# Maximum likelihood inference of pathogen population size history from a phylogeny

Xavier Didelot[1,*] and Erik M Volz[2]

[1] School of Life Sciences and Department of Statistics, University of Warwick, United Kingdom

[2] Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, United Kingdom

[*] Corresponding author. Tel: 0044 (0)2476 572827. Email: xavier.didelot@gmail.com

# ABSTRACT

Inference of effective population size from genomic data can provide unique information about demographic history, and when applied to pathogen genetic data can also provide insights into epidemiological dynamics. Non-parametric models for population dynamics combined with molecular clock models which relate genetic data to time have enabled phylodynamic inference based on large sets of time-stamped genetic sequence data. The theory for non-parametric inference of effective population size is well-developed in the Bayesian setting, but here we develop a frequentist approach based on non-parametric latent process models of population size dynamics. We appeal to statistical principles based on out-of-sample prediction accuracy in order to optimize parameters that control shape and smoothness of the population size over time. We demonstrate the flexibility and speed of this approach in a series of simulation experiments and apply the models to genetic data from several pathogen data sets.

# INTRODUCTION

Past fluctuation in the size of a population are reflected in the genealogy of a sample of individuals from that population. For example, under the coalescent model, two distinct lines of ancestry coalesce (ie find a common ancestor) at a rate that is inversely proportional to the effective population size at any given time (Kingman 1982; Griffiths and Tavare 1994; Donnelly and Tavare 1995). More coalescent events are therefore likely when the population size is small compared to when the population size is large. This causal effect of population size on genealogies can be reversed in an inferential framework to recover past population size dynamics from a given pathogen genealogy. This approach to inference of past demographic changes was first proposed 20 years ago (Pybus et al. 2000, 2001; Strimmer and Pybus 2001) and has been fruitfully applied to many disease systems (Pybus and Rambaut 2009; Ho and Shapiro 2011; Baele et al. 2016).

Population size analysis is often performed within the Bayesian BEAST framework (Suchard et al. 2018; Bouckaert et al. 2019) which jointly infers a phylogeny and demographic history from genetic data. Here we focus on an alternative approach in which the dated phylogeny is inferred first, for example using treedater (Volz and Frost 2017), TreeTime (Sagulenko et al. 2018) or BactDating (Didelot et al. 2018), and demography is investigated on the basis of the phylogeny. Although potentially less powerful, this approach has the advantage of scalability to very large sequence data sets and allows more focus on models and assumptions involved in the demographic inference itself as previously noted in studies following the same post-processing strategy (Lan et al. 2015; Karcher et al. 2017; Volz and Didelot 2018). However, some of the methodology and results we describe here should be applicable in a joint inferential setting too.

The reconstruction of past population size dynamics is usually based on a non-parametric model, since the choice of any parametric function for the past population size would cause restrictions and be hard to justify in many real-life applications (Drummond et al. 2005; Ho and Shapiro 2011). However, even if a non-parametric approach offers a lot more flexibility than a parametric one, it does not fully circumvent the question of which demographic model to use as the basis of inference. For example, the *skygrid* model considers that the log population size is piecewise constant, with values following a Gaussian Markov chain, in which each value is normally distributed around neighbouring values and standard deviation determined by a precision hyperparameter (Gill et al. 2013). This model can be

justified as the discretisation of a continuous *skyride* model in which the log population size is ruled by a Brownian motion (Minin et al. 2008). Alternatively, the *skygrowth* model is a similar Gaussian Markov chain on the growth rate of the population size (Volz and Didelot 2018). Both models can be conveniently extended to explore the association between population size dynamics and covariate data (Gill et al. 2016; Volz and Didelot 2018).

The *skygrid*, *skygrowth* or other similar models can be assumed when performing the inference of the demographic function, and the effect of this model choice has not been formally investigated. Furthermore, these discretised non-parametric models require to select the number of pieces in the demographic function, the location of boundaries between pieces, and the prior expectation for the difference from one piece to another, all of which can have significant effect on the inference results. Here we propose several statistical procedures to automatically select the best values for some of these variables. In particular, the parameter controlling the smoothness of the population size function is usually assumed to have an arbitrary non-informative prior distribution in a Bayesian inferential setting (Minin et al. 2008; Gill et al. 2013), whereas we show here that it can be selected using a frequentist statistical approach based on out-of-sample prediction accuracy. We tested the effect of these procedures on simulated datasets, where the correct demographic function is known and can be used to assess the relative value of inference under various conditions. We also reanalysed real datasets from recent studies on viral and bacterial infectious diseases, and show that the new methods can lead to improved epidemiological insights.

# MATERIALS AND METHODS

## Non-parametric Models

Let the demographic function $N_e(t)$ be piecewise linear with $R$ pieces of equal lengths $h$. Let $\gamma_i$ denote the log of the effective population size in the $i$-th piece. In the *skygrid* model (Gill et al. 2013), the values of $\gamma_i$ follow a Gaussian Markov chain, with the conditional distribution of $\gamma_{i+1}$ given $\gamma_i$ equal to:

$$\gamma_{i+1} \sim \mathcal{N}(\gamma_i, h/\tau) \tag{1}$$

By contrast, the *skygrowth* model (Volz and Didelot 2018) is defined using the effective population size growth rates $\rho_i$ which are assumed constant in each interval and are equal to:

$$\rho_i = \frac{\exp(\gamma_{i+1}) - \exp(\gamma_i)}{h\exp(\gamma_i)} \tag{2}$$

These growth rate values form a Gaussian Markov chain, with:

$$\rho_{i+1} \sim \mathcal{N}(\rho_i, h/\tau) \tag{3}$$

We also define a new model which we call *skysigma* based on the values $\sigma_i$ of the second order differences of the log of the effective population size:

$$\sigma_i = (\gamma_{i+1} - \gamma_i) - (\gamma_i - \gamma_{i-1}) = \gamma_{i+1} - 2\gamma_i + \gamma_{i-1} \tag{4}$$

Once again we consider a Gaussian Markov chain in which:

$$\sigma_{i+1} \sim \mathcal{N}(\sigma_i, h/\tau) \tag{5}$$

Each of the models above defines a demographic function $N_e(t)$ from which the likelihood of the genealogy $\mathcal{G}$ can be calculated. Let $n$ denote the number of tips in $\mathcal{G}$, let $s_{1:n}$ denote the dates of

the leaves and $c_{1:(n-1)}$ denote the dates of the internal nodes. Let $A(t)$ denote the number of extant lineages at time $t$ in $\mathcal{G}$ which is easily computed as the number of leaves dated after $t$ minus the number of internal nodes dated after $t$:

$$A(t) = \sum_{i=1}^{n} \mathbb{1}[s_i > t] - \sum_{i=1}^{n-1} \mathbb{1}[c_i > t] \tag{6}$$

This quantity is important because in the coalescent model, each pair of lineages coalesces at rate $1/N_{\mathrm{e}}(t)$, so that the total coalescent rate at time $t$ is equal to:

$$\lambda(t) = \begin{cases} \frac{A(t)(A(t)-1)}{2N_{\mathrm{e}}(t)}, & \text{if } A(t) \geq 2 \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The full likelihood of the coalescent process is therefore computed as (Griffiths and Tavare 1994; Donnelly and Tavare 1995):

$$L(\mathcal{G}|N_{\mathrm{e}}(t)) = \exp\left(-\int_{-\infty}^{\infty} \mathbb{1}[A(t) \geq 2] \frac{A(t)(A(t)-1)}{2N_{\mathrm{e}}(t)} \mathrm{d}t\right) \prod_{i=1}^{n-1} \frac{1}{N_{\mathrm{e}}(c_i)} \tag{8}$$

This computation is straightforward for the models considered here where the demographic function $N_{\mathrm{e}}(t)$ is piecewise constant.

## Selection of the precision parameter

The precision parameter (also called the 'smoothing' parameter) $\tau$ controls how much subsequent values of the Markov $N_{\mathrm{e}}$ will vary when data is uninformative. The selection of this parameter is therefore shaped by competing aims of maximizing explanatory power while reducing overfitting. In frequentist statistics, a standard approach to selecting smoothing parameters is to minimize out-of-sample prediction error. Here, we pursue a $k$-fold cross-validation strategy where genealogical data is partitioned into $k$ sets, $k-1$ of which are used for fitting, and the last one is used for prediction. This

procedure is equivalent to maximizing the following objective function:

$$f(\tau) = \prod_{j=1}^{k} P(\mathcal{G} \setminus X_j | \hat{N}_{\mathrm{e}}(X_j, \tau)), \tag{9}$$

where $\hat{N}_{\mathrm{e}}(X_j, \tau)$ is the maximum likelihood estimates of $N_{\mathrm{e}}$ on the partial data $X_j \subset \mathcal{G}$ and assuming the precision parameter is $\tau$. In this case $X_{j=1:k}$ represents a subset of the sample times and internal node times of the genealogy $\mathcal{G}$.

This is a standard formulation of the cross-validation method, but the implementation depends on how genealogical data is partitioned. We use the strategy of discretizing the coalescent likelihood (Equation 8) into intervals bordered by the time of nodes (tips $s_i$ or internal nodes $c_i$ of the tree) and/or the $R-1$ times when the piecewise-constant $N_{\mathrm{e}}$ changes value. Given $R-1$ change points, $n$ tips, and $n-1$ internal nodes of $\mathcal{G}$, there are $R+2n-3$ intervals $(\iota_1, \cdots, \iota_{R+2n-3})$. Each cross-validation training set is formed by taking a staggered sequence of intervals and collecting the genealogical data contained in each, so that $X_k = \{\iota_j | \mathrm{modulo}(j,k) \neq 0\}$.

## Selection of the grid resolution

Before any of the non-parametric models described above can be fitted, the number $R$ of pieces in the piecewise demographic function needs to be specified. Setting $R$ too low may lead to an oversimplified output that does not capture all the information on past population changes suggested by the genealogy, whereas setting $R$ too high can lead to overfitting. We therefore propose to use well established statistical methods to select the optimal value of $R$. First the model is fitted for multiple proposed values of $R$, and then for each output we compute the Akaike information criterion (AIC), which is equal to:

$$\mathrm{AIC}_R = 2R - 2\log(L_R) \tag{10}$$

where $L_R$ is the maximum value of the likelihood when using $R$ pieces. The value of $R$ giving the smallest value of $\mathrm{AIC}_R$ is selected. We also implemented the Bayesian information criterion (BIC), which are respectively equal to:

$$\mathrm{BIC}_R = R\log(n-1) - 2\log(L_R) \tag{11}$$

However, we found that the BIC was often overly conservative in the choice of the resolution as previously noted (Kuha 2004; Weakliem 1999), and therefore we focus here in the use of AIC.

## Simulation of testing data

In order to test the accuracy of our methodology, we implemented a new simulator of coalescent genealogies given sampling dates and a past demographic function $N_{\mathrm{e}}(t)$. When the demographic function is constant, the simulation of coalescent genealogies is equivalent to simulating from a homogeneous Poisson process, in which the waiting times from one event to the next are exponentially distributed. To extend this to the situation where the demographic function is non-constant requires to simulate from an equivalent non-homogeneous Poisson process. The approach we used to achieve this is to consider a homogeneous Poisson process with a population size $N_{\mathrm{m}}$ which is lower than any value of $N_{\mathrm{e}}(t)$, ie $\forall t, N_{\mathrm{e}}(t) \geq N_{\mathrm{m}}$. We simulate this process using exponential waiting times, but filter an event happening at time $t$ according to the ratio $N_{\mathrm{m}}/N_{\mathrm{e}}(t)$. Specifically, we draw $u \sim \mathrm{Unif}(0, 1)$ and if $u < N_{\mathrm{m}}/N_{\mathrm{e}}(t)$ the event is accepted and otherwise rejected. The resulting filtered Poisson process simulates from the non-homogeneous Poisson process as required (Ross 2014). The disadvantage of this approach over other methods of simulations is that there may be many rejections if $N_{\mathrm{e}}(t)$ takes small values so that $N_{\mathrm{m}}$ needs to be small too. This is therefore not the most efficient method of simulation. However, efficiency of simulation is not important for our purpose here, and this method has the advantage to avoid the computation of integrals on the $N_{\mathrm{e}}(t)$ function which other methods would require.

## Implementation

We implemented the simulation and inference methods described in this paper into a new R package entitled *mlesky* which is available at `https://github.com/emvolz-phylodynamics/mlesky` . If multiple CPU cores are available, these resources are exploited within the procedure of selection of the smoothing parameter where the computation can be split between the different cross values in the cross-validation. Multicore processing is also applied in the procedure of selection of the grid resolution where computation can be split between different values of the number $R$.

# RESULTS

## Application to simulated dataset with constant population size

A dated phylogeny was simulated with 200 tips sampled at regular intervals between 2000 and 2020, and a constant past population size function $N_e(t) = 20$ (Figure 1). To illustrate the importance of the resolution $R$ and precision $\tau$ parameters, we inferred the demographic function under the *skygrid* model (cf Equation 1) for a grid of values with $R \in \{5, 20, 50\}$ and $\tau \in \{1, 10, 20\}$ (Figure 2). The results look quite different depending on the parameters used, and in particular when $R$ is large and $\tau$ is small, fluctuations in the population size are incorrectly inferred. When applying the AIC procedure to this dataset, the correct value of $R = 1$ was inferred for which the parameter $\tau$ becomes irrelevant. In these conditions the effective population size was estimated to be 19.65 with confidence interval ranging from 17.10 to 22.57 which includes the correct value of 20 used in the simulation. We repeated the AIC procedure for 100 different phylogenies all which had been simulated under the same constant population size conditions described above. For 65 of these phylogenies the AIC procedure selected $R = 1$, with the third quartile falling on $R = 3$ and only one simulation giving $R > 10$.

## Application to simulated dataset with sinusoidal population size

Next we simulated a dated phylogeny with the same number and dates of the tips as before, but using a demographic function $N_e(t)$ that was sinusoidal with minimum 2 and maximum 22, with period 6.28 years. Figure 3 shows both the demographic function used and the resulting simulated phylogeny. We attempted to reconstruct the demographic function based on the phylogeny under the three models *skygrid*, *skygrowth* and *skysigma* described in Equations 1, 3 and 5, respectively. For each model the precision parameter $\tau$ was optimised using our new cross-validation procedure and the number of pieces was set to be $R = 20$ for ease of comparison. The results obtained in these conditions were very similar under the three models (Figure 4). This suggests that when the precision parameter is optimised using the cross-validation method, the choice between these three models becomes relatively unimportant. The choice of using one model rather than another is therefore mostly guided by the presence of covariate data and whether these are expected to correlate with the effective population

size directly or some other function of it such as the population growth rates (Gill et al. 2016; Volz and Didelot 2018).

## Application to simulated dataset with bottleneck in population size

We simulated another dated phylogeny with the same and dates of the tips as before, but using a bottleneck function for $N_e(t)$ which was equal to 10 at all times except between 2005 and 2010 when it was equal to 1. Figure 5 shows both this bottleneck function and the phylogeny simulated accordingly. We reconstructed the demographic function using the *skygrid* model. The lowest value of the AIC was obtained for $R = 14$, and the precision parameter was optimised using the cross-validation procedure to $\tau = 0.87$. The inferred demographic function is shown in Figure 5, where the bottleneck between 2005 and 2010 has been accurately detected.

## Application to HIV dataset

We analyzed 399 HIV-1 sequences from Senegal between 1990 and 2014 (Nascimento et al. 2020). All sequences are subtype CRF02_AG. We used treedater (Volz and Frost 2017) to reconstruct a dated phylogeny (Figure 6). This phylogeny has a common ancestor around 1972 and the number of lineages through time having rapid change in the early 1980s when the HIV epidemic was expanding. We applied the AIC procedure to determine the optimal number of pieces to be used for the demographic function, which was found to be $R = 35$. The optimal value of the precision parameter was determined using the cross-validation procedure, and found to be $\tau = 4.40$. The demographic function inferred using these values of $R$ and $\tau$ is shown in Figure 6. The whole analysis took less than 30 seconds on a standard laptop computer.

## Application to cholera dataset

We applied our methodology to a previously described collection of 260 genomes from the seventh pandemic of *Vibrio cholerae* (Didelot et al. 2015). A genealogy was estimated in this previous study using an early version of BactDating (Didelot et al. 2018), and it is reproduced in Figure 7. We applied

the AIC procedure to determine that the demographic function would be modelled using $R = 16$ pieces. The precision parameter was optimised to a value of $\tau = 1.84$ using the cross-validation procedure. The whole analysis took less than 20 seconds on a standard laptop computer. The result is shown in Figure 7. A first peak was detected in the 1960s, followed by a second peak in the 1970s and finally a third peak in the 1990s. This demographic function follows closely on the previously described three "waves" of cholera spreading globally from the Bay of Bengal (Mutreja et al. 2011; Didelot et al. 2015; Weill et al. 2017). However, these three waves had previously been described based on phylogeographic reconstructions of the spread of the pandemic around the world. The fact that we found a similar wave pattern in our analysis which did not include any information about the geographical origin of the genomes provides further support for the validity of this phylodynamic reconstruction.

# DISCUSSION

Non-parametric phylodynamic inference of population size dynamics is usually carried out in a Bayesian framework (Minin et al. 2008; Gill et al. 2013; Volz and Didelot 2018). Here we presented methods for performing such inference in a frequentist setting with a particular view towards model selection and reducing over-fitting. Optimal smoothing can be obtained in a natural way using standard cross-validation methods, and the optimal resolution of the discretised demographic function is achieved using the well-established AIC criterion. This approach can be advantageous when prior distributions are difficult to design or results are sensitive to arbitrarily chosen priors. Methods based on likelihood maximization is also quite fast and scalable to data sets much larger than is conventionally studied with Bayesian methods, and the selection of smoothing parameters does not require arbitrarily chosen hyperparameters. Conventional AIC metrics also alleviate the difficulty of model selection. In most of our simulations, we find relatively little difference in our estimates when parameterizing the model in terms of $N_{\mathrm{e}}(t)$ (Equation 1), or the growth rate of $N_{\mathrm{e}}(t)$ (Equation 3) or its second order variation of $N_{\mathrm{e}}(t)$ (Equation 5).

Our methodology assumed that a dated phylogeny has been previously reconstructed from the genetic data. It is therefore well suited for the post-processing analysis of the outputs from *treedater* (Volz and Frost 2017) or *TreeTime* (Sagulenko et al. 2018). A key assumption of our method, as with its Bayesian counterparts, is that all samples in the phylogeny come from a single population ruled by a unique demographic function. To ensure that this is indeed the case, complementary methods are emerging that can test for the presence or asymmetry or hidden population structure in dated phylogenies (Dearlove and Frost 2015; Volz et al. 2020).

Past variations in the effective population size of a pathogen population can reveal key insights into past epidemiological dynamics and help make predictions about the future. It is important to note that the effective population size is not generally equal to or even proportional to the number of infections over time (Volz et al. 2009; Dearlove and Wilson 2013). On the other hand, the growth rate of the effective population size (Equation 2) can be used to estimate the basic reproduction number over time $R(t)$ which represents the average number of secondary infections caused by an infected individual (Volz et al. 2013; Volz and Didelot 2018). Having good estimates of this quantity is especially important for assessing the effect of infectious disease control measures (Fraser 2007), and phylodynamic approaches

provide a useful complementary approach to more traditional methods of estimation based on case report data (Cori et al. 2013).

# References

Baele G, Suchard MA, Rambaut A, Lemey P. 2016. Emerging concepts of data integration in pathogen phylodynamics. Syst. Biol. 00:1–24.

Bouckaert R, Vaughan TG, Fourment M, Gavryushkina A, Heled J, Denise K, Maio ND, Matschiner M, Ogilvie H, Plessis L, et al. (11 co-authors). 2019. BEAST 2.5 : An Advanced Software Platform for Bayesian Evolutionary Analysis. PLoS Comput. Biol. 15:e1006650.

Cori A, Ferguson NM, Fraser C, Cauchemez S. 2013. A new framework and software to estimate time-varying reproduction numbers during epidemics. Am. J. Epidemiol. 178:1505–12.

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

Dearlove BL, Frost SDW. 2015. Measuring Asymmetry in Time-Stamped Phylogenies. PLoS Comput. Biol. 11:e1004312.

Didelot X, Croucher NJ, Bentley SD, Harris SR, Wilson DJ. 2018. Bayesian inference of ancestral dates on bacterial phylogenetic trees. Nucleic Acids Res. 46:e134.

Didelot X, Pang B, Zhou Z, McCann A, Ni P, Li D, Achtman M, Kan B. 2015. The Role of China in the Global Spread of the Current Cholera Pandemic. PLoS Genet. 11:e1005072.

Donnelly P, Tavare S. 1995. Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. 29:401–21.

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

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

Gill MS, Lemey P, Bennett SN, Biek R, Suchard MA. 2016. Understanding Past Population Dynamics : Bayesian Coalescent-Based Modeling with Covariates. Syst. Biol. 65:1041–1056.

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

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

Ho SYW, Shapiro B. 2011. Skyline-plot methods for estimating demographic history from nucleotide sequences. Mol. Ecol. Resour. 11:423–434.

Karcher MD, Palacios JA, Lan S, Minin VN. 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.

Kuha J. 2004. AIC and BIC: Comparisons of assumptions and performance. Sociol. Methods Res. 33:188–229.

Lan S, Palacios JA, Karcher M, Minin VN, Shahbaba B. 2015. An efficient Bayesian inference framework for coalescent-based nonparametric phylodynamics. Bioinformatics. 31:3282–3289.

Minin VN, Bloomquist EW, Suchard MA. 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol. Biol. Evol. 25:1459–1471.

Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, Croucher NJ, Choi SY, Harris SR, Lebens M, et al. (21 co-authors). 2011. Evidence for several waves of global transmission in the seventh cholera pandemic. Nature. 477:462–465.

Nascimento FF, Baral S, Geidelberg L, Mukandavire C, Schwartz SR, Turpin G, Turpin N, Diouf D, Diouf NL, Coly K, et al. (15 co-authors). 2020. Phylodynamic analysis of HIV-1 subtypes B, C and CRF 02_AG in Senegal. Epidemics. 30.

Pybus OG, Charleston MA, Gupta S, Rambaut A, Holmes EC, Harvey PH. 2001. The Epidemic Behavior of the Hepatitis C Virus. Science. 292:2323–2325.

Pybus OG, Rambaut A. 2009. Evolutionary analysis of the dynamics of viral infectious disease. Nat. Rev. Genet. 10:540–50.

Pybus OG, Rambaut A, Harvey PH. 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics. 155:1429–1437.

Ross SM. 2014. Introduction to probability models. Academic press.

285 Sagulenko P, Puller V, Neher RA. 2018. TreeTime: Maximum likelihood phylodynamic analysis. Virus
286 Evol. 4:vex042.

287 Strimmer K, Pybus OG. 2001. Exploring the Demographic History of DNA Sequences Using the
288 Generalized Skyline Plot. Mol. Biol. Evol. 18:2298–2305.

289 Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. 2018. Bayesian phylogenetic
290 and phylodynamic data integration using BEAST 1.10. Virus Evol. 4:vey016.

291 Volz EM, Didelot X. 2018. Modeling the Growth and Decline of Pathogen Effective Population Size
292 Provides Insight into Epidemic Dynamics and Drivers of Antimicrobial Resistance. Syst. Biol.
293 67:719–728.

294 Volz EM, Frost SDW. 2017. Scalable relaxed clock phylogenetic dating. Virus Evol. 3:vex025.

295 Volz EM, Koelle K, Bedford T. 2013. Viral Phylodynamics. PLoS Comput. Biol. 9:e1002947.

296 Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SDW. 2009. Phylodynamics of
297 infectious disease epidemics. Genetics. 183:1421–30.

298 Volz EM, Wiuf C, Grad YH, Frost SDW, Dennis AM, Didelot X. 2020. Identification of hidden
299 population structure in time-scaled phylogenies. Syst. Biol. 69:884–896.

300 Weakliem DL. 1999. A critique of the Bayesian information criterion for model selection.

301 Weill Fx, Domman D, Njamkepo E, Tarr C, Rauzier J, Fawal N, Keddy KH, Salje H, Moore S,
302 Mukhopadhyay AK, et al. (15 co-authors). 2017. Genomic history of the seventh pandemic of
303 cholera in Africa. Science. 789:785–789.

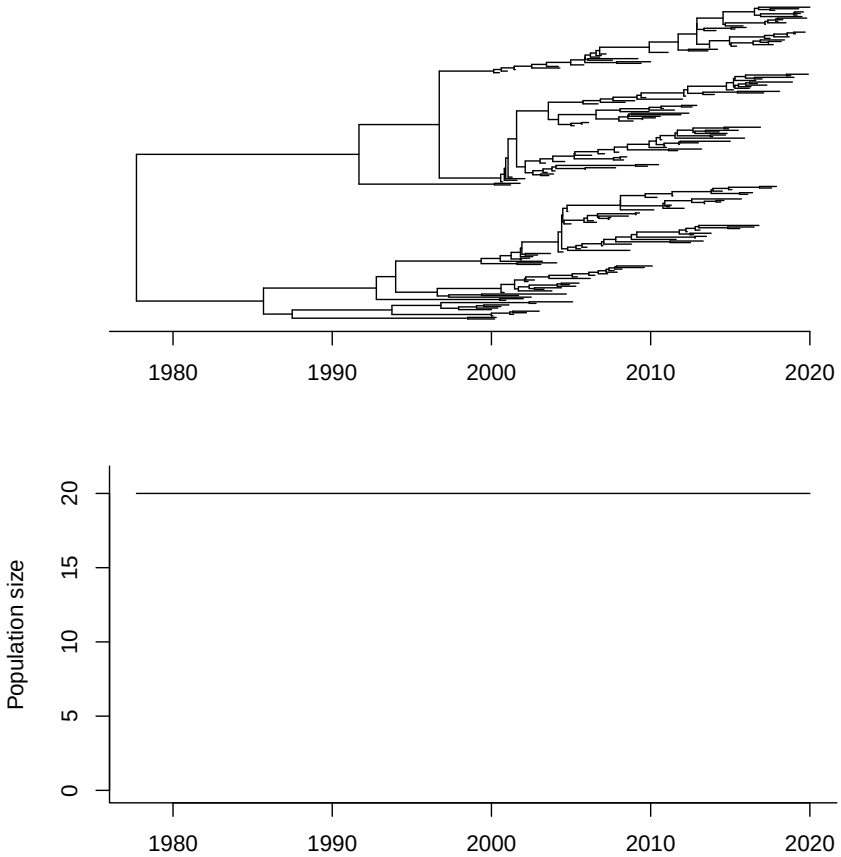


Figure 1: Simulated dataset using a constant past population size function.

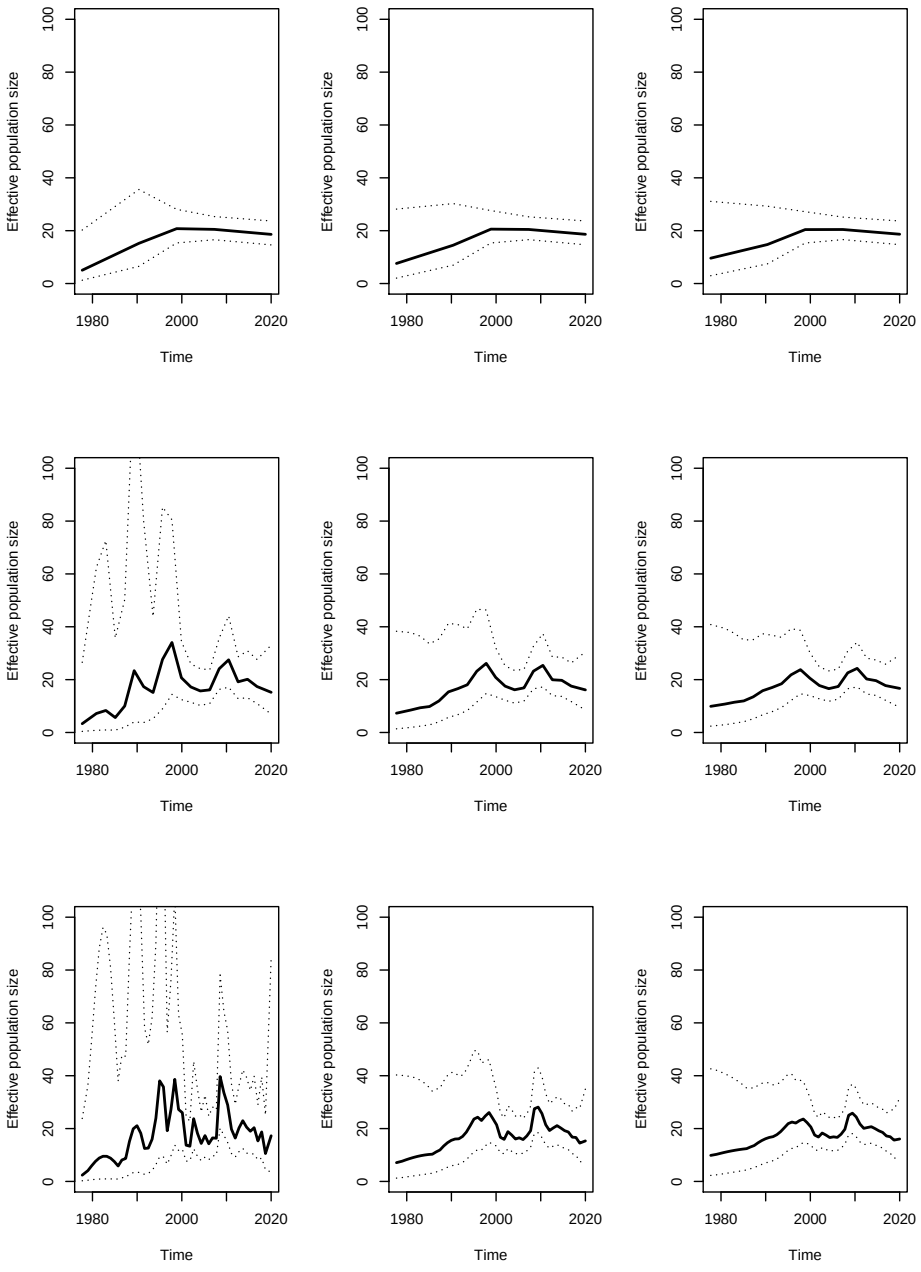


Figure 2: Result on simulated dataset shown in Figure 1 using the skyline model, from top to bottom $R = 5, 20, 50$ and from left to right $\tau = 1, 10, 20$.

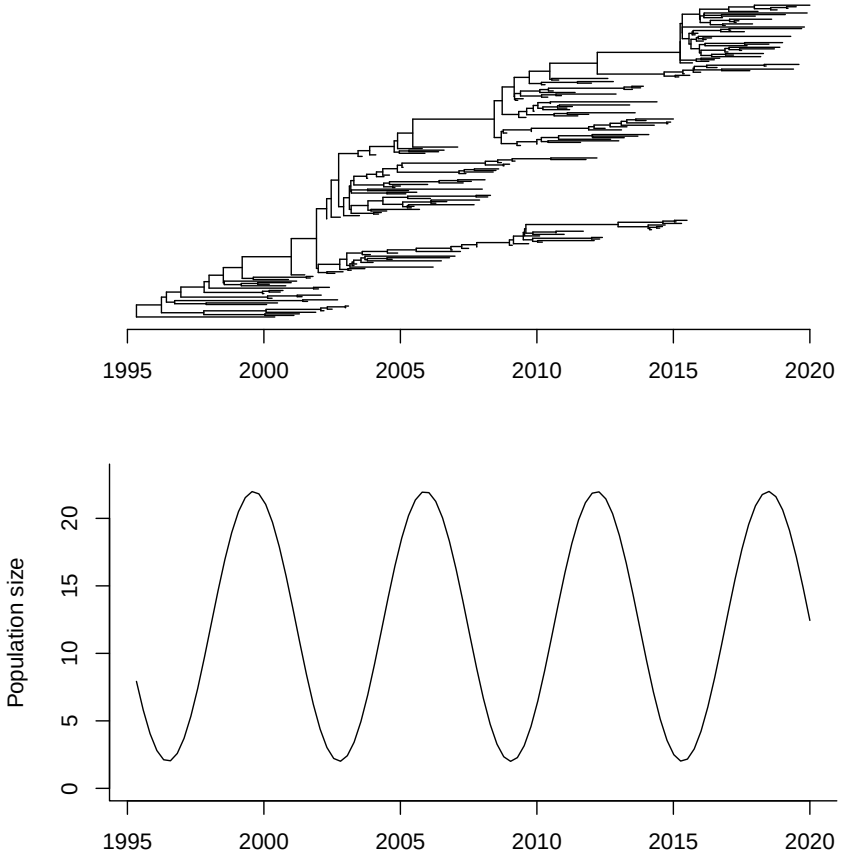


Figure 3: Simulated data using a sinusoidal past population size function.

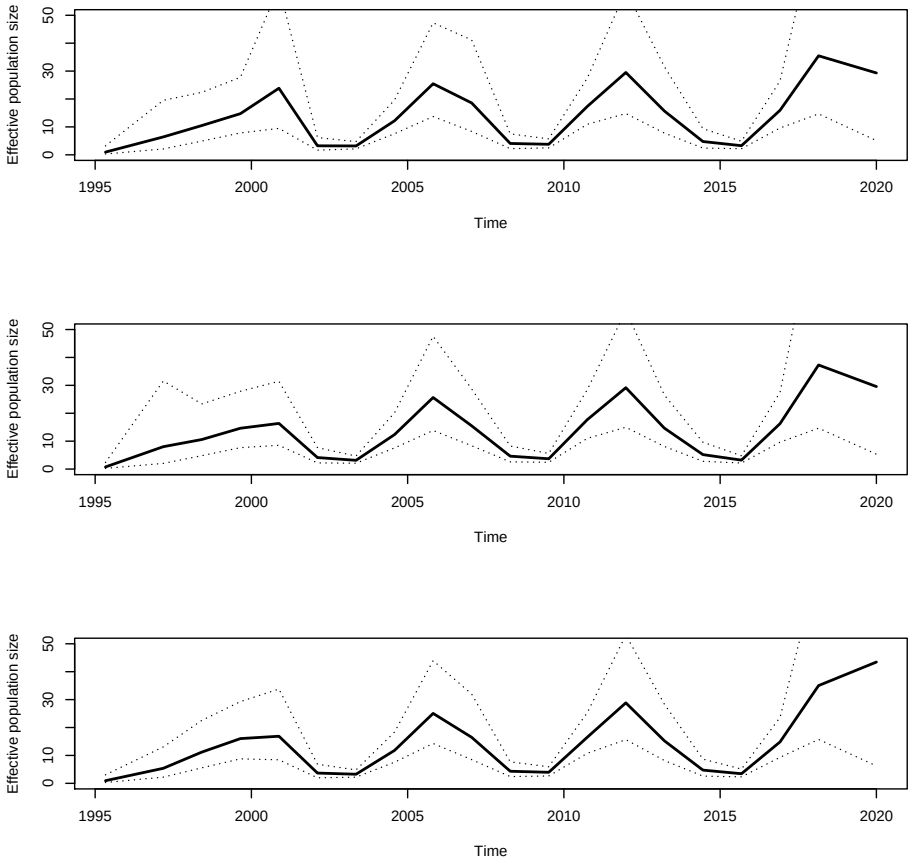


Figure 4: Result of applying the three different models to the phylogeny shown in Figure 3.

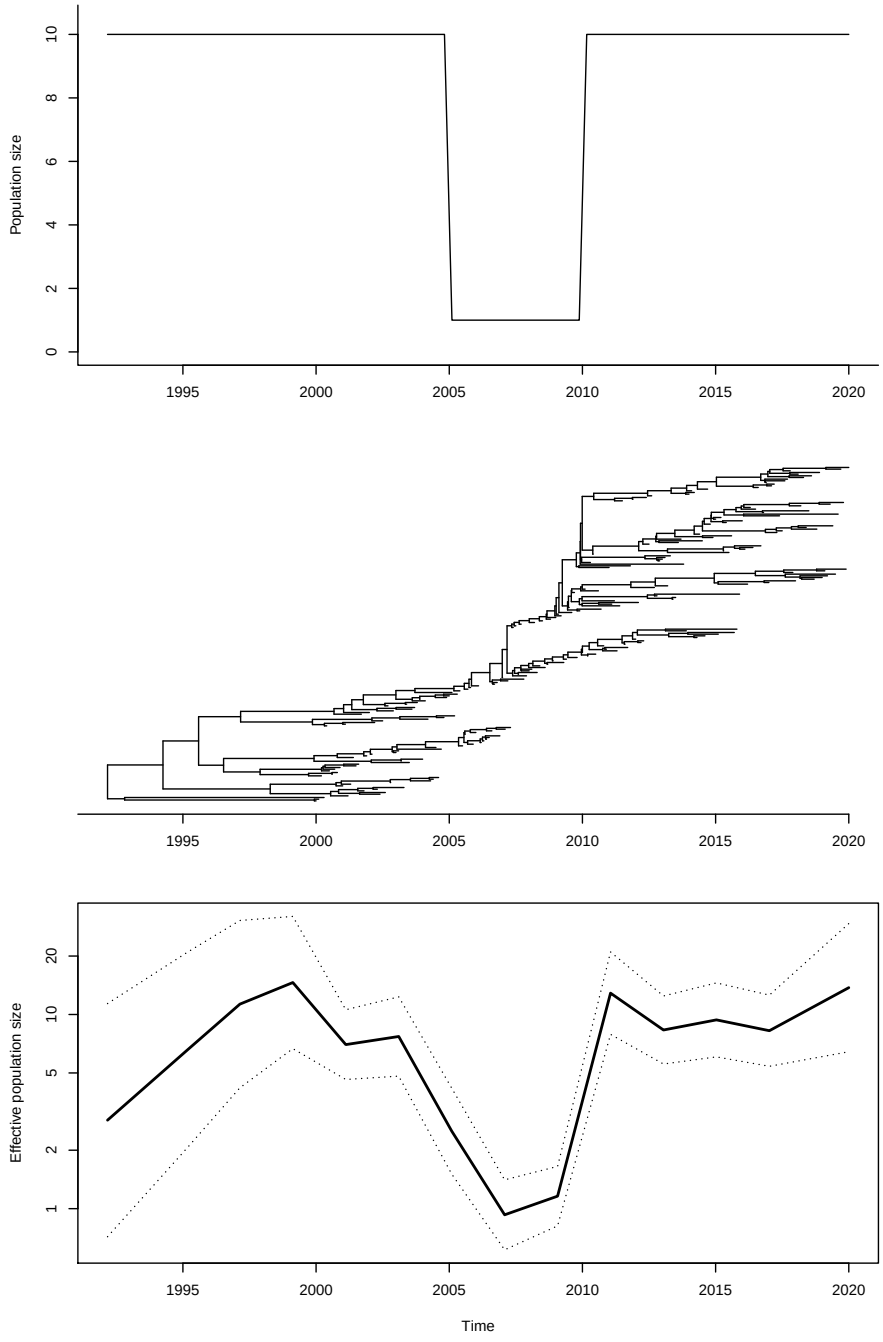


Figure 5: Demographic function (top), phylogeny (middle) and inferred demographic function (bottom) for a simulated dataset under a bottleneck model.

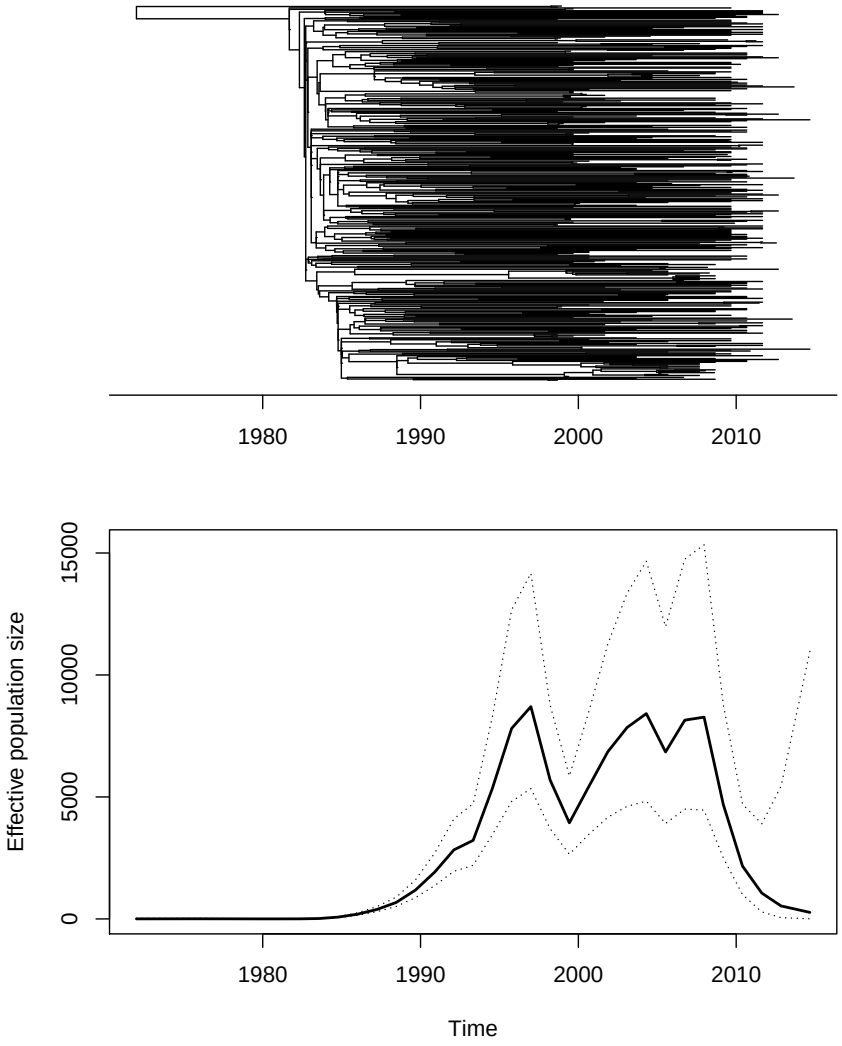


Figure 6: Analysis of the HIV dataset. Top: Dated phylogeny used as the starting point of past population size inference. Bottom: Demographic function reconstructed based on the phylogeny above.

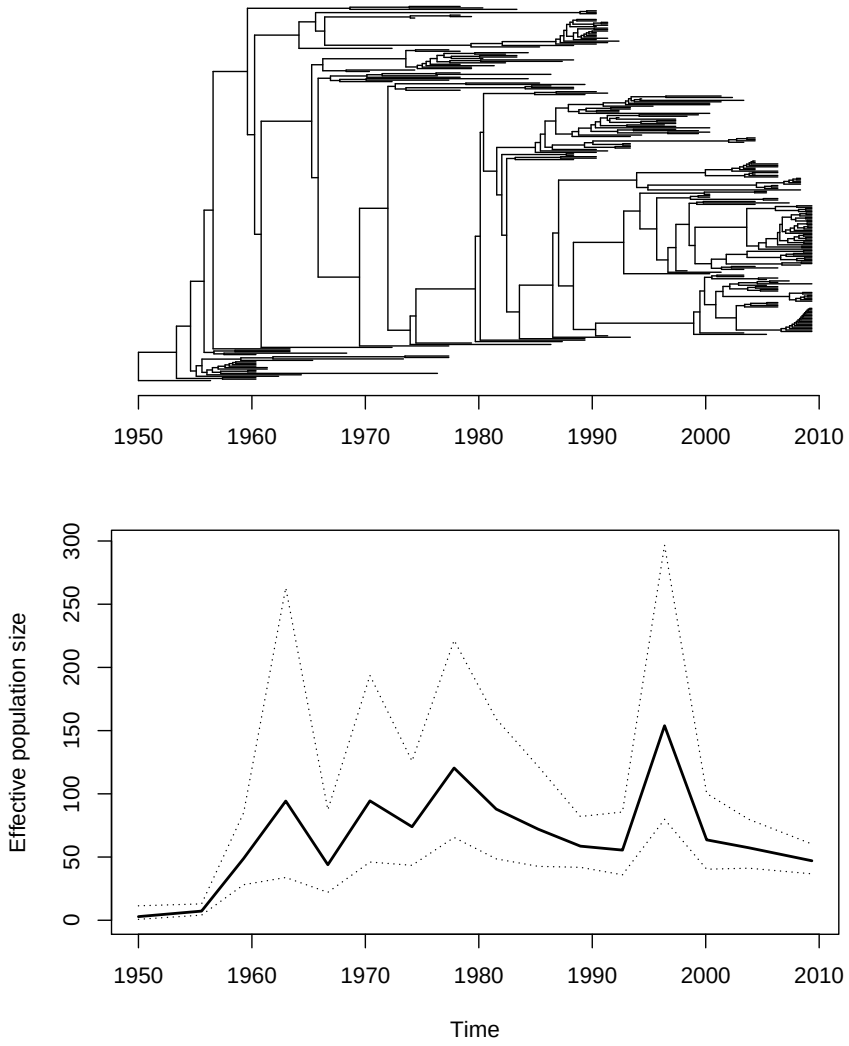


Figure 7: Analysis of the seventh pandemic of *Vibrio cholerae*. Top: Dated phylogeny used as the starting point of past population size inference. Bottom: Demographic function reconstructed based on the phylogeny above.