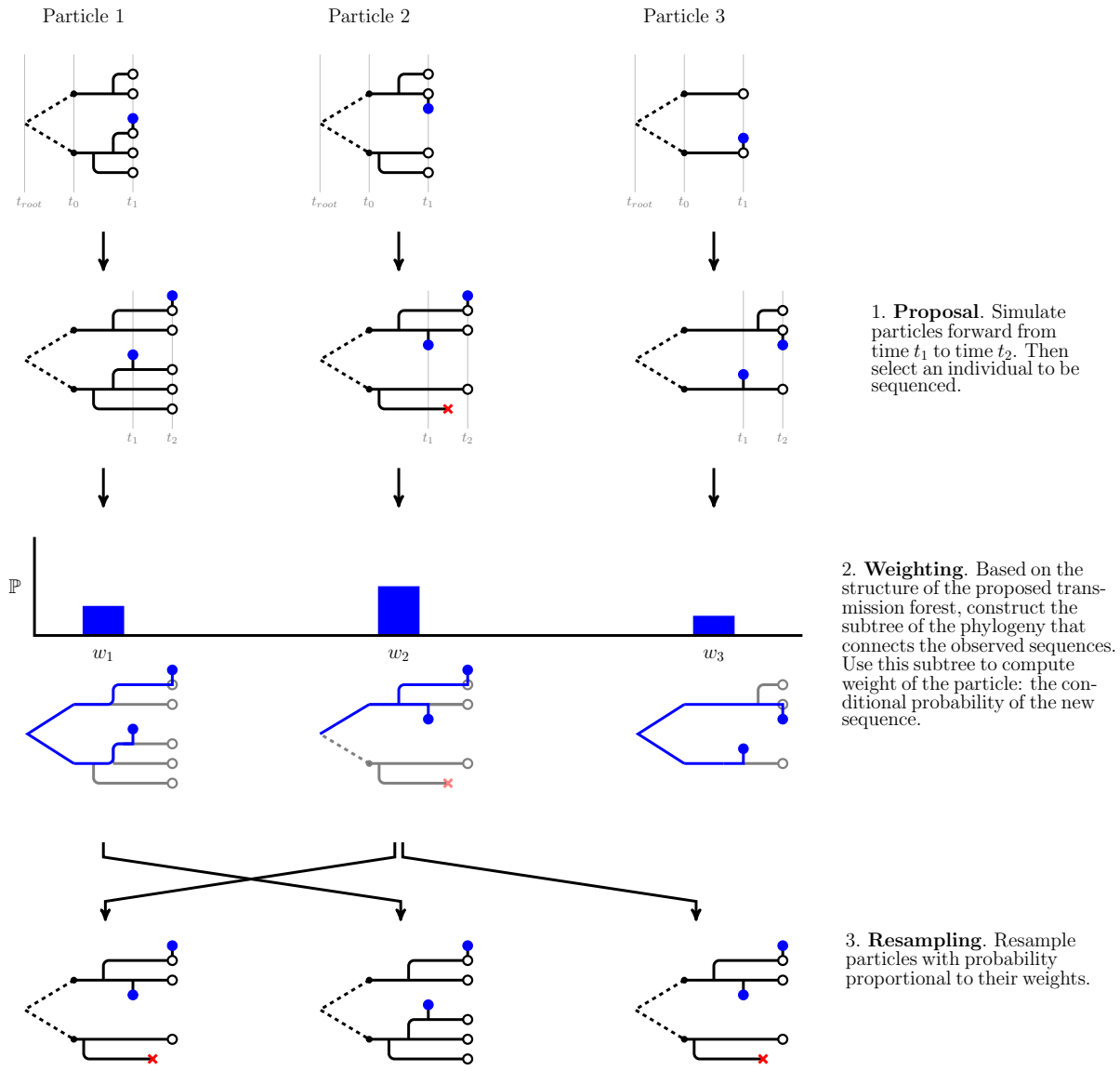

Figure 3.7: A schematic of the particle filter. Here, we show steps to run the filter from the first sequence to the second. Transmission forests are shown in black and phylogenies that connect observed sequences, $\widetilde{\mathcal{P}}(t)$, are shown in blue. Observed sequences are depicted as blue dots. This schematic shows how the algorithm uses *just-in-time* construction of state variables to ease computational costs. Although the model describes how $\mathcal{P}(t)$ relates to $\mathcal{T}(t)$ across all branches of the transmission tree, the algorithm only constructs the subtree of the phylogeny needed to connect the observations (and therefore evaluate conditional probabilities of sequences). Note that in our implementation of the particle filter we introduce additional procedures in the proposal and weighting steps. These procedures, which are detailed below, allow for more accurate assessment of a particle's weight (through hierarchical sampling) and estimation of the conditional probability of a sequence under a relaxed clock. In our current implementation (Algorithm 3), assimilation of each data point is followed by systematic resampling (Arulampalam et al., 2002; Douc et al., 2005); future developments may aim to increase efficiency further using alternative resampling schemes.

3.7.2 Maximization of the likelihood via iterated filtering

The particle filter provides access to the likelihood surface, but it does not provide an efficient way to maximize the likelihood. A closely related class of algorithms, iterated filtering, allows for maximizing the likelihood. Iterated filtering incorporates perturbation of unknown parameters into the particle filter. Repeatedly passing the filter over the data while shrinking the size of the perturbations allows the parameters to converge to their maximum likelihood estimates. The setup here, with the use of just-in-time construction of unobserved states, does not perfectly match the framework used to develop iterated filtering by Ionides et al. (2015). However, the basic iterated filtering approach of perturbing parameters and filtering repeatedly can be applied, and can be assessed on its empirical success at maximizing the likelihood.

3.7.3 Computational Structure

One way our algorithms differ from a standard SMC approach is that each particle maintains a latent state comprising of tree structures that reach back to t_{root}. As the algorithm incorporates each additional data point its memory requirement grows. From a practical perspective, the necessity of maintaining a deep structure in the particles presents challenges for writing a computationally feasible implementation of the algorithm. We developed several innovations to meet the computational challenges posed by numerically integrating over tree space. In this section, we give an overview of key components of our implementation that contributed to numerical tractability. For details, see the source code at `https://github.com/kingaa/genpomp` (to be archived at datadryad.org).

3.7.3.1 Data structures and their relationship to model specification

Our implementation holds two tree structures in memory for each particle: (1) $\mathcal{T}(t)$, the transmission tree, and (2) $\widetilde{\mathcal{P}}(t)$, the subtree of $\mathcal{P}(t)$ that connects all sequences observed up to time t. We represent $\mathcal{T}(t)$ as a vector of nodes, where each node contains the index of its mother, a timestamp, and the index of the genetic lineage with which it is associated (if any). Although the model of the latent state includes the full phylogeny of the pathogen, $\mathcal{P}(t)$, our algorithms only need to keep a subtree of the phylogeny, $\widetilde{\mathcal{P}}(t)$, in memory. We also represent $\widetilde{\mathcal{P}}(t)$ as a vector of nodes. However, nodes of $\widetilde{\mathcal{P}}(t)$ require more memory than the nodes of $\mathcal{T}(t)$. In addition to the information in a transmission tree node, each node of $\widetilde{\mathcal{P}}(t)$ contains the indices of the node's daughters, an array of probabilities, and an evolutionary edge length. These additional components allow for computing the likelihood of observing the sequences at the tips of $\widetilde{\mathcal{P}}(t)$.

Our implementation provides a set of functions that allow for specifying the model via forward-in-time simulation of the latent state. These functions provide access to the latent state and allow for modifying the latent state by branching lineages in $\mathcal{T}(t)$, terminating leaves in $\mathcal{T}(t)$, etc. Our code does not provide access to $\widetilde{\mathcal{P}}(t)$. Instead, internal functions update the structure of $\widetilde{\mathcal{P}}(t)$ as necessary (detailed in the following section on just-in-time construction of state variables). The structure of $\widetilde{\mathcal{P}}(t)$ is in part determined by the molecular clock model. Our current implemention supports strict molecular clock models and relaxed molecular clocks with gamma distributed edge lengths (as we use in this paper). Alternative models for $\mathcal{P}(t)$ are possible, and the plug-and-play structure of our algorithms allows the user to explore a wide range of alternative models.

3.7.3.2 Just-in-time construction of state variables

Although the model of the latent process includes the full phylogeny of the pathogen, $\mathcal{P}(t)$, for the purposes of computation we need only store $\widetilde{\mathcal{P}}(t)$ in memory. In our implementation, we add new edges to $\widetilde{\mathcal{P}}(t)$ at the time of measurement; it is not until a sequence is placed on a lineage of $\mathcal{T}(t)$ that we have enough information to update $\widetilde{\mathcal{P}}(t)$. We call this approach *just-in-time* construction of state variables because simulation of part of the state is postponed until the last moment. An alternative approach would include simulation of $\mathcal{P}(t)$ in tandem with the transmission forest. Then, when a sequence is attached to $\mathcal{T}(t)$ the necessary components of $\mathcal{P}(t)$ to relate the new sequence to all previously observed sequences would be guaranteed to be present. When the transmission forest is large relative to the phylogeny such an approach would be costly in both computation and memory.

3.7.3.3 A hierarchical sampling scheme

We developed a hierarchical sampling scheme to allow for scaling the effective number of particles while holding only a fraction of the effective number of particles in memory. This sampling scheme allows for holding J particles in memory while approaching effective sample sizes approaching JK, where J is the number of base particles and K is the number of nested particles. In this hierarchical scheme, we split the proposal into two steps: (1) proposal of the transmission forest, and (2) proposal of the location of the sampled sequence on the transmission forest. Each of J particles first proposes a transmission forest. Then each of the J particles calculates the likelihood of the observed sequence for K possible locations of the observed sequence (Fig. 3.8). One of the K nested particles is kept, sampled with weight proportional to its conditional likelihood, and the remaining $K-1$ particles are discarded. The weight of the surviving particle is the average of the

conditional likelihoods of the K nested particles.

3.7.3.4 A Monte Carlo procedure for the relaxed molecular clock

As we have no closed-form expression for the conditional probability of an observed sequence under a relaxed clock, we estimate this probability via simulation. Fig. 3.9 shows how we incorporate this Monte Carlo procedure into our SMC framework. We generate L instances of the subtree of the phylogeny that connects all previously observed sequences up to time t, $\widetilde{\mathcal{P}}(t)$. We then augment each subtree with an edge to accommodate the new sequence. The length of this edge is gamma distributed as described above. When connecting the new edge to the existing phylogeny, there are two cases: either the edge connects at the root or the new edge splits an existing edge. In the case of a split edge, we allocate edge length to either side of the split according to a beta distribution. This procedure maintains gamma distributed edge lengths. Having constructed the phylogeny connecting all sequences up to the new sequence, we then use the peeling algorithm (Felsenstein, 1981) to compute the conditional probability of the new sequence. The average of the conditional probability given each of the L subtrees is an estimate of the conditional probability of the new sequence under a relaxed clock.

3.7.3.5 Parallelization

We used openMP (Dagum and Menon, 1998) to parallelize the algorithm at the level of a single machine to reduce runtimes. In particular, we parallelized the outer loop of the hierarchical sampling scheme described above. Each processor handles one base particle at a time. The cost in memory for n processors handling J particles with a nested sample size of K is therefore at worst $J + nK$, as each processor may have at most K additional particles in memory.

3.7.4 A model of HIV transmission: computation of the measurement model

Each diagnosis event consists of a diagnosis time and, possibly, an associated genetic sequence. In the case where the diagnosis event has no sequence, the measurement model is only the conditional density of the diagnosis time. When there is an associated sequence, it is the product of the conditional density of the diagnosis time and the conditional probability of the genetic sequence.

We compute the conditional density of a diagnosis time as follows. We decompose the density into two terms: (1) The probability of no diagnosis over the last interdiagnosis

interval: $\exp\left(-\sum_{k=0}^{2}\Lambda_k\right)$ where Λ_k is the cumulative hazard of a diagnosis from class I_k, $k \in \{0,1,2\}$. That is, $\Lambda_k = \rho_k \sum_{r=1}^{R} \delta_r N_{I_k,r}$, where, ρ_k is the diagnosis rate for class I_k, δ_r is length of the r^{th} subinterval in the interdiagnosis interval over which the count of class I_k, $N_{I_k,r}$, is constant, and (2) the hazard of a diagnosis at the time of diagnosis: $\sum_{k=0}^{2}\rho_k N_{I_k}$. The conditional density of a diagnosis time is the product of these two quantities, and is therefore a mixture of a probability and a density. To compute the first, each particle accumulates the person-years of undiagnosed individuals over the last diagnosis interval (Fig. 3.10). The second is easily computed given the number of each class of undiagnosed individual at the time of diagnosis.

The conditional probability of a genetic sequence is the probability of observing that sequence given the latent state of the system and all previously observed sequences. Our Monte Carlo approach for computing this probability under a relaxed clock is detailed in the 'Computational Structure' section.

3.7.5 Data analysis methods: the sequence data

We preprocessed the sequence data following Volz et al. (2013) to facilitate comparision with that work. We excluded poor quality sequences and recombinant sequences, and accounted for known sources of selection. We first aligned all sequences to the reference sequence for the pol gene of HIV subtype-B. We then masked known drug resistant sites, as specified in the Stanford database of HIV drug resistance (Bennett et al., 2009). We used the program HyPhy (Pond et al., 2005) to identify the type of each sequence and then excluded recombinant sequences and nonsubtype-B sequences. Many individuals in the dataset have multiple sequences. To limit the complexity of the problem, we chose to keep only first available sequences that were collected within one year of diagnosis. Our methods could, in principle, allow for multiple sequences from each individual. However, this extension has not yet been implemented. We took the time of diagnosis as the time of sequencing – for most sequences this is a reasonable approximation. Poor quality sequencing often manifests as sequences with clipped ends. We therefore considered the length of a sequence as a proxy for quality, and we excluded sequences whose concatenated length was shorter than 1100 base pairs.

3.7.6 Data analysis methods: selecting a subepidemic

The Michigan Department of Community Health (MDCH) maintains an extensive dataset on HIV positive individuals living in the state of Michigan. This dataset stretches

back to the beginnings of the HIV epidemic in the United States, and includes over 30,000 diagnoses and nearly 9,000 genetic sequences. Analysis of the full dataset is beyond the scope of our current implementation. Further developments, possibly including preliminary splitting of the full phylogeny into clusters, will be necessary to apply our methods to larger-scale situations. We therefore selected a subset of the cases based on a number of clinical covariates. We chose to focus on the young, black, MSM, subepidemic, which has been of recent concern in Detroit and elsewhere in the USA (Maulsby et al., 2014). In selecting this subset, one of our goals was to choose a well-defined subpopulation. We selected records of individuals from the the MDCH dataset that met the following criteria: black, MSM, known not to be an intravenous drug user, and diagnosed in one of 10 counties that comprise the Detroit Metropolitan Area. For this subpopulation, the distribution of the age at diagnosis through time shows striking patterns. In particular, it there is evidence for cohorts of infected individuals that may be clusters of transmission within the young, black, MSM community. We selected a cohort from this population that may represent such a cluster of transmission: individuals that were between the ages of 19 and 28 inclusive in the year 2011 (a span of 10 years) and were diagnosed between 1 January 1999 and 31 December 2011 (Fig. 3.5). We selected this particular cohort of individuals because it contains what appears to be a pulse of transmission, and because it coincides with when we have high rates of sampling for the genetic sequence data. Counts of individuals diagnosed between 1 January 1999 and 31 December 2003 were used to determine initial conditions (detailed in the supplement). We fit models to data from 1 January 2004 to 31 December 2011. This portion of the cohort has 709 diagnoses and 253 primary genetic sequences. We subsampled the genetic sequences, randomly selecting 100 sequences to keep in the analysis. For the current implementation of our methodology, and in the context of this HIV model, 100 sequences was around the limit of computational tractability.

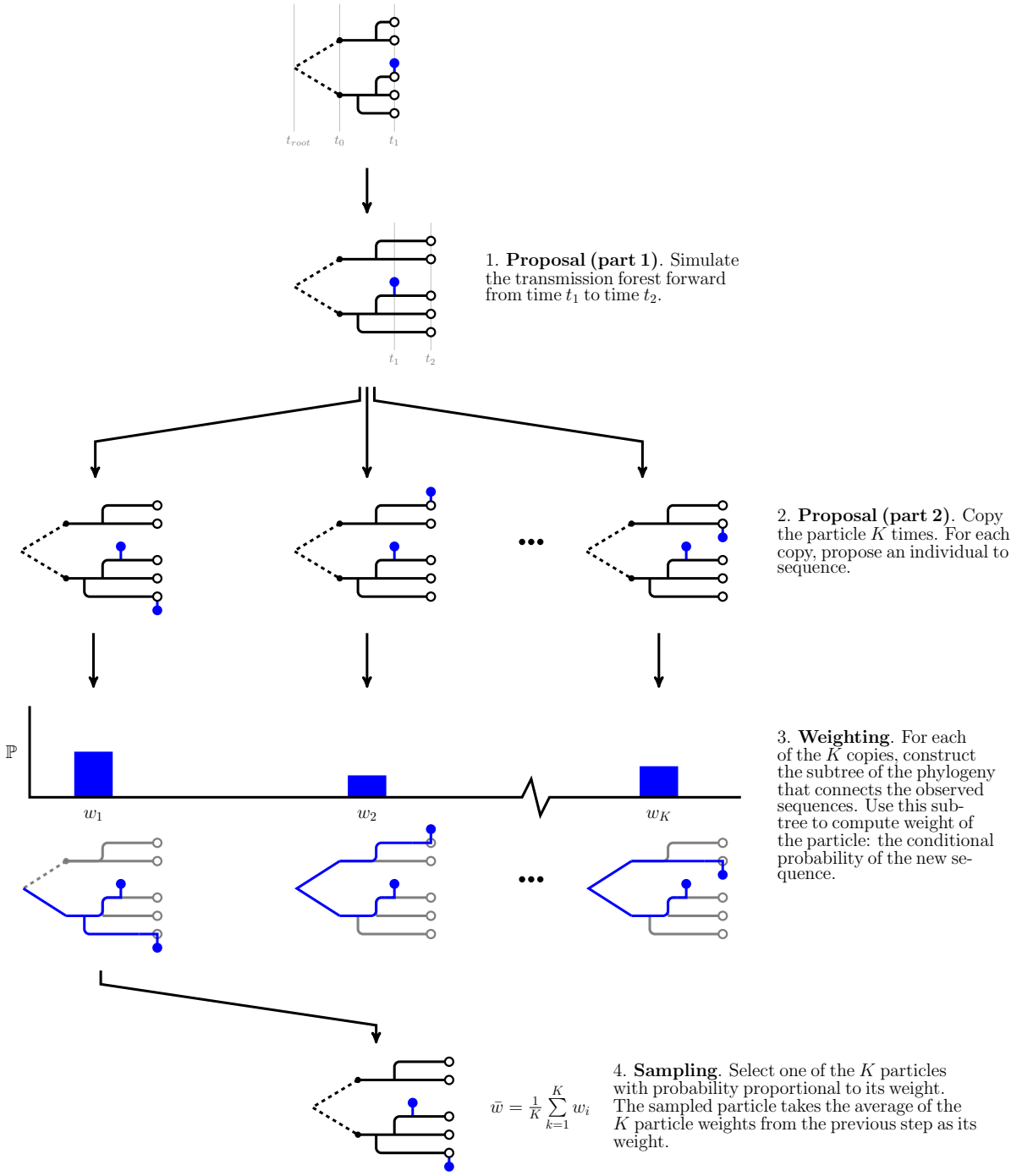

Figure 3.8: A schematic of our hierarchical sampling scheme. In this scheme, we split the proposal into two steps: (1) simulation of the transmission forest and (2) selecting an eligible individual to be sequenced. When each particle is expensive, it may pay to invest more effort in evaluating the conditional probability of a sequence given the latent state. This procedure is easily nested within the simpler form of the particle filter shown in Fig. 3.7. In turn, one can add additional Monte Carlo steps to the weighting step in this procedure to evaluate the conditional probability of a sequence under a relaxed clock (see Fig. 3.9).

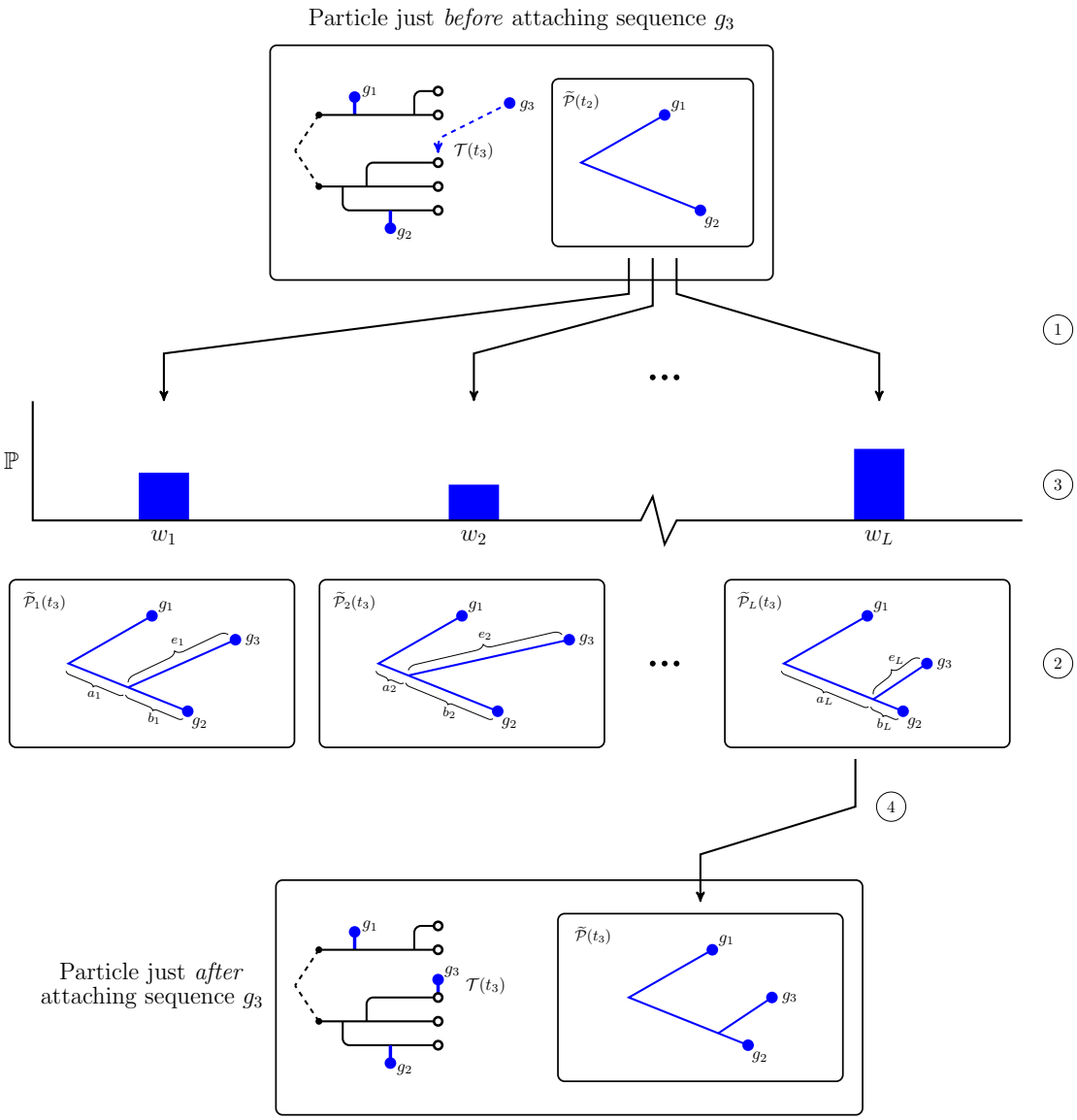

Figure 3.9: A schematic showing our Monte Carlo approach to estimate the conditional probability of a sequence under a relaxed clock. Note that this procedure only modifies the subtree of the phylogeny that joins the sequences, $\widetilde{\mathcal{P}}(t)$. At the top, we show a particle just before attaching a new sequence. In this case, the particle has already incorporated two sequences, and the location of the third sequence on the transmission forest has already been selected. First, we make L copies of $\widetilde{\mathcal{P}}(t_2)$, the subtree of the phylogeny that connects all sequences observed up to time t_2 (at ①). For each of these phylogenies we propose an attachment site and an edge length for sequence g_3 (at ②). The edge length of the edge subtending sequence g_3, e_ℓ, is drawn from a Gamma distribution parameterized as described in the text. We split the edge between the root and sequence g_2 according to a Beta distribution into two lengths, a_ℓ and b_ℓ; this procedure preserves Gamma distributed edge lengths for two components of the split edge. Then, for each proposed phylogeny, we use the peeling algorithm to compute the conditional probability of sequence g_3 (at ③). Finally, we sample one of these proposed phylogenies with probability proportional to its weight (at ④). The unsampled proposals are discarded and the particle takes the average of the conditional probabilities as its weight.

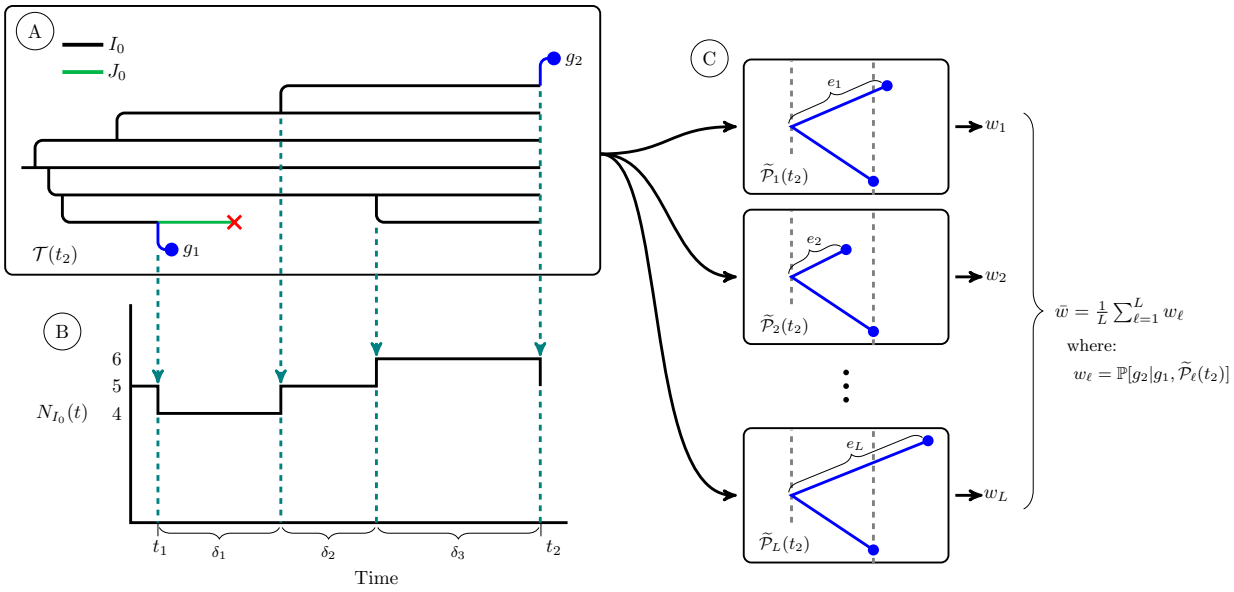

Figure 3.10: A schematic of quantities used in calculation of the conditional density of a diagnosis and the conditional probability of a genetic sequence. At Ⓐ we show a simulated transmission tree. For simplicity, this tree only has individuals of class I_0 and class J_0. Dashed arrows fall from events in the transmission tree that change the count of I_0 individuals in the population. At Ⓑ we show a plot of the trajectory of the I_0 class. This plot shows the quantities we use to calculate the cumulative hazard of diagnosis for the I_0 class, Λ_0, over an interval of time from t_1 to t_2. We first subdivide the time interval into R subintervals over which the number of I_0 individuals is constant (indicated with dashed lines). We let the number of I_0 individuals in the r^{th} subinterval be $N_{I_0,r}$. The cumulative hazard of diagnosis is then: $\Lambda_0 = \rho_0 \sum_{r=1}^{R} \delta_r N_{I_0,r}$. The cumulative hazards of diagnosis for the other two classes of undiagnosed individuals are computed in the same fashion. At Ⓒ we show the set of L subtrees of the phylogeny that we use to numerically estimate the conditional probability of sequence g_2 under our relaxed clock model. The ℓ^{th} subtree is constructed by augmenting $\widetilde{\mathcal{P}}(t_1)$ with a new edge with length e_ℓ drawn from a gamma distribution parameterized as described in the text. For each of these L subtrees we use the peeling algorithm to compute $w_\ell = \mathbb{P}[g_2|g_1, \widetilde{\mathcal{P}}_\ell(t_2)]$, the conditional probability of observing sequence g_2 given sequence g_1 and the structure of $\widetilde{\mathcal{P}}_\ell(t_2)$. The average of these conditional probabilities is a numerical estimate of the conditional probability of g_2 under our relaxed clock model. For simplicity, here we do not show the case in which the edge length of g_2 splits an existing edge; this case requires a beta bridge to apportion the length of the split edge so as to maintain gamma distributed edge lengths. For this more complicated case, see Fig. 3.9.

3.8 Supplementary Material

Supplementary materials are available online at Molecular Biology and Evolution (http://www.mbe.oxfordjournals.org/). The supplement provides a formal specification of the class of models described in the New Approaches section, technical details on the algorithms we developed to maximize and evaluate the likelihood of these models, and additional details concerning the data analysis presented in the main paper. The supplement provides a formal specification of the class of models described in the New Approaches section, technical details on the algorithms we developed to maximize and evaluate the likelihood of these models, and additional details concerning the data analysis presented in the main paper. The source code for our software implementation of the SMC algorithms is available at https://github.com/kingaa/genpomp (to be archived at datadryad.org).

3.9 Acknowledgments

Data on the the HIV epidemic in Detroit were provided by the Michigan Department of Community Health under a data sharing agreement that received IRB approval. We acknowledge James Koopman and Mary-Grace Brandt for their help in giving us access to these data and for discussions on HIV epidemiology. We are grateful for the support of the Genome Sciences Training Program at the University of Michigan and the following grants: NSF-DMS 1308919, NIH 1-U54-GM111274-01, and NIH 1-U01-GM-110712-01.

CHAPTER IV

Inferring transmission in a simulated hospital outbreak

4.1 Abstract

Over the last decades, Vancomycin-resistant Enterococcus (VRE) has emerged as a common pathogen in hospitals worldwide. The epidemiology of VRE transmission and the factors promoting persistence of VRE in hospitals are not yet well understood. Open questions remain about the role of antibiotics, both in modulating the transmission rate of VRE and in facilitating the growth of preexisting low levels of VRE colonization. The role of heathcare workers and the role of the hospital environment in transmission are also active areas of research. A potential barrier to a deeper understanding of VRE epidemiology is the inability to adequately model all the essential elements in the system. In principle, mechanistic models of transmission and VRE emergence, informed by whole genome sequences of the pathogen and surveillance data, have the potential to generate new insights into the roles of different drivers of VRE dynamics. However, developing the methodology to fit such models to detailed patient-level data is nontrivial. In this chapter, we first develop a model of VRE transmission on two wards. In a study on simulated data, we demonstrate that `genPomp`, equipped with a targeted proposal, can successfully recover parameters of within and between-ward transmission. Lessons from this study will serve as a foundation for later work with data from real outbreaks. We conclude with a discussion of possible ways forward for fitting a data from an outbreak at the NIH clinical center.

4.2 Introduction

Since its emergence in 1986 in Europe, Vancomycin-resistant Enterococcus (VRE) has spread to become a common pathogen in hospitals around the world (O'Driscoll and Crank, 2015). Multiple factors may underlie the success of VRE in the hospital. A number of studies have found a link between antibiotic usage and risk of VRE colonization (Gouliouris

et al., 2018; Peel et al., 2012; Papadimitriou-Olivgeris et al., 2014). A limitation of these studies is that they are unable to assess the relative roles of two likely mechanistic explanations for how antibiotics promote the success of VRE. One possibility is that antibiotic usage affects the transmission rate of VRE, either through increasing susceptibility of uncolonized patients or through increasing the infectiousness of colonized patients. Another possibility is that antibiotic usage facilitates the emergence of preexisting low-levels of infection. In outbreaks of VRE, both of these mechanisms may be a factor. Understanding the frequency with which these two mechanisms occur could be critical in developing effective control measures.

The mode of transmission of VRE is also an important area of research. Both environmental contamination and healthcare workers are thought to play important roles in transmission. A study exploring persistence times of different strains of *Enterococcus faecium* found the bacteria could persist on dry surfaces for at least a week and sometimes as long as 4 months (Wendt et al., 1998). Furthermore, another study found that prior room occupancy by a VRE positive patient is a risk factor for VRE colonization (Drees et al., 2008). A controlled study of the effectiveness of gloving found that use of gloves reduced hand carriage of VRE by healthcare workers (Tenorio et al., 2001). This study found that ungloved workers had significant rates of carriage of VRE on their hands, indicating they could facilitate transmission.

Fitting mechanistic models of transmission to hospital outbreak data may have the potential to estimate the relative importance of different factors driving VRE colonization. In particular, whole genome sequences may be able to inform our understanding of different mechanisms of VRE transmission and emergence. In Chaper III we proposed that `genPomp` could be particularly useful for studying hospital outbreaks of drug resistant bacteria. Hospital outbreaks are often relatively small in scale, therefore stochasticity may play an important role in shaping their transmission dynamics. By design, GenPOMP models explicitly specify the nature of stochasticity, both in process and in measurement. Furthermore, hospital outbreaks may offer detailed patient level data, such as surveillance test results and locations of patients through time. By using `genPomp` we can easily incorporate these additional sources of information as covariates or data to more efficiently leverage information in the pathogen genetic sequences.

However, using `genPomp` to fit models to data from hospital outbreaks poses the challenge of particle filtering on data rich with individual-level information. Fitting an individual-based model to such data using a particle filter is challenging because standard forward simulators will, with high probability, propose latent states that are entirely incompatible with the data. For example, we may interpret a pathogen genetic sequence sampled from

a patient as indication of a true positive with no error. When a particle arrives at an observed genetic sequence and the latent state does not include the sequenced individual in the proposed transmission tree, then that particle is assigned a weight of zero and is subsequently culled at the resample step. When this mismatch occurs in all particles for a single data point, then filtering fails. If the data and model are such that filtering failures are pervasive then inference becomes impossible. To combat this phenomenon, we developed a targeted proposal scheme, described in detail in Appendix B.

In this Chapter, through a study on a small simulated outbreak of VRE, we show that it is possible to use `genPomp`, coupled with a targeted proposal, to infer rates of within and between ward transmission from surveillance data and from pathogen genetic sequence data. As in the study in Chapter III, we examine how genetic data may revise our understanding of the system. We conclude by discussing future directions for fitting extensions of this model to data from real hospital outbreaks. In building up to an analysis of data on the scale of a full hospital, further simulation work will likely be necessary.

4.3 Methods

This study focuses on a small, simulated outbreak unfolding on two hospital wards. We ask the following questions. Can we estimate how much transmission occurs within each ward? Can we estimate how much transmission occurs between the two wards? How much does surveillance data inform inference of transmission? How much do the pathogen sequence data inform inference of transmission? Finally, how does inference using both datatypes compare to either used singly?

To explore these questions, we simulated culture tests and genetic sequences from a yearlong hospital outbreak. We then estimated grid-based likelihood profiles for two parameters of interest: β_w, the within ward rate of transmission, and β_b, the between ward rate of transmission. The form of the model, how we simulated, and how we estimated profiles are described below.

4.3.1 A model of transmission on a hospital with two wards

In this section we describe a stochastic, individual-based model of transmission of VRE on a hospital with two wards. This model belongs to the class of partially observed Markov process models described in Chapter III, Section 3.3. The latent state of the system at time t is a Markov process with three components: $X(t) = \big(\mathcal{T}(t),\ \mathcal{P}(t),\ \mathcal{U}(t)\big)$. Here, $\mathcal{T}(t)$ is the transmission forest, $\mathcal{P}(t)$ is the pathogen phylogeny, and $\mathcal{U}(t)$ is itself a Markov process describing the state of each individual in the population at time t. The set of possible states

an individual may take is $\{\mathrm{S}_1, \mathrm{I}_1, \mathrm{S}_2, \mathrm{I}_2\}$, where S indicates the individual is susceptible, I indicates the individual is infected, and the subscript indicates whether the individual is on ward 1 or ward 2. The probabilities of state changes for each individual over an interval of duration δ are given by

$$\begin{aligned} \mathbb{P}\big[\mathrm{S}_1 \to \mathrm{I}_1\big] &= \delta\lambda_1(t) + o(\delta), \\ \mathbb{P}\big[\mathrm{S}_2 \to \mathrm{I}_2\big] &= \delta\lambda_2(t) + o(\delta), \\ \mathbb{P}\big[\mathrm{I}_1 \to \mathrm{S}_1\big] &= \delta\gamma + o(\delta), \\ \mathbb{P}\big[\mathrm{I}_2 \to \mathrm{S}_2\big] &= \delta\gamma + o(\delta), \end{aligned} \tag{4.1}$$

where

$$\begin{aligned} \lambda_1(t) &= \beta_w N_{\mathrm{I}_1}(t) + \beta_b N_{\mathrm{I}_2}(t), \\ \lambda_2(t) &= \beta_b N_{\mathrm{I}_1}(t) + \beta_w N_{\mathrm{I}_2}(t), \end{aligned} \tag{4.2}$$

and $N_{\mathrm{I}_1}(t)$ and $N_{\mathrm{I}_2}(t)$ are the number of infected individuals on wards 1 and 2, respectively. We assume that there is no movement between wards, no death, and no entry or exit from the hospital over the time period studied. A schematic of the model is shown in Figure 4.1.

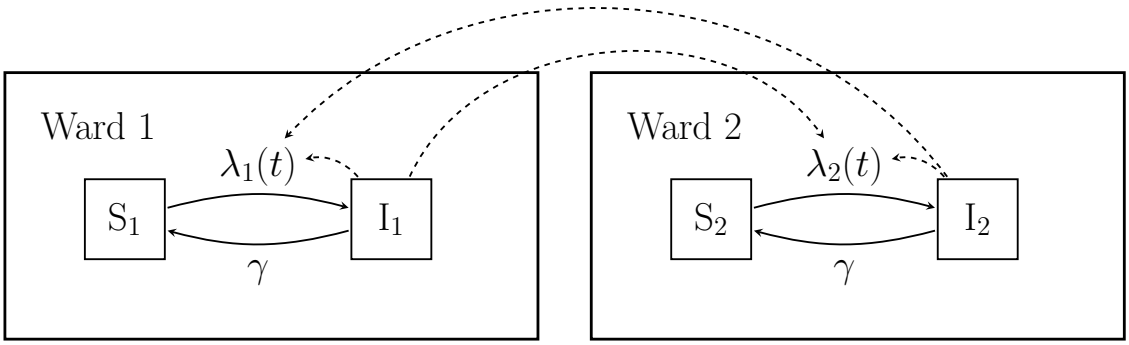

Figure 4.1: A schematic of a model of transmission on a hospital with two wards. Individuals recover at rate γ and become infected at rate $\lambda_1(t)$ and $\lambda_2(t)$ on wards 1 and 2, respectively. Dashed lines indicate that individuals in the infected class of both wards contribute to the time-varying overall rate of transmission in each ward.

We assume that the topology of the pathogen phylogeny, $\mathcal{P}(t)$, maps directly onto that of the transmission tree. That is, $\mathcal{P}(t)$ has the same branching pattern as $\mathcal{T}(t)$; each edge and each node in $\mathcal{P}(t)$ has a corresponding edge or node in $\mathcal{T}(t)$. We let the edge lengths of $\mathcal{P}(t)$ differ from those of $\mathcal{T}(t)$ so as to allow for heterogeneity in the rate of molecular evolution. We assume the edge lengths of $\mathcal{P}(t)$ are Gamma distributed, with expected

value equal to the corresponding edge length in $\mathcal{T}(t)$ and variance proportional to that edge length. That is, if L is the length of an edge of $\mathcal{P}(t)$ corresponding to an edge of length D in $\mathcal{T}(t)$, we assume that $L|D$ is Gamma distributed with $\mathbb{E}[L|D=d] = d$ and $\text{Var}[L|D=d] = \sigma\, d$. Having specified the structure of $\mathcal{P}(t)$, the choice of the time-reversible molecular substitution model determines the joint distribution of the sequences at the tips of $\mathcal{P}(t)$. For this study, we used the TN93 model of molecular evolution (Tamura and Nei, 1993), which is fully specified by the following rate matrix:

$$Q = \begin{bmatrix} * & \beta\pi_T & \beta\pi_C & \alpha_R\pi_G \\ \beta\pi_A & * & \alpha_Y\pi_C & \beta\pi_G \\ \beta\pi_A & \alpha_Y\pi_T & * & \beta\pi_G \\ \alpha_R\pi_A & \beta\pi_T & \beta\pi_C & * \end{bmatrix}$$

4.3.2 Simulation

We simulated an outbreak from the model described above using parameters specified in Table 4.1. We implemented this model in `genPomp` and used an exact method (Gillespie, 1977) to simulate one epidemic. Each ward contained 20 individuals and each ward was initialized with one VRE positive individual. We conditioned on the times of measurement: each individual was cultured once a month over the one year period. If the individual was VRE positive at the time of culturing, and the individual had never been sequenced before, then individual was sequenced. A visualization of the latent states of individuals and the observed culture tests is shown in Figure 4.2.

Table 4.1: Parameters values used in simulating the outbreak on two wards.

Parameter	**Interpretation**	**Value**
β_w	Within-ward transmission coefficient	0.0005 day^{-1}
β_b	Between-ward transmission coefficient	0.0001 day^{-1}
λ	Rate of infection from the community to the hospital	0 day^{-1}
γ	Rate of recovering from VRE+ to VRE-	0.01 day^{-1}
ρ_{cul}	Probability of a false positive culture	0.01
ϕ_{cul}	Probability of a false negative culture	0.1
t_p	Time of the polytomy	-365 days
$N_{S_1}(t_0)$	Number of susceptible individuals at time t_0 on ward 1	19
$N_{I_1}(t_0)$	Number of VRE+ individuals at time t_0 on ward 1	1
$N_{S_1}(t_0)$	Number of susceptible individuals at time t_0 on ward 2	19
$N_{I_1}(t_0)$	Number of VRE+ individuals at time t_0 on ward 2	1
β	Rate of transversions	0.0003 day^{-1}
α_Y	Rate of transitions between purines	0.003 day^{-1}
α_R	Rate of transitions between pyrimidines	0.001 day^{-1}
π_A	Equilibrium frequency of adenine	0.25
π_G	Equilibrium frequency of guanine	0.25
π_C	Equilibrium frequency of cytosine	0.25
π_T	Equilibrium frequency of thymine	0.25
σ	Relaxation of the molecular clock with respect to edges	0.01 day
δ_{fixed}	The initial component of the sequence stem	0.001 day
δ_{prop}	Proportion of time since infection to add to the sequence stem	0.05

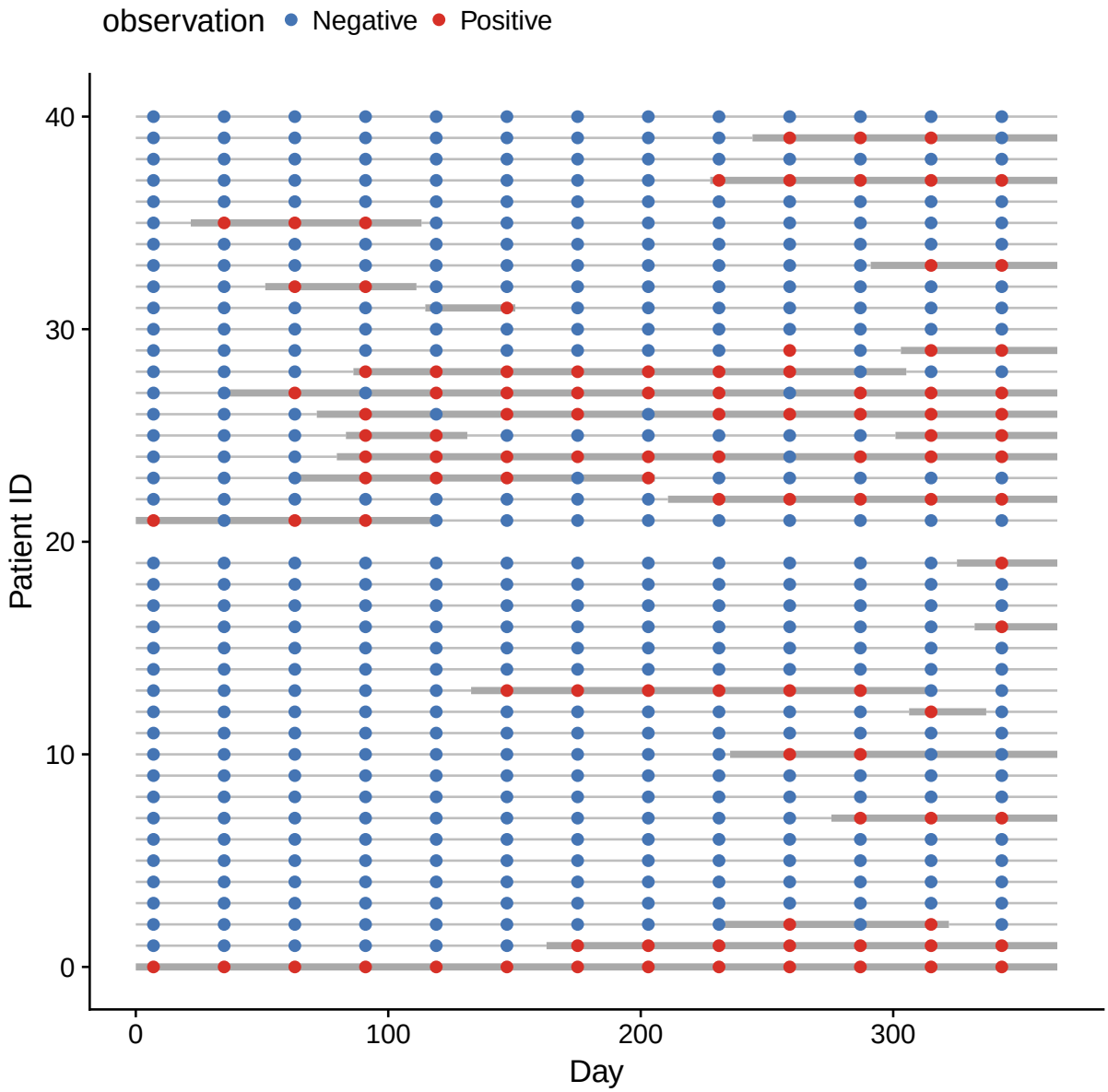

Figure 4.2: A simulated epidemic on two wards showing the latent states of individuals and their culture test results. Each row represents a patient. Thick grey lines indicate the patient is infected and thin grey lines indicate the patient is susceptible. Patient IDs 0-19 are on ward 1 and patient IDs 21-40 are on ward 2.

4.3.3 Inference

We estimated grid-based likelihood profiles for two parameters: β_w, the within ward rate of tranmission, and β_b, the between ward rate of transmission. We estimated these grid-based profiles in two ways: (1) using only the culture data and (2) using both the culture

data and the genetic sequence data together. To estimate the profiles we used the following procedure. We fixed all other parameters at their known true values and estimated a two-dimensional likelihood surface by running the particle filter 10 times for each parameter value pair. We used the targeted proposal when filtering over both datatypes as well as when filtering using either alone. We then averaged these 10 likelihood evaluations (on the natural scale) to obtain a likelihood estimate for each grid point. The grid spanned a range of values that encompassed the known parameter values for β_w and β_b. To obtain grid-based profiles, we took the maximum likelihood obtained for each value of the focal parameter. Differences in curvature and location of the maximum in the profiles obtained using the cultures alone or both datatypes together allows us to see how each datatype shapes inference.

4.4 Results

Grid-based profiles of β_w and β_b reveal that how well this inference approach captures the true parameter values depends on which datatypes are used (Figure 4.3). In the case of using only the culture data, we see curvature about the truth in the profile for β_w. In contrast, the profile over β_b does not rule out low values of the parameter; in the culture data alone there is only information on the upper bound of the between-ward rate of transmission. When using both datatypes we see that both profiles show curvature about the truth. Futhermore, the curvature for profiles computed using both datatypes is greater than than for those computed using only the culture data.

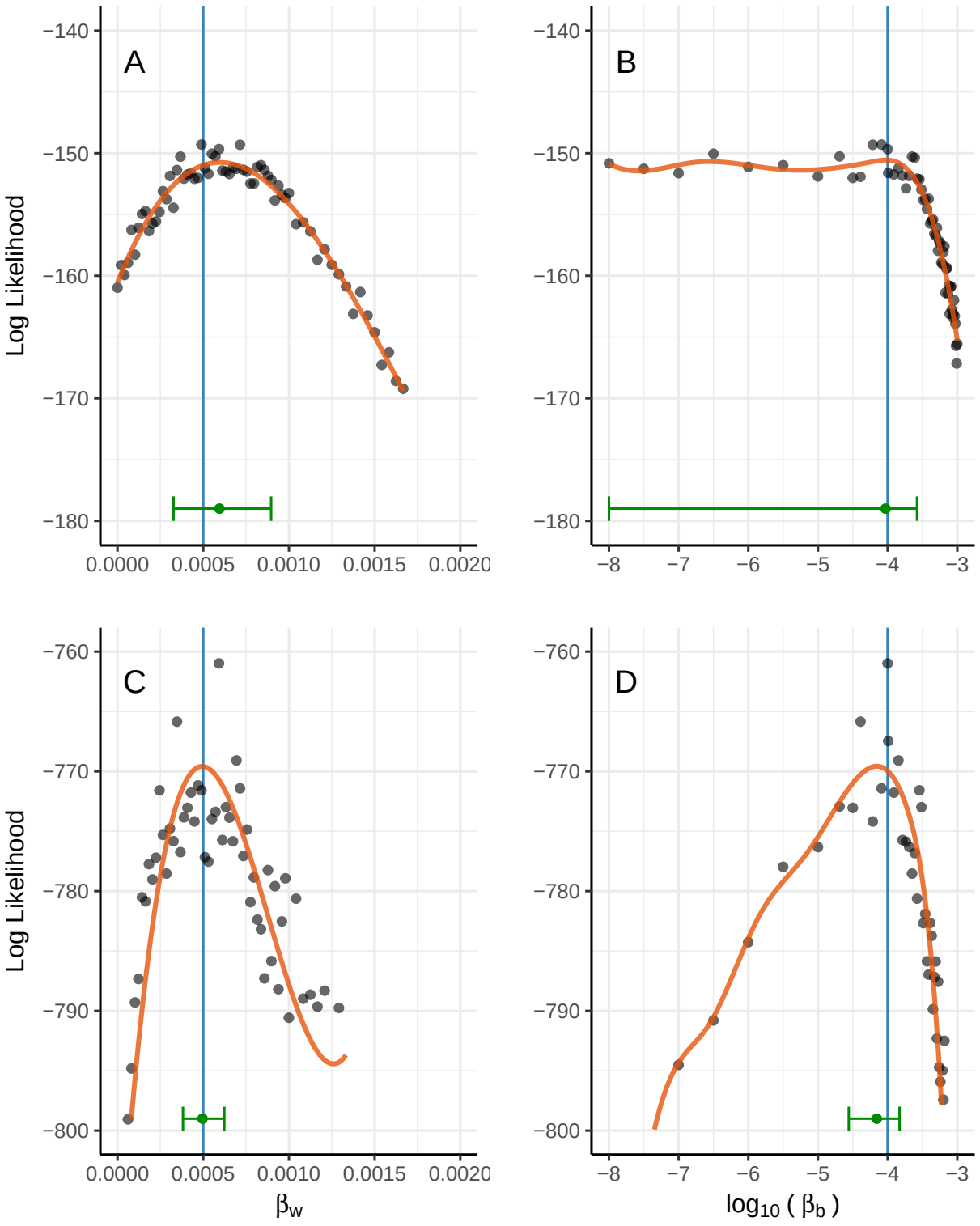

Figure 4.3: Grid-based likelihood profiles over β_w and β_b. Panels A and B show profiles computed using only the culture data. Panels C and D show profiles computed using both the culture and the sequence data. Scales are the same in all panels; the y-axes only differ by vertical shifts. True parameter values are shown in blue, smoothed profiles in red, and point estimates and confidence intervals in green.

4.5 Discussion

In the study on simulated data described above, we made a number of simplifying assumptions, some of which take us far from reality. Nevertheless, this exercise does provide useful lessons and will serve as a stepping stone toward an analysis of data from real outbreaks. In particular, this study shows that the particle filter with the targeted proposal appears to work well both when fitting the culture data alone and when fitting both datatypes together; variance in the likelihood estimate is small enough allow for inference on the parameters of interest. Another important lesson is that the sequence data are crucial for providing information about the rate of transmission between the two wards. The profile of β_b using only the culture data cannot rule out low values of the parameter. An analysis using both datatypes together yields the strongest results; the culture data inform the size of the epidemic such that the sequence data are able to provide additional information about the transmission rate between wards.

In future iterations of the model, the unrealistic assumptions of this simulation study can be modified to accommodate the complexity of real data. Some of the assumptions, for example those which have to do with movement and death, will naturally enter the model as covariates when fitting real data. In the next sections, we outline a way forward for fitting data from an outbreak of VRE at the NIH clinical center.

4.6 Future Directions

In future work, we will fit models to an outbreak of VRE at the NIH clinical center that took place over a span of about 4 years. The dataset consists of patient location data, surveillance tests, antibiotic treatment, timing of contact isolation procedures, and whole genome sequences. We have the location of patients at the level of ward for all individuals that pass through the hospital over the time period of the epidemic. Over the 4 year period, 7480 patients pass through the hospital. There are 12 wards in the NIH clinical center. For 833 patients we have some type of surveillance test, either a culture test or a PCR tests. Many of these patients are have multiple surveillance tests through time. For 44 individuals we have records of antibiotic treatment, the timing of contact isolation, and whole genome sequences.

4.6.1 Fitting a genetically-defined cluster

The whole genome sequences of VRE fall into three major clusters (Figure 4.4). These groupings correspond to three pulsed field gel electrophoresis (PFGE) subtypes of VRE

that likely comprise separate transmission histories. Figure 4.4 also reveals finer structure in this dataset that indicates there may be information on transmission history. Sequences that are close in genetic distance belong to individuals who often, but not always, overlap in time and space. Our first goal is to fit models to the largest of the three subgroups, the 29 sequences that constitute the large cluster in the lower left of the distance matrix shown in Figure 4.4. These 29 individuals comprise the most common PFGE subtype observed in the outbreak. We will fit the model described in Section 4.6.2 to the outbreak defined by this large cluster of sequences. When fitting models to the outbreak of this subtype, because the protocol was to sequence all cultures that were in this PFGE subtype, we will treat all positive cultures (and PCR tests) that were not from sequenced individuals as negative tests.

4.6.2 A more complicated model of VRE transmission

This model described in this section is designed to estimate the rate of transmission within ward, between wards, and from the community to the hospital. It also has the potential to examine the efficacy of contact isolation procedures. This model is an extension of the two ward model described in Section 4.3.1. The models of $\mathcal{T}(t)$ and $\mathcal{P}(t)$ are unchanged as is our choice of the model of molecular evolution. The modifications to bring this model closer to reality involve changes to $\mathcal{U}(t)$ and the new possibility of infection from a source outside the hospital.

We model the transmission of VRE on twelve hospital wards, W_j, $j \in \{0, 1, ..., 12\}$. Outside the hospital is denoted by $j = 0$. Figure 4.5 depicts the structure of the model. We use an individual-based model in which the state of each patient consists of the patient's VRE status (+ or -), the location of the patient (in one of the twelve wards or outside the hospital), and whether or not the patient is under contact isolation procedures. Both the locations of patients and contact isolation procedures are observed, so in fitting the model these two elements of each patient's state enter as covariates. The set of possible states an individual may take is $\{\mathrm{S}_j^n, \mathrm{S}_j^c, \mathrm{I}_j^n, \mathrm{I}_j^c\}$, where S indicates the individual is VRE negative, I indicates the individual is VRE positive, c indicates the individual is under contact isolation, n indicates that the individual is not under contact isolation, and j indicates the individual is in ward j. The probabilities of state changes for each individual over an interval of

duration δ are given by

$$\begin{aligned}
\mathbb{P}\left[\mathrm{S}_j^n \to \mathrm{I}_j^n\right] &= \delta\beta_j(t) + o(\delta), \\
\mathbb{P}\left[\mathrm{S}_j^c \to \mathrm{I}_j^c\right] &= \delta\phi\beta_j(t) + o(\delta), \\
\mathbb{P}\left[\mathrm{I}_j^n \to \mathrm{S}_j^n\right] &= \delta\gamma + o(\delta), \\
\mathbb{P}\left[\mathrm{I}_j^c \to \mathrm{S}_j^c\right] &= \delta\gamma + o(\delta),
\end{aligned} \tag{4.3}$$

where

$$\beta_j(t) = \lambda + \beta_w[N_{\mathrm{I}_j^n}(t) + \psi N_{\mathrm{I}_j^c}(t)] + \beta_b \sum_{k=1,k\neq j}^{12} [N_{\mathrm{I}_k^n}(t) + \psi N_{\mathrm{I}_k^c}(t)], \tag{4.4}$$

$\phi \in [0,1]$, and $\psi \in [0,1]$. Here, ϕ represents the effect of contact isolation susceptibility, with $\phi = 1$ indicating no protection and $\phi = 0$ indicating full protection. On the other hand, ψ represents the effect of contact isolation on infectiousness, with $\psi = 1$ indicating no reduction of transmission and $\psi = 0$ indicating complete prevention. In practice, contact isolation procedures are usually employed to combat infectiousness. We include the effect of contact isolation on susceptibility for completeness. We assume that the intensity of within-ward transmission, β_w, is the same for all wards. Similarly, we assume that the intensity of between-ward transmission, β_b, is the same between all wards. We assume that the rate of infection from the community to the hospital, λ, is constant. The recovery rate, γ, does not depend on contact isolation status. The above equations specify probabilities of each possible event that may change the infection status of an individual. All other event types, which include movement and change in contact isolation status, are observed directly.

4.6.3 Testing the feasibility of inference

One question is whether the plan outlined above is feasible. A challenge with fitting this dataset is that a large part of the system is unobserved. There are many people, for example, for whom we observe only movement data. These individuals may act as links in hidden chains of transmission. Also, this dataset is much larger than that of the simulated two ward example above; we have not yet shown through simulation that we can scale to a system of this size.

We can test whether fitting models to a dataset of this scale is a reasonable exercise. For example, we could condition on the observed covariates of the dataset, including the movement of all patients, and then simulate data from a small epidemic on the scale of the epidemic we aim to fit. We can then test whether we can recover the known values of parameters of interest. This would be a similar exercise to the two ward example described above, but with the aim of understanding the limits of inference with a simulated example

that is much closer to that of the real data. Depending on the results of this exploration, we can either proceed with an analysis of the full dataset or possibly scale down to model transmission on a subset of the wards.

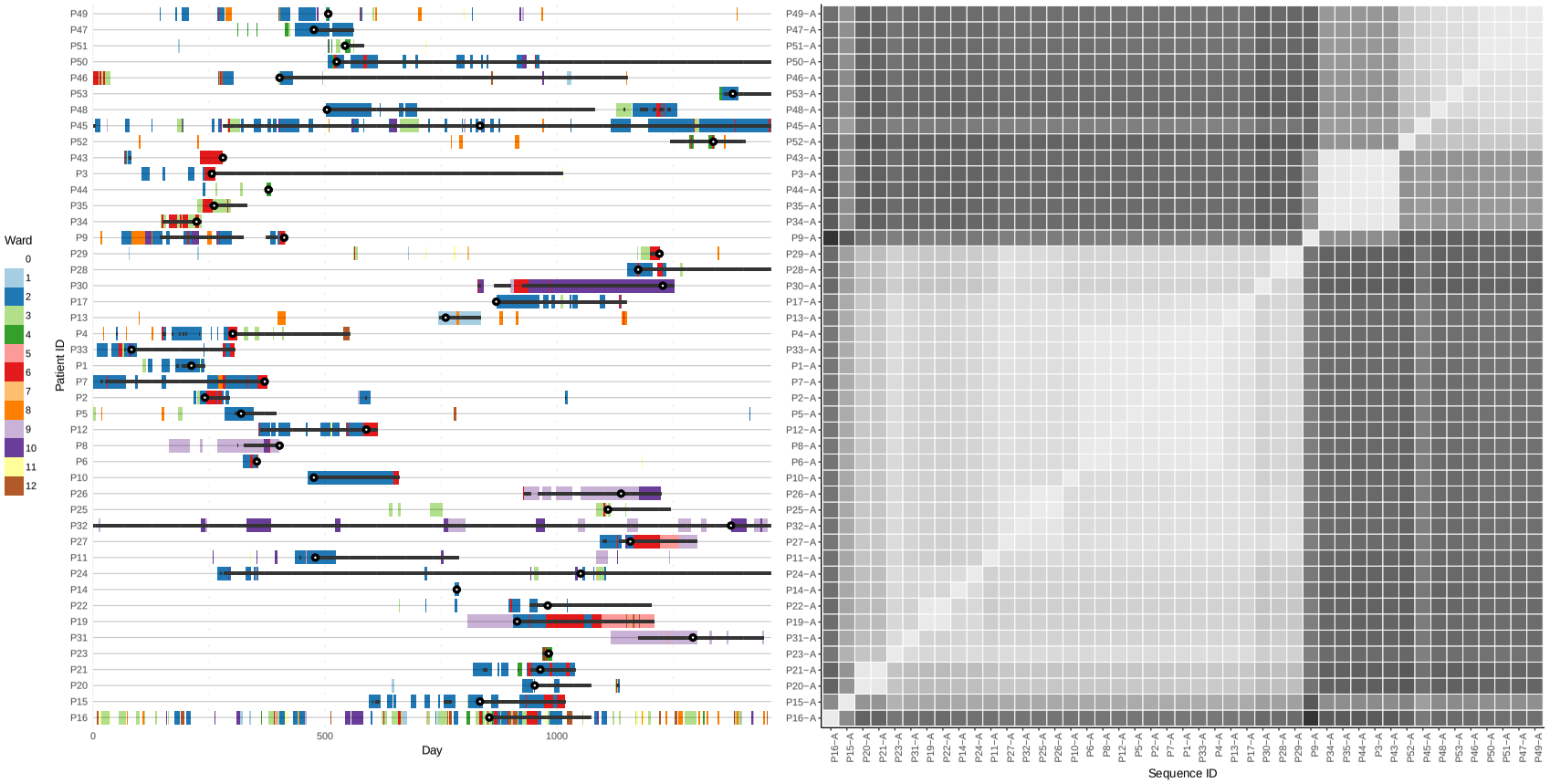

Figure 4.4: On the left, the locations of sequenced individuals through time. On the right, a distance matrix of the whole genome sequences of VRE. Sequences are lined up with their corresponding patient traces on the left. The distance metric is the Hamming distance, with lighter colors indicating fewer differences and darker colors indicating more differences. Black dots with white circles indicate the time of sequencing. Thick horizontal lines indicate time periods when patients are under contact isolation procedures. Sequences that are genetically closely related, often but not always, belong to patients that cluster in time and space. This plot shows that the data contain structure that may be informative about transmission history.

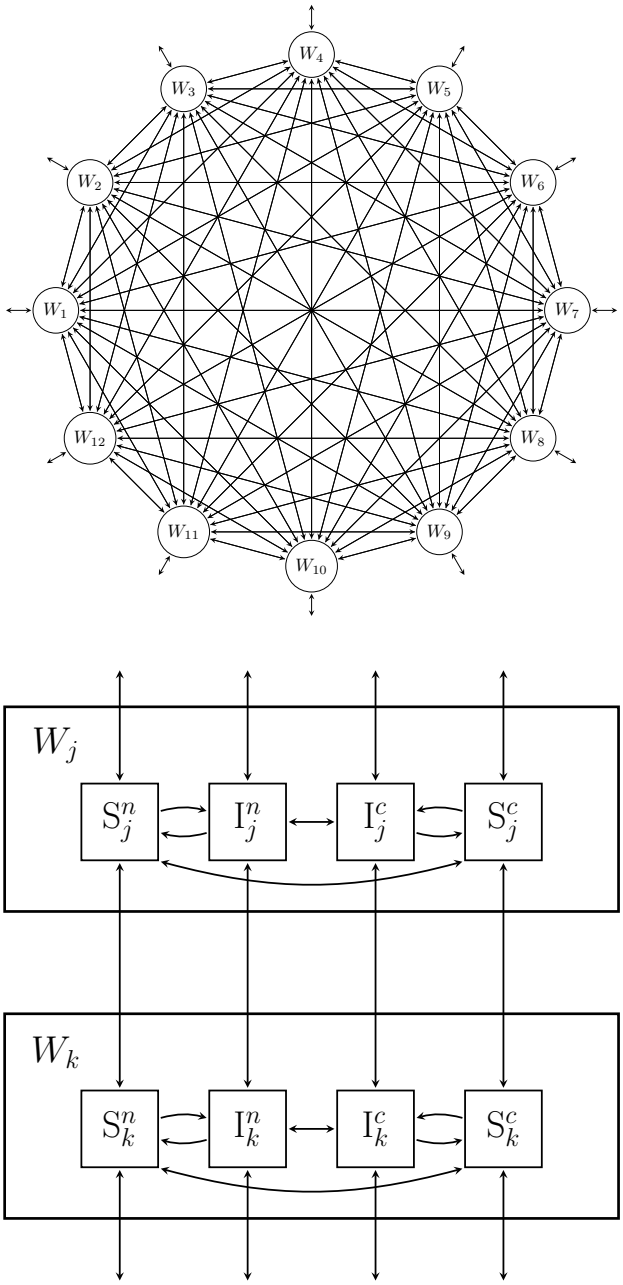

Figure 4.5: A model of VRE transmission on a hospital with 12 wards. The top schematic shows the connectivity between wards in the study. We directly observe patient movements between wards as well as movement into and out of the hospital. The model allows for transmission both within and between wards. The lower schematic shows the pairwise connectivity between any two wards j and k and the possible state changes of individuals.

CHAPTER V

Conclusion

In the years since Grenfell et al. (2004) presented a phylodynamic framework for understanding how various processes shape phylogenies, the field of phylodynamic inference has made significant steps forward. However, two-stage inference techniques have been the favored approach of many scientific studies. In this thesis we demonstrated one way in which two-stage inference may falter. We then proposed a flexible new method for joint inference and demonstrated its utility in two studies with very different types of data.

In Chapter II, we performed a simulation study to explore the strengths and weakness of a two-stage inference approach proposed by Rasmussen et al. (2011). In this study we used `genPomp` to simulate epidemics from a seasonal SIR model. We then assessed the quality of inference of the Rasmussen approach when given the true phylogeny versus a phylogeny estimated using BEAST. The key result from this study was that errors in phylogenetic reconstruction may drive bias in two-stage phylodynamic inference. This potential consequence of two-stage inference demonstrates the need for methodology for joint inference of the transmission model and the pathogen phylogeny.

In Chapter III we presented the central work of this thesis, a simulation-based, statistically efficient method for joint inference of disease dynamics and pathogen phylogeny via maximum likelihood. In Appendix A we proved that the class of algorithms for estimating and maximizing the likelihood are valid sequential Monte Carlo algorithms. To test the feasibility of our approach, we performed a study on simulated data. This study showed that our algorithms provide access to the likelihood surface of a dynamic model fit to genetic data. Grid-based likelihood profiles revealed that the known true values of stage-specific infectiousness parameters can be recovered when using both diagnosis and pathogen genetic data. We then used our methodology to study stage-specific infectiousness of HIV in a subepidemic in the young black MSM population in Detroit, MI. In this data analysis, we demonstrated one cycle of the iterative process of formulating and fitting a mechanistic model. Our results showed that the form of the model was likely too rigid to allow for

properly leveraging information in the genetic sequences on stage-specific infectiousness.

One possible direction of future work is to modify the model to allow for a wider range of dynamics for the infected population. Figure 5.1 shows one model we may explore. In this new model, we extend the model fit in Chapter III by adding a susceptible pool. By allowing for a susceptible pool this model has the flexibility to generate dynamics other than exponential growth or decay. Our hypothesis is that this model will allow for pulses and lulls in transmission that may better correspond to observed clusters of transmission in HIV epidemics. Flexibility in the possible dynamics of the model may allow for more efficient use of subtle information in the genetic sequences on stage-specific infectiousness.

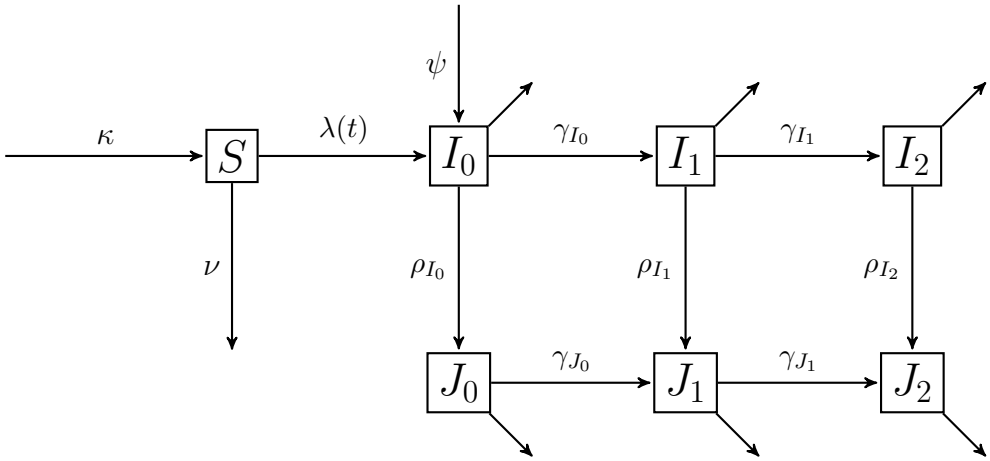

Figure 5.1: A state-space model that includes both a susceptible pool and the infected population. The I_k classes and J_k classes again represent the infected populations, with the top row representing undiagnosed individuals (I_k, $k \in \{0, 1, 2\}$), and the bottom row representing diagnosed individuals (J_k, $k \in \{0, 1, 2\}$). The ρ_{I_k} are rates of diagnosis, and the γ_{I_k} and γ_{J_k} are rates of disease progression. Unlabeled arrows out of infected states represent the combined flow out of the infected population due to death and emigration. ψ is a constant rate of infection from a source (or sources) outside the population. Three rates modulate the size of the susceptible class, S: κ, a constant rate of inflow; ν, a constant rate of outflow; and λ, the rate of new infections. $\lambda(t)$ is state dependent: $\lambda(t) = N_S(t)[\beta_{I_0} N_{I_0}(t) + \beta_{I_1} N_{I_1}(t) + \beta_{I_2} N_{I_2}(t) + \beta_{J_0} N_{J_0}(t) + \beta_{J_1} N_{J_1}(t) + \beta_{J_2} N_{J_2}(t)]$.

In Chapter IV, we developed a model of transmission of VRE in a hospital setting. In a simulation study, we showed that pathogen genetic sequences and surveillance data may contain information on between-ward and within-ward transmission rates. Development of a targeted proposal to combat particle depletion due to perfectly observed states was an essential step in making this study feasible. Although several assumptions of the simulated example were unrealistic, this initial study serves as a test of underlying code and will inform future efforts at fitting real data. We concluded by outlining steps toward scaling

up to an analysis of real data from an outbreak of VRE at the NIH clinical center as well as steps for testing the feasibility of our approach.

It is possible that current computational limitations of `genPomp` render an analysis of the full VRE dataset infeasible. For this particular dataset, it may be possible to reduce the problem by focusing on a handful of wards on which the majority of transmission occurs. Also, work with simulated data indicated that knowledge of initial conditions plays a significant role in the Monte Carlo variance of the likelihood estimate. Another possibility is to make strong assumptions about the number of individuals initially infected on each ward. Constraining initial conditions may reduce the Monte Carlo variance of the likelihood estimate to levels suitable for inference.

Tailoring a genPOMP model to fit detailed, patient-level data required additional work beyond a straightforward application of our algorithms. This work may seem a relatively high cost to pay. However, it is not clear that other methods for fitting mechanistic models would be able to fit this type of data at all. For example, coalescent-based approaches are currently infeasible for highly structured populations. Approximations derived from birth-death models would face similar challenges when faced with fitting a compartmental model with a large number of compartments. While the `genPomp` approach may involve additional careful work, it may be the most promising way forward in developing a system for incorporating multiple individual-level datatypes into a single analysis. Furthermore, although in Chapter IV we focused on specific steps to fit a particular outbreak of VRE, the model in this chapter could be applied, or easily modified, to fit data from other hospital outbreaks (either of VRE or of other drug resistant bacteria). The targeted proposal solves a general problem that may be an issue in other similar datasets. As stochasticity may play a large role in small outbreaks, it will be important to study multiple outbreaks to replicate findings.

5.1 Other ideas

One of the great strengths of the `genPomp` approach is its flexibility to entertain a wide class of models. In this thesis, we have only begun to explore the scope of this methodology. In this section, we outline potential future directions for development of `genPomp`:

1. One of the three research areas highlighted by Grenfell et al. (2004) was to "establish how epidemic and metapopulation disease dynamics modulate selective forces [...] to drive long-term phylogenetic patterns". In the current implementation of `genPomp` we ignore the role of selection. However, one could imagine a scheme to incorporate selection into a GenPOMP model. One way to do so would be to allow the measure-

ment of sequences to affect the state. The fitness of a sequence measured from an individual could modify that individual's transmission rate; individuals with high fitness sequences would be more likely to transmit and those with low fitness sequences less likely. Such an approach could borrow methodology from the work of Luksza and Lässig (2014), which developed a model of the fitness of flu to predict which strains would be successful in the next season. Alternatively, one could use methods from the field of antigenic cartography (Smith et al., 2004) to assess the fitness of sequences.

2. Real-time analysis of sequences, such as publicly available sequences of influenza, has the potential to dramatically shorten the timescale on which scientific analyses inform public health efforts and shape vaccine development. Researchers have already created such a system. For example, the program **nextflu** performs real-time analysis and visualization of influenza sequences (Neher and Bedford, 2015). This is an open source project, and in principle it would be possible to integrate `genPomp` into this environment.

3. It would be interesting to explore how well `genPomp` can estimate transmission trees. In this thesis, our primary concern has been estimation of parameters of the epidemiological model. However, there are many contexts in which the transmission tree may be of greater interest. In hospital outbreaks, for example, we may be interested in the probabilities of particular paths of transmission as opposed to population level rates of transmission. There are existing methods for estimating transmission trees from pathogen sequence data (e.g. the Bayesian methodology proposed by Didelot et al. (2014)). Although `genPomp` was designed with another goal in mind, it could be used for estimating transmission trees that are consistent with models of disease dynamics. A comparison of the ability of `genPomp` to estimate transmission trees with current state-of-the-art methodology could be worthwhile.

5.2 Final remarks

The work in this thesis represents a significant advance in developing likelihood-based methodology for joint inference of disease dynamics and phylogeny. The flexibility of `genPomp`, both in its ability to fit a wide range of models and in its ability to easily incorporate multiple data types, is a major strength of the method. Both in Chapter III and in Chapter IV we demonstrated that constructing likelihood profiles using multiple datatypes allowed for leveraging information in the genetic data to achieve stronger inference. Furthermore, construction of profiles with or without different datatypes allowed for

evaluating the influence of different sources of information. In future work, this capability will be useful both in model development and in design of data collection protocols.

A limitation of `genPomp` is that the computational cost of a full data analysis is currently quite high. Scaling `genPomp` to tackle large problems will require further algorithmic development. In cases where joint inference using `genPomp` is infeasible due to computational constraints, various two-stage methods may be applicable. Which alternative method is most suitable may be an empirical question. As we demonstrated in Chapter II, `genPomp` can play a role in assessing the effectiveness of other methods under various scenarios. In this chapter, we used `genPomp` to simulate genetic sequences of pathogens consistent with a specific mechanistic model of transmission and with a particular model of sequence evolution. This same approach can be used to study the strengths and weaknesses of methods other than that proposed in Rasmussen et al. (2011). Without a method for likelihood-based joint inference, such as `genPomp`, it can be difficult to assess how much is lost in using methods that rely on approximations.

As computational power improves, use of `genPomp` on problems of modest size will be more and more feasible. In the work here, we have only begun to explore the potential of the `genPomp` approach. By uniting models of disease transmission and pathogen evolution in a consistent fashion, with the capability to fit models of arbitrary complexity, the work of this thesis provides a strong foundation for diverse avenues of future study.

APPENDICES

APPENDIX A

Supplementary Material for GenPomp: A flexible method for joint phylodynamic inference

This section is published as supplementary material a paper published in *Molecular Biology and Evolution* in 2017 (DOI: https://doi.org/10.1093/molbev/msx124). Refer to the beginning of Chapter III for a description of the contributions of the authors.

A.1 The GenPOMP model: linking infectious disease dynamics with genetic data

We define a class of models that describes an environment within which our general software implementation can be described. We aim at sufficient generality to represent the breadth of applicability of our methodology and the key methodological innovations, yet including enough details to describe the specific data analysis in the main text.

Data consist of n^* genetic sequences of a pathogen. We use a convention that $j{:}k$ denotes the arithmetic sequence $(j, j+1, \ldots, k)$, so that the entire collection of genetic sequence data can be written as

$$(g_1^*, g_2^*, \ldots, g_{n^*}^*) = g_{1:n^*}^*.$$

We use asterisks to denote data, to distinguish these from quantities arising in the model. The times at which the sequenced samples are collected are also data, and the total number of sequences, n^*, will be modeled as the outcome of a random process rather than some fixed quantity. We write the genetic sequence times as

$$(t_1^*, t_2^*, \ldots, t_{n^*}^*) = t_{1:n^*}^*.$$

We suppose that the data are collected in a time interval

$$\mathbb{T} = [t_o, t_{\text{end}}],$$

with $t_0 \le t_1^* < t_2^* < \cdots < t_{n^*}^* \le t_{\text{end}}$. Note that we allow multiple observations at the same time: such ties can be resolved arbitrarily in the ordering of the t_n^*. For simplicity, we exclude the possibility of such simultaneous observations in the following. If no sequence is available for the diagnosis at some time t_n^*, we set $g_n^* = \text{NA}$. Otherwise, we suppose the collection of sequences $g_{1:n^*}^*$ consist of aligned sequences of length L, i.e., $g_n^* \in \{A, C, T, G\}^L$.

Here, we do not include the possibility of additional clinical or epidemiological measurements available at diagnosis, though an extension to allow this is fairly straightforward. Further, we consider that only a consensus pathogen sequence is available from each host, so we ignore the possibility of extracting information from data on pathogen genetic diversity within hosts. Nevertheless, our framework can account for sequencing error and differences between observed and transmitted pathogen populations.

The partially observed Markov process (POMP) model consists of a latent, unobservable, Markov process $\{X(t), t \in \mathbb{T}\}$ and an observable process $\{Y(t), t \in \mathbb{T}\}$. $X(t)$ takes values in a set $\mathbb{X}$ and $Y(t)$ takes values in a set $\mathbb{Y}$. A POMP model for genetic data, which we call a GenPOMP, is required to have the following structure. $\{Y(t)\}$ consists of a collection of random number N of diagnosis times, denoted $T_{1:N}$, and corresponding sequences $G_{1:N}$. The observed outcomes are $N = n^*$ and $(T_n, G_n) = (t_n^*, g_n^*)$ for $n \in 1{:}n^*$. We adhere to a convention that random variables are denoted by upper case letters; the corresponding lower case letters are used for possible values of the random variable, and asterisks denote the actual data for observable variables; blackboard bold typeface is used for sets.

Recall that, in the main text, we wrote $X(t) = \big(\mathcal{T}(t), \mathcal{P}(t), \mathcal{U}(t)\big)$ where $\mathcal{T}(t)$ is a *transmission forest* and $\mathcal{P}(t)$ is a *phylogeny*. Here, it is convenient to take a different, but functionally equivalent, approach. We do not require that $X(t)$ itself contains $\mathcal{T}(t)$ and $\mathcal{P}(t)$, but we do require that $\{X(u), t_0 \le u \le t\}$ is sufficient to construct $\mathcal{T}(t)$ and $\mathcal{P}(t)$. This additional layer of abstraction lets us define the GenPOMP model without having to explicitly construct the processes $\{\mathcal{T}(t), t \in \mathbb{T}\}$ and $\{\mathcal{P}(t), t \in \mathbb{T}\}$.

The set $\mathbb{X}$ should describe the state of each individual in a study population. The study population is supposed to contain a finite number of individuals drawn from a countable collection of individuals who could potentially enter the study population. We suppose these potential individuals are labeled with values in the natural numbers, $\mathbb{N} = \{1, 2, 3, \dots\}$, and so collections of individuals in the study population take values in the set $\mathbb{H}$ consisting of all

finite subsets of $\mathbb{N}$. We suppose there is a random process $\{H(t), t \in \mathbb{T}\}$, with $H(t)$ taking a value in $\mathbb{H}$ corresponding to the identities of all individuals in the study population at time t. Formally, we suppose that $H(t)$ is constructed from $X(t)$ via a suitable function mapping $\mathbb{X}$ to $\mathbb{H}$. We suppose that each individual in the study population has a state in a set $\mathbb{S}$. For a simple compartment model, $\mathbb{S}$ could be finite or countable, however, we also allow for the possibility of continuous real-valued state variables. In particular, we will later define a random clock process governing the rate of pathogen evolution within each individual infected host. To keep track of the state of each member of the study population, we suppose that the state of any individual i in the study population at time t is given by a random variable $X_i(t)$, constructed from $X(t)$ via a suitable function mapping $\mathbb{X}$ to $\mathbb{S}$. A canonical way to do this is to take

$$\mathbb{X} = \bigcup_{h \in \mathbb{H}} \mathbb{S}^h, \tag{A.1}$$

for which an element $\left(s_{i_1}, s_{i_2}, \ldots, s_{i_k}\right) \in \mathbb{X}$ is interpreted to mean that the study population is $\{i_1, i_2, \ldots, i_k\} \in \mathbb{H}$ and individual i_j is in state $s_j \in \mathbb{S}$. Our definition of the study population is the collection of individuals being modeled, and so the state of individuals outside the study population is necessarily undefined. In order to define $\{X_i(t), t \in \mathbb{T}\}$ as a stochastic process, one can formally define an additional state $\odot$ and set $X_i(t) = \odot$ when $i \notin H(t)$. Note that, in general, $\{X_i(t), t \in \mathbb{T}\}$ does not inherit the Markov property from $\{X(t), t \in \mathbb{T}\}$. If individual state transitions occur as an independent Markovian process once that individual is infected (as is the case in our HIV example) then $\{X_i(t), t \in [t_i, t_{\text{end}}]\}$ has a conditional Markov property given $i \in H(t_i)$.

The state process may, in general, need to include other components in addition to $\{X_i(t), i \in H(t)\}$. For example, $X(t)$ may include dynamic variables affecting the entire population, such as environmental or sociological processes. For the remainder of this article, the specific construction in equation (A.1) suffices, but that is not essential to our approach. If $\mathbb{S}$ is countable then $\mathbb{X}$, given by (A.1), is also countable and $\{X(t)\}$ is a Markov chain. Otherwise, $\{X(t)\}$ is a more general Markov process.

Some basic properties of individuals characterize the model as a disease transmission system, and these are required to construct the evolutionary process model for the pathogen. This leads us to define functions that return properties about the state of an individual, and we call these *query functions*. This notation differs from usual compartment models, where each individual is modeled as residing in a single compartment. We write properties as functions of $X(t)$, rather than components of $X(t)$, to keep applicability to a broad class of population models. As long as the required query functions can be defined for a population model, the statistical methodology developed will apply, giving the scientist

considerable flexibility in the specification of the model.

We require that an individual's state, i.e., its value in $\mathbb{S}$, can describe whether that individual is infected and infectious. We represent this requirement by supposing that there is a query function

$$Q_I : \mathbb{S} \to \{0, 1\}$$

defined as,

$$Q_I(s) = \begin{cases} 1 & \text{if } s \text{ is an infected state,} \\ 0 & \text{if } s \text{ is an uninfected state.} \end{cases}$$

To link the model to diagnosis data, we require that a state in $\mathbb{S}$ determines whether an individual is diagnosed while part of the study population. Specifically, we suppose there is a query function

$$Q_D : \mathbb{S} \to \{0, 1\}$$

such that

$$Q_D(s) = \begin{cases} 1 & \text{if } s \text{ is a state for individuals diagnosed as infected while in the study population,} \\ 0 & \text{otherwise.} \end{cases}$$

We then define

$$D(t) = \sum_{i \in H(t)} Q_D\big(X_i(t)\big),$$

which counts the number of individuals diagnosed while in the study population, by time t. This counting process (i.e., a non-decreasing integer-valued process) is relevant for relating the model to the data on the study population. Note that $D(t)$ does not count the number of clinically diagnosed individuals in the study population at time t, which would require a different accounting for the possibility of immigration and emigration of diagnosed individuals.

Now, we define the set of infected states to be

$$\mathbb{I} = \big\{ s \in \mathbb{S} : Q_I(s) = 1 \big\}.$$

We suppose that the state contains information about the identify of the infector, and we do this by requiring the existence of a query function

$$Q_L : \mathbb{I} \to \mathbb{N} \cup \{0\}$$

defined such that

$$Q_L(s) = \begin{cases} j & \text{if } s \text{ is infected by individual } j \text{ within the study population,} \\ 0 & \text{if } s \text{ is infected by an infector outside the study population.} \end{cases}$$

The capability to construct the query function $Q_L(s)$ requires that the identity of the infector is stored in the state variable at the point of infection, so it is available later as part of the state of the infectee. Information on the identity of the infector is not usually required for a compartment model, but is useful when working with genetic data in order to track lineages of the pathogen.

The evolutionary process of the pathogen genome within an individual in the host populations is modeled using a relaxed molecular clock, meaning that standard molecular models for evolution are applied on a stochastically perturbed timescale. It has become established that the usual models for molecular evolution fit sequence data better if one allows such fluctuations in the rate of evolution (Drummond et al., 2006). To implement a relaxed clock, we construct a random process on each edge of the transmission tree. This process scales calendar time to evolutionary time, the latter meaning a modified timescale on which the evolutionary rate is constant. We therefore require the existence of a query function

$$Q_\Gamma : \mathbb{I} \to \mathbb{R}$$

returning the relaxed evolutionary clock time corresponding to evolution of a transmissible pathogen population within an infected individual. Specifically, $Q_\Gamma(s)$ represents the random, individual-specific, clock time for the evolutionary process that separates the host's transmissible pathogen population from the rest of the pathogen community when the host is in state $s \in \mathbb{I}$. For an individual based model in which an individual is infected within the study population, this corresponds to the evolutionary process within the host subsequent to infection. Immigrant pathogens require additional assumptions on how they relate genetically to pathogens already circulating in the study population. Conditional on the randomly perturbed molecular clock, pathogen evolution is usually specified by a general time-reversible Markov model.

We also suppose the existence of a query function

$$Q_\Delta : \mathbb{I} \to \mathbb{R}$$

which returns the relaxed evolutionary clock time separating the measurable pathogen population from the transmissible host population within an infected individual. If and when

an individual gives rise to a pathogen genetic sequence within the dataset, this clock time adds to the clock time $Q_\Gamma(s)$ in determining the probability distribution of the measured sequence.

The separation of the pathogen evolutionary process into transmitted and untransmitted mutations has multiple interpretations. The choice of primary interpretation has consequences for the appropriate model specification of the branch separating the measurement node v from the transmission tree. The plausibility of these different interpretations will depend on the biological system under investigation.

(B1) Measurement error. Sequencing error could be modeled by an arbitrary evolution-like process on the branch separating the measured sequence from the transmissible sequence.

(B2) Transmissible versus measurable strains. The measured sequence may reflect the dominant strain reproducing most competitively within the host. It is conceivable that much of the diversity resulting from within-host evolution may lead to pathogens which are non-viable or non-competitive for between-host transmission. The evolutionary branch corresponding to the measurement event could represent this dead-end evolution, leaving the main body of the transmission tree to represent evolution of a transmissible strain.

(B3) Within-host diversity. A strain transmitted subsequent to sequencing could be more similar to an ancestral strain than to the sequenced strain by chance, due to within-host genetic variation, even without appealing to a phenomenon such as (B2). The measurement branch permits such behavior, so may help to adjust for unmodeled within-host pathogen genetic diversity.

Other model-specific quantities can be defined by additional query functions, but are not essential components of a GenPOMP model. For example, epidemiological models commonly consider the number of susceptible or removed individuals. Also, having defined an appropriate query function for a category of individuals, one can define a process counting such individuals. For example, to complement the query function Q_I for infected individuals, we can define a process

$$I(t) = \sum_{i \in H(t)} Q_I\big(X_i(t)\big)$$

counting the number of infected individuals in the study population. We can also write the

size of the study population at time t as,

$$P(t) = |H(t)| = \sum_{i \in H(t)} 1.$$

Our framework therefore incorporates the structure of arbitrary compartment models (Bretó et al., 2009) represented at the level of compartment membership for each individual.

The history of the query functions for infected individuals,

$$\big\{Q_I\big(X_i(u)\big), Q_D\big(X_i(u)\big), Q_L\big(X_i(u)\big), Q_\Gamma\big(X_i(u)\big), Q_\Delta\big(X_i(u)\big) : t_0 \leq u \leq t\big\}, \tag{A.2}$$

is sufficient to construct the transmission forest, $\mathcal{T}(t)$, and phylogeny, $\mathcal{P}(t)$, described in the New Approaches section of the main text. Formally, for (A.2) and elsewhere, we extend the query functions to take an undefined value, denoted by $\odot$, when the argument is outside the defined domain. To specify the measurement process model, recall that the measurement process $\{Y(t)\}$ consists of an increasing sequence of diagnosis times $\{T_n\}$ associated with the diagnosis counting process $\{D(t)\}$, together with a collection of genetic sequences $\{G_n\}$. We suppose that the sequences $\{G_n\}$ are modeled as a continuous-time Markov chain on $\mathcal{P}(t)$. The probability distribution of the genetic sequence G_n at time T_n, conditional on $\{X(t), t \leq t_n\}$ and $G_{1:n-1}$, therefore depends on $\mathcal{P}(t)$ and $G_{1:n-1}$. If a genetic sequence for the diagnosis at time T_n is not available, we assign G_n the value NA. We suppose this occurs with probability $1 - p_G$, independently of $\{X(t)\}$.

We have defined the GenPOMP model so that the pathogen genetic sequence arises only in the measurement model. No genetic sequences are included in the state process, or its particle representation. Our approach is consistent with viewing the genetic evolutionary model as a principled way to define and evaluate a statistical metric between genetic sequences that respects the tree structure of the evolutionary process and has the property that similar sequences are more likely to come from closely related pathogens. A measurement model satisfying these criteria and providing a statistical fit to the data need not be judged on the details of its biological strengths and weaknesses if the microevolutionary processes are not the focus of the investigation. The individual, stochastic molecular clocks determining the rate of evolution within each host are included in the latent process component of the GenPOMP model to facilitate Monte Carlo integration over the distribution of these clocks, as described in Section A.2.

The definition of a GenPOMP model given here is general and abstract. The population model $\{X(t)\}$ corresponds to an arbitrary individual-based Markovian model constrained to include concepts of transmission of a pathogen and measurement of pathogen genetic

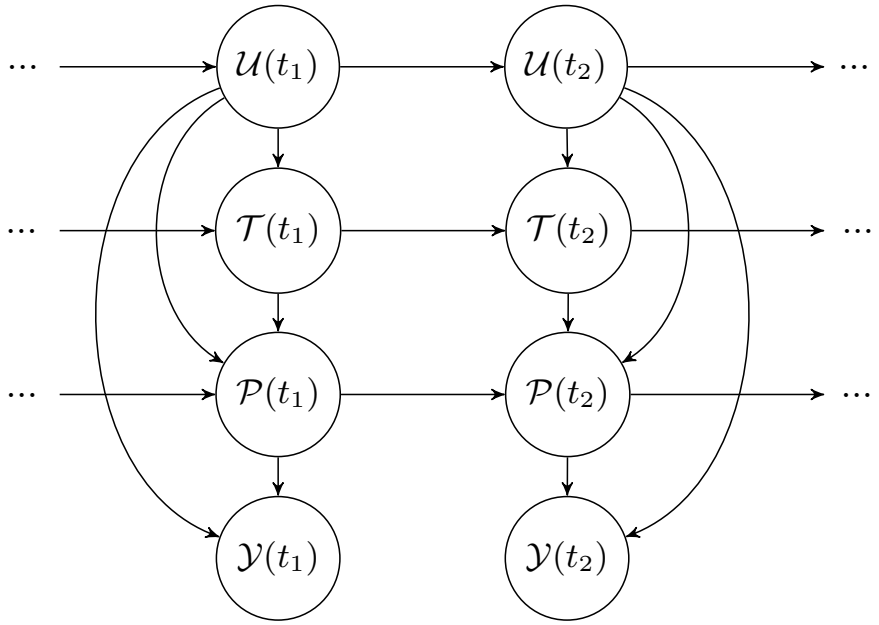

Figure A.1: A directed acyclical graph representation of dependencies among GenPOMP model components.

sequences. The measurement model is constrained to be based on a Markovian evolutionary process, but this is standard in current models used for phylodynamic inference. Our methodological approach applies to this general GenPOMP model class, subject to being able to simulate from the individual-based model and compute the rate at which individual hosts provide a pathogen sequence. The Markovian assumption is convenient algorithmically. In one sense, it is not fundamentally a limitation since non-Markovian models may be approximated by Markovian models with additional state variables. In another sense, it is a practical limitation since increasing the size of the state space increased the computational effort required.

A.2 A GenSMC algorithm for filtering the GenPOMP model

We develop a sequential Monte Carlo (SMC) approach for the framework of Section A.1. We will use the name GenSMC to describe an SMC algorithm for GenPOMP models. As an instance of SMC, the basic principles and theoretical foundation for GenSMC follows from the general theory of SMC (Liu, 2001). However, GenPOMP models have a particular structure that places particular demands on a GenSMC algorithm. Many variations are possible on our GenSMC algorithm, but demonstration of one successful GenSMC algorithm

provides a foundation and motivation for future improvements. Our GenSMC approach is presented as pseudocode in Algorithm 3, which is an expanded version of Algorithm 1 in the main text. We proceed to define the notation that will be required.

To construct our algorithm, we specify a concrete class of GenPOMP models. Let $\{X(t), t \in \mathbb{T}\}$ be a latent GenPOMP process with the form

$$X(t) = \Big(\Phi(t), \Psi(t), \Gamma(t), \Delta(t), D(t)\Big), \tag{A.3}$$

having components $\Phi(t)$, $\Psi(t)$, $\Gamma(t)$, $\Delta(t)$ and $D(t)$ defined as follows:

$\{D(t)\}$ records diagnosis events within the study population, as defined in Section A.1. We suppose that no diagnoses occur simultaneously, so $\{D(t)\}$ is a *simple* counting process. Therefore, we can model $\{D(t)\}$ via a conditional intensity process $\rho\big(\Phi(t), \Psi(t)\big)$ such that

$$\begin{aligned}
\mathbb{P}\big[D(t+\delta) - D(t) = 0 \,|\, \Phi(t), \Psi(t)\big] &= 1 - \delta\rho\big(\Phi(t), \Psi(t)\big) + o(\delta), \\
\mathbb{P}\big[D(t+\delta) - D(t) = 1 \,|\, \Phi(t), \Psi(t)\big] &= \delta\rho\big(\Phi(t), \Psi(t)\big) + o(\delta), \\
\mathbb{P}\big[D(t+\delta) - D(t) > 1 \,|\, \Phi(t), \Psi(t)\big] &= o(\delta),
\end{aligned}$$

where $o(\delta)$ denotes a function $f : [0, \infty) \to \mathbb{R}$ satisfying $\lim_{\delta \to 0} f(\delta)/\delta = 0$. Here, $\rho(X(t))$ is called the diagnosis rate.
$\{\Psi(t)\}$ is a piecewise constant process which records a list of the identity labels of individuals diagnosed by time t.

$\{\Phi(t)\}$ contains everything else in the GenPOMP model, so is essentially arbitrary within the general requirements of a GenPOMP model. We suppose that observation events are also recorded in the state process; specifically, the observation counting process $\{D(t)\}$ is a function of $\{\Phi(t)\}$ which gives rise to observation times $\{T_1, T_2, \dots\}$ at which the genetic measurements $\{G_1, G_2, \dots\}$ are made.

$\{\Gamma(t)\}$ is a list of the relaxed clock process for all the interior edges of the transmission tree, i.e., $\Gamma(t) = \{Q_\Gamma(X_i(t)), i \in \mathbb{N}\}$ where Q_Γ is defined in Section A.1.

$\{\Delta(t)\}$ is a list of the relaxed clock process for the terminal branches of the transmission tree, i.e., $\Delta(t) = \{Q_\Delta(X_i(t)), i \in \mathbb{N}\}$ where Q_Δ is defined in Section A.1.

The relaxed clock processes affect the micro-evolution of the pathogen, but in our model the genetic process has no consequence for the transmission dynamics: the genetic sequence is simply a marker, and the genetic models we use are models for neutral evolution. A

consequence of this is that the relaxed clock processes only have to be evaluated when needed to compute the conditional probability mass function for attaching a new genetic sequence to the genetic tree. If these components of the latent process can be computed when needed, there is no need to continually update them. Our computational strategy to take advantage of this is called a just-in-time representation and is formally described in Section A.3.4. Informally, the just-in-time representation is the tool that lets us define the latent GenPOMP model as a continuous-time Markov process while updating the relaxed clock processes at diagnosis times, when needed. To simulate the GenPOMP model forward in time using a just-in-time representation, we need to be able to evaluate the relaxed clock process over arbitrary time intervals, and also split the evolutionary time over a branch of the transmission tree if a new measurement divides this branch. An example of a Markovian clock with these properties is the Gamma process.

We will show that the relaxed clock processes $\{\Gamma(t)\}$ and $\{\Delta(t)\}$ can be represented by two processes $\{U(t)\}$ and $\{V(t)\}$ which generate the evolutionary clocks that are necessary to evaluate the likelihood of the sequences. The processes $\{U(t)\}$ and $\{V(t)\}$ are constant except at diagnosis times, and so are fully specified by the discrete processes $U_{0:N}$ and $V_{0:N}$, with $U_n = U(T_n)$ and $V_n = V(T_n)$. The construction of $\{U(t), V(t)\}$ is an instance of just-in-time variables, as discussed further in Section A.3.4. Therefore, for the purposes of Algorithm 3, it is convenient to replace the representation in equation (A.3) with an equivalent representation,

$$X(t) = \big(\Phi(t), \Psi(t), U(t), V(t), D(t)\big), \tag{A.4}$$

The construction of $\{U(t)\}$ and $\{V(t)\}$ is described in Figure A.2

Algorithm 3 is written using discrete time steps corresponding to the sequence of observation times, together with the start and end times of the interval $\mathbb{T}$. It convenient to define

$$t_0^* = t_0, \qquad t_{n^*+1}^* = t_{\text{end}},$$

so that $\mathbb{T} = [t_0^*, t_{n^*+1}^*]$. $\{\Psi(t)\}$ is fully specified by its values at the discrete set of observation times, and so we define a process $\{\Psi_n\}$ with

$$\Psi_n = \Psi(t_n^*).$$

To provide a discrete time representation of $\{\Phi(t)\}$, we write

$$\Phi_n = \{\Phi(t), t_{n-1}^* \leq t \leq t_n^*\},$$

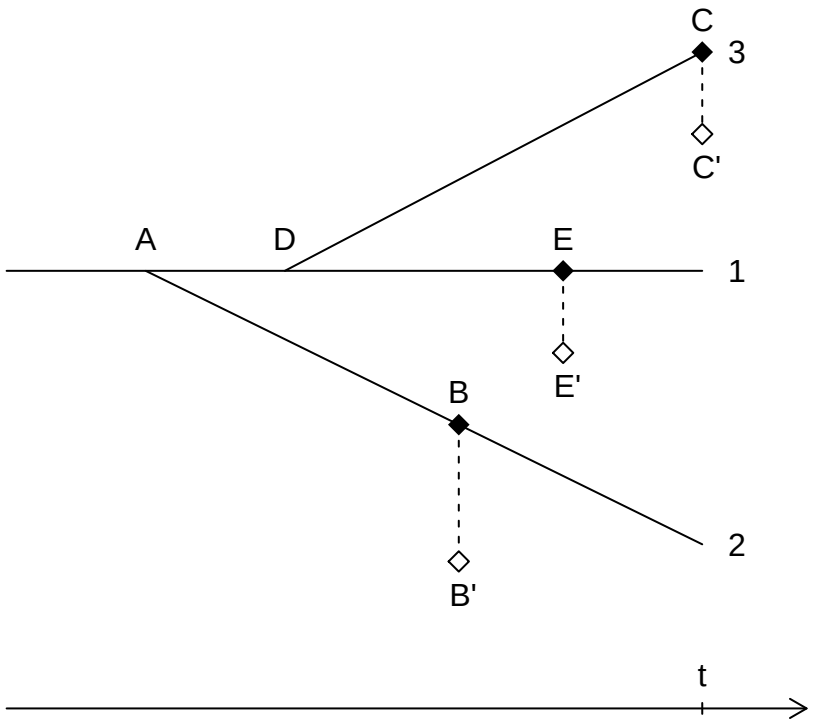

Figure A.2: The diagram represents the transmission tree for a particle where individual 1 infected individual 2 at time $A < t$ and individual 3 at time $D < t$. Sequences are collected at times B, C and E. Measured but untransmitted sequence mutations occur along BB', CC' and EE'. For this particle, we know that the sequence at time B corresponds to individual 2, and the sequence at time E belongs to individual 1. Suppose we then wish to evaluate the probability of the new sequence at time t conditional on it belonging to individual 3, as shown on the diagram. From the previous observed sequences, assigned to B' and E', this particle has already been assigned evolutionary clock times for the segments AB' and AE'. To place the new sequence at C', we first generate a new clock process for the segment DC', which is represented by the variable $U^P_{n,jkl}$ in step 8 of Algorithm 3. Then, we split the evolutionary clock time for AE' into AD and DE', in a way that is consistent with the corresponding calendar times and the stochastic evolutionary clock process. This computation is represented by the variable $V^P_{n,jklm}$ in step 10 of Algorithm 3.

for $n = 1, \ldots, n^* + 1$, with $\Phi_0 = \Phi(t_0^*)$. Similarly, we write

$$D_n = \{D(t), t_{n-1}^* \le t \le t_n^*\}.$$

Diagnosis events are modeled as perfectly observed, almost tautologically. We write $d^*(t)$ for the observed value of $D(t)$, defined as

$$d^*(t) = \sup\{n : t_n^* \le t\}.$$

Also, we write d_n^* for the observed value of D_n. Perfectly observed components of the latent process of a POMP model require special attention in sequential Monte Carlo algorithms, and so Algorithm 3 uses the targeted proposal developed in Section A.3.2 to handle the diagnosis process.

Hierarchical sampling (described in Section A.3.3) is carried out in Algorithm 3 over the components $\Phi(t)$ and $\Psi(t)$ in (A.3), as well as over the components U_n and V_n in the just-in-time representation of $\{\Gamma(t)\}$ and $\{\Delta(t)\}$.

The pseudocode for Algorithm 3 adopts a space-saving convention that index j always ranges over $1:J$, index k ranges over $1:K$, index l ranges over $1:L$, and index m ranges over $1:M$. Thus, for example, line 6 of Algorithm 3 has an implicit loop over $j \in 1:J$ and $k \in 1:K$.

If $g_n^* = \text{NA}$ then $w_2(n, j, k, l, m)$ is defined to be the probability of not recording a genetic sequence at diagnosis. In this case, steps 7 to 16 are not necessary: it suffices to take $K = 1$, with U_n and V_n being undefined. This special case is omitted from Algorithm 3 for simplicity.

To implement Algorithm 3, we require code to generate initial values, and to simulate the dynamic model for all the hierarchical layers conditional on the diagnosis events. Specifically, we require simulators for

$$f_{\Phi_0,\Psi_0}(\phi_0, \psi_0), \tag{A.5}$$

$$f_{\Phi_n|\Phi_{n-1},\Psi_{n-1},D_n}\big(\phi_n|\phi_{n-1}, \psi_{n-1}, d_n^*\big), \tag{A.6}$$

$$f_{\Psi_n|\Phi_n,\Psi_{n-1}}\big(\psi_n|\phi_n, \psi_{n-1}\big), \tag{A.7}$$

$$f_{U_n|U_{n-1},V_{n-1},\Phi_{0:n},\Psi_{0:n}}\big(u_n|u_{n-1}, v_{n-1}, \phi_{0:n}, \psi_{0:n}\big), \tag{A.8}$$

$$f_{V_n|V_{n-1},U_n,\Phi_{0:n},\Psi_{0:n}}\big(v_n|v_{n-1}, u_n, \phi_{0:n}, \psi_{0:n}\big). \tag{A.9}$$

We then require code to evaluate the diagnosis rate,

$$\rho\big(\phi(t), \psi(t)\big) \tag{A.10}$$

as well as the genetic measurement model,

$$f_{G_n|G_{1:n-1},\Phi_{0:n},\Psi_{0:n},U_{0:n},V_{0:n}}\big(g_n^* \,|\, g_{1:n-1}^*, \phi_{0:n}, \psi_{0:n}, u_{0:n}, v_{0:n}\big). \tag{A.11}$$

All the densities in (A.5–A.11) may additionally depend on a parameter vector θ.

Algorithm 3: GenSMC

input: dynamic model simulators listed in (A.5–A.9) and observation model evaluators (A.10, A.11); sequences, $g^*_{1:n^*}$; observation times $t^*_{1:n^*}$; initial time, t^*_0; terminal time, $t^*_{n^*+1} = t_{\text{end}}$; number of particles, J; number of hierarchical samples, K, L, M.

simulate $\big(\Phi^F_{0,j}, \Psi^F_{0,j}\big) \sim f_{\Phi_0,\Psi_0}(\phi_0, \psi_0)$ and set $U^F_{0,j} = V^F_{0,j} = 0$

for n in $1{:}n^*$ **do**

  propose transmission process: $\Phi^P_{n,j}(t) \sim f_{\Phi_n|\Phi_{n-1},\Psi_{n-1},D_n}\big(\phi_n|\Phi^F_{n-1,j}, \Psi^F_{n-1,j}, d^*_n\big)$

  set $w_1(n,j) = \exp\big\{-\int_{t^*_{n-1}}^{t^*_n} \rho\big(\Phi^P_{n,j}(t), \Psi^P_{n-1,j}(t)\big)\, dt\big\}\, \rho\big(\Phi^P_{n,j}(t^*_n), \Psi^P_{n-1,j}(t^*_n)\big)$

  set $\Phi^P_{0:n,j} = (\Phi^F_{0:n-1,j}, \Phi^P_{n,j})$

  propose attachment site: $\Psi^P_{n,jk} \sim f_{\Psi_n|\Phi_n,\Psi_{n-1}}\big(\psi_n|\Phi^P_{n,j}, \Psi^F_{n-1,j}\big)$

  set $\Psi^P_{0:n,jk} = (\Psi^F_{0:n-1,j}, \Psi^P_{n,jk})$

  evolution on the new branch:
  $U^P_{n,jkl} \sim f_{U_n|U_{n-1},V_{n-1},\Phi_{0:n},\Psi_{0:n}}\big(u_n|U^F_{n-1,j}, V^F_{n-1,j}, \Phi^P_{0:n,j}, \Psi^P_{0:n,jk}\big)$

  set $U^P_{0:n,jkl} = (U^F_{0:n-1,j}, U^P_{n,jkl})$

  evolution on the split branch:
  $V^P_{n,jklm} \sim f_{V_n|V_{n-1},U_n,\Phi_{0:n},\Psi_{0:n}}\big(v_n|V^F_{n-1,j}, U^P_{n,j}, \Phi^P_{0:n,j}, \Psi^P_{0:n,jk}\big)$

  set $V^P_{0:n,jklm} = (V^F_{0:n-1,j}, V^P_{n,jklm})$

  set $w_2(n,j,k,l,m) =$
  $f_{G_n|G_{1:n-1},\Phi_{0:n},\Psi_{0:n},U_{0:n},V_{0:n}}\big(g^*_n \,|\, g^*_{1:n-1}, \Phi^P_{0:n,j}, \Psi^P_{0:n,jk}, U^P_{0:n,jkl}, V^P_{0:n,jklm}\big)$

  weights: $w(n,j,k,l,m) = w_1(n,j)\, w_2(n,j,k,l,m)$

  set $w(n,j,k) = (1/LM) \sum_{l=1}^{L} \sum_{m=1}^{M} w(n,j,k,l,m)$

  resample: select index $(l',m')(j,k)$ with probability $\frac{w(n,j,k,l,m)}{w(n,j,k)}$

  set $w(n,j) = (1/K) \sum_{k=1}^{K} w(n,j,k)$

  resample: select index $k'(j)$ with probability $\frac{w(n,j,k)}{w(n,j)}$

  set $w(n) = (1/J) \sum_{j=1}^{J} w(n,j)$

  resample: select indices $j'(j)$ with probability $\frac{w(n,j)}{w(n)}$

  set $\Phi^F_{0:n,j} = \Phi^P_{0:n,j'(j)}$ and $\Psi^F_{0:n,j} = \Psi^P_{0:n,j'(j)\,k'(j')}$

  set $U^F_{0:n,j} = U^P_{0:n,j'(j)\,k'(j')\,l'(j',k')\,m'(j',k')}$ and $V^F_{0:n,j} = V^P_{0:n,j'(j)\,k'(j')\,l'(j',k')\,m'(j',k')}$

  conditional log likelihood estimate: $\hat{\ell}_{n|1:n-1} = \log w(n)$

end

simulate $\Phi^P_{n^*+1,j}(t) \sim f_{\Phi_{n^*+1}|\Phi_{n^*},\Psi_{n^*},D_{n^*+1}}\big(\phi(t)|\Phi^F_{n^*,j}, \Psi^F_{n^*,j}, d^*_{n^*+1}\big)$

set $w(n^*{+}1,j) = \exp\big\{-\int_{t^*_{n^*}}^{t_{\text{end}}} \rho\big(\Phi^P_{n^*+1,j}(t)\big)\, dt\big\}$

conditional log likelihood: $\hat{\ell}_{n^*+1|1:n^*} = \log\big\{(1/J) \sum_{j=1}^{J} w(n^*{+}1,j)\big\}$

output: log likelihood estimate: $\hat{\ell} = \sum_{n=1}^{n^*+1} \hat{\ell}_{n|1:n-1}$, and filtered state estimates

complexity: $\mathcal{O}(J\,K\,L\,M\,n\,\log n)$, assuming the transmission forest is balanced

A.2.1 The implementation of GenSMC in the `genPomp` program

Many computational issues arise in effective implementation of a GenSMC method such as Algorithm 3. Data structures are needed to keep track of the individuals in the study population, and the genetic relationships between pathogens in different hosts. Efficient implementation of all these computations, including use of a multi-processor computing environment, is necessary to work on problems of a practical scientific scale. The record of our implementation is the open-souce code for the `genPomp` program that we developed to carry out inference for GenPOMP models, available at `https://github.com/kingaa/genpomp`. The accuracy of `genPomp` has been successfully tested against exact analytic calculations for some very small scale situations, and against the `pomp` package (King et al., 2016a) for situations where no diagnoses lead to genetic sequences.

There is a substantial difference in the level of abstraction between the formal mathematical representation of a GenPOMP model in Algorithm 3 and the practical implementation in `genPomp`. One could write more pseudocode to bridge this gap, but that is beyond the scope of this article. We have focused instead on the foundational task of understanding how Algorithm 3 fits in with the theory and practice of SMC.

A.2.2 Extending GenSMC to infer unknown parameters: The GenIF algorithm

Sequential Monte Carlo (SMC) algorithms such as Algorithm 3 produce a Monte Carlo approximation to the likelihood of the model, but do not directly provide estimates of unknown parameters. A substantial literature has emerged on using SMC as a basis for statistical inference (Kantas et al., 2015). Iterated filtering (Ionides et al., 2006, 2015) uses SMC, together with parameter perturbations, to maximize the likelihood function. Iterated filtering has demonstrated effectiveness on various nonlinear models arising in infectious disease transmission studies (Ionides et al., 2015, and references therein). We developed an adaptation of Algorithm IF2 of Ionides et al. (2015), which we call GenIF as an abbreviation of *iterated filtering for GenPOMP models*. Our implementation of this GenIF algorithm is included within the `genPomp` program, as described fully in the source code. Conceptually, and computationally, GenIF is a simple extension to GenSMC. GenIF carries out multiple iterations of Algorithm 3 (GenSMC) adding perturbations to the candidate values of unknown parameters. GenSMC selects particles consistent with the data, and so allowing particles to have diversity in their parameters values naturally selects for parameter values consistent with the data. The theory and practice of iterated filtering focuses on using this phenomenon, with multiple SMC iterations having perturbations of decreasing

magnitude, to maximize the likelihood. Previous iterated filtering theory does not encompass the just-in-time variables employed by GenSMC. In the context of GenIF, this means that the current theoretical justification of IF2 (Ionides et al., 2015) does not perfectly apply when we carry out inference for the molecular evolution parameters. Heuristically, however, the principle of iterated filtering still applies, and we rely on empirical results to confirm that maximization performance is satisfactory.

Algorithms that permit numerically satisfactory likelihood maximization and likelihood evaluation provide a foundation for carrying out likelihood-based statistical inference. Profile likelihood methods can be used for obtaining confidence intervals, and likelihood ratio tests or Akaike's information criterion can be used for model selection.

A.2.3 Scalability of GenSMC

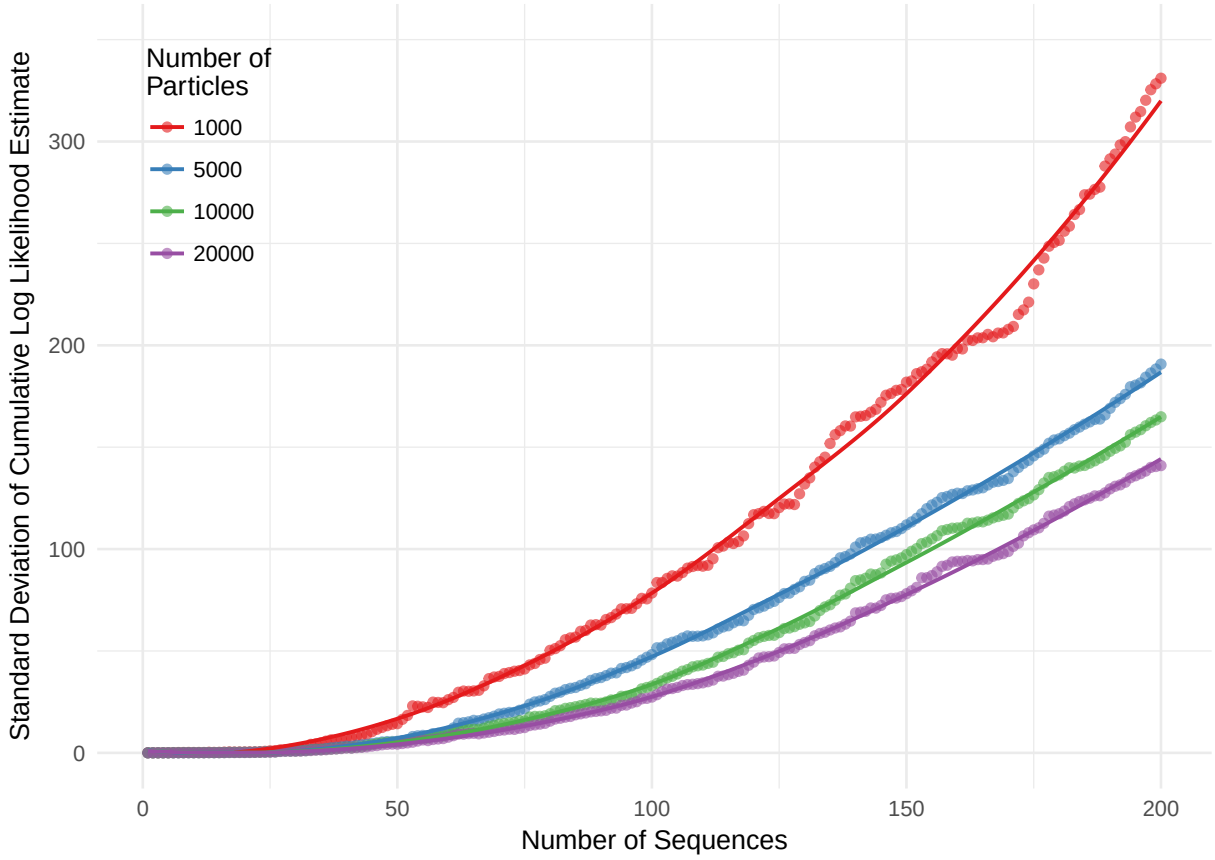

Figure A.3: Results from an experiment exploring how the standard deviation of the log likelihood estimate scales with both the number of sequences fit and the number of particles used. We ran the particle filter at the truth 80 times for each number of particles (1000, 5000, 10,000 and 20,000) on a simulated dataset of 200 sequences.

To explore the scalability of our GenSMC implementation, we performed the following experiments using simulated data. We first simulated an epidemic conditional on observing 200 sequences. We then ran the particle filter at the truth using 1000, 5000, 10,000, and 20,000 particles. For each number of particles we used we ran 80 particle filters. Finally, for each sequence, we computed the standard deviation of the cumulative log likelihood estimate across the 80 filtering evaluations. This computation yields a measure of the variability in the log likelihood estimate if one were to stop filtering at each sequence. The results from this experiment provide a controlled assessment of how Monte Carlo variance scales as the number of sequences grows. The standard deviation of the log likelihood estimate remains relatively small up to around 25 sequences (Figure A.3). An interpretation of this is that placing early sequences on the growing phylogenetic tree is relatively easy. It can become harder to find trees with appropriate places to attach later sequences, leading to increasing Monte Carlo variance. Monte Carlo variance is expected to grow as the size of a computational problem increases, but we did not find a rapid exponential growth. The peeling algorithm for computing the likelihood of the genetic sequences conditional on the phylogeny was typically the largest computational component, though not for all regions of parameter space.

A.3 A theoretical derivation of the GenSMC algorithm

To derive and justify GenSMC (Algorithm 3) for the GenPOMP model, we work up in stages from a simple and standard SMC algorithm. Initially working in discrete time, we start in Section A.3.1 by writing an SMC algorithm that allows for general dependence between the latent process and the observation process. Then, we consider a useful class of targeted proposal distributions in Section A.3.2. We add hierarchical layers of resampling in Section A.3.3. In Section A.3.4, we consider a *just-in-time* approach to construction of state variables which can have their creation postponed until necessary. In Section A.3.5, we move these developments into the context of continuous time models. Putting these components together, we obtain Algorithm 3.

A.3.1 A basic SMC algorithm

Consider a model consisting of a latent stochastic process $X_{0:N} = (X_0, X_1, \ldots, X_N)$ and an observable process $Y_{1:N} = (Y_1, Y_2, \ldots, Y_N)$. In this setting, N corresponds to the number of discrete time points, differing from the notation of Section A.1. Data consist of a sequence $y^*_{1:N} \in \mathbb{Y}^N$, modeled as a realization of $Y_{1:N}$. We suppose X_n and Y_n take values in measurable spaces $\mathbb{X}$ and $\mathbb{Y}$, and we require the existence of a joint density $f_{X_{0:N},Y_{1:N}}$ on

$\mathbb{X}^{N+1} \times \mathbb{Y}^N$. Conditional densities are denoted using subscripts, for example, the density of Y_n given $Y_{1:n-1}$ and $X_{0:n}$ is written as

$$f_{Y_n|X_{0:n},Y_{1:n-1}}(y_n \,|\, x_{0:n}, y_{1:n-1}). \tag{A.12}$$

In a standard POMP model, $\{X_n\}$ is a latent Markov process and the conditional distribution of Y_n depends only on X_n (Bretó et al., 2009). In the context of GenPOMP, we require the marginal Markov property for the latent process,

$$f_{X_n|X_{0:n-1}}(x_n \,|\, x_{0:n-1}) = f_{X_n|X_{n-1}}(x_n \,|\, x_{n-1}). \tag{A.13}$$

but we allow a general form for the measurement model in equation (A.12), where the conditional distribution of the nth observation can depend on the entire histories of the latent process and the observation process. SMC techniques for POMP models can be extended to this more general dependence structure (Liu, 2001). A basic SMC algorithm is outlined in Algorithm 4. This is essentially the basic bootstrap filter algorithm of Gordon et al. (1993), generalized to allow for the dependence on the history of the process in (A.12). Notationally, for Algorithm 4 we set $\mathscr{X}_n = X_{0:n}$ and use superscripts F and P to denote particles representing the filtering and prediction distributions respectively. We use systematic resampling in place of multinomial resampling (Arulampalam et al., 2002; Douc et al., 2005).

Algorithm 4: A basic Sequential Monte Carlo (SMC) algorithm:

input: simulator for $f_{X_n|X_{n-1}}(x_n \,|\, x_{n-1})$; simulator for $f_{X_0}(x_0)$; evaluator for $f_{Y_n|X_{1:n},Y_{1:n-1}}(y^*_n \,|\, x_{1:n}, y^*_{1:n-1})$; data, $y^*_{1:N}$; number of particles, J.

initialize filter particles: simulate $X^F_{0,j} \sim f_{X_0}(x_0)$ for j in $1:J$

initialize particle filter history: $\mathscr{X}^F_{0,j} = X^F_{0,j}$

for n in $1:N$ **do**

  prediction simulation: $X^P_{n,j} \sim f_{X_n|X_{n-1}}\big(x_n|X^F_{n-1,j}\big)$ for j in $1:J$.

  history of the prediction: $\mathscr{X}^P_{n,j} = \big(\mathscr{X}^F_{n-1,j}, X^P_{n,j}\big)$

  evaluate weights: $w(n,j) = f_{Y_n|X_{0:n},Y_{1:n-1}}(y^*_n|\mathscr{X}^P_{n,j}, y^*_{1:n-1})$ for j in $1:J$

  normalize weights: $\tilde{w}(n,j) = w(n,j)/\sum_{m=1}^{J} w(n,m)$

  apply systematic sampling to select indices $k_{1:J}$ with $\mathbb{P}\{k_j = m\} = \tilde{w}(n,m)$.

  resample: set $X^F_{n,j} = X^P_{n,k_j}$ and $\mathscr{X}^F_{n,j} = \mathscr{X}^P_{n,k_j}$ for j in $1:J$

  estimate conditional log likelihood: $\hat{\ell}_{n|1:n-1} = \log\big(J^{-1}\sum_{m=1}^{J} w(n,m)\big)$

end

output: log likelihood estimate, $\hat{\ell} = \sum_{n=1}^{N} \hat{\ell}_{n|1:n-1}$; filter sample, $\mathscr{X}^F_{n,1:J}$, for n in $0:N$.

Computational resources are an issue for GenPOMP models, since the spaces $\mathbb{X}$ and $\mathbb{Y}$ are both large. Furthermore, the dependence on the history in (A.12) leads to additional computational requirements for both memory and numerical operations. Careful implementation of SMC is therefore necessary to make the approach practical. We therefore proceed to develop extensions of Algorithm 4 that are necessary to improve numerical tractability for GenPOMP models.

To understand Algorithm 4, and subsequently extend it, we write out an algebraic justification of the prediction and filtering steps. For a general latent process $X_{0:N}$ and observable process $Y_{1:N}$ modeling data $y^*_{1:N}$ collected at times $t_{1:N}$, assuming (A.12) and (A.13), the *prediction identity* is

$$f_{X_{0:n}|Y_{1:n-1}}(x_{0:n} \,|\, y^*_{1:n-1}) = f_{X_n|X_{n-1}}(x_n \,|\, x_{n-1}) f_{X_{0:n-1}|Y_{1:n-1}}(x_{0:n-1} \,|\, y^*_{1:n-1}) \tag{A.14}$$

The SMC interpretation of (A.14) is that $f_{X_{0:n-1}|Y_{1:n-1}}(x_{0:n-1} \,|\, y^*_{1:n-1})$ is represented by a collection of J filter particles $\mathscr{X}^{F,j}_{n-1}$, $j = 1, \ldots, J$. Algorithm 4 corresponds to a basic version of SMC in which particle j has a time t_n value generated from $f_{X_n|X_{n-1}}(x_n \,|\, X^F_{n-1,j})$ to give rise to a time t_n prediction particle $X^P_{n,j}$. $X^P_{n,j}$ inherits its history from $X^F_{n-1,j}$ and so $\mathscr{X}^P_{n,j} = \big(\mathscr{X}^F_{n-1,j}, X^P_{n,j}\big)$.

A general *filtering identity* is

$$f_{X_{0:n}|Y_{1:n}}(x_{0:n} \,|\, y^*_{1:n}) = \left[\frac{f_{Y_n|X_{0:n},Y_{1:n-1}}(y^*_n \,|\, x_{0:n}, y^*_{1:n-1})}{f_{Y_n|Y_{1:n-1}}(y^*_n \,|\, y^*_{1:n-1})}\right] f_{X_{0:n}|Y_{1:n-1}}(x_{0:n} \,|\, y^*_{1:n-1}). \tag{A.15}$$

The SMC interpretation of (A.15) is that observation y^*_n requires the prediction particle $\mathscr{X}^P_{n,j}$ representing $f_{X_{0:n}|Y_{1:n-1}}(x_{0:n} \,|\, y^*_{1:n-1})$ to be given a weight proportional to $f_{Y_n|X_{0:n},Y_{1:n-1}}(y^*_n \,|\, \mathscr{X}^P_{n,j}, y^*_{1:n-1}$ The denominator on the right hand side of (A.15) is an irrelevant constant for computing the normalized weights. However, this denominator is approximated in Algorithm 4 as the normalizing constant, giving a Monte Carlo estimate of the nth term in a factorization of the likelihood of the data,

$$f_{Y_{1:N}}(y^*_{1:N}) = f_{Y_0}(y^*_0) \prod_{n=1}^{N} f_{Y_n|Y_{1:n-1}}(y^*_n \,|\, y^*_{1:n-1}). \tag{A.16}$$

For a discrete time representation of a simple GenPOMP model, Algorithm 4 might be directly applicable. For example one can take X_n to correspond to all the information about individuals in the population at time n, so that $X_{0:n}$ includes the transmission tree. We could also suppose that $X_{0:n}$ includes information on who would get sequenced if there are observed sequences—but not how many sequences were observed, which is part of the

measurement. For example, at each time point t_n, the state could contain a permutation listing the order in which eligible individuals are sequenced. This construction may appear somewhat contrived, and we proceed to relax it by allowing part of the latent process to be fully observed and therefore also be part of the measurement process. Regardless of that issue, evaluation of $f_{Y_n|X_{0:n},Y_{1:n-1}}(y_n^*|x_{0:n}, y_{1:n-1}^*)$ involves evaluating the likelihood of a phylogeny, which can be computed efficiently by a peeling algorithm, together with term for the probability of the sequence being collected.

A.3.2 A targeted SMC approach with a partial plug-and-play property

Some models of interest may have the feature that the event of obtaining a measurement has an appreciable consequence for the latent dynamics. HIV, for example, has the features that sequencing of the pathogen typically occurs at diagnosis. The fraction of infections which are sequenced is high, and diagnosis plays an important role in transmission dynamics both through changes in sexual contact behavior and reduced infectivity due to antiviral drugs. For HIV, it is therefore natural to consider models where sequencing events correspond to transitions of an individual between states and therefore correspond to a perfectly observed component of the latent process. This kind of situation needs some extra care, since $f_{Y_n|X_{0:n},Y_{1:n-1}}(y_n^*|\mathscr{X}_{n,j}^P, y_{1:n-1}^*)$ in Algorithm 4 becomes zero for every draw of $\mathscr{X}_{n,j}^P$ which is not consistent with y_n^*. The standard SMC approach to this is to allow for the possibility of targeted SMC proposal distributions, not necessarily the "vanilla" choice $f_{X_n|X_{n-1}}$. Suppose the proposal distribution for the SMC algorithm is $q_n(x_n|x_{n-1}, y_n^*)$, which is permissible since the proposal distribution is in general allowed to depend on any past, current or future observations. This corresponds to rewriting (A.14) as

$$f_{X_{0:n}|Y_{1:n-1}}(x_{0:n} \,|\, y_{1:n-1}^*) = \left[\frac{f_{X_n|X_{n-1}}(x_n \,|\, x_{n-1})}{q_n(x_n|x_{n-1}, y_n^*)}\right] q_n(x_n|x_{n-1}, y_n^*) f_{X_{0:n-1}|Y_{1:n-1}}(x_{0:n-1} \,|\, y_{1:n-1}^*), \tag{A.17}$$

which is interpreted to mean that the targeted SMC proposal particle $\mathscr{X}_{n,j}^P$, with $X_{n,j}^P$ drawn from $q_n(x_n|X_{n-1,j}^F, y_n^*)$, must be given a weight $f_{X_n|X_{n-1}}(X_{n,j}^P \,|\, X_{n-1,j}^F)\big\{q_n(X_{n,j}^P \,|\, X_{n-1,j}^F, y_n^*)\big\}^{-1}$ corresponding to the ratio in square brackets in (A.17).

A special case of a targeted proposal arises in the situation where part of the state variable is perfectly observed. To describe this situation, suppose we can partition the latent and observable processes as,

$$X_n = (A_n, B_n), \tag{A.18}$$

$$Y_n = (B_n, C_n), \tag{A.19}$$

with the data being $(b^*_{1:N}, c^*_{1:N})$. The prediction identity in (A.17) can then be written as

$$f_{A_n,B_n,X_{0:n-1}|Y_{1:n-1}}(a_n, b^*_n, x_{0:n-1} \,|\, y^*_{1:n-1})$$
$$= \left[\frac{f_{A_n,B_n|X_{n-1}}(a_n, b^*_n \,|\, x_{n-1})}{q_n(a_n|x_{n-1}, y^*_n)}\right] q_n(a_n|x_{n-1}, y^*_n) f_{X_{0:n-1}|Y_{1:n-1}}(x_{0:n-1} \,|\, y^*_{1:n-1}). \tag{A.20}$$

Then, to obtain the filtering distribution $f_{A_n,X_{1:n-1}|B_n,Y_{1:n-1}}(a_n, x_{0:n-1} \,|\, b^*_n, y^*_{1:n-1})$ one normalizes the weighted particle representation of $f_{A_n,B_n,X_{0:n-1}|Y_{1:n-1}}(a_n, b^*_n, x_{1:n-1}|y^*_{1:n-1})$ in (A.20), with the normalizing constant being the conditional likelihood, $f_{B_n|Y_{1:n-1}}(b^*_n|y^*_{1:n-1})$. A particular target choice of interest in (A.20) is

$$q_n(a_n|x_{n-1}, y^*_n) = f_{A_n|X_{n-1}}(a_n \,|\, x_{n-1}). \tag{A.21}$$

(A.20) becomes

$$f_{A_n,B_n,X_{0:n-1}|Y_{1:n-1}}(a_n, b^*_n, x_{0:n-1} \,|\, y^*_{1:n-1})$$
$$= \left[f_{B_n|A_n,X_{n-1}}(b^*_n|a_n, x_{n-1})\right] f_{A_n|X_{n-1}}(a_n \,|\, x_{n-1}) f_{X_{0:n-1}|Y_{1:n-1}}(x_{0:n-1} \,|\, y^*_{1:n-1}). \tag{A.22}$$

On the component of the state space that is not perfectly observed, the proposal in (A.21) is *plug-and-play* (Bretó et al., 2009; He et al., 2010) meaning that the algorithm needs only a simulation from $f_{A_n|X_{n-1}}(a_n \,|\, x_{n-1})$. However, we require numerically tractable evaluation of the importance sampling weight

$$f_{B_n|A_n,X_{n-1}}(b^*_n|a_n, x_{n-1}),$$

arising from the identity (A.22), and so we describe the algorithm as *partially plug and play*.

Using a targeted proposal typically leads to algorithms without the plug-and-play property. Here, we work with situations where $f_{B_n|A_n,X_{n-1}}(b^*_n|a_n, x_{n-1})$ is tractable, even if the complete transition density of (A_n, B_n) is intractable. Thus, $f_{A_n|X_{n-1}}(a_n \,|\, x_{n-1})$ can be specified in a fairly arbitrary way.

Example 1. B_n might be the number of diagnoses at time n, which might have a Poisson or negative binomial distribution conditional on A_n.

Example 2. Writing the number of sequenced diagnoses at time n by D^S_n, unsequenced diagnoses by D^U_n, and infected individuals by I_n, we might have $B_n = (D^S_n, D^U)$ and $A_n = I_n$. The joint distribution of D^S_n, D^U_n and $I_n - D^S_n - D^U_n$ might be multinomial given I_n.

Example 3. B_n might describe the race or age group of diagnosed individuals as well as whether they were sequenced.

A.3.3 SMC with hierarchical sampling

For computational considerations, it may be preferable to maintain J filtering particles and generate K prediction particles from each, rather than maintaining JK filtering particles. Computation of the K prediction particles can be localized on a single core of multi-processor hardware, and the memory usage may scale with J rather than JK.

In the context of Algorithm 4, extended to include the general proposal distribution of Section A.3.2, we write $\{X^P_{n,jk}, k \in 1:K\}$ for K draws from $q_n(x_n|X^F_{n-1,j}, y^*_n)$ for each value of j. We compute the weights in the second layer of the hierarchy by

$$w_{n,jk} = f_{X_n|X_{n-1}}\left(X^P_{n,jk} \,|\, X^F_{n-1,j}\right)\left[q_n\left(X^P_{n,j} \,|\, X^F_{n-1,j}, y^*_n\right)\right]^{-1}.$$

We then define $X^F_{n,j}$ to be a draw from $\{X^P_{n,jk}, k \in 1:K\}$ with probability proportional to $w_{n,jk}$, with the history $\mathscr{X}^F_{n,j}$ being constructed accordingly. We then assign $\mathscr{X}^F_{n,j}$ a weight

$$w_{n,j} = \frac{1}{K}\sum_{k=1}^{K} w_{n,jk}. \tag{A.23}$$

The filter particles $\{\mathscr{X}^F_{n,j}, j = 1, \ldots, J\}$ can be again resampled with weight proportional to $w_{n,j}$ if so desired. Resampling each layer of the hierarchy one at a time gives an approach that we call *staggered resampling*. It might sometimes be preferable to resample J particles from all JK particles $\{X^P_{n,jk}, j = 1:J, k = 1:K\}$ with weights $w_{n,jk}$. This process, resampling two or more layers of the hierarchy at the same time, we call *simultaneous resampling*. The staggered resampling in (A.23) can have computational advantages in terms of memory: one never needs to keep all JK particles in memory simultaneously. Also, staggered resampling is convenient in a multi-processor computational environment, where the computations for the first layer of the hierarchy can be split across processors and the second layer can be computed without any need for communication between processors.

Another motivation for hierarchical sampling arises when one can separate the generation of the prediction particle into a computationally expensive step followed by a cheap step. Heuristically, if the particles are large and computationally expensive, one wants to ensure that a particle does not get culled due to a single unfortunate draw from a proposal distribution. A component of the proposal distribution that is computationally expensive but not critical for the particle weight should be carried out relatively few times. By con-

trast, a component of the proposal distribution that is computationally cheap but critical for the particle weight, and hence for the survival of the particle, should be carried out relatively many times. For this motivation, there may be no compelling reason to carry out staggered resampling, in which case simultaneous resampling should be preferred. Both hierarchical sampling possibilities can arise in different parts of a single algorithm, potentially giving rise to several layers of sampling and resampling.

SMC with hierarchical sampling fits within the general theory of SMC (Naesseth et al., 2015), and theory exists to guide a good sampling structure (Skinner et al., 1989). In practice, however, preliminary experimentation is a good guide. Hierarchical resampling receives diminishing returns for increasing values of K, since since J is the basic Monte Carlo sample size which asymptotically justifies the Monte Carlo approach. Moderate values of $K > 1$ can have compelling practical advantages, which can be quantified by evaluating the variance of the Monte Carlo likelihood estimate.

A.3.4 Just-in-time evaluation of some state variable components

In equation A.3, our GenPOMP model included state processes $\{\Gamma(t)\}$ and $\{\Delta(t)\}$ which have no role in the dynamics, meaning that they do not affect the infinitesimal transition probabilities for $\{\Phi(t)\}$ and $\{\Psi(t)\}$ but do affect the measurements. If the measurements depend only on some subset or combination of these state variables, it is computationally desirable to generate the required subsets or combinations only when needed. Carrying out this computational shortcut, which we call just-in-time generation, does not change the model under consideration so long as the required variables are properly constructed at the time they become necessary. Two advantages to just-in-time state variable generation are

1. There may be state variables which, on some event of positive probability, have no effect on the measured components of the system. These state variables can be omitted when carrying out inferences on the rest of the system.

2. The sampling of these variables, and consequent resampling of particles, occurs only when information on the just-in-time variables arrives. In combination with hierarchical sampling (Section A.3.3), trying multiple copies of the just-in-time variables for each particle can help to prevent particles being lost due to a single unfortunate draw of a random variable.

To formalize the definition of just-in-time variables, we suppose that X_n can be split into two parts, written as

$$X_n = (\Phi_n, \Upsilon_n).$$

We say that $\Xi_n = h_n(X_{0:n})$ gives a just-in-time representation of Υ_n if

$$f_{Y_n|Y_{1:n-1},X_{0:n}}(y_n \,|\, y_{1:n-1}, x_{0:n}) = f_{Y_n|Y_{1:n-1},\Phi_{0:n},\Xi_{0:n}}\big(y_n \,|\, y_{1:n-1}, \phi_{0:n}, \xi_{0:n}\big), \tag{A.24}$$

where $\xi_n = h_n(x_{0:n})$. If we can evaluate (A.24) and simulate draws from $f_{\Phi_n,\Xi_n|\Phi_{0:n-1},\Xi_{n-1}}$, then we can effectively replace Υ_n by Ξ_n in an SMC method such as Algorithm 4. A particular case, arising in the just-in-time replacement of $\big(\Gamma(t), \Delta(t)\big)$ by $\big(U(t), V(t)\big)$ in Algorithm 3, occurs when the dynamics of $\{\Phi_n\}$ do not depend on $\{\Upsilon_n\}$, i.e.,

$$f_{\Phi_n|X_{0:n-1}}(\phi_n \,|\, x_{0:n-1}) = f_{\Phi_n|\Phi_{0:n-1}}(\phi_n \,|\, \phi_{0:n-1}). \tag{A.25}$$

In this case, implementing a just-in-time scheme requires that we can draw from $f_{\Xi_n|\Phi_{1:n},\Xi_{n-1}}$ and we can evaluate the density in equation (A.24). In practice, Ξ_0 may be a trivial random variable, since there is no observation at t_0, but this is not necessary for the just-in-time construction.

The utility of just-in-time evaluation depends in part on the reduction of dimension in replacing Ξ_n by Υ_n. For example, nothing is gained by the just-in-time representation $\Xi_n = \Upsilon_n$.

A.3.5 Moving from discrete time to continuous time

Continuous time Markov population models can be approximated in discrete time by a Markov chain (Bretó et al., 2009) using a stochastic Euler method. A continuous time measurement model can similarly be discretized to match the time steps of the Euler approximation. For a continuous time latent process model, suppose that $\{X(t), t \in \mathbb{T}\}$ is a right continuous stochastic process taking values in $\mathbb{X}$. We suppose that the continuous-time measurement process $\{Y(t)\}$ consists of a counting process, $\{D(t)\}$, together with a sequence of measurements $\{G_1, G_2, \dots\}$ where G_n occurs at time $T_n = \inf\{t : D(t) \geq n\}$. This notational setup is based on Section A.1, but we no not require any of the additional structure of a GenPOMP model at this point. We write $t_1^* < t_2^* < \cdots < t_{n^*}^*$ for the observation times of the data, $g_{1:n^*}^*$. Here, we suppose that $D(t)$ is part of $X(t)$ and, specifically, is represented by the observed component $B(t)$ in the decomposition

$$X(t) = \big(A(t), B(t)\big)$$

corresponding to a continuous-time version of equation (A.18). This situation arises in GenPOMP models when $\{D(t)\}$ counts diagnosis events for a disease transmission model $\{X(t)\}$, as in Section A.1. Suppose that the rate of observation events at time t does not

depend on the measurement process $\{Y_n : t_n \leq t\}$ given the current state process $X(t)$, i.e.,

$$\mathbb{P}[D(t+\delta) - D(t) = 0 \,|\, X(t), \{Y_s, s \leq t\}] = 1 - \rho(A(t))\,\delta + o(\delta), \quad \text{(A.26)}$$

$$\mathbb{P}[D(t+\delta) - D(t) = 1 \,|\, X(t), \{Y_s, s \leq t\}] = \rho(A(t))\,\delta + o(\delta). \quad \text{(A.27)}$$

Then, dividing the interval $(t^*_{n-1}, t^*_n]$ into subintervals of width δ and taking $\delta \to 0$, the limit of discrete approximations using (A.26) and (A.27) corresponds to a combined weight from evaluating (A.23) in each of the $1/\delta$ subintervals with no measurement followed by one subinterval with a measurement, i.e.,

$$\lim_{\delta \to 0} \left\{ \prod_{m=1}^{(t^*_n - t^*_{n-1})/\delta} \Big(1 - \rho\big(A(t^*_{n-1} + m\delta)\big)\Big) \right\} \times \rho\big(A(t^*_k)\big) f_{G_k|G_{1:n-1},T_{1:n},X_{0:n}}(g^*_n \,|\, g^*_{1:n-1}, t^*_{1:n}, x_{0:n})$$

$$= \exp\left\{ - \int_{t^*_{n-1}}^{t^*_n} \rho\big(A(s)\big)\, ds \right\} \times \rho\big(A(t^*_n)\big) \times f_{G_k|G_{1:n-1},T_{1:n},X_{0:n}}(g^*_n \,|\, g^*_{1:n-1}, t^*_{1:n}, x_{0:n}). \quad \text{(A.28)}$$

Note that one can view the first two terms of the product in equation (A.28) as a density with respect to Poisson counting measure.

A.4 Details of the HIV model used in the main text

In this section we provide additional details that describe the HIV model used in the main text. As the system is Markovian, we can fully specify the model by defining probabilities of each possible change to the state of the system given the current state over an interval of time δ. There are three types of events that change the state of system, each in a fundamentally different way:

1. An individual changes class. This event modifies an existing lineage on a transmission tree.

2. An individual in the study population infects a new individual. This event adds a new lineage to an existing transmission tree.

3. An individual outside the study population infects a new individual. This event seeds a new transmission tree consisting of a single individual. The genetic tree associated with with this new transmission tree joins all other genetic trees at the polytomy.

We define probabilities for the first two types of events from an individual-based perspective. Recall that the state of any individual i at time t is given by a random process $\{X_i(t)\}$.

The probabilities of class changes for each individual over an interval of time δ are given by

$$\begin{aligned}
\mathbb{P}\big[I_0 \to I_1\big] &= \delta\gamma_{I_0} + o(\delta),\\
\mathbb{P}\big[I_0 \to J_0\big] &= \delta\rho_{_0} + o(\delta),\\
\mathbb{P}\big[I_1 \to I_2\big] &= \delta\gamma_{I_1} + o(\delta),\\
\mathbb{P}\big[I_1 \to J_1\big] &= \delta\rho_{_1} + o(\delta),\\
\mathbb{P}\big[I_2 \to J_2\big] &= \delta\rho_{_2} + o(\delta),\\
\mathbb{P}\big[J_0 \to J_1\big] &= \delta\gamma_{J_0} + o(\delta),\\
\mathbb{P}\big[J_1 \to J_2\big] &= \delta\gamma_{J_1} + o(\delta),\\
\mathbb{P}\big[s \to \odot\big] &= \delta(\mu_s + \phi) + o(\delta).
\end{aligned} \tag{A.29}$$

Above, μ_s is a state-dependent death rate for an individual in state $s \in \mathbb{S} = \{I_0, I_1, I_2, J_0, J_1, J_2\}$, $X_i(t) = \odot$ if individual i is not in the study population at time t, and ϕ is a constant rate of emigration from the study population. The probability that an infected individual from inside the population gives rise to a new infection is,

$$\mathbb{P}\big[\text{the i}^{th} \text{ individual infects a new individual in } [t, t+\delta] |\ X_i(t) = s\big] \ = \ \delta\varepsilon_s + o(\delta),$$

where ε_s is the infectiousness of an individual in state s. The probability that an infected individual from outside the population gives rise to a new infection is,

$$\mathbb{P}\big[\text{an infection occurs from outside the study population in } [t, t+\delta]\big] \ = \ \delta\psi + o(\delta).$$

Note that this last probability, in contrast to those before, is not defined on a per capita basis. Also note that all new infections start in class I_0; this model does not allow immigration of later stage infected (or diagnosed) individuals into the population.

This model closely resembles a model from a recent phylodynamic analysis of the Detroit HIV epidemic Volz et al. (2013), but differs in key ways. First, whereas Volz et al. (2013) modeled incidence as a smooth, deterministic function, we model incidence mechanistically as a function of the states of individuals in the system. Second, instead of using a system of deterministic ordinary differential equations to model counts of individuals in each state, our model incorporates stochasticity into the process of state transitions.

A.4.1 Initial values for the HIV model

The *initial value* for a GenPOMP model is $X(t_0)$. In general, the initial value can be treated as an unknown parameter vector which can be estimated using our GenIF

methodology. There may be only limited information about these parameters in the data, but that is not a major problem for constructing profile likelihood estimates on other parameters of interest. However, a more parsimonious modeling approach is to set $X(t_0)$ to be a suitable function of the values of the dynamic parameters. For example, under a stationarity assumption for the dynamic system, one might set $X(t_0)$ to be a random draw from the stationary distribution or some mean value approximation to this. Our HIV model is not stationary, since we follow an age-cohort, but nevertheless we decided to initialize at plausible values given the dynamic parameters rather than estimate additional parameters. Further investigation could relax this assumption.

Part of the specification of $X(t_0)$ involves determining the genetic relationship assumed between infections that do not occur in the study population during the modeled period. The time t_0 at which we start modeling the population does not have to match the time at which we start to observe it. We could, for example, have zero sequencing probability before some time point. However, for our HIV model, these two times coincide. In the context of this HIV model, this component of the initial value involves determining the depth of the assumed polytomy, quantified by the time $t_{\text{root}} < t_0$ at which all trees in the transmission forest are modeled as meeting in the phylogenetic tree.

We carried out the following construction of the initial values of the membership of each compartment. We first note that the total number of diagnosed individuals is a perfectly observed quantity. By selecting a cohort, we have the advantage of working with a well-defined subpopulation. Over the time period from 2000 to 2012 we know exactly how many individuals were diagnosed. The MDCH dataset only has gene sequences between 2004 and 2012, so we decided to set $t_0 = 2004$. By 2004, the cohort grew to have 42 diagnosed individuals. Our aim in specifying initial counts is to apportion these 42 individuals to the three different classes of diagnosed individuals and populate the three unobserved states (the undiagnosed individuals) with counts. We assume no deaths over this period of four years. We constructed initial counts for each class by calculating under some additional assumptions under which these values become numerically tractable. First, we made the approximation that all rates of flow, with the exception of $h(t)$, are fixed at a current parameter estimate. Further, we suppose that $h(t)$ is constant at some fixed value,

$$h(t) = h_0,$$

ignoring the dependence of $h(t)$ on the state of the system. We then approximate the initial state by setting up and solving differential equations representing a deterministic solution to the model equation, formally equivalent to requiring the system of equations (A.29) to

hold in expectation. We fixed all rates of flow except $h(t_0)$ as described in the main text. Then, if the study cohort begins with all counts at zero in 2000, there is only one possible h_0 for which the total number of diagnoses in this approximating model matches the observed total number of diagnoses. We then solve for this value of h_0 and in doing so we obtain the counts in each compartment. Trajectories for the six states and their final values after four years are shown in Figure A.4. This approach to setting initial counts is not self-consistent with the model, as the model assumes that the rate of new infections is dependent on the state of the system, or with the timing of diagnoses observed in the four years leading up to the start of filtering. This simple way of setting the initial conditions is a starting point. Exploring the effect of initial conditions on model fits could be an area of future work.

We treated the time of the polytomy as an initial value parameter, with each particle starting with its own polytomy time. In this way, the polytomy time fits naturally into the iterated filtering maximization routines.

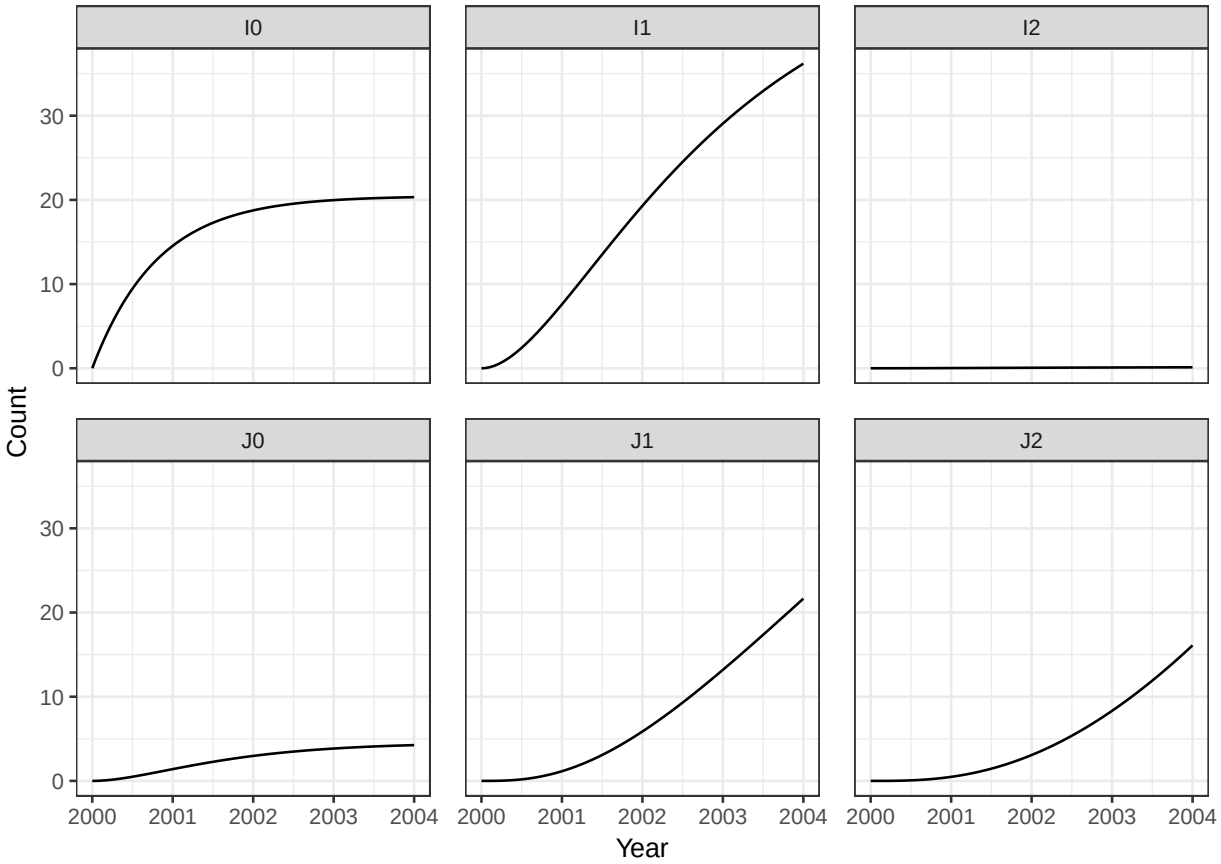

Figure A.4: Trajectories of counts of each class of infected individuals over four years prior to $t_0 = 2004$ when assuming a constant rate of new infections, all flows between and out of compartments as specified in the main text, and zero individuals initially in the cohort. We used the resulting counts in 2004 as the initial values for the data analysis.

A.4.2 Algorithmic parameters used for the numerical results

The choice of algorithmic parameters can affect the numerical efficiency of the GenSMC and GenIF algorithms. For large computations, when Monte Carlo variability is an appreciable component of parameter uncertainty, this can have an effect on the quality of the resulting statistical inferences. In Table A.1 we supply the algorithmic parameters that we used in the simulation study (for GenSMC) and in the data analysis (for both GenIF and GenSMC). We selected J, K, L and M such that Monte Carlo uncertainty on parameter estimates and confidence intervals was tolerable (Ionides et al., 2016) and such that runtimes were not prohibitively long.

Three of the algorithmic parameters are only used in GenIF: the random walk standard deviation, σ_{rw}, the cooling factor, α_c, and the number of GenIF iterations, I. Together, these parameters determine the extent to which GenIF shrinks the diameter of the parameter swarm. In the GenIF algorithm, perturbation of parameters over which we are maximizing occurs for each particle just before the proposal step. We perturb the parameters by multiplying each by a random deviate from a log normal distribution with mean one and standard deviation $\sigma_{rw}\alpha_c^i$, where $i \in \{0, 1, ..., I-1\}$ is the iteration of GenIF. This choice of perturbation is appropriate for nonnegative parameters. Although our framework allows for a different random walk standard deviation for each parameter, in this case we found that the same random walk standard deviation for all parameters was effective, and we report this value in Table A.1.

The algorithmic parameters in Table A.1 together with the source code at `https://github.com/kingaa/genpomp` are sufficient to reproduce the methodology we apply in our analysis. The HIV sequence data we analyzed are not publicly available, in accordance with our data use agreement with Michigan Department of Community Health.

Table A.1: Algorithmic parameters used in the simulation study and the data analysis.

Algorithmic parameter	**Description**	**Simulation Study**		**Data Analysis**			
		GenSMC		**GenIF**		**GenSMC**	
		Diagnosis data only	**Diagnosis data and genetic sequences**	**Diagnosis data only**	**Diagnosis data and genetic sequences**	**Diagnosis data only**	**Diagnosis data and genetic sequences**
J	Number of particles	10000	60000	10000	10000	10000	10000
K	Number of attachment sites per sequence	-	5	-	10	-	10
L	Number of relaxed clock gamma samples per attachment site	-	10	-	10	-	10
M	Number of relaxed clock beta samples per gamma	-	1	-	1	-	1
α_c	Cooling factor	-	-	0.95	0.95	-	-
σ_{rw}	Random walk standard deviation	-	-	0.01	0.01	-	-
I	Number of GenIF iterations	-	-	50	30	-	-

APPENDIX B

A targeted proposal to combat particle depletion due to perfect measurement

We derived this correction factor with the help of F. M. G. Magpantay.

B.1 Background

Perfect measurement of unobserved states may degrade efficiency of the particle filter. In particular, when it becomes highly unlikely that a particle will propose an unobserved state consistent with a perfect measurement, severe particle depletion may render inference infeasible.

In the context of hospital outbreak data, in which we have detailed data at the individual level, we may interpret measurement of a pathogen genetic sequence as indication of a true positive with no possibility of error. This is a perfect measurement in that the infection status of the individual is observed without error. Therefore, whenever a particle arrives at a pathogen genetic sequence, a sequence known to belong to a particular individual, if that individual is uninfected in the particle's proposed latent state then the particle is assigned a weight of zero and will be culled in the resample step. If prevalence is relatively low, parameter regimes that yield case counts consistent with the data may also be highly unlikely to recapitulate the particular pattern of perfectly observed infections via forward simulation. In this scenario, filtering failures are so common that inference is not possible.

One way to combat depletion due to perfect measurement is to use a targeted proposal. Instead of simulating the infection process in the usual fashion, we simulate conditional on the known censored infection times of a subset of individuals. Note that we do not perfectly observe the time of infection, rather we perfectly observe that some individuals are infected by certain times. Below, we describe the untargeted proposal, how we modify

this proposal to simulate conditional on censored observations of infection times, and how we correct for this targeted proposal in calculation of the likelihood.

B.1.1 The untargeted proposal

In the standard setting, when we simulate using the Gillespie algorithm, we assume constant transition rates and independence. Let $\{M, N\} \subset \mathbb{Z}_{>0}$, $\{\lambda_1, \ldots, \lambda_N\} \subset \mathbb{R}_{>0}$, $\{\xi_1, \ldots, \xi_M\} \subset \mathbb{R}_{>0}$, and let the set of random variables representing the times to the next event

$$\{X_1, \ldots, X_N, Y_1, \ldots, Y_M\}$$

be independent. Also define $\Lambda = \sum_{n=1}^{N} \lambda_n$ and $\Xi = \sum_{m=1}^{M} \xi_m$. If we assume that each X_n is exponentially distributed with parameter λ_n and each Y_m is exponentially distributed with parameter ξ_m then

$$F_{X_n}(t) = 1 - e^{-\lambda_n t} \tag{B.1}$$

$$f_{X_n}(t) = \lambda_n e^{-\lambda_n t} \tag{B.2}$$

$$F_{Y_m}(t) = 1 - e^{-\xi_m t} \tag{B.3}$$

$$f_{Y_m}(t) = \xi_m e^{-\xi_m t} \tag{B.4}$$

We split the random variables into two classes in anticipation of modifying the distribution of proposal times for a subset of the events in the case of the targeted proposal. Under the above assumptions, we can compute the density that a given event occurs first.

Suppose that X_ν occurs first for some $\nu \in \{1, \ldots, N\}$.

$$\begin{aligned}
&\mathbb{P}\big(\{X_\nu \in [t, t+\Delta t]\} \cap \{X_n > t + \Delta t,\, \forall n \neq \nu\} \cap \{Y_m > t + \Delta t,\, \forall m\}\big) \\
&= \mathbb{P}(X_\nu \in [t, t+\Delta t]) \prod_{n=1, n\neq\nu}^{N} \mathbb{P}(X_n > t + \Delta t) \prod_{m=1}^{M} \mathbb{P}(Y_m > t + \Delta t) && \text{(B.5)} \\
&= f_{X_\nu}(t)\Delta t \prod_{n=1, n\neq\nu}^{N} (1 - F_{X_n}(t)) \prod_{m=1}^{M} (1 - F_{Y_m}(t)) + o(\Delta t) && \text{(B.6)} \\
&= \lambda_\nu e^{-\lambda_\nu t}\Delta t \prod_{n=1, n\neq\nu}^{N} e^{-\lambda_n t} \prod_{m=1}^{M} e^{-\xi_m t} + o(\Delta t) && \text{(B.7)} \\
&= (\Delta t)\lambda_\nu e^{-\Lambda t} e^{-\Xi t} + o(\Delta t) && \text{(B.8)}
\end{aligned}$$

Suppose that Y_μ occurs first for some $\mu \in \{1, \ldots, M\}$.

$$\begin{aligned}
&\mathbb{P}\big(\{Y_\mu \in [t, t+\Delta t]\} \cap \{X_n > t+\Delta t,\, \forall n\} \cap \{Y_m > t+\Delta t,\, \forall m \neq \mu\}\big) \\
&= \mathbb{P}(Y_\mu \in [t, t+\Delta t]) \prod_{n=1}^{N} \mathbb{P}(X_n > t+\Delta t) \prod_{m=1, m\neq\mu}^{M} \mathbb{P}(Y_m > t+\Delta t) && \text{(B.9)} \\
&= f_{Y_\mu}(t)\Delta t \prod_{n=1}^{N} (1 - F_{X_n}(t)) \prod_{m=1, m\neq\mu}^{M} (1 - F_{Y_m}(t)) + o(\Delta t) && \text{(B.10)} \\
&= \xi_\mu e^{-\xi_\mu t}\Delta t \prod_{n=1}^{N} e^{-\lambda_n t} \prod_{m=1, m\neq\mu}^{M} e^{-\xi_m t} + o(\Delta t) && \text{(B.11)} \\
&= (\Delta t)\xi_\mu e^{-\Lambda t} e^{-\Xi t} + o(\Delta t) && \text{(B.12)}
\end{aligned}$$

B.1.2 A targeted proposal

Let $\{t_1^*, \ldots, t_M^*\} \subset \mathbb{R}_{>0}$ be observed times of sequencing, i.e., times by which individuals are known to be infected. One way to simulate such that these M individuals are sure to be infected by their time of sequencing is to allow for the rate of individual m becoming infected to increase as the simulation approaches t_m^*. We let the rate that individual m becomes infected at time $t \in [0, t_m^*]$ be

$$\phi_m(t) = \frac{1}{t_m^* - t}$$

A rate of this form implies a uniform distribution on the time of infection of individual m over the interval $[0, t_m^*]$. Under these changing rates for a subset of individuals, we again assume independence of the set of random variables representing event times

$$\{X_1, \ldots, X_N, \tilde{Y}_1, \ldots, \tilde{Y}_M\}$$

where each X_n is exponentially distributed with parameter λ_n and each $\tilde{Y}_m$ is uniform on the interval $[0, t_m^*]$. The distributions of the Y_m random variables are therefore:

$$F_{\tilde{Y}_m}(t) = \frac{t}{t_m^*}, \tag{B.13}$$

$$f_{\tilde{Y}_m}(t) = \frac{1}{t_m^*}. \tag{B.14}$$

Also define $P(t) = \prod_{m=1}^{M} \big(1 - F_{\tilde{Y}_m}(t)\big) = \prod_{m=1}^{M} \big(1 - \frac{t}{t_m^*}\big)$. Under this targeted proposal we can again compute the density that a given event occurs first.

Suppose that X_ν occurs first for some $\nu \in \{1, \ldots, N\}$.

$$\begin{aligned}
&\mathbb{P}\big(\{X_\nu \in [t, t+\Delta t]\} \cap \{X_n > t+\Delta t,\, \forall n \neq \nu\} \cap \{\tilde{Y}_m > t+\Delta t,\, \forall m\}\big) \\
&= \mathbb{P}(X_\nu \in [t, t+\Delta t]) \prod_{n=1, n\neq\nu}^{N} \mathbb{P}(X_n > t+\Delta t) \prod_{m=1}^{M} \mathbb{P}(\tilde{Y}_m > t+\Delta t) && \text{(B.15)} \\
&= f_{X_\nu}(t)\Delta t \prod_{n=1, n\neq\nu}^{N} (1 - F_{X_n}(t)) \prod_{m=1}^{M} (1 - F_{\tilde{Y}_m}(t)) + o(\Delta t) && \text{(B.16)} \\
&= \lambda_\nu e^{-\lambda_\nu t}\Delta t \prod_{n=1, n\neq\nu}^{N} e^{-\lambda_n t} \prod_{m=1}^{M} \big(1 - \frac{t}{t_m^*}\big) + o(\Delta t) && \text{(B.17)} \\
&= (\Delta t)\lambda_\nu e^{-\Lambda t} P(t) + o(\Delta t) && \text{(B.18)}
\end{aligned}$$

Suppose that $\tilde{Y}_\mu$ occurs first for some $\mu \in \{1, \ldots, M\}$.

$$\begin{aligned}
&\mathbb{P}\big(\{\tilde{Y}_\mu \in [t, t+\Delta t]\} \cap \{X_n > t+\Delta t,\, \forall n\} \cap \{\tilde{Y}_m > t+\Delta t,\, \forall m \neq \mu\}\big) \\
&= \mathbb{P}(\tilde{Y}_\mu \in [t, t+\Delta t]) \prod_{n=1}^{N} \mathbb{P}(X_n > t+\Delta t) \prod_{m=1, m\neq\mu}^{M} \mathbb{P}(\tilde{Y}_m > t+\Delta t) && \text{(B.19)} \\
&= f_{\tilde{Y}_\mu}(t)\Delta t \prod_{n=1}^{N} (1 - F_{X_n}(t)) \prod_{m=1, m\neq\mu}^{M} (1 - F_{\tilde{Y}_m}(t)) + o(\Delta t) && \text{(B.20)} \\
&= \frac{\Delta t}{t_\mu^*} \prod_{n=1}^{N} e^{-\lambda_n t} \prod_{m=1, m\neq\mu}^{M} \big(1 - \frac{t}{t_m^*}\big) + o(\Delta t) && \text{(B.21)} \\
&= \frac{\Delta t}{t_\mu^*} e^{-\Lambda t} \frac{P(t)}{1 - \frac{t}{t_\mu^*}} + o(\Delta t) && \text{(B.22)} \\
&= (\Delta t) \frac{e^{-\Lambda t} P(t)}{t_\mu^* - t} + o(\Delta t) && \text{(B.23)}
\end{aligned}$$

We can simulate from the targeted proposal using the first reaction method. That is, we find the first reaction that would occur among the X_n independently (by drawing $u_n \sim$ uniform(0,1) and setting $x_n = \frac{1}{\lambda_n} \ln\big(\frac{1}{u_n}\big)$, or by doing the standard Gillespie step here) and the first reaction to occur among the $\tilde{Y}_m$ (by drawing a $v_m \sim$ uniform(0,1) and setting $\tilde{y}_m = t_m^* v_m$) and compare to determine which reaction actually occurs first. After executing the event that happens first, we update rates appropriately and then iterate.

B.1.3 The correction term

When filtering, we need to correct the particle weight by the ratio of the density of the event under the untargeted proposal to the density under the targeted proposal. The relevant ratios taken in the limit that $\Delta t \to 0$ are

$$\frac{\mathbb{P}\big(\{X_\nu \in [t, t+\Delta t]\} \cap \{X_n > t+\Delta t,\, \forall n \neq \nu\} \cap \{Y_m > t+\Delta t,\, \forall m\}\big)}{\mathbb{P}\big(\{X_\nu \in [t, t+\Delta t]\} \cap \{X_n > t+\Delta t,\, \forall n \neq \nu\} \cap \{\tilde{Y}_m > t+\Delta t,\, \forall m\}\big)} \to \frac{e^{-\Xi t}}{P(t)}, \tag{B.24}$$

$$\frac{\mathbb{P}\big(\{Y_\mu \in [t, t+\Delta t]\} \cap \{X_n > t+\Delta t,\, \forall n\} \cap \{Y_m > t+\Delta t,\, \forall m \neq \mu\}\big)}{\mathbb{P}\big(\{\tilde{Y}_\mu \in [t, t+\Delta t]\} \cap \{X_n > t+\Delta t,\, \forall n\} \cap \{\tilde{Y}_m > t+\Delta t,\, \forall m \neq \mu\}\big)} \to \frac{\xi_\mu e^{-\Xi t}(t^*_\mu - t)}{P(t)} \tag{B.25}$$

As filtering proceeds, we carry in each particle a weight which is the ratio of the density of proposed events under the model to the density of proposed events under the targeted proposal. Whenever an event happens, we update the particle's weight by multiplying by the appropriate ratio given above. There are two instances when simulation pauses and no events have occurred: when the next event time proposed by the Gillespie algorithm either lies beyond the time of the next data point or lies beyond the time of an observed change in the system (movement, change in isolation status, or any other covariate that one may choose to condition on). In these instances we update the particle weight by the ratio of the probability of no event happening under the untargeted proposal versus the probability of no event happening under the targeted proposal. For some interval of time of length t over which no event occurs, the ratio of these probabilities is:

$$\frac{\mathbb{P}\big(\{X_n > t,\, \forall n\} \cap \{Y_m > t,\, \forall m\}\big)}{\mathbb{P}\big(\{X_n > t,\, \forall n\} \cap \{\tilde{Y}_m > t,\, \forall m\}\big)} = \frac{e^{-\Xi t}}{P(t)} \tag{B.26}$$

B.2 A simple test of the targeted proposal

Here we construct a simple scenario to test that the procedure described above correctly estimates the likelihood. In this scenario there is only one individual. This individual is known to be uninfected at time t_0 and subsequently tests positive for a culture at time t^*_1. Parameters of the model fit are given in Table B.1. We estimate the likelihood using both the untargeted proposal and using the targeted proposal. This scenario is simple, so we can compute the true likelihood by hand. Figure B.1 shows the empirical distribution

Table B.1: Parameters of the transmission model used in the simple test of the targeted proposal approach.

Parameter	**Interpretation**	**Value**
β_w	Within-ward transmission coefficient	0 day^{-1}
λ	Rate of infection from the community to the hospital	0.1 day^{-1}
γ	Rate of recovering from VRE+ to VRE-	0 day^{-1}
ρ_{cul}	Probability of a false positive culture	0
ϕ_{cul}	Probability of a false negative culture	0.01
t_0	Time to start filtering t_0	0 day
$N_S(t_0)$	Number of susceptible individuals at time t_0	1
$N_I(t_0)$	Number of infected individuals at time t_0	0

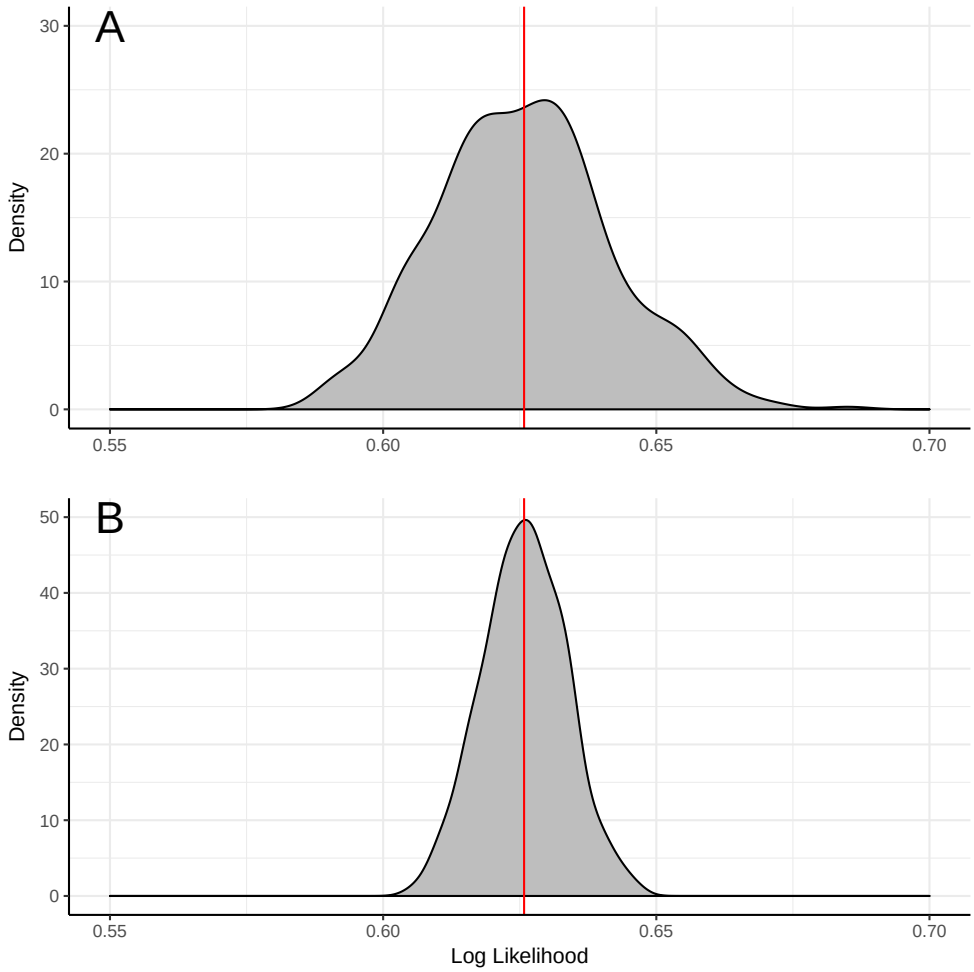

Figure B.1: A comparison of the distribution of likelihood estimates computed using the untargeted proposal (Panel A) and using the targeted proposal (Panel B). The true likelihood is shown in red. Both estimators generate a distribution centered around the truth, but the targeted proposal produces a distribution with a smaller variance.

BIBLIOGRAPHY

REFERENCES

Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T. (2002). A tutorial on particle filters for online nonlinear, non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing* **50,** 174 – 188.

Baele, G., Suchard, M. A., Rambaut, A., and Lemey, P. (2017). Emerging concepts of data integration in pathogen phylodynamics. *Syst Biol* **66,** e47–e65.

Bennett, D. E., Camacho, R. J., Otelea, D., Kuritzkes, D. R., Fleury, H., Kiuchi, M., Heneine, W., Kantor, R., Jordan, M. R., Schapiro, J. M., Vandamme, A.-M., Sandstrom, P., Boucher, C. A. B., van de Vijver, D., Rhee, S.-Y., Liu, T. F., Pillay, D., and Shafer, R. W. (2009). Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. *PLoS ONE* **4,** 1–8.

Biek, R., Pybus, O. G., Lloyd-Smith, J. O., and Didelot, X. (2015). Measurably evolving pathogens in the genomic era. *Trends in Ecology & Evolution* **30,** 306 – 313.

Boskova, V., Bonhoeffer, S., and Stadler, T. (2014). Inference of epidemiological dynamics based on simulated phylogenies using birth-death and coalescent models. *PLoS Comput Biol* **10,** e1003913.

Bouchard-Côté, A., Sankararaman, S., and Jordan, M. I. (2012). Phylogenetic inference via sequential Monte Carlo. *Systematic Biology* **61,** 579–593.

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

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

Dagum, L. and Menon, R. (1998). OpenMP: an industry standard API for shared-memory programming. *Computational Science & Engineering, IEEE* **5,** 46–55.

Del Moral, P. (2004). Feynman-Kac formulae: genealogical and interacting particle systems with applications. New York: Springer-Verlag.

Didelot, X., Gardy, J., and Colijn, C. (2014). Bayesian inference of infectious disease transmission from whole-genome sequence data. *Molecular Biology and Evolution* **31,** 1869–1879.

Douc, R., Cappé, O., and Moulines, E. (2005). Comparison of resampling schemes for particle filtering. In *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005*, pages 64–69. IEEE.

Drees, M., Snydman, D. R., Schmid, C. H., Barefoot, L., Hansjosten, K., Vue, P. M., Cronin, M., Nasraway, S. A., and Golan, Y. (2008). Prior environmental contamination increases the risk of acquisition of vancomycin-resistant enterococci. *Clinical Infectious Diseases* **46,** 678–685.

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

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

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., Pybus, O. G., Rambaut, A., Forsberg, R., and Rodrigo, A. G. (2003). Measurably evolving populations. *Trends in Ecology & Evolution* **18,** 481 – 488.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17,** 368–376.

Fisher, R., Bennett, J., and Bennett, H. (1999). *The Genetical Theory of Natural Selection: A Complete Variorum Edition*. OUP Oxford.

Frost, S. D., Pybus, O. G., Gog, J. R., Viboud, C., Bonhoeffer, S., and Bedford, T. (2015). Eight challenges in phylodynamic inference. *Epidemics* **10,** 88–92.

Giardina, F., Romero-Severson, E. O., Albert, J., Britton, T., and Leitner, T. (2017). Inference of transmission network structure from hiv phylogenetic trees. *PLoS Comput Biol* **13,** e1005316.

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B., and Suchard, M. A. (2013). Improving bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol Biol Evol* **30,** 713–24.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry* **81,** 2340–2361.

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

Gouliouris, T., Warne, B., Cartwright, E. J. P., Bedford, L., Weerasuriya, C. K., Raven, K. E., Brown, N. M., Török, M. E., Limmathurotsakul, D., and Peacock, S. J. (2018). Duration of exposure to multiple antibiotics is associated with increased risk of vre bacteraemia: a nested case-control study. *J Antimicrob Chemother* .

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L., Daly, J. M., Mumford, J. A., and Holmes, E. C. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303,** 327–332.

Hall, M. D., Woolhouse, M. E. J., and Rambaut, A. (2016). The effects of sampling strategy on the quality of reconstruction of viral population dynamics using bayesian skyline family coalescent methods: A simulation study. *Virus Evol* **2,** vew003.

He, D., Ionides, E. L., and King, A. A. (2010). Plug-and-play inference for disease dynamics: Measles in large and small towns as a case study. *Journal of the Royal Society Interface* **7,** 271–283.

Ho, S. Y. W. and Duchêne, S. (2014). Molecular-clock methods for estimating evolutionary rates and timescales. *Molecular Ecology* **23,** 5947–5965.

Ionides, E. L., Bretó, C., and King, A. A. (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the USA* **103,** 18438–18443.

Ionides, E. L., Breto, C., Park, J., Smith, R. A., and King, A. A. (2016). unpublished data, https://arxiv.org/abs/1612.02710, last accessed 21 December 2016.

Ionides, E. L., Nguyen, D., Atchadé, Y., Stoev, S., and King, A. A. (2015). Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. *Proceedings of the National Academy of Sciences of the USA* **112,** 719—724.

Kantas, N., Doucet, A., Singh, S. S., Maciejowski, J., and Chopin, N. (2015). On particle methods for parameter estimation in state-space models. *Statistical Science* **30,** 328–351.

Karcher, M. D., Palacios, J. A., Lan, S., and Minin, V. N. (2016). phylodyn: an R package for phylodynamic simulation and inference. *Molecular Ecology Resources* .

Kenah, E., Britton, T., Halloran, M. E., and Longini Jr, I. M. (2016). Molecular infectious disease epidemiology: survival analysis and algorithms linking phylogenies to transmission trees. *PLoS Comput Biol* **12,** e1004869.

King, A., Nguyen, D., and Ionides, E. (2016a). Statistical inference for partially observed markov processes via the r package pomp. *Journal of Statistical Software, Articles* **69,** 1–43.

King, A. A., Nguyen, D., and Ionides, E. L. (2016b). Statistical inference for partially observed markov processes via the R package pomp. *Journal of Statistical Software* **69,** 1–43.

Kingman, J. F. C. (1982a). The coalescent. *Stochastic Processes and their Applications* **13,** 235 – 248.

Kingman, J. F. C. (1982b). On the genealogy of large populations. *Journal of Applied Probability* **19,** 27–43.

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

Kühnert, D., Stadler, T., Vaughan, T. G., and Drummond, A. J. (2014). Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death sir model. *J R Soc Interface* **11,** 20131106.

Kühnert, D., Stadler, T., Vaughan, T. G., and Drummond, A. J. (2016). Phylodynamics with migration: A computational framework to quantify population structure from genomic data. *Mol Biol Evol* **33,** 2102–16.

Lau, M. S., Marion, G., Streftaris, G., and Gibson, G. (2015). A systematic Bayesian integration of epidemiological and genetic data. *PLoS Comput Biol* **11,** e1004633.

Lepage, T., Bryant, D., Philippe, H., and Lartillot, N. (2007). A general comparison of relaxed molecular clock models. *Molecular Biology and Evolution* **24,** 2669–2680.

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.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer, New York.

Luksza, M. and Lässig, M. (2014). A predictive fitness model for influenza. *Nature* **507,** 57–61.

Lythgoe, K. A. and Fraser, C. (2012). New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels. *Proceedings of the Royal Society B: Biological Sciences* **279,** 3367–3375.

Maulsby, C., Millett, G., Lindsey, K., Kelley, R., Johnson, K., Montoya, D., and Holtgrave, D. (2014). HIV among black men who have sex with men (MSM) in the United States: a review of the literature. *AIDS and Behavior* **18,** 10–25.

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,** 1459–71.

Moran, P. A. P. (1958). Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society* **54,** 60–71.

Naesseth, C. A., Lindsten, F., and Schön, T. (2015). Nested sequential Monte Carlo methods. In *Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6-11 July, 2015*, volume 37, pages 1292–1301. Journal of Machine Learning Research.

Neher, R. A. and Bedford, T. (2015). nextflu: real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics* **31,** 3546–8.

O'Driscoll, T. and Crank, C. W. (2015). Vancomycin-resistant enterococcal infections: epidemiology, clinical manifestations, and optimal management. *Infect Drug Resist* **8,** 217–30.

Paige, B., Wood, F., Doucet, A., and Teh, Y. W. (2014). Asynchronous anytime sequential Monte Carlo. *Advances in Neural Information Processing Systems* **27,** 3410–3418.

Papadimitriou-Olivgeris, M., Drougka, E., Fligou, F., Kolonitsiou, F., Liakopoulos, A., Dodou, V., Anastassiou, E. D., Petinaki, E., Marangos, M., Filos, K. S., and Spiliopoulou, I. (2014). Risk factors for enterococcal infection and colonization by vancomycin-resistant enterococci in critically ill patients. *Infection* **42,** 1013–22.

Peel, T., Cheng, A. C., Spelman, T., Huysmans, M., and Spelman, D. (2012). Differing risk factors for vancomycin-resistant and vancomycin-sensitive enterococcal bacteraemia. *Clin Microbiol Infect* **18,** 388–94.

Pond, S. L. K., Frost, S. D. W., and Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21,** 676–679.

Poon, A. F. Y. (2015). Phylodynamic inference with kernel ABC and its application to HIV epidemiology. *Molecular Biology and Evolution* .

Posada, D. and Crandall, K. A. (2001). Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1). *Molecular Biology and Evolution* **18,** 897–906.

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

Rasmussen, D. A., Kouyos, R., Günthard, H. F., and Stadler, T. (2017). Phylodynamics on local sexual contact networks. *PLoS Comput Biol* **13,** e1005448.

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

Rasmussen, D. A., Volz, E. M., and Koelle, K. (2014). Phylodynamic inference for structured epidemiological models. *PLoS Computational Biology* **10,** e1003570.

Ratmann, O., Donker, G., Meijer, A., Fraser, C., and Koelle, K. (2012). Phylodynamic inference and model assessment with approximate Bayesian computation: Influenza as a case study. *PLoS Computational Biology* **8,** e1002835.

Romero-Severson, E., Skar, H., Bulla, I., Albert, J., and Leitner, T. (2014). Timing and order of transmission events is not directly reflected in a pathogen phylogeny. *Molecular Biology and Evolution* .

Skinner, C., Holt, D., and Smith, T. (1989). *Analysis of complex surveys*. Wiley series in probability and mathematical statistics. Wiley.

Smith, D. J., Lapedes, A. S., de Jong, J. C., Bestebroer, T. M., Rimmelzwaan, G. F., Osterhaus, A. D. M. E., and Fouchier, R. A. M. (2004). Mapping the antigenic and genetic evolution of influenza virus. *Science* **305,** 371–6.

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,** 1119–27.

Stadler, T. (2010). Sampling-through-time in birth-death trees. *J Theor Biol* **267,** 396–404.

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

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.

Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* **10,** 512–526.

Tenorio, A. R., Badri, S. M., Sahgal, N. B., Hota, B., Matushek, M., Hayden, M. K., Trenholme, G. M., and Weinstein, R. A. (2001). Effectiveness of gloves in the prevention of hand carriage of vancomycin-resistant enterococcus species by health care workers after patient care. *Clin Infect Dis* **32,** 826–9.

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

Vaughan, T. G., Kühnert, D., Popinga, A., Welch, D., and Drummond, A. J. (2014). Efficient bayesian inference under the structured coalescent. *Bioinformatics* **30,** 2272–2279.

Volz, E. M. and Didelot, X. (2018). Modeling the growth and decline of pathogen effective population size provides insight into epidemic dynamics and drivers of antimicrobial resistance. *Syst Biol* .

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

Volz, E. M., Ionides, E. L., Romero Severson, E., Brandt, M., Mokotoff, E., and Koopman, J. S. (2013). 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. (2013). Viral phylodynamics. *PLoS Computational Biology* **9,** e1002947.

Volz, E. M., Kosakovsky Pond, S. L., Ward, M. J., Leigh Brown, A. J., and Frost, S. D. W. (December 2009). Phylodynamics of infectious disease epidemics. *Genetics* **183,** 1421–1430.

Wendt, C., Wiesenthal, B., Dietz, E., and Rüden, H. (1998). Survival of vancomycin-resistant and vancomycin-susceptible enterococci on dry surfaces. *Journal of Clinical Microbiology* **36,** 3734–3736.

Wright, S. (1931). Evolution in mendelian populations. *Genetics* **16,** 97–159.