Modeling the growth and decline of pathogen effective population size provides insight into epidemic dynamics and drivers of antimicrobial resistance

Erik M. Volz[1] and Xavier Didelot[1]

[1] *Department of Infectious Disease Epidemiology, Imperial College London,*

Norfolk Place, W2 1PG, United Kingdom

Corresponding author: Erik Volz, Department of Infectious Disease Epidemiology, Imperial College London, Norfolk Place, W2 1PG, United Kingdom; E-mail: e.volz@imperial.ac.uk

Abstract

Non-parametric population genetic modeling provides a simple and flexible approach for studying demographic history and epidemic dynamics using pathogen sequence data. Existing Bayesian approaches are premised on stationary stochastic processes which may provide an unrealistic prior for epidemic histories which feature extended period of exponential growth or decline. We show that non-parametric models defined in terms of the growth rate of the effective population size can provide a more realistic prior for epidemic history. We propose a non-parametric autoregressive model on the growth rate as a prior for effective population size, which corresponds to the dynamics expected under many epidemic situations. We demonstrate the use of this model within a Bayesian phylodynamic inference framework. Our method correctly reconstructs trends of epidemic growth and decline from pathogen genealogies even when genealogical data is sparse and conventional skyline estimators erroneously predict stable population size. We also propose a regression approach for relating growth rates of pathogen effective population size and time-varying variables that may impact the replicative fitness of a pathogen. The model is applied to real data from rabies virus and *Staphylococcus aureus* epidemics. We find a close correspondence between the estimated growth rates of a lineage of methicillin-resistant *S. aureus* and population-level prescription rates of β-lactam antibiotics. The new models are implemented in an open source R package called *skygrowth* which is available at `https://mrc-ide.github.io/skygrowth/`.

(Keywords: phylodynamics, effective population size, growth rate, *skygrowth* , antimicrobial resistance, MRSA)

Non-parametric population genetic modeling has emerged as a simple, flexible, popular and powerful tool for interrogating genetic sequence data to reveal demographic history (Ho and Shapiro 2011). This approach has proved especially useful for analysis of pathogen sequence data to reconstruct epidemic history and such models are increasingly incorporated into surveillance systems for infectious diseases (Volz et al. 2013). The most commonly used techniques are derivatives of the original *skyline* coalescent model, which describes the evolution of effective population size as a piecewise constant function of time (Pybus et al. 2000). The basic *skyline* model is prone to overfitting and estimating drastic fluctuations in effective population size, so that numerous approaches were subsequently developed for smoothing population size trajectories. Initial approaches to smoothing *skyline* estimators were based on aggregating adjacent coalescent intervals within a maximum likelihood framework (Strimmer and Pybus 2001). Subsequent development has largely focused on Bayesian approaches where a more complex stochastic diffusion process provides a prior for the evolution of a piecewise-constant function of effective population size (Drummond et al. 2005). Non-parametric Bayesian approaches are now the most popular approach for phylodynamic inference and such approaches have illuminated the epidemic history of numerous pathogens in humans and animals (Ho and Shapiro 2011).

To date, all Bayesian non-parametric models have assumed that the effective population size (or its logarithm) follows a stationary stochastic process such as a Brownian motion (Minin et al. 2008; Palacios and Minin 2013). The choice of a stationary process as prior can have large influence on size estimates especially when genealogical data is sparse and uninformative. Genealogies often provide very little information about effective population size near the present (or most recent sample), especially in exponentially increasing populations (de Silva et al. 2012). In such cases, *skyline* estimators with Brownian motion priors on the effective population size may produce estimates which stabilize at a constant level even when the true size is increasing or decreasing exponentially. We argue that in many situations, a more realistic prior can be defined in terms of the growth rate of the effective population size. Below, we describe such a prior based on a simple autoregressive stochastic process defined on the growth rate of effective population size.We show how this prior can lead to substantially different estimates and argue that these estimates are more accurate in many situations. When genealogical data is sparse, our model will retain the growth rate learned from other parts of the genealogy and will correctly capture trends of exponential growth or decline. Even though our approach is non-parametric, we consider its relationship with parametric models of epidemic population genetics to show that our estimates of growth rates of pathogen effective population size are often likely to correspond to growth rates of an infectious disease epidemic.

Smoothing effective population size trajectories using a prior on growth rates also has important advantages when incorporating non-genetic covariate data into phylodynamic inference (Baele et al. 2016). Recent work has focused on refining effective population size estimates using both the times of sequencing sampling (Karcher et al. 2016) or using environmental data which are expected to correlate with size estimates, such as independent epidemic size estimates based on non-genetic data (Gill et al. 2016). Existing statistical models have assumed that the effective population size has a linear or log-linear relationship with temporal covariates. However in many cases, a more realistic model would specify that the growth rate of effective population size is correlated with covariates, as when for example an environmental variable impacts the replicative fitness of a pathogen. We provide a similar extension of previous *skyride* models with covariate data (Gill et al. 2016) to show how such data can be used to test hypotheses concerning their effect and, when a significant effect exists, to refine estimates of both the growth rates and the effective population sizes.

We illustrate the potential advantages of our growth rate model using a rabies virus dataset that has been thoroughly studied using previous phylodynamic methods (Biek et al. 2007; Gill et al. 2016). In particular, we show how our model correctly estimates a recent decline in epidemic size whereas previous models mistakenly predict a stabilisation of the epidemic prevalence. We also apply our methodology to a genomic dataset of methicilin-resistant *Staphylococcus aureus* that had not formally been analysed using phylodynamic methods (Uhlemann et al. 2014). We show how time series on prescription rates of β-lactam antibiotics correlate strongly with growth and decline of the effective population size, revealing the impact of antibiotic use on the emergence and spread of resistant bacterial pathogens.

Methods and Materials

We model effective population size through time as a first order autoregressive stochastic process on the growth rate. This provides an intuitive link between the growth rate of effective population size of pathogens and epidemic size as well as the reproduction number of the epidemic. We further show how to incorporate time-varying environmental covariates into phylodynamic inference.

Previous Bayesian non-parametric phylodynamic models

Several non-parametric phylodynamic models have been proposed based on Brownian motion (BM) processes and the Kingman coalescent genealogical model (Kingman 1982). In particular, the Bayesian non-parametric *skyride* model uses a BM prior to smooth trajectories of the logarithm of the effective population size (Minin et al. 2008). Let $\gamma(t) = \log(\mathrm{Ne}(t))$ denote the logarithm of the effective population size as a function of time. The BM prior is defined as:

$$\gamma(t + \mathrm{d}t) \sim \gamma(t) + \mathcal{N}(0, \mathrm{d}t/\tau) \tag{1}$$

where τ is an estimated precision parameter, for which an uninformative Gamma prior is typically used.

This BM prior has been adapted and applied in a variety of ways to enable statistical inference. In the *skygrid* model (Gill et al. 2012), time is discretized and γ is defined to be a piecewise constant function of time over a grid with time increments h, and the value γ_i is estimated for each interval i. Time intervals do not in general correspond to coalescent times in the genealogy. In this case, the BM prior is computed over increments of γ:

$$p(\gamma_{1:m}|\tau) \propto \prod_{i=1}^{m-1} p(\gamma_{i+1} - \gamma_i|\tau) \tag{2}$$

where

$$p(\gamma_{i+1} - \gamma_i|\tau) = \sqrt{\frac{\tau}{2\pi h}} e^{-\frac{\tau}{2h}(\gamma_{i+1}-\gamma_i)^2}$$

The genealogical data takes the form $\mathcal{G} = (c_{1:(n-1)}, s_{1:n})$ where c and s are respectively ordered coalescent times (internal nodes of the genealogy) and sampling times (terminal nodes of the genealogy). In the coalescent framework, the sampling times are usually considered to be fixed, so that $p(s) = 1$ and $p(\mathcal{G}) = p(c|s)$. Alternatively, in some variations of this model, a prior $p(s|\mathrm{Ne})$ is also provided for the sequence of sampling times, making this approach similar to but more flexible than sampling-birth-death-models (Karcher et al. 2016; Volz and Frost 2014).

Given a genealogy, the posterior distribution of the parameters τ and $\gamma_{1:m}$ is decomposed as:

$$p(\gamma_{1:m}, \tau|\mathcal{G}) \propto p(\mathcal{G}|\gamma_{1:m})p(\gamma_{1:m}|\tau)p(\tau) \tag{3}$$

The second term is given by Equation 2 and the last term by the prior on τ. To assist with the definition of the first term, we first denote $A(t)$ to be the number of extant lineages at time t:

$$A(t) = \sum_{i=1}^{n} I(s_i > t) - \sum_{i=1}^{n-1} I(c_i > t) \tag{4}$$

where $I(x)$ is an indicator function equal to one when x is true and equal to zero otherwise. The probability density of the genealogical data given the population size history $\gamma_{1:m}$ is then equal to (Griffiths and Tavare 1994):

$$\begin{aligned} p(\mathcal{G}|\gamma_{1:m}) = \prod_{i=1}^{2n-2} \Bigg(& I(t_i \in c_i) \frac{\binom{A(t_i)}{2}}{\mathrm{Ne}(t_{i+1})} e^{-\int_{t_i}^{t_{i+1}} -\binom{A(t_i)}{2} \frac{1}{\mathrm{Ne}(t)} \mathrm{d}t} \\ & + (1 - I(t_i \in c_i)) e^{-\int_{t_i}^{t_{i+1}} -\binom{A(t_i)}{2} \frac{1}{\mathrm{Ne}(t)} \mathrm{d}t} \Bigg) \end{aligned} \tag{5}$$

where $t_{1:(2n-1)} = c_{1:(n-1)} \cup s_{1:n}$ is the set union of sample and coalescent times in descending order.

Relationship between the growth rate of effective population size and epidemic properties

Several recent studies have investigated the relationship between the effective population size of a pathogen and the number of infected hosts (Koelle et al. 2011; Dearlove and Wilson 2013; Rosenberg and Nordborg 2002). A simple link between these quantities does not exist, since the relationship depends on how incidence and epidemic size change through time (Volz et al. 2009), population structure (Volz 2012), and complex evolution of the pathogen within hosts (Didelot et al. 2016; Volz et al. 2017). Under idealized situations, there is however a simple relationship between the growth rate of effective population size and the growth rate of an epidemic (Frost and Volz 2010; Volz et al. 2013).

Let $Y(t)$ and $\beta(t)$ denote the number of infected hosts and per-capita transmission rate, respectively, as functions of time. Note that $\beta(t)$ may depend on the density of susceptible individuals in the population, as in the common susceptible-infected-removed (SIR) model, in which case $\beta(t) \propto S(t)/N$ (Allen 2008). The coalescent rate for an infectious disease epidemic was previously derived under the assumption that within-host effective population size is negligible and that super-infection does not occur (Volz et al. 2009; Frost and Volz 2010):

$$\lambda(t) = \binom{A(t)}{2} \frac{2\beta(t)}{Y(t)} \qquad (6)$$

Equating this rate with the coalescent rate under the coalescent model $\lambda(t) = \binom{A(t)}{2}/\mathrm{Ne}(t)$ (Kingman 1982) yields the following formula for the effective population size:

$$\mathrm{Ne}(t) = \frac{Y(t)}{2\beta(t)} \qquad (7)$$

Differentiating with respect to time (denoting with a dot superscript) yields:

$$\dot{\mathrm{Ne}}(t) = \frac{\dot{Y}(t)}{2\beta(t)} - \frac{\dot{\beta}(t)Y(t)}{2(\beta(t))^2} \qquad (8)$$

Note that in general the growth rate of the effective population size does not correspond to the growth rate of Y, however if the per-capita transmission rate is constant ($\dot{\beta} = 0$), we have $\dot{\mathrm{Ne}} = \dot{Y}/(2\beta) \propto \dot{Y}$. Thus, we expect that over phases of the epidemic where per-capita transmission rates are nearly constant there will be close correspondence between the growth or decline of the effective population size and the growth or decline of the unobserved number of infected hosts. This condition is often satisfied near the beginning of an outbreak which has an exponential phase. It is also often satisfied towards the end of epidemics when the epidemic size is decreasing at a constant exponential rate.

The basic reproduction number R_0 describes the expected number of transmission events caused by a single infected individual in an otherwise susceptible population. By extension, we can define $R(t)$ as the expected number of transmissions by an infected host infected at time t (Fraser 2007). Assuming that all infected individuals are equally infectious (as is the case for example in the SIR model), we have that during periods when the epidemic growth rate is constant, each infected individual transmits at rate $\beta(t) = R(t)/\psi$ where ψ is the mean duration of infections. With these definitions, the number of infections $Y(t)$ varies according to the following differential equation:

$$\dot{Y}(t) = Y(t)\frac{R(t) - 1}{\psi} \tag{9}$$

Combining Equations 7, 8 and 9 leads to the following approximate estimator for the reproduction number through time:

$$\hat{R}(t) = 1 + \psi\frac{\dot{\mathrm{Ne}}(t)}{\mathrm{Ne}(t)} \tag{10}$$

This estimator makes use of the quantity $\dot{\mathrm{Ne}}(t)/\mathrm{Ne}(t)$ which will be estimated in our model below. Equation 10 is likely to be a good estimator over periods of the epidemic where per-capita transmission rates are invariant. A special case of this occurs at the start of an epidemic, in which case Equation 10 can be used to estimate the basic reproduction number R_0, as previously noted (Pybus 2001).

A growth rate prior for effective population size

We propose a model in which the growth rate of the effective population size is an autoregressive process with stationary increments. This growth rate is defined as:

$$\rho(t) = \frac{\dot{\mathrm{Ne}}(t)}{\mathrm{Ne}(t)} \tag{11}$$

Note that $\rho(t)$ is a real-valued quantity, with negative and positive values respectively indicating an increase and decrease in the effective population size. In particular, if the population is exponentially growing or declining from $t = 0$ then we have $\mathrm{Ne}(t) = \mathrm{Ne}(0)\exp(\rho t)$ so that $\rho(t) = \rho$ at every time $t \geq 0$. More generally, we model $\rho(t)$ using a BM process: $\rho(t) \sim \mathrm{BM}(\tau)$ (cf Equation 1). To facilitate statistical inference, we work with a discretized time axis with m intervals of length h as in the *skygrid* model (Gill et al. 2013). We define the growth rate in time interval i as:

$$\rho_i = \frac{\mathrm{Ne}_{i+1} - \mathrm{Ne}_i}{h\mathrm{Ne}_i} \tag{12}$$

We use the following approximate model for $p(\rho_{i+1}|\rho_i)$:

$$\rho_{i+1} \sim \rho_i + \mathcal{N}(0, h/\tau) \tag{13}$$

Note that Equation 12 implies $\rho_i \in (-1/h, \infty)$ since Ne cannot decline below zero, whereas the approximate model in Equation 13 assumes support on the entire real line. We have found performance with this approximate model to be superior to exact models on the log transformation of Ne provided that h is small.

With the above definitions, the prior density of a sequence $\rho_{1:m}$ is defined in terms of the increments:

$$p(\rho_{1:m}|\tau) \propto \prod_{i=1}^{m-2} p(\rho_{i+1} - \rho_i|\tau) \tag{14}$$

where

$$p(\rho_{i+1} - \rho_i|\tau) = \sqrt{\frac{\tau}{2\pi h}} e^{-\frac{\tau}{2h}(\rho_{i+1}-\rho_i)^2}$$

This equation can be compared with the *skygrid* density, Equation 2.

Incorporating covariates into phylodynamic inference

A simple model was recently proposed for incorporating time-varying covariates into phylodynamic inference with *skygrid* models (Gill et al. 2016). Suppose we observe q covariates at m time points denoted $X = (X_{1:m,1:q})$, and such that observation times correspond to the grid used in the phylodynamic model. The following linear model for the marginal distribution of γ with covariate vector $\alpha_{1:q}$ was proposed:

$$p(\gamma_i | X, \alpha_{1:q}, \epsilon) \sim \mathcal{N}(\alpha_0 + X_{i,1:q}\alpha_{1:q}, \epsilon) \tag{15}$$

where α_0 is the expected mean of γ without covariate effects.

This implies, along with the BM model, the following marginal distribution of the increments:

$$p(\gamma_{i+1} - \gamma_i | X, \alpha_{1:q}, \tau, \epsilon) \sim \mathcal{N}(X_{i+1,1:q}\alpha_{1:q} - X_{i,1:q}\alpha_{1:q}, h/\tau + 2\epsilon) \tag{16}$$

When covariates are likely to be associated with growth rates of the effective population size instead of the logarithm of the effective population size, we can analogously define the density of increments of ρ:

$$p(\rho_{i+1} - \rho_i | X, \alpha_{1:q}, \tau, \epsilon) \sim \mathcal{N}(X_{i+1,1:q}\alpha_{1:q} - X_{i,1:q}\alpha_{1:q}, h/\tau + 2\epsilon) \tag{17}$$

When fitting this model, we drop ϵ for simplicity (as in Gill et al. 2016), and estimate a single variance parameter τ.

Inference and software implementation

Our growth rate model is implemented in an open-source R package called *skygrowth* , available from `https://mrc-ide.github.io/skygrowth/`, and which includes both maximum a posteriori (MAP) and Bayesian Markov Chain Monte Carlo (MCMC) methods for model fitting.

The MCMC procedure uses a Gibbs-within-Metropolis algorithm that alternates between sampling the growth rate vector $\rho_{1:m}$ and sampling of the precision parameter τ. Metropolis-Hastings sampling is also performed for regression coefficients $\alpha_{1:q}$ if covariate data is provided with univariate normal proposals. The elements of $\rho_{1:m}$ are sampled in sequence (from past to present), and multiple Gibbs iterations (by default one hundred) are performed before updating other parameters using Metropolis-Hastings steps.

Maximum a posteriori (MAP) is used as a starting point for the MCMC. The MAP estimator alternates between optimisation of $\gamma_{1:m}$ using gradient descent (*BFGS* in R, Goldfarb 1970) and univariate optimisation of τ until convergence in the posterior is observed. Approximate credible intervals are provided for the MAP estimator based on curvature of the posterior around the optimum.

Results

Simulations

We evaluated the ability of the *skygrowth* model to infer epidemic trends by simulating partially-sampled genealogies from a stochastic individual-based susceptible-infected-recovered (SIR) model. Simulated data were generated using the BEAST2 package MASTER (Vaughan and Drummond 2013), and code to reproduce simulated results is available at `https://github.com/emvolz/skygrowth-experiments`. The *skygrowth* model was also compared to *skygrid* model as implemented in the *phylodyn* R package (Karcher et al. 2016, 2017) which estimates effective population size using a fast approximate Bayesian non-parametric reconstruction (BNPR). The SIR model was density dependent with a reaction rate $\beta S(t)I(t)$ of generating new infections. Figure 1 shows results of a single simulation with $R_0 = 1.3$ and 10,000 initial susceptible individuals. Additional simulations are shown in supporting Figure S1. Estimates with *skygrowth* were obtained using the MCMC algorithm and an Exponential(0.1) prior on the precision parameter. We report the posterior means from both *skygrowth* and *skygrid* BNPR. Genealogies were reconstructed by samping 200 or 1000 infected individuals at random from the entire history of the epidemic. In this scenario, both the *skygrowth* and *skygrid* models reproduce the true epidemic trend, capturing both the rate of initial exponential increase, the time of peak prevalence, and the rate of epidemic decline. However, when sampling only 200 lineages (Figure 1A), the genealogy contains relatively little information about later epidemic dynamics, and the *skygrid* estimates revert to a stationary prior producing an unrealistic levelling-off of Ne. Estimates using the *skygrid* BNPR model were highly similar to results using an exact MCMC algorithm for sampling the posterior also included in the *phylodyn* package.

While the results in Figure 1A and B suggest that Ne(t) can serve as a very effective proxy for epidemic size, the degree of correspondence will depend on details of the epidemic model as discussed in the Methods section. Figure 1C and supporting Figure S2 shows a scenario where estimates of $N_e(t)$ capture the initial rate of exponential growth but fail to estimate the time of peak epidemic prevalence, and the *skygrid* model also fails to detect that the epidemic ever decreases. This scenario was based on a higher $R_0 = 5$ and only 2,000 initially susceptible individuals, such that almost all hosts are eventually infected and the rate of epidemic decline predominantly reflects the host recovery rate. This is easily understood using the formula Ne(t) $\propto I(t)/S(t)$ (cf. Equation 7). When R_0 is large, $S(t)$ will change drastically over the course of the epidemic. In the later stages, almost all hosts have been infected so that $1/S(t)$ is large, producing correspondingly large effective population sizes.

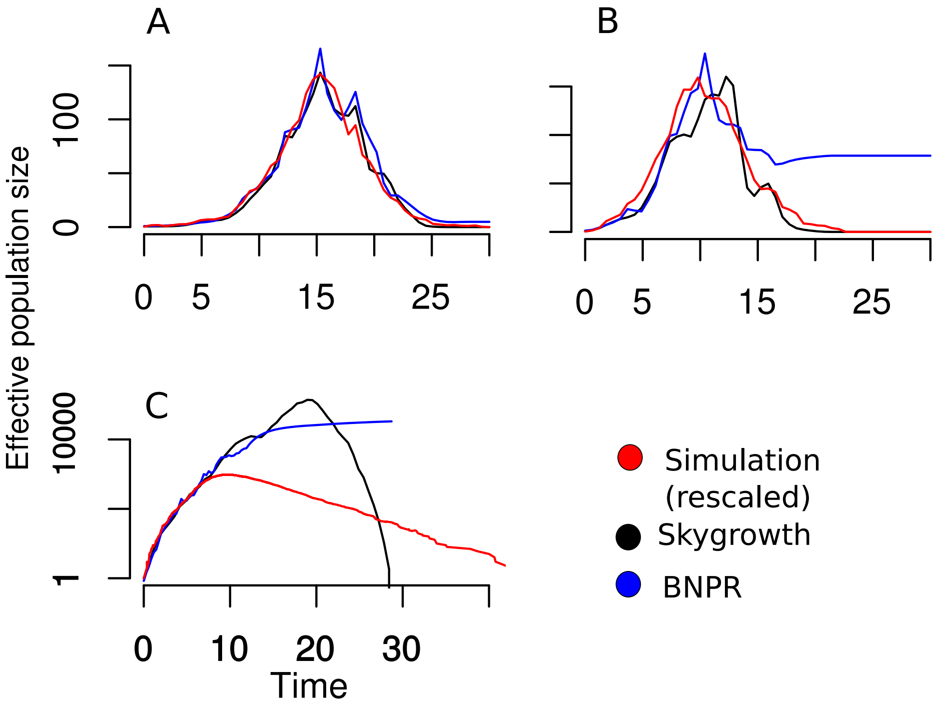

Figure 1: Comparison of effective population size estimates using the *skygrowth* and *skygrid* models applied to data from a susceptible-infected-recovered simulated epidemic. Effective population size estimates are also compared to the number of infected hosts through time under a linear rescaling (red). A. Estimates using a SIR model and simulated genealogy with 1000 sampled lineages and $R_0 = 1.3$. B. Estimates using a SIR model and simulated genealogy with 200 sampled lineages and $R_0 = 1.3$. C. Estimates using a SIR model and simulated genealogy with 200 sampled lineages and $R_0 = 5$.

Rabies virus

An epidemic of rabies broke out in the late 1970s in the North American raccoon population, following the emergence of a host-adapted variant of the virus called RRV. By the end of the 1990s, this outbreak had spread to a vast geographical area including all Northeast and mid-Atlantic US states (Childs et al. 2000). A sample of 47 RRV isolates has been sequenced in a previous study (Biek et al. 2007), and BEAST (Drummond et al. 2012) was used to reconstruct a dated phylogenetic tree. A standard *skyline* analysis (Drummond et al. 2005) was performed, which visually suggested a correlation between the inferred effective population size (Ne) and the monthly area newly affected by RRV (hereafter denoted V), but without attempting to quantify the strength or significance of this association.

This data was recently reanalysed using the *skygrid* model with covariates (Gill et al. 2016). No significant association was found between Ne and V, but the authors noted that since V is the newly affected area, V would be expected to be associated with a change in Ne rather than Ne itself. Since the *skyride* method is focused on Ne, like all previous phylodynamic methods, the authors considered the cumulative distribution of V and showed that this is slightly associated with Ne (with a 95% credible interval of [0.18-2.86] on the covariate effect size, Gill et al. 2016). However, this approach is not fully satisfactory. In particular, since V is always positive, the cumulative distribution of V is always increasing, whereas Ne is in principle equally likely to increase or decrease over time. Furthermore both V and its cumulative distribution were considered on a logarithm scale, so that the latter flattens over time by definition.

A more natural solution is to keep the covariate V untransformed, and investigate its association with the growth rate $\rho(t)$ rather than Ne(t) as implemented in our methodology (Figure 2). For this analysis we used exactly the same dated phylogeny as previously published (Biek et al. 2007) (reproduced in Supporting Figure S3). When the covariate was not used (red results in Figure 2), the growth rate was inferred to be positive but declining progressively to zero from 1973 to ∼1983, then stable around zero up to ∼1990, followed by a period of positive growth until ∼2000, after which the growth rate decreased below zero. This implies that the effective population size increased from 1973 to ∼1983, then was stable until ∼1990, increased to a peak in ∼1997 and afterwards decreased. Two waves of spread have therefore been inferred as in previous analyses (Biek et al. 2007; Gill et al. 2016), with the first one starting in the 1970s and ending in ∼1983 and the second one lasting from ∼1990 to ∼1997.

Unfortunately the covariate data V starts in September 1978 and therefore does not cover the first wave. However, the covariate data shows that the epidemic was spreading very quickly between 1992 and 1997, much faster than before or after these dates, and this timing corresponds fairly precisely to the second wave of spread. When the covariate data was integrated into phylodynamic inference, the covariate effect size was found to be statistically significant but only slightly so, with a large 95% credible interval for the covariate effect size of [0.03-4.61] and posterior mean of 1.09. The reconstructed growth rate and effective population size when using the covariate data (blue results in Figure 2) were compatible with results without covariate data. Using additional informative data tightens the credible interval as would be expected, except in the second wave during which the covariate data suggests higher values for both the growth rate and effective population size. The mean posterior growth rate reached a value of about 2.5 per year in the 1990s (Figure 2) and the average generation time of raccoon rabies has previously been estimated to be around 2 months (Biek et al. 2007). We can use Equation 10 to infer a reproduction number of $R = 1.4$, slightly higher than a previous estimate around $R = 1.1$ based on the same data (Biek et al. 2007).

One of the main novel findings of our analysis is that we found a significant decline of the effective population size of raccoon rabies post-2000, whereas previous phylodynamic

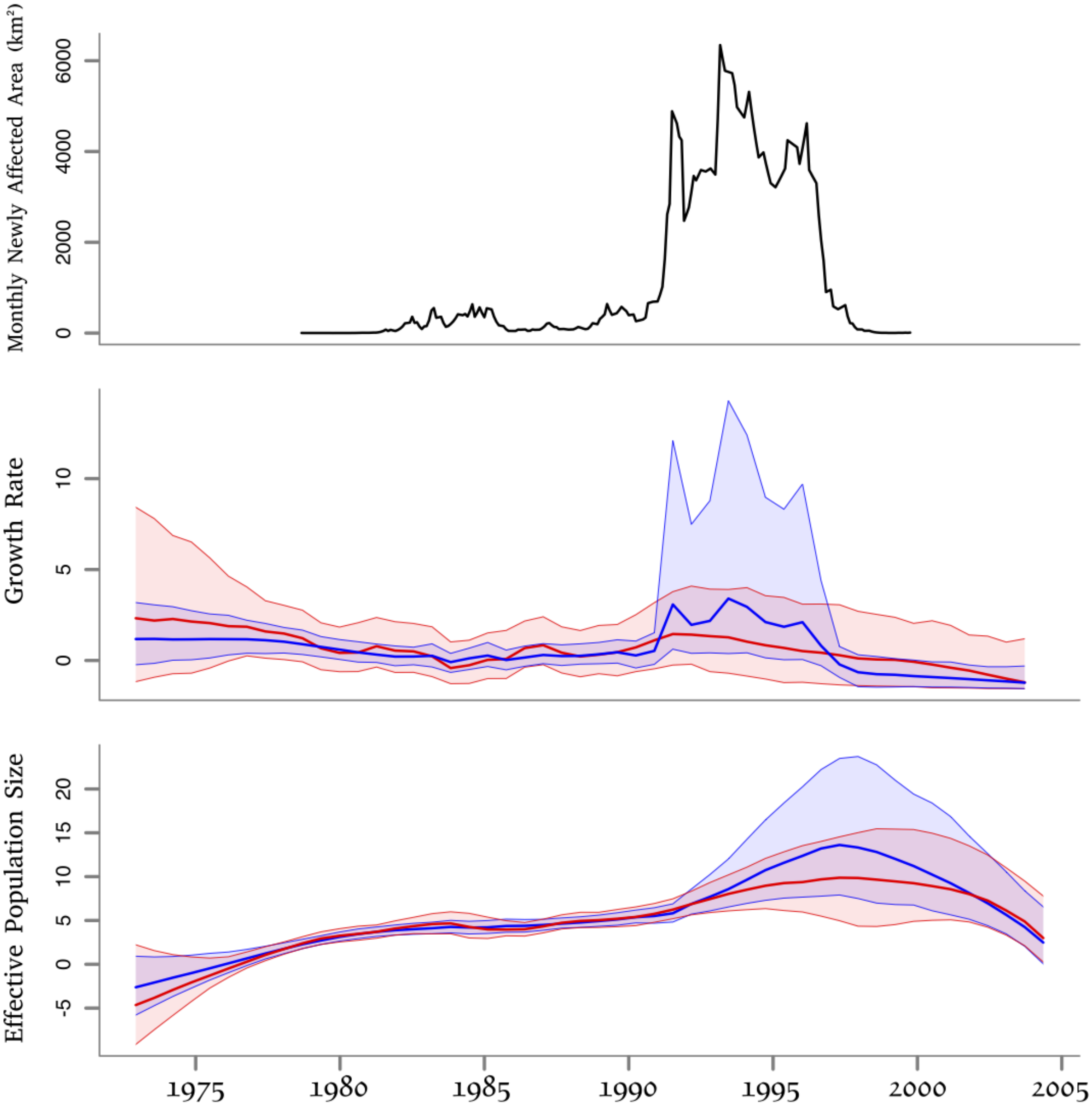

Figure 2: Results on the rabies application. Top: covariate data, representing the area in km^2 newly affected by rabies recorded monthly between September 1978 and October 1999. Middle: growth rate estimates. Bottom: log effective population size estimates. The middle and bottom plots show results without (red) and with (blue) the use of the covariate data, and with a solid line indicating posterior means and shaded areas indicating the 95% credible regions.

studies based on the same data found this to be constant (Biek et al. 2007; Gill et al. 2016). Previous methods consider a Brownian motion on the logarithm of Ne, which results in a strong prior that Ne is constant in recent time. By contrast, our model results in the growth rate being a-priori constant, so that the clear decline in growth rate started in the mid-1990s is likely to have continued to the point that the growth rate became negative and Ne declined. Our result is in good agreement with CDC surveillance that shows a clear decline in rabid raccoons after the peak in the mid-1990s (Monroe et al. 2016).

Staphylococcus aureus USA300

Staphylococcus aureus is a bacterium that causes infections ranging from mild skin infections to life-threatening septicaemia. In the 1980s and 1990s, several variants of *S. aureus* have emerged that are resistant to methicilin and other β-lactam antibiotics, and collectively called methicilin-resistant *S. aureus* (MRSA) (Chambers and Deleo 2009). MRSA are well known as a leading cause of hospital infections worldwide, but the MRSA variant called USA300 differs from most others by causing infections mostly in communities rather than hospitals. USA300 was first reported in 2000, and has since spread throughout the USA and internationally (Tenover and Goering 2009). A recent study sequenced the genomes from 387 isolates of USA300 sampled from New York between 2009 and 2011, and reconstructed phylogeographic spread that frequently involved transmission within households (Uhlemann et al. 2014).

The USA300 phylogenetic tree (Uhlemann et al. 2014) was dated using a previously described method (Didelot et al. 2012) and a clock rate of ∼3 substitutions per year for USA300 (Uhlemann et al. 2014; Alam et al. 2015). We analysed the resulting dated phylogeny (Supporting Figure S4) using our phylodynamic methodology (Figure 3). We initially performed this analysis without the use of any covariate data (red results in Figure 3) and found that the growth rate had been around zero up until 1985, after which it steadily increased until ∼1995, and subsequently decreased almost linearly, becoming negative in ∼2002 and continuing to decrease afterwards. The effective population size was accordingly found to have been very small until the mid-1990s, to have peaked in ∼2002 and to have declined since. These results are in very good agreement with a phylodynamic analysis of USA300 performed using a traditional *skyline* plot on a different genomic dataset (Glaser et al. 2016) as well as USA300 incidence trends (Planet 2017). However, the causes for the recent decline in USA300 are still unclear (Planet 2017). Declines in other MRSA lineages were recently described (Ledda et al. 2017) and have been attributed to improved hospital infection control measures, but this does not apply to the community-associated USA300 lineage.

We hypothesized that the dynamics of USA300 may be driven by the consumption of β-lactams in the USA, and we therefore gathered data on this from three different sources covering respectively the periods between 1980 and 1992 (McCaig and Hughes 1995), between 1992 and 2000 (McCaig et al. 2003) and between 2000 and 2012 (CDDEP 2017). There was an overlap of one year between the first and second, and between the second and third of these sources, which was used to scale data for consistency between the three sources. Specifically, values from the second source were scaled so that the 2000 value is equal to the one in the third source, and values from the first source were then scaled so that the 1992 value is equal to the one in the second source. The rescaled data is therefore measured as in the third source, namely in standard units of β-lactams (ie narrow-spectrum and broad spectrum penicilins plus cephalosporins) consumed per 1000 population in the USA (CDDEP 2017). This data show that the consumption of β-lactams almost doubled between 1980 and 1991, and subsequently decreased to reach around 2010 levels comparable to the early 1980s (Figure 3). These trends on β-lactams consumption therefore appear to be very similar to the ones observed for the USA300 growth rate without the use of covariates (red results in Figure 3). To confirm this observation, we repeated our

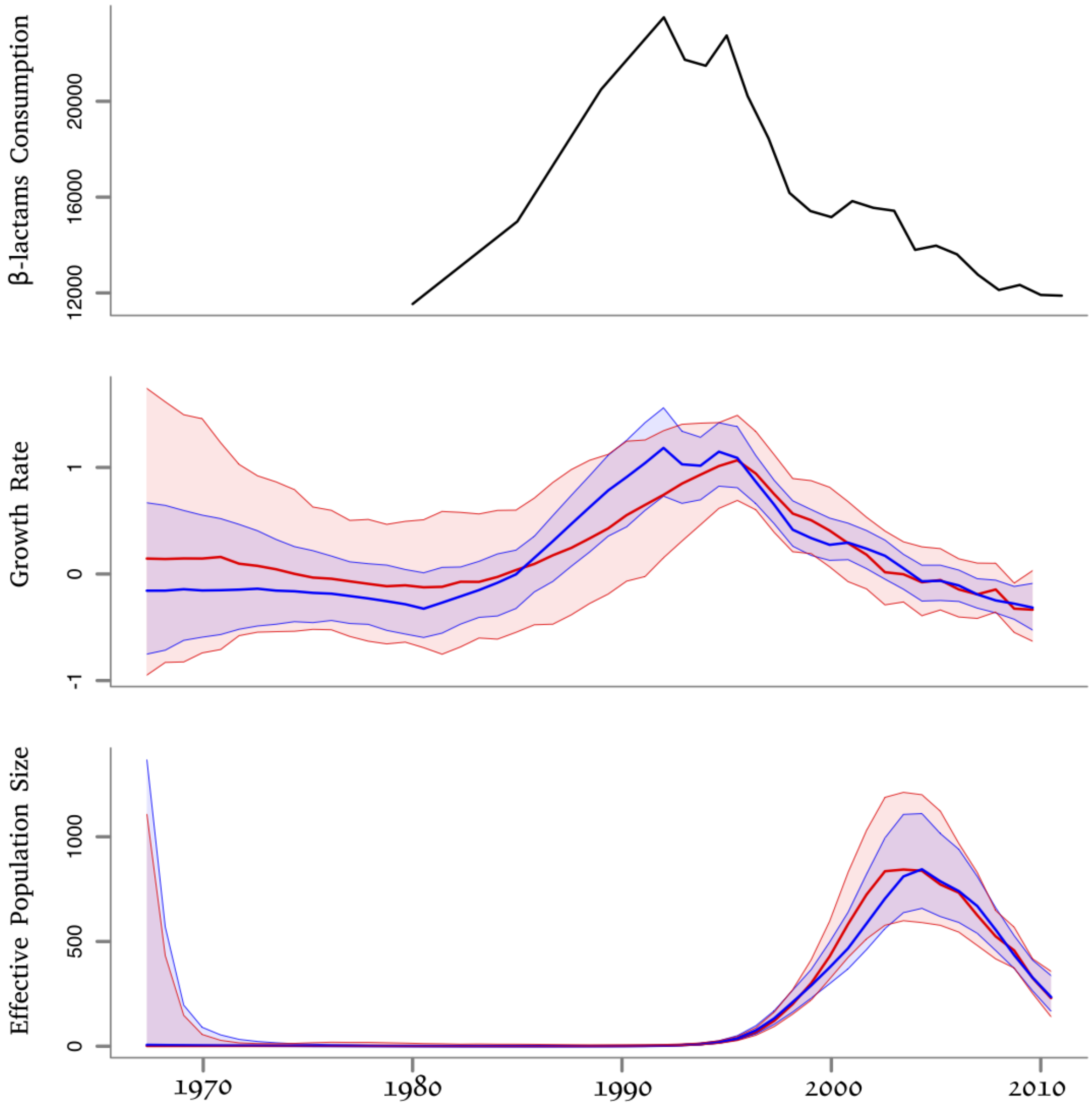

Figure 3: Results on the USA300 application. Top: covariate data, representing the consumption of β-lactams between 1980 to 2012 in the USA, measured in standard units per 1000 population. Middle: growth rate estimates. Bottom: log effective population size estimates. The middle and bottom plots show results without (red) and with (blue) the use of the covariate data, and with a solid line indicating posterior means and shaded areas indicating the 95% credible regions.

phylodynamic analysis with integration of the β-lactam use as a covariate (blue results in Figure 3). We found that the covariate was significantly associated with growth rate, with a mean posterior effect of 0.48 and 95% credible interval [0.18-0.71]. The growth rate dynamics inferred when using covariate data was almost identical to those inferred without the use of covariate data, except for a clear reduction of the width of the intervals which reflects the gain in information when combining two independent types of data.

Our analysis therefore suggests that the rise in β-lactams consumption in the 1980s was responsible for the emergence of the highly successful USA300 lineage. From the mid-1990s, the use of β-lactams has declined, both due to an overall reduction in antibiotic use and a diversification of the type of antibiotics prescribed (McCaig et al. 2003; CDDEP 2017), and the growth rate of USA300 has consequently decreased. Importantly, the consumption of antibiotics is expected to be associated with the growth rates of resistant bacterial pathogens, rather than with their effective population sizes, which here is not at all correlated with the covariate (Figure 3). Amongst pairs of individuals thought to have infected one another within households, the distribution of genomic distance had a mean of 4 substitutions (Uhlemann et al. 2014), and this represents on average twice the number of substitutions occurring during an infection when accounting for within-host diversity (Didelot et al. 2012, 2014, 2016). Given that the molecular clock rate of USA300 is approximately 3 substitutions per year (Uhlemann et al. 2014; Alam et al. 2015), the average duration of infections in this outbreak is around eight months. In the first half of the 1990s, the growth rate peaked around 1 per year (Figure 3) and using Equation 10 we estimate that the reproduction number was around $R = 1.6$, which is in good agreement with the recent estimate $R = 1.5$ for MRSA in the US population (Hogea et al. 2014). The fact that this estimate is only modestly above the minimum threshold of $R = 1$ required for outbreaks to take place could help explain why the USA300 is declining, even though β-lactams are still widely used. The consumption level may have lowered below the threshold caused by the fitness cost of resistance, as previously discussed for other resistant bacteria (Whittles et al. 2017; Dingle et al. 2017).

Discussion

Many environmental covariates, particularly those with a mechanistic influence on replicative fitness of pathogens, are closely related to the growth rate of epidemic size but not necessarily related to absolute epidemic size. We have found that these relationships can be inferred from random samples of pathogen genetic sequences by relating environmental covariates to the growth rate of the effective population size. This enables the estimation of the fitness effect of environmental covariates as well as the prediction of future epidemic dynamics should conditions change. We have found a clear and highly significant relationship between the growth and decline of community-associated MRSA USA300 and the population-level prescription rates of β-lactam antibiotics (Figure 3). This relationship is not apparent when comparing antibiotic usage directly with the effective population size of MRSA USA300. Our methodology focused on growth rate is therefore well suited to investigate the drivers of antibiotic resistance, compared to previous phylodynamic methods focused on the effective population size.

The *skygrowth* model can provide a more realistic prior for many infectious disease epidemics where the growth rate of epidemic size is likely to approach stationarity as opposed to the absolute effective population size. Conventional *skyride* and *skygrid* models are prone to erroneously estimating a stable effective population size when genealogical data is uninformative, as for example when estimating epidemic trends in the latter stages of SIR epidemics (Figure 1). The *skygrowth* model will correctly predict epidemic decline in this situation. Moreover, under ideal conditions, the estimated growth rate can be related to the reproduction number of an epidemic, and the *skygrowth* model provides a simple non-parametric estimator of the reproduction number through time given additional information about the natural history of infection (Equation 10). Caution should be exercised when using the effective population size as a proxy for epidemic size, as the relationship between the two is complex (cf. Simulation results). In general, there will be close correspondence between the growth of epidemic size and growth of effective population size during periods where the growth rate is relatively constant.

The methods presented here can be applied more generally to evaluate the role of antibiotic stewardship, vaccine campaigns, or other public health interventions on epidemic growth rates. Some environmental covariates, such as independent prevalence estimates, may be more closely related to effective population size rather than growth rates, and future work is indicated on the development of regression models in terms of both statistics. More complex stochastic models can also be considered, such as processes with both autoregressive and moving average components. A variety of mathematical models have been developed to explain de novo evolution of antimicrobial resistance as a function of population-level antimicrobial usage (Bonhoeffer et al. 1997; Austin et al. 1999; Spicknall et al. 2013; Whittles et al. 2017), and an important direction for future work will be the development of parametric and semi-parametric structured coalescent models (Volz 2012) that can be applied to bacterial phylogenies featuring a mixture of antibiotic sensitive and resistant lineages. This methodology will allow us to estimate key evolutionary parameters, such as the fitness cost and benefit of resistance, or the rate of mutation from sensitive to resistant status, which are needed to make well informed recommendations on resistance control strategies.

*

References

Alam, M. T., T. D. Read, R. A. Petit, S. Boyle-Vavra, L. G. Miller, S. J. Eells, R. S. Daum, and M. Z. David. 2015. Transmission and microevolution of USA300 MRSA in U.S. households: Evidence from whole-genome sequencing. MBio 6:1–10.

Allen, L. 2008. An introduction to stochastic epidemic models. Pages 81–130 *in* Math. Epidemiol. vol. 1945 of *Lecture Notes in Mathematics*. Springer Berlin Heidelberg.

Austin, D. J., K. G. Kristinsson, and R. M. Anderson. 1999. The relationship between the volume of antimicrobial consumption in human communities and the frequency of resistance. PNAS 96:1152–1156.

Baele, G., M. A. Suchard, A. Rambaut, and P. Lemey. 2016. Emerging concepts of data integration in pathogen phylodynamics. Systematic biology 66:e47–e65.

Biek, R., J. C. Henderson, L. a. Waller, C. E. Rupprecht, and L. a. Real. 2007. A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. Proc Natl Acad Sci USA 104:7993–8.

Bonhoeffer, S., M. Lipsitch, and B. R. Levin. 1997. Evaluating treatment protocols to prevent antibiotic resistance. Proc Natl Acad Sci U S A 94:12106–12111.

CDDEP. 2017. The Center for Disease Dynamics Economics and Policy. ResistanceMap, available at http://resistancemap.cddep.org/, accessed July 2017.

Chambers, H. F. and F. R. Deleo. 2009. Waves of resistance: Staphylococcus aureus in the antibiotic era. Nat. Rev. Microbiol. 7:629–41.

Childs, J. E., A. T. Curns, M. E. Dey, L. A. Real, L. Feinstein, O. N. Bjornstad, and J. W. Krebs. 2000. Predicting the local dynamics of epizootic rabies among raccoons in the United States. Proc. Natl. Acad. Sci. 97:13666–13671.

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

Dearlove, B. and D. Wilson. 2013. Coalescent inference for infectious disease: meta-analysis of hepatitis C. Philos. Trans. R. Soc. B Biol. Sci. 368:20120314.

Didelot, X., D. W. Eyre, M. Cule, C. L. C. Ip, M. A. Ansari, D. Griffiths, A. Vaughan, L. O'Connor, T. Golubchik, E. M. Batty, P. Piazza, D. J. Wilson, R. Bowden, P. J. Donnelly, K. E. Dingle, M. Wilcox, A. S. Walker, D. W. Crook, T. E. Peto, and R. M. Harding. 2012. Microevolutionary analysis of Clostridium difficile genomes to investigate transmission. Genome Biol 13:R118.

Didelot, X., J. Gardy, and C. Colijn. 2014. Bayesian inference of infectious disease transmission from whole genome sequence data. Mol. Biol. Evol. 31:1869–1879.

Didelot, X., A. S. Walker, T. E. Peto, D. W. Crook, and D. J. Wilson. 2016. Within-host evolution of bacterial pathogens. Nat. Rev. Microbiol. 14:150–162.

Dingle, K. E., X. Didelot, T. P. Quan, D. W. Eyre, N. Stoesser, T. Golubchik, R. M. Harding, D. J. Wilson, D. Griffiths, A. Vaughan, and Others. 2017. Effects of control interventions on Clostridium difficile infection in England: an observational study. Lancet Infect. Dis. 17:411–421.

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

Drummond, A. J., M. A. Suchard, D. Xie, and A. Rambaut. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969–1973.

Fraser, C. 2007. Estimating individual and household reproduction numbers in an emerging epidemic. PLoS One 2:e758.

Frost, S. D. W. and E. M. Volz. 2010. Viral phylodynamics and the search for an 'effective number of infections'. Phil. Trans. R. Soc. B 365:1879–1890.

Gill, M. S., P. Lemey, S. N. Bennett, R. Biek, and M. A. Suchard. 2016. Understanding Past Population Dynamics : Bayesian Coalescent-Based Modeling with Covariates. Syst. Biol. 65:1041–1056.

Gill, M. S., P. Lemey, N. R. Faria, A. Rambaut, B. Shapiro, and M. A. Suchard. 2012. Improving bayesian population dynamics inference: a coalescent-based model for multiple loci. Molecular biology and evolution 30:713–724.

Gill, M. S., P. Lemey, N. R. Faria, A. Rambaut, B. Shapiro, and M. A. Suchard. 2013. Improving bayesian population dynamics inference: A coalescent-based model for multiple loci. Mol. Biol. Evol. 30:713–724.

Glaser, P., P. Martins-Simões, A. Villain, M. Barbier, A. Tristan, C. Bouchier, L. Ma, M. Bes, F. Laurent, D. Guillemot, T. Wirth, and F. Vandenesch. 2016. Demography and intercontinental spread of the USA300 community-acquired methicillin-resistant Staphylococcus aureus lineage. MBio 7:1–11.

Goldfarb, D. 1970. A family of variable-metric methods derived by variational means. Mathematics of Computation 24:23–26.

Griffiths, R. and S. Tavare. 1994. Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. B Biol. Sci. 344:403–410.

Ho, S. Y. and B. Shapiro. 2011. Skyline-plot methods for estimating demographic history from nucleotide sequences. Molecular ecology resources 11:423–434.

Hogea, C., T. Van Effelterre, and C. J. Acosta. 2014. A basic dynamic transmission model of Staphylococcus aureus in the US population. Epidemiol. Infect. 142:468–478.

Karcher, M. D., J. A. Palacios, T. Bedford, M. A. Suchard, and V. N. Minin. 2016. Quantifying and mitigating the effect of preferential sampling on phylodynamic inference. PLoS computational biology 12:e1004789.

Karcher, M. D., J. A. Palacios, S. Lan, and V. N. Minin. 2017. phylodyn: an R package for phylodynamic simulation and inference. Mol. Ecol. Resour. 17:96–100.

Kingman, J. 1982. The coalescent. Stoch. Process. their Appl. 13:235–248.

Koelle, K., O. Ratmann, D. A. Rasmussen, V. Pasour, and J. Mattingly. 2011. A dimensionless number for understanding the evolutionary dynamics of antigenically variable RNA viruses. Proc. R. Soc. B Biol. Sci. 278:3723–3730.

Ledda, A., J. R. Price, K. Cole, M. J. Llewelyn, A. M. Kearns, D. W. Crook, J. Paul, and X. Didelot. 2017. Re-emergence of methicillin susceptibility in a resistant lineage of Staphylococcus aureus. J. Antimicrob. Chemother. 72:1285–1288.

McCaig, L. F., R. E. Besser, and J. M. Hughes. 2003. Antimicrobial drug prescription in ambulatory care settings, United States, 1992-2000. Emerg. Infect. Dis. 9:432–7.

McCaig, L. F. and J. M. Hughes. 1995. Trends in antimicrobial drug prescribing among office-based physicians in the United States. J. Am. Med. Assoc. 273:214–219.

Minin, V. N., E. W. Bloomquist, and M. A. Suchard. 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Molecular biology and evolution 25:1459–1471.

Monroe, B., P. Yager, J. Blanton, M. Birhane, A. Wadhwa, L. Orciari, B. Petersen, and R. Wallace. 2016. Rabies surveillance in the United States during 2014. J. Am. Vet. Med. Assoc. 248:777–788.

Palacios, J. A. and V. N. Minin. 2013. Gaussian process-based bayesian nonparametric inference of population size trajectories from gene genealogies. Biometrics 69:8–18.

Planet, P. J. 2017. Life After USA300: The Rise and Fall of a Superbug. J. Infect. Dis. 215:S71–S77.

Pybus, O. G. 2001. The Epidemic Behavior of the Hepatitis C Virus. Science 292:2323–2325.

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

Rosenberg, N. A. and M. Nordborg. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat. Rev. Genet. 3:380–90.

Spicknall, I. H., B. Foxman, C. F. Marrs, and J. N. S. Eisenberg. 2013. A modeling framework for the evolution and spread of antibiotic resistance: Literature review and model categorization. Am. J. Epidemiol. 178:508–520.

Strimmer, K. and O. G. Pybus. 2001. Exploring the demographic history of dna sequences using the generalized skyline plot. Molecular Biology and Evolution 18:2298–2305.

Tenover, F. C. and R. V. Goering. 2009. Methicillin-resistant Staphylococcus aureus strain USA300: Origin and epidemiology. J. Antimicrob. Chemother. 64:441–446.

Uhlemann, A.-C., J. Dordel, J. R. Knox, K. E. Raven, J. Parkhill, M. T. G. Holden, S. J. Peacock, and F. D. Lowy. 2014. Molecular tracing of the emergence, diversification, and transmission of S. aureus sequence type 8 in a New York community. Proc. Natl. Acad. Sci. U. S. A. 111:6738–43.

Vaughan, T. G. and A. J. Drummond. 2013. A stochastic simulator of Birth–Death master equations with application to phylodynamics. Mol. Biol. Evol. 30:1480–1493.

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

Volz, E. M. and S. D. W. Frost. 2014. Sampling through time and phylodynamic inference with coalescent and birth – death models. J. R. Soc. Interface 11:20140945.

Volz, E. M., K. Koelle, and T. Bedford. 2013. Viral Phylodynamics. PLoS Comput. Biol. 9:e1002947.

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

Volz, E. M., E. Romero-Severson, and T. Leitner. 2017. Phylodynamic Inference across Epidemic Scales. Mol. Biol. Evol. 34:1276–1288.

Whittles, L., P. White, and X. Didelot. 2017. Estimating the fitness cost and benefit of cefixime resistance in neisseria gonorrhoeae to inform prescription policy: a modelling study. PLoS Medicine .