Statistical Science
2022, Vol. 37, No. 2, 162–182
https://doi.org/10.1214/22-STS853

Statistical Challenges in Tracking the Evolution of SARS-CoV-2

Lorenzo Cappello, Jaehee Kim, Sifan Liu and Julia A. Palacios

Abstract. Genomic surveillance of SARS-CoV-2 has been instrumental in tracking the spread and evolution of the virus during the pandemic. The availability of SARS-CoV-2 molecular sequences isolated from infected individuals, coupled with phylodynamic methods, have provided insights into the origin of the virus, its evolutionary rate, the timing of introductions, the patterns of transmission, and the rise of novel variants that have spread through populations. Despite enormous global efforts of governments, laboratories, and researchers to collect and sequence molecular data, many challenges remain in analyzing and interpreting the data collected. Here, we describe the models and methods currently used to monitor the spread of SARS-CoV-2, discuss long-standing and new statistical challenges, and propose a method for tracking the rise of novel variants during the epidemic.

Key words and phrases: Phylodynamics, genetic epidemiology, coalescent, Bayesian nonparametrics, birth-death processes, SIR models.

1. INTRODUCTION

In the last couple of years, we have witnessed an unprecedented global effort to collect and share SARS-CoV-2 molecular data and sequences. This effort has resulted in more than ten million molecular sequences being available for download in public repositories such as GISAID (Shu and McCauley, 2017) and GenBank today. These viral RNA sequences are *consensus*[1] sequences of about 30,000 nucleotides isolated from biological samples, such as nasal swabs, from infected individuals. Analyses of viral molecular sequences provide evidence of human-to-human transmission and allow the investigations of SARS-CoV-2 origins (Andersen et al., 2020, Boni et al., 2020). Moreover, they are routinely used to investigate outbreaks (MacCannell et al., 2021, Deng et al., 2020), track the speed and spread of viral transmission across the world (Hadfield et al., 2018), and monitor the evolution of new variants (Volz et al., 2021a).

The field of phylodynamics of infectious diseases, also referred to as molecular epidemiology, aims to understand disease dynamics by joint modeling of evolutionary, immunological, and epidemiological processes (Grenfell et al., 2004, Volz, Koelle and Bedford, 2013). It is assumed that these processes shape the underlying viral phylogeny of a sample of molecular sequences at a *locus*. Under models of *neutral evolution*, it is assumed that a process of *substitutions* is superimposed along the branches of the phylogeny, generating the observed variation in the sample of molecular sequences. More complex evolutionary models consider the effects of other types of *mutations* and sources of variation, such as *recombination* and *selection* (Wakeley, 2009).

A viral phylogeny is a timed bifurcating tree that represents the ancestral history of a sample at a locus (Figure 2(A)). This viral phylogeny can be obtained by maximum parsimony methods or by maximum likelihood from observed molecular sequences (Felsenstein, 2004). In the case of maximum likelihood, a model of substitutions (or mutations) is required. In phylodynamics, however, the study usually does not end at a single phylogeny. The aim is to understand the evolutionary and epidemiological forces that shape the phylogeny. To this end, the phylogeny is typically assumed to be the realization of either a birth–death-sampling process (BDSP) (Stadler and Bonhoeffer, 2013) or a coalescent process (CP) (Kingman, 1982a, Rodrigo and Felsenstein, 1999). In the context

Lorenzo Cappello is Assistant Professor, Departments of Economics and Business, Universitat Pompeu Fabra, 08005, Spain (e-mail: lorenzo.cappello@upf.edu). Jaehee Kim is Assistant Professor, Department of Computational Biology, Cornell University, Ithaca, New York 14853, USA (e-mail: jaehee.kim@cornell.edu). Sifan Liu is a Ph.D. student, Department of Statistics, Stanford University, Stanford, California 94305, USA (e-mail: sfliu@stanford.edu). Julia A. Palacios is Assistant Professor, Departments of Statistics and Biomedical Data Sciences, Stanford University, Stanford, California 94305, USA (e-mail: juliapr@stanford.edu).

[1]A glossary in the appendix explains the terms in italics that not all statisticians may be familiar with.

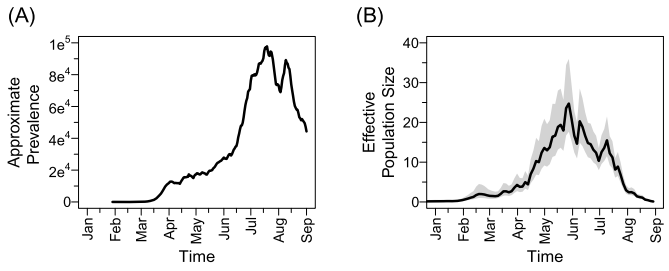

FIG. 1. *Phylodynamic analysis of SARS-CoV-2 sequences in California in* 2020. (*A*) 10*-days cumulative sum of the daily number of new cases in California*. (*B*) *Posterior median of the effective population size* (*black line*) *and* 95% *credible region* (*gray area*). *Model and data details appear in the Appendix*.

of disease dynamics, the BDSP is parameterized by the transmission rate $(\lambda(t))_{t\geq 0}$ and recovery rate $(\mu(t))_{t\geq 0}$, all of which are parameters of interest in epidemiology and public health. The CP is parameterized by the effective population size $(N_e(t))_{t\geq 0}$, a measure of relative genetic diversity over time that serves as a proxy of the growth and decline in the number of infections over time. For example, Figure 1(B) shows the estimated effective population size of SARS-CoV-2 in California in the first nine months of 2020, together with panel (A) that shows the 10-day cumulative number of new cases; this quantity is a proxy to the number of active cases at each day.

It is possible to link epidemiological compartmental models, such as the susceptible-infected-recovered (SIR) model, to phylogenies via the CP (Volz and Frost, 2014, Boskova, Bonhoeffer and Stadler, 2014). With the simplest SIR model, the coalescent effective population size $(N_e(t))_{t\geq 0}$ is expressed in terms of the number of infections over time, transmission rate, and the number of susceptible individuals. More complex population dynamics and compartmental models can also be incorporated into the CP framework (Volz and Siveroni, 2018). We refer to the general class of such models as the CP–EPI. In this paper, we survey current methods and challenges for estimating epidemiological parameters from the BDSP and the CP–EPI frameworks and their applications in studying the evolution and epidemic spread of SARS-CoV-2.

1.1 Motivations

Although the field of phylodynamics has advanced in recent years, it has been recognized that there are still many challenges in using sequence data to infer disease dynamics. In Frost et al. (2015), the authors stated the following challenges: (1) modeling of more complex evolutionary processes such as recombination, selection, within-host evolution, population structure, and stochastic population dynamics; (2) modeling that accounts for the sequences sampling design and/or the lack of a well-designed sampling strategy, (3) joint modeling of phenotypic and genetic data, and (4) computation. We have subsequently seen advances in solving some of these challenges, such as modeling of recombination (Müller, Kistler and Bedford, 2022) and stochastic population dynamics (Stadler et al., 2013, Volz and Siveroni, 2018), incorporation of more complex sampling scenarios (Karcher et al., 2016, 2020, Parag, du Plessis and Pybus, 2020, Cappello and Palacios, 2021), and joint modeling of epidemiological and genetic data (Li, Grassly and Fraser, 2017, Tang et al., 2019, Zarebski et al., 2021, Featherstone et al., 2021). However, even in the simplest evolutionary model, inference involves integration over the high dimensional space of phylogenies. This is usually achieved via Markov chain Monte Carlo (MCMC) methods, making inference computationally intractable for large sample sizes.

Apart from the existing challenges, the pandemic presented us with new statistical challenges. Here, we focus our discussion on four challenges: (1) scalability, (2) phylodynamic hypotheses testing, (3) adaptive modeling of the sampling process and (4) interpretability of model parameters.

Current phylodynamic implementations are computationally incapable of analyzing the amount of SARS-CoV-2 sequences available; researchers are forced to subsample available data and to sacrifice model complexity. In Section 3, we focus on the scalability of Bayesian phylodynamic methods. We provide an overview of current practices for analyzing SARS-CoV-2, recent advances in Bayesian computation and the particular challenges in applying such advances in phylodynamics.

The continual rise of new SARS-CoV-2 variants with putative higher transmissibility, demands for novel strategies for statistical hypotheses tests that not only rely on molecular data but also on the sampling process of sequences and phenotypic information from the host and the pathogen. In Section 4, we provide an overview of current practices for testing higher transmissibility of variants of concern and provide a new semi-parametric model that allows for this testing.

Increasing the interpretability of model parameters is becoming one of the most important challenges in phylodynamic inference. Meaningful parameterization often requires more complex modeling and inferential challenges. In Section 5, we provide an overview of phylodynamic methods that aim to infer prevalence and other epidemiological parameters from molecular sequences and count data, and highlight some future directions in the field. Finally, heterogeneous strategies of molecular sequence collection demands for adaptive phylodynamic methods that properly account for this heterogeneity. In Section 6, we discuss recent advances in temporal modeling of the sampling process of molecular sequences. Section 7 concludes with a discussion of encompassing themes that have emerged in the paper.

2. BACKGROUND

Neutral models of evolution typically assume that the tree topology and the branching times (or coalescent times) are independent. In the next two sections, we will summarize the two most popular models on phylogenies used in phylodynamics.

2.1 Coalescent process (CP)

A retrospective probability model on phylogenies is the standard coalescent. The standard coalescent was initially proposed as the limiting stochastic process of the ancestry of n samples chosen uniformly from a large population of $N \gg n$ individuals undergoing simple forward dynamics (Kingman, 1982a, 1982b). It was later extended to variable population sizes (Slatkin and Hudson, 1991, Griffiths and Tavare, 1994) and heterochronous sampling (Felsenstein and Rodrigo, 1999). Here, we consider these extensions, and assume that samples are obtained at times $\mathbf{y} = (y_1, \ldots, y_n)$, with y_i denoting the sampling time of the ith sample. Coalescent models have been reviewed extensively (Rosenberg and Nordborg, 2002, Marjoram and Tavaré, 2006, Tavaré, 2004, Berestycki, 2009, Wakeley, 2009, 2020) and we refer the reader to those references for further details.

The space of phylogenies is the product space $\mathcal{G}_n = \mathcal{T}_n \times \mathbb{R}^{n-1}$ of discrete ranked and labeled tree topologies $\mathcal{T}_n$ and of vectors of coalescent times $\mathbf{t} = (t_2, \ldots, t_n)$, where t_k indicates the $(n-k+1)$th time two lineages have a common ancestor, when proceeding backwards in time from the tips to the root (Figure 2(A)). The coalescent density of the phylogeny is:

$$(1) \quad p(\mathbf{g} \mid N_e(t)) = \exp\left(-\int_0^\infty \frac{C(t)}{N_e(t)}\,dt\right) \prod_{k=2}^{n} \frac{1}{N_e(t_k)},$$

where $C(t) = \frac{A(t)(A(t)-1)}{2}$, termed the coalescent factor, is a combinatorial factor of the number of extant lineages $A(t) = \sum_{i=1}^{n} I(y_i > t) - \sum_{k=2}^{n} I(t_k > t)$. Here, the density is parameterized by $(N_e(t))_{t\geq 0} := N_e$ that denotes the effective population size (EPS). In the CP, the rate of coalescence, which is when two lineages meet a common ancestor, is inversely proportional to the EPS. That is, going backwards in time, a long waiting time for the first coalescence indicates large EPS during that period of time. Under population dynamics following a *Wright–Fisher model*, at time t, $N_e(t) = N(t)/N(0)$ is the relative census population size (Tavaré, 2004). Under more general population dynamics, the EPS is usually interpreted as a relative measure of genetic diversity as it might not depend linearly on the census population size (Wakeley and Sargsyan, 2009).

2.2 Birth-Death-Sampling Process (BDSP)

In the BDSP (Stadler et al., 2013), the population dynamics follows an inhomogeneous birth-death Markov process forward in time in which a birth represents a transmission event, and a death represents the event in which the individual either recovers, becoming noninfectious, or dies. The process starts with a single infected individual at time $t = 0$. At time t, a transmission occurs with rate $\lambda(t)$ and an individual becomes noninfectious with rate $\mu(t)$. Given that we observed a fraction of the population, BDSP requires the definition of a sampling process. The sampling process selects single lineages according to a Poisson process with rate $\psi(t)$, and/or in bulk at predetermined fixed time points with a sampling probability $\rho(t)$ of each lineage. That is, a fraction $\rho(t)$ of the pool of infected individuals at time t is selected uniformly at random to be in the sample. Figure 2(B) depicts a full realization of the process in which only black tips are sampled to form the sampled phylogeny.

Current implementations of the BDSP (Stadler et al., 2013, Bouckaert et al., 2019) assume all rates are piecewise constant functions with jumps at $u_1 < \cdots < u_{p-1}$ (marked by green dotted lines in Figure 2(B)) and denoted in vector form by $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$, $\boldsymbol{\psi}$, and $\boldsymbol{\rho}$, where the ith element corresponds to the rate during $[u_{i-1}, u_i)$ $(i = 1, \ldots, p)$. Further, s tips are sequentially sampled at $y_1 < \cdots < y_s$ (marked by red dotted lines in Figure 2(B)), and additionally, m_i lineages are sampled in bulk at each time u_i with the sampling probability ρ_i per lineage, resulting in $n = s + \sum_{i=1}^{p} m_i$ total samples. In the example of Figure 2(B), no sequences are sampled in bulk at time u_1 $(m_1 = 0)$, while two sequences are sampled in bulk at time u_2 $(m_2 = 2)$. The $n-1$ branching times of the n samples are denoted by $x_1 < \cdots < x_{n-1}$ (marked by blue dotted lines in Figure 2(B)), and let n_i be the number of infected

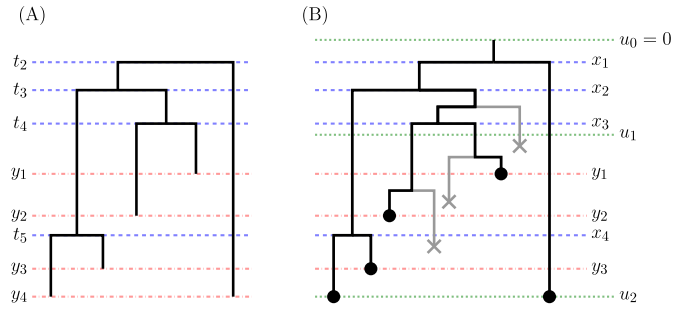

FIG. 2. *Example of a phylogeny. (A) Example of a phylogeny realization from the CP with $n = 5$. t_i's and y_i's indicate coalescent times and sampling times, respectively. For details, see Section 2.1. (B) An example phylogeny from the BDSP that started at $t = u_0$ and ended at u_2. The filled circles represent sampled lineages and the crosses indicate extinct lineages. At each time interval $[u_{i-1}, u_i)$, the rate parameters are assumed to be constant. The branching times are denoted by x_k. The lineages are sampled under two sampling schemes: sequentially at times y_k or in bulk at times u_i. For details, see Section 2.2.*

individuals in the sampled phylogeny at time u_i excluding newly sampled lineages in bulk at u_i. For example, in Figure 2(B), $n_1 = 4$ and $n_2 = 0$. The likelihood of the sampled phylogeny is

$$
\begin{aligned}
&p(\mathbf{g} \mid \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\psi}, \boldsymbol{\rho}, \boldsymbol{t}) \\
(2) \quad &= \underbrace{q_1(0)}_{\text{trans. at the root}} \underbrace{\prod_{i=1}^{n-1} \lambda_{I(x_i)} q_{I(x_i)}(x_i)}_{\text{trans. at internal nodes}} \underbrace{\prod_{i=1}^{s} \frac{\psi_{I(y_i)}}{q_{I(y_i)}(y_i)}}_{\text{seq. sampling trans.}} \\
&\quad \times \underbrace{\prod_{i=1}^{p} \left(\frac{\rho_i}{q_i(u_i)} \right)^{m_i}}_{\text{bulk. sampling trans.}} \underbrace{\prod_{i=1}^{p-1} \left((1-\rho_i) q_{i+1}(u_i) \right)^{n_i}}_{\text{no trans. among } n_i \text{ extant lineages}},
\end{aligned}
$$

where $I(t) = i$ $(i = 1, \ldots, p)$ for $t \in [u_{i-1}, u_i)$ and 0 otherwise. $q_i(t)$ denotes the density of the per-lineage dwelling time in $[u_{i-1}, u_i)$, that is, the density that a lineage at time $t \in [u_{i-1}, u_i)$ evolves as observed in the tree, with $q_i(u_i) = 1$. The explicit expression for $q_i(t)$ is parametrized in terms of $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$, $\boldsymbol{\psi}$, and $\boldsymbol{\rho}$, and it appears in the Supplementary Material of Stadler et al. (2013). We further note that Equation (2) assumes the two types of sampling schemes. When sequences are sequentially sampled, it is assumed that sequences become noninfectious immediately after and hence, their contribution to the likelihood becomes $\frac{\psi_{I(y_i)}}{q_{I(y_i)}(y_i)}$ at each sampling time y_i. Similarly, sequences sampled in bulk become noninfectious and hence their contribution to the likelihood becomes $\frac{\rho_i}{q_i(u_i)}$ at each bulk sampling effort at time u_i. However, sequences that are not sampled in bulk at time u_i (with probability $1 - \rho_i$) remain infectious in the sampled phylogeny and contribute to the likelihood with factors $q_{i+1}(u_i)$.

Equation (2) is the result of a series of papers. Thompson (1975) showed that in the case of constant birth and death rates, the branching times of the tree with tips consisting of only present-day individuals, conditioned on the time at the root, are i.i.d. Nee, May and Harvey (1994) and Gernhard (2008) showed that the same result can be obtained when conditioning on the number of tips. The fact that Equation (2) can be obtained as a completely observed Markovian process is the result of Stadler (2009, 2010), who showed that the BDSP can be interpreted as a birth-death process with reduced rates and complete sampling. Finally, Stadler et al. (2013) extended the result under piece-wise constant birth, death and sampling rates. Here, branching times are no longer i.i.d. but remain independent. We note however that increasing the number of parameters increases the risk of model nonidentifiability and runaway behavior of parameter estimates.

Extensions to the BDSP include the flexibility of modeling the probability $r(t)$ of a sampled lineage to become effectively noninfectious immediately following the sampling event, and the modeling of multi-type birth and death events, accounting for population structure (Scire et al., 2020). A more general framework unifying existing BDSP models has been recently proposed by MacPherson et al. (2021).

2.3 Bayesian Phylodynamic Inference

Phylogenies are usually not observed; the CP or the BDSP density is used as prior on the phylogeny in order to infer phylodynamic parameters, denoted by $\boldsymbol{\theta}$, such as N_e or transmission rate $(\lambda(t))_{t \geq 0}$. Let D denote the observed molecular sequences sampled at times $\mathbf{y}$. In the phylodynamic generative model, phylodynamic parameters stochastically dictate the shape of the phylogeny; given a phylogeny, a process of substitutions is superimposed along the branches of the phylogeny that generates observed data. The target posterior distribution is the augmented posterior $P(\mathbf{g}, \boldsymbol{\theta}, \mathbf{Q} \mid D, \mathbf{y})$, where $\mathbf{Q}$ denotes substitution parameters, such as the global mutation rate and *transition* and *transversion* substitution rates between nucleotide bases. The number of substitution parameters of different substitution models can vary extensively. Yang (2014) provides a comprehensive reference of different mutation models used in phylodynamics.

3. SCALABILITY

The posterior distribution $P(\mathbf{g}, \boldsymbol{\theta}, \mathbf{Q} \mid D, \mathbf{y})$ is usually approximated via Markov chain Monte Carlo (MCMC). Mixing of Markov chains in the high dimensional space of phylogenetic trees and model parameters is challenging, mostly because the posterior distributions on these discrete-continuous state spaces are highly multimodal (Whidden and Matsen IV, 2015). State-of-the-art algorithms, such as those implemented in BEAST (Suchard et al., 2018) and BEAST2 (Bouckaert et al., 2019), exploit GPUs (Ayres et al., 2012) and multi-core CPUs to run multiple MCMC chains in parallel, and carefully designed transition kernels to improve the mixing.

A parallel tempering method proposed by Altekar et al. (2004) apply the Metropolis-coupled MCMC (MC^3) method in which multiple chains are run in parallel and "heated". Here, the posterior term in the acceptance ratio is raised to a power (temperature). After a certain number of iterations, two chains are selected to swap states, encouraging them to explore the parameter space and prevent them from getting stuck in a peak. Müller and Bouckaert (2020) improve upon this MC^3 method by choosing the temperatures adaptively.

Despite these efforts, current methods can only be applied to hundreds or few thousands of samples and thus have limited applicability to pandemic-size datasets. The main bottleneck in these algorithms is the exploration

of the space of phylogenetic trees. Under the substitution models typically used for phylodynamic inference, all phylogenies with n tips have nonzero likelihood, and Markov chains on the space of phylogenetic tree topologies are known to mix in polynomial time (Simper and Palacios, 2020).

In the rest of this section, we first summarize some of the most popular pipelines recently used for phylodynamic analyses of SARS-CoV-2 sequences. Then we review some of the recent advances towards scalable phylodynamic inference.

3.1 Practices in Analyzing SARS-CoV-2 Data

Lacking a method or software capable of dealing with the number of available sequences, researchers usually resort to different types of approximations: (1) partition available data into subsets and analyze each subset independently (Lemey et al., 2020, Volz et al., 2021a), or (2) analyze a subsample selected at random from the set of available sequences (Choi, 2020, Müller et al., 2021), or (3) estimate a single MLE phylogeny from subsampled sequences, for example, the phylogeny available and periodically updated in Nextstrain (Hadfield et al., 2018), or obtain an MLE phylogeny directly with fast implementations such as TreeTime (Sagulenko, Puller and Neher, 2018) and IQ-TREE (Minh et al., 2020); phylodynamic parameters are then inferred from the fixed phylogeny (van Dorp et al., 2020, Maurano et al., 2020, Dellicour et al., 2021). However, these approaches have their limitations. Conducting analyses with only a subset of the data may increase estimates uncertainty and reduce the time interval of estimation. The lattter occurs because larger sample sizes will take longer to meet a common ancestor in expectation (Wakeley, 2009). Estimation of evolutionary parameters from a fixed estimated genealogy is also known to underestimate uncertainty (Palacios et al., 2014).

Examples of the largest scale analyses have been Volz et al. (2021a), who include approximately 27,000 sequences and du Plessis et al. (2021), who study 50,887 SARS-CoV-2 genomes in the UK. du Plessis et al. (2021) divide the full dataset into five smaller datasets according to whether the samples carry one of five groups of mutations. The authors then estimate the five phylogenies with an approximate MLE method, where they employ an approximate likelihood in lieu of an exact one. The five MLE phylogenies are then analyzed separately. Phylogenies obtained with MLE methods cannot be readily used to infer evolutionary parameters in a CP framework if they are multifurcating trees. The MLE phylogeny results in a multifurcating tree—a tree with nodes that directly subtend more than two children—when multiple lineages have the same likelihood of descending from the same parent. This is a common situation in SARS-CoV-2 applications. To infer EPS and to sample from the phylogenetic posterior distribution, the authors sample over the set of binary trees compatible with a given multifurcating tree. Here, a binary phylogeny is compatible with a multifurcating phylogeny if the latter can be obtained by removing internal nodes from the binary phylogeny. Let $\mathbf{g}_{\mathrm{MLE}}$ denote the estimated MLE phylogeny. The authors then approximate the posterior distribution $P(\mathbf{g} \prec \mathbf{g}_{\mathrm{MLE}}, N_e, \mathbf{Q} \mid D, \mathbf{g}_{\mathrm{MLE}})$ while constraining the posterior exploration to binary phylogenies that are compatible with the MLE phylogeny. Although this method does not account for all phylogenetic uncertainty, it does stochastically resolve multifurcating into bifurcating events. In du Plessis et al. (2021), the authors identified the eight largest transmission lineages (from the five empirical posterior phylogenetic distributions) and compared sequence frequencies over time across the eight lineages and their geographic dispersion in order to understand the different patterns of transmissions. Here, a transmission lineage corresponds to a subtree whose inferred origin occurred out of UK but with subsequent inferred local transmission within the UK. The authors showed that lineages introduced prior to their national lockdown tended to be larger and more dispersed. Volz et al. (2021a) first estimate the MLE phylogeny, then identify on the MLE phylogeny several clades (clusters) of interest. Finally, phylodynamic analyses are conducted on each cluster of samples independently.

3.2 Recent Advances

In the following, we review some computationally efficient approaches for Bayesian phylogenetic inference, including approximate MCMC, online algorithms, and parallel algorithms. While some of the described attempts are promising, they are not yet readily applicable to the type of questions researchers have tried to address in the pandemic. We expect to see many statistical developments in this area in the years to come.

3.2.1 *Sequential Monte Carlo*. Sequential Monte Carlo (SMC) methods (also called particle filters) are a set of algorithms used to approximate posterior distributions; See Chopin and Papaspiliopoulos (2020) for an introduction. SMC-based algorithms have been used to approximate the posterior of phylogenies and mutation parameters through particle MCMC (Bouchard-Côté, Sankararaman and Jordan, 2012, Wang, Bouchard-Côté and Doucet, 2015).

Recently, Wang, Wang and Bouchard-Côté (2020) proposed to approximate the joint posterior of phylogeny and mutation parameters with a fully SMC approach based on annealed importance sampling (Neal, 2001). Here, at each iteration, the SMC algorithm maintains k phylogenetic trees and substitution parameters (particles) with their corresponding weights. The k particles are updated

according to traditional Markov chain moves, and acceptance probabilities are based on a likelihood raised to a power (temperature) according to a fixed temperature schedule. A great promise of SMC methods is the possibility to be naturally extended to the online setting. We discuss some proposals in the following subsection.

3.2.2 *Online methods*. During an outbreak or epidemic, sequencing data often come in sequentially. Redoing the analysis whenever a new sequence becomes available is time-consuming. Thus, it is desirable to have an online algorithm that can update the inference using new sequences without having to start the analysis from the beginning. Both Dinh, Darling and Iv (2018) and Fourment et al. (2018), propose online SMC algorithms, which updates the particles and weights when a new sample is added. Again, these methods target phylogeny and mutation parameters and requires further work in order to incorporate the SMC approach in phylodynamics.

Gill et al. (2020) propose a distance-based method that adds a new sample to the current sampled phylogeny in the last iteration, simultaneously updating the phylogeny, phylodynamic and evolutionary parameters. The Markov chain is then resumed with the newly added sample. This method is applicable to phylodynamic analysis and is implemented in BEAST. Lemey et al. (2020) recently applied this method to update a previous analysis of SARS-CoV-2 sequence data with newly acquired samples.

3.2.3 *Variational Bayes* (*VB*). VB (Jordan et al., 1999, Hoffman et al., 2013, Blei, Kucukelbir and McAuliffe, 2017) is a popular alternative to MCMC methods for approximating posterior distributions. Given a class of parametric distributions, VB finds the distribution in the class closest to the target posterior distribution in the sense of Kullback–Leibler (KL) divergence. So the problem of approximating the posterior distribution is recast as an optimization problem, which tends to be faster than classic MCMC. The challenge of applying variational methods to phylogenetics is to choose a sufficiently flexible class of distributions for the tree topologies. Zhang and Matsen IV (2019) introduce the variational distribution $Q_{\phi,\psi}(\mathcal{T}, q) = Q_\phi(\mathcal{T})Q_\psi(q \mid \mathcal{T})$ for phylogenetics, where Q_ϕ is the distribution over tree topologies and Q_ψ is the distribution over branch lengths. They take Q_ϕ to be the subsplit Bayesian network, which is defined as the product of conditional probabilities at each internal node (split) from the root to the leaves, and in such a way that the transition probabilities only depend on the parent-child pair, and not on the particular node in the tree. The branch lengths distributions Q_ψ are chosen to be independent log-normal distributions. The number of parameters grows with the number of samples and it can potentially be computationally expensive for large sample sizes. We note that both the SMC and VB methods are only designed for inferring the phylogeny and mutation parameters. It demands further work to apply them to estimate the phylodynamic parameters like effective population size.

3.2.4 *Divide-and-conquer.* Divide-and-conquer MCMC is an attractive strategy in which the full dataset is partitioned into several subsets; each subposterior—posterior given the subset—is then approximated by running independent MCMC chains, and the subposteriors are then combined to estimate the full posterior (Huang and Gelman, 2005, Neiswanger, Wang and Xing, 2013, Srivastava et al., 2015). However, most of these algorithms rely on the crucial assumption that the subsets are mutually independent. This assumption is violated because molecular sequences share ancestral history (or transmission), modeled by the phylogeny.

4. TESTING IN PHYLODYNAMICS

In the previous section, we described a challenge researchers face while inferring the phylogeny and phylodynamic (coalescent or birth-death) parameters. Inference of these parameters is commonly an intermediate step to address other scientific questions.

In the current pandemic, we have witnessed a surge of novel variants that have caused public health concern (Volz et al., 2021b, Davies et al., 2021). A significant focus of SARS-CoV-2 research has been the study of whether specific mutations (variants) impact viral properties, such as transmissibility, virulence, and the ability to increase disease severity. While it is often possible to study cell infectivity in animal models and to study *in vitro* whether a mutation is associated with changes in viral phenotypes, determining whether it leads to significant differences in viral transmission or disease response relies on observational data from both, the pathogens and the hosts. These data often consist of molecular sequences, epidemiological and clinical data. These types of statistical analyses are challenging because although an increase in frequency is a signal of selective advantage, observed increase can also be the product of many other factors such as multiple introductions and human behaviors. In this section, we restrict our attention to two types of analyses designed to test whether there are significant differences in transmissibility between a variant of concern (VOC) and a non-VOC. The first type is solely based on molecular data, and the second type utilizes molecular data paired with phenotypic traits and clinical data.

4.1 Detecting Higher Transmissibility Relying Solely on Molecular Data

4.1.1 *Practices in analyzing SARS-CoV-2 data*. A simple and popular strategy for estimating the growth rate of the VOC and non-VOC populations consists of modeling

the sampling times of sequences solely (ignoring molecular data) (Volz et al., 2021a, 2021b, Davies et al., 2021, Trucchi et al., 2021). This is commonly done through a logistic growth model which assumes that after a phase of initial growth, the growth rate decreases as the population size approaches its maximum size. Data, in this case, consists of counts of genomes belonging to the VOC and the non-VOC over time, with counts binned into weeks. This type of analysis simply models the proportion of VOC sequences over the total number of collected sequences over time, rather than estimating the EPS.

Phylodynamic models have been applied to estimate the effective population size (EPS) of several VOCs and non-VOCs from molecular samples. It is assumed that the non-VOC spread through the population and accumulated variation before the VOC appeared in the population. If the VOC confers higher transmissibility, its population should increase at a faster rate than that of the non-VOC in multiple locations around the world. Moreover, comparisons between the two growth rates should be based on VOC and non-VOC samples sharing the same environmental factors, such as public policies and temporal seasons, to control for possible confounders.

Volz et al. (2021a) stress the need to observe repeated independent introductions of each variant and follow their trajectories. The authors analyzed molecular data collected in the UK during the first six months of 2020 to test whether the VOC (D614G substitution) had selective advantages. The authors first obtained a global MLE phylogeny, together with inferred location of phylogenetic branches, in order to identify UK clusters. Then, each UK cluster was labeled VOC (sequences carrying 614G) and non-VOC (sequences carrying 614D). These clusters included one or a small number of introductions of the virus in the UK. The authors then estimated EPS growth rates for each cluster and compared the empirical distributions of the point estimates of the clusters of VOC and non-VOC growth rates and found no significant difference.

Other studies have also identified multiple introductions for estimating VOC growth rates. For example, Davies et al. (2021) considered introductions across different countries. A challenge with this type of analyses lies in the detection of independent introductions. It is unclear how ignoring phylogenetic uncertainty affects the definition of introductions and estimation of EPS, and whether introductions in different locations can be treated as independent. Volz et al. (2021b) performed a phylodynamic case-control study which consisted in selecting 100 random samples of 1000 sequences with the VOC (alpha variant) paired with another 1000 non-VOC sequences. Those sequences were matched by the week and the location of the collection. The random samples were selected with weights proportional to the number of reported cases per week and local authority in the UK and hence, expected to be representative of the UK. The 200 phylogenies were estimated via MLE, and 200 EPS trajectories were inferred from each tree in order to obtain two bootstrap distributions of VOC and non-VOC EPSs. In their study, the comparison of the two EPS distributions supported an increase in the transmissibility of the VOC.

In the two phylodynamic studies discussed, the EPSs are estimated independently for the two populations. We argue that this approach might be suboptimal because the two trajectories may be correlated. In addition, both methods assume piece-wise constant growth rates and report averages across time, that is, the variation over time of the growth rate and their uncertainty quantification are completely lost in the comparison between the two populations. In the next section, we discuss a simple hierarchical model that jointly models the two lineages so that the difference in growth rates is easily interpretable.

4.1.2 *A simple model to test for population growth.* Assume that we are provided with the two phylogenies g_0 and g_1 of the non-VOC and the VOC, respectively. We can model the two phylogenies as conditionally independent given a shared baseline EPS denoted by N_e. More specifically, we assume g_0 is a realization of a CP with parameter N_e and g_1 is a realization of a CP with parameter αN_e^{β}. The model is parsimonious, describing the relative rate of growth of the non-VOC population to that of the VOC-population with a single parameter β: $\beta = 1$ indicates that the growth rate of the EPS in the two groups is identical, $\beta > 1$ indicates that the growth in genetic diversity of the VOC is larger than the non-VOC. Note that β can also take negative values. The parameter $\alpha > 0$ is a scaling parameter that allows to adjust for the fact that VOC and non-VOC EPSs could have different absolute sizes. However, it is not time-varying, so different values of α are not informative for how the growth rate between the two populations differ: for a fixed β, $0 < \alpha < 1$ indicates that the VOC EPS is smaller than that of the non-VOC, a value of $\alpha > 1$ indicates the opposite.

This simple model is highly interpretable, with a single parameter, β, quantifying the change in transmissibility of the VOC relative to the non-VOC. One can choose the preferred prior on N_e, such as a Gaussian Markov random field (GMRF) (Minin, Bloomquist and Suchard, 2008), a Gaussian process (Palacios and Minin, 2013), and the Horseshoe Markov random field (Faulkner et al., 2020). While we would like to approximate the posterior distribution $P(N_e, \alpha, \beta \mid g_0, g_1)$, this is computationally very demanding with traditional sampling-based methods like MCMC. The integrated nested Laplace approximation (INLA) (Rue, Martino and Chopin, 2009) is a highly competitive approximation available for latent Gaussian models, allowing us to approximate the marginal posterior distributions within seconds. The accuracy of this approximation in phylodynamics has been studied in Lan

et al. (2015). We provide a publicly available implementation of the following model in `phylodyn` available in https://github.com/JuliaPalacios/phylodyn:

$$
\begin{aligned}
g_0 \mid N_e, \mathbf{y}_0 &\sim \text{Coalescent with EPS} \quad N_e, \\
g_1 \mid N_e, \alpha, \beta, \mathbf{y}_1 &\sim \text{Coalescent with EPS} \quad \alpha N_e^{\beta}, \\
\log N_e \mid \tau &\sim \text{GMRF}, \\
\tau \mid a_0, b_0 &\sim \text{Gamma}(a_0, b_0), \\
\log \alpha \mid \sigma_0^2 &\sim \mathcal{N}(0, \sigma_0^2), \\
\beta \mid \sigma_1^2 &\sim \mathcal{N}(0, \sigma_1^2).
\end{aligned} \tag{3}
$$

For parsimony, we are ignoring the discretization of $\log N_e$. Model (3) enforces a strict parametric relationship between the two EPSs. While this may be too restrictive, we argue that it is a reasonable price to pay for the sake of interpretability and parsimony. We illustrate the methodology by applying the model to SARS-CoV-2 sequences collected in Washington state at the beginning of the epidemic.

Application to SARS-CoV-2 sequences in Washington state. We randomly selected 100 sequences with the D codon (non-VOC) and 100 sequences with the G codon (VOV) in position 614, from the 4356 publicly available sequences in GISAID (Shu and McCauley, 2017) collected in Washington state between January 1, 2020 and June 30, 2020. We analyzed the two samples independently and obtained the two phylogenies (Figure 3) by summarizing the two corresponding posterior distributions obtained with `BEAST2` (Bouckaert et al., 2019). Details of model and MCMC parameters are located in the Appendix.

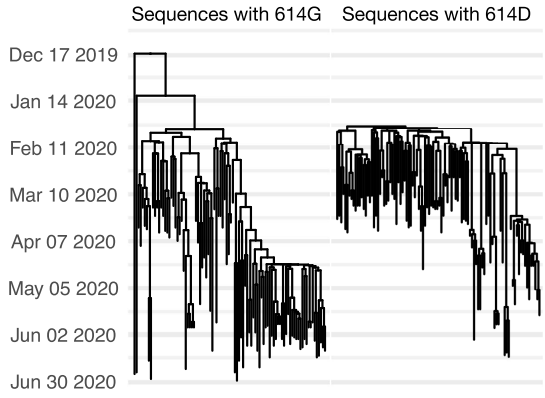

FIG. 3. *Phylogenies of the G and D variants inferred in Washington state. Phylogenies are the maximum clade credibility trees obtained from posterior distributions estimated with BEAST (Appendix). Each tree is generated from* 100 *sequences chosen at random among those collected in Washington state between January* 1, 2020 *to June* 30, 2020. *The left tree includes sequences with G in the codon position* 614 *of the viral spike protein. The right tree includes sequences with D in the codon position* 614. *By* 2021, *the G type dominated the pandemic.*

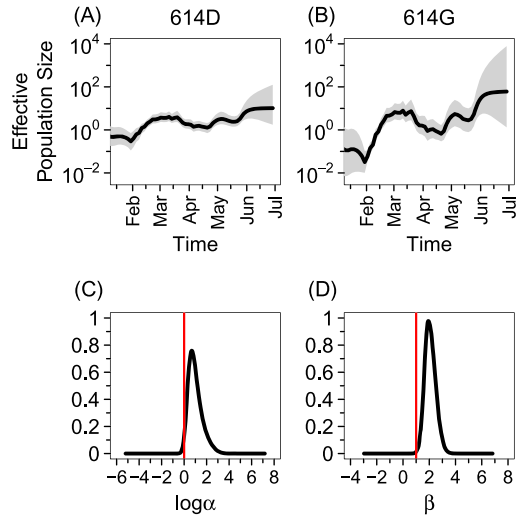

FIG. 4. *Effective population sizes of D and G variants in Washington State. Panels (A–B) depicts posterior mean of* N_e *and* αN_e^{β}, *the effective population size trajectories of the D and the G variants respectively. Shaded areas represent* 95% *BCIs. Panel (C) depicts estimated posterior distribution of* $\log\alpha$ *and panel (D) depicts estimated posterior distribution of* β. *Red lines indicate the values of* $\log\alpha$ *and* β *under the hypothesis that both variants share the same effective population size trajectory.*

Panels (A–B) of Figure 4 depict the posterior medians (solid lines) and 95% BCIs of N_e (shaded areas) obtained by fitting model (3). Panel (C) depicts the estimated posterior distribution of α and panel (D) depicts the estimated posterior distribution of β. We note that INLA provides approximation to marginal posteriors of $\log\alpha$ and β but not jointly. In the random subsample considered, sequences with the D variant generally have earlier collection dates than sequences with the G variant. This is consistent with the general observation that the G variant progressively replaced the D variant (Hadfield et al., 2018). The main parameter of interest is β. The posterior distribution has mean 2.08 and 95% credible region (1.35, 2.97). It is well above 1, suggesting that EPS growth is more pronounced among sequences having the G variant. The impact of $\beta \approx 2$ is evident in the first two panels of Figure 4, where the EPS of G grows at a higher rate than that of the control group. The posterior median of α is 2.46, with 95% Bayesian credible region (1.12, 13.35). As mentioned above, this does not indicate difference in growth rate, but a vertical shift of the baseline EPS.

A benefit of the model described here is the flexible nonparametric prior placed on N_e. Panels (A–B) of Figure 4 suggest that parametric models would not reasonably approximate the trajectory: for this dataset, our estimates indicate that N_e fluctuates in the period considered. The goal of the analysis is inferring the parameter β. Hence, we argue that the best possible fit in modeling N_e is necessary. A future development includes the infer-

ence of the proposed model parameters from molecular data directly.

4.2 Combining Molecular Sequence Data and Other Types of Data

We now examine the situation when viral molecular data are matched with host clinical data and we are interested in testing an association between clinical traits such as disease severity and transmission history. For example, does the variant of concern affect disease severity? We first describe some current practices in testing for such associations.

4.2.1 *Practices in analyzing SARS-CoV-2 data.* If one is studying a VOC, the variant naturally partitions the hosts into two groups: individuals carrying the VOC and those carrying the non-VOC. Here, one can resort to standard statistical tests for detecting changes in mean or distribution (Volz et al., 2021a, Volz et al., 2021b, Davies et al., 2021, Leung et al., 2021). For example, Volz et al. (2021a) study the mutation in codon position 614 (D and G mutations) and analyze the difference in several response variables such as disease severity and age using a Mann-Whitney U-test. One related approach that tests for overall correlation between phenotypes and shared ancestry (transmission structure) that accounts for phylogenetic uncertainty is the BaTS test (Parker, Rambaut and Pybus, 2008). This method relies on simple statistics such as the parsimony score (Fitch, 1971) and the association index (Wang et al., 2001). In particular, the parsimony score is the minimum number of trait value changes at internal nodes needed to be consistent with observed traits at the tips. A strong trait-phylogeny association would imply small number of changes. The parsimony score is then calculated for every tree in the posterior distribution and its posterior distribution reported. The association index is calculated as a weighted average frequency of the least common trait across all internal nodes in the phylogeny, with low values indicating strong phylogeny-trait association.

However, the situation is more challenging when the candidate VOC has not yet been identified. For example, Zhang et al. (2020) estimated a phylogeny and used it to identify two major clades, the authors then characterized these two clades, that is, they identified which mutations differentiate them, and tested for association with clinical data. Here the choice of which clades to pick and compare is somewhat arbitrary.

4.2.2 *Recent advances.* Behr et al. (2020) recently proposed treeSeg, a method for testing multiple hypotheses of association between a response variable and the phylogeny tree structure. A key feature in treeSeg is to formulate the testing problem as a multiscale change-point problem along the hierarchy defined by a given phylogeny. The test statistic is based on a sequence of likelihood ratio values, and the change-point detection methodology is based on the SMUCE estimator (Frick, Munk and Sieling, 2014). This method was recently applied to test an association between the inferred phylogeny from SARS-CoV-2 sequences collected in Santa Clara County, California in 2020, and disease severity (Parikh et al., 2021). The authors did not find any significant association.

One statistical challenge in applying treeSeg to phylodynamics is that it ignores uncertainty in the tree estimation. If a subtree is found to have an association to the response, we can assess uncertainty in the subtree formation (independent of treeSeg analysis) by an estimate of the subtree posterior probability or the subtree bootstrap support (Efron, Halloran and Holmes, 1996). A more integral approach is an open problem.

Another situation arises when we are interested in assessing phenotypic correlations among traits (Felsenstein, 1985, Grafen, 1989, Pagel, 1994). Here, multiple traits are modeled as stochastic processes evolving along the branches of the phylogeny; for example, as Markov chains (Pagel, 1994), or as multivariate Brownian motion (Felsenstein, 1985, Huelsenbeck and Rannala, 2003, Felsenstein, 2005, Felsenstein, 2012, Cybis et al., 2015). Despite their relevance in understanding viral evolution and drug development, computation is the main limitation preventing the widespread use of this methods' class. Zhang et al. (2021) is a recent attempt to make inference more scalable. They introduce an algorithm based on recent advances in the MCMC literature (the Bouncy particle sampler (Bouchard-Côté, Vollmer and Doucet, 2018a)). However, the implementation of the methodology seems quite involved preventing broader applicability.

5. PHYLODYNAMIC INFERENCE OF EPIDEMIOLOGICAL PARAMETERS

Epidemiological parameters are often estimated from case count time series; these estimates, however, can be biased due to delays and errors in reporting. Sequence data provide complementary information that can be used for estimating critical epidemiological parameters within a phylodynamic framework. Formal model integration of the CP and epidemiological compartmental models establishes a link between the EPS of pathogens and the underlying number of infected individuals. Equivalently, in the forward-in-time BDSP model, parameters such as the rate of transmission and effective reproduction number can be directly inferred from molecular data.

5.1 Phylodynamic Inference Relying Solely on Molecular Data

5.1.1 *Phylodynamic inference with CP–EPI.* While a linear relationship between the viral EPS and the disease prevalence exists at *endemic equilibrium*, such simple correspondence is not valid in general (Koelle and Rasmussen, 2012). The CP–EPI provides a probability model of a phylogeny in terms of epidemiological parameters by linking the EPS trajectories to a mechanistic epidemic model (Volz et al., 2009). The infectious disease population dynamics can be modeled as a CTMC whose state space is the vector of occupancies in compartments corresponding to disease states. However, the transition probability becomes intractable even for the simplest SIR model (Tang et al., 2019). One way to mitigate the computational issue is to deterministically model the disease dynamics (Kermack, McKendrick and Walker, 1927); we term such model as a deterministic CP–EPI.

Volz et al. (2009) developed a theoretical basis for the deterministic CP–EPI. In the particular case of SIR dynamics, the population is divided into compartments. At time t, the state is $\{S(t), I(t), R(t)\}$, of susceptible, infected and recovered individuals respectively. In this context, the phylogeny represents the ancestry of a sample of infected individuals in the population. Let $A(t)$ denote the number of lineages ancestral to the sample in the phylogeny at time t. The probability that a transmission event at time t corresponds to a transmission event ancestral to the sample is $\binom{A(t)}{2}/\binom{I(t)}{2}$, since out of all $\binom{I(t)}{2}$ infected pairs, only $\binom{A(t)}{2}$ pairs occur within the ancestors of the sample. Denoting the total number of new infections at time t by $f(t)$, the rate of coalescence is

$$\lambda_A(t) = f(t)\frac{\binom{A(t)}{2}}{\binom{I(t)}{2}} \approx \binom{A(t)}{2}\frac{2f(t)}{I^2(t)}. \tag{4}$$

Assuming a per capita transmission rate $\beta(t)$, $f(t) = \beta(t)S(t)I(t)$ is the number of transmissions per unit time (the incidence of infection). The population dynamics of compartments, $\{S(t), I(t), R(t)\}$, is governed by an initial state and a system of ordinary differential equations. Recall that $\lambda_A(t) = \binom{A(t)}{2}/N_e(t)$ is the coalescence rate in the standard CP, we then get at time t

$$N_e(t) = \frac{I^2(t)}{2f(t)} = \frac{I(t)}{2\beta(t)S(t)}. \tag{5}$$

The initial CP–EPI model has been extended to incorporate serial sampling, population structure, time- and state-dependent rate parameters, and a large class of epidemic processes (Volz, 2012, Volz and Siveroni, 2018). In Volz and Siveroni (2018), the authors assumed that recovery rate and number of susceptible individuals are known; the transmission rate is modeled as a straight line with normal prior on the slope parameter and lognormal prior on the intercept parameter. Inference is performed via MCMC in BEAST2 (Bouckaert et al., 2019). We note that in the SIR models, not all parameters are identifiable. We usually need to assume known values of some parameters and very informative priors (See Louca et al., 2021 for further details).

The deterministic CP–EPI has provided a computationally efficient framework for studying the evolution and pathogenesis of SARS-CoV-2 via estimating $R_0(t)$ at the beginning of the pandemic (Volz et al., 2020, Geidelberg et al., 2021), fine-scale spatiotemporal community-level transmission rate variation (Moreno et al., 2020), and the effects of control measures on epidemic spread (Miller et al., 2020, Ragonnet-Cronin et al., 2021).

So far, we have ignored within-host evolution, that is, we have assumed that pathogen diversity within a host is negligible. It can be shown that Equation (4) is a limiting case of a more general model (Dearlove and Wilson, 2013, Volz, Romero-Severson and Leitner, 2017), which relaxes many assumptions from the previous derivation, such as negligible evolution within host. In the metapopulation CP–EPI, which is based on the metapopulation CP (Wakeley and Aliacar, 2001), each deme corresponds to a single infected host and can be reinfected more than once. Within each host, there is a nonnegligible pathogen population size, and the within-host coalescence does not occur immediately following an infection. Further, during an inter-host transmission, nonnegligible genetic diversity can be transmitted across hosts. Due to its complexity, the current metapopulation CP–EPI model assumes constant rate parameters and deterministic disease dynamics.

As empirical evidence of reinfection and of the effects of within-host diversity on patient disease severity and transmissibility mounts for SARS-CoV-2 (Tillett et al., 2021, Al Khatib et al., 2020, San et al., 2021), it is becoming apparent that developing computationally tractable methods that incorporate both time-varying parameters and stochasticity into the metapopulation CP–EPI framework is an important future direction in the field.

While the deterministic CP–EPI is computationally efficient, epidemiological dynamics are inherently stochastic, with both demographic and environmental stochasticity playing important roles in disease dynamics. The deterministic epidemic model can lead to overconfident estimations when the disease prevalence is low or the population size is small, or when fitting models to long-term data, as the effects of stochasticity accumulate over time (Popinga et al., 2015). The CP–EPI with the stochastic epidemic model, which we term as the stochastic CP–EPI, is better suited for addressing important epidemiological questions, such as the early-stage behavior of an epidemic, the outbreak size distribution, and the extinction probability and expected duration of the epidemic, while accounting for the uncertainties in the estimations (Britton, 2010). We fitted the stochastic CP–EPI model with SIR dynamics to the same data from California used

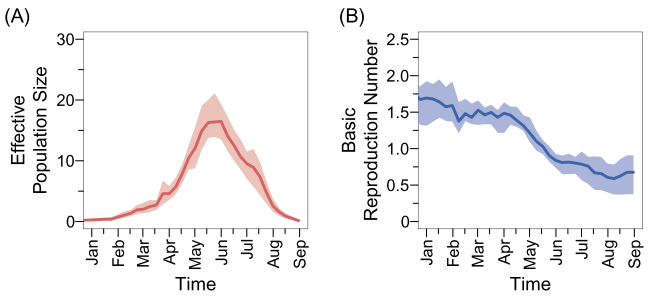

FIG. 5. *The stochastic CP–EPI analysis of SARS-CoV-2 sequences in California in* 2020. *The stochastic CP–EPI implementation of* Tang et al. (2019) *was used to infer the effective population size* N_e (*A*) *and the basic reproduction number* R_0 (*B*) *from the same molecular sequences of California of Figure* 1(*B*). *Posterior medians are indicated by bold lines and* 95% *credible regions by shaded areas*.

to generate Figure 1(B). We used the implementation of Tang et al. (2019) to infer N_e and R_0 depicted in Figure 5. Details can be found in the Appendix.

5.1.2 *Phylodynamic inference with BDSP.* The BDSP (Stadler et al., 2013) introduced in Section 2.2 has been extended to incorporate population structure (Kühnert et al., 2016) and it has been applied for inferring $R_e(t)$ early in the SARS-CoV-2 pandemic in Europe (Nadeau et al., 2021, Hodcroft et al., 2021). The BDSP, however, requires specification of the sampling probabilities, and its misspecification results in biased estimates, as demonstrated in inferring R_0 from the SARS-CoV-2 data in the northwest USA (Featherstone et al., 2021). This is because, in the BDSP, sampling times provide information about the whole population dynamics (Volz and Frost, 2014).

An important distinction between BDSP models and the CP–EPI models is that the BDSP model is parameterized in terms of birth, death, and sampling rates, however it does not directly model the number of infected individuals and the number of recovered individuals over time. The CP–EPI, instead, directly models the number of individuals in each compartment, together with birth and death rates. We note that the BDSP has been extended to infer the prevalence trajectory from molecular sequences and case count data (Section 5.2.2).

5.2 Phylodynamic Inference Relying on Molecular Data and Disease Count Data

When fitting mechanistic population dynamic models, integrating multiple sources of information, particularly time series surveillance data, with molecular sequence data, can improve phylodynamic inference of epidemic model parameters. This subsection describes extensions of CP–EPI and BDSP to incorporate both data sources.

5.2.1 *Phylodynamic inference with CP–EPI.* Rasmussen, Ratmann and Koelle (2011) employed PMCMC for Bayesian inference under the stochastic CP–EPI, from both a fixed phylogeny and time series incidence data. In their implementation, they allow the transmission and recovery rates to vary in time. Unfortunately, inference is computationally expensive due to the high-dimensional parameter space. Other extensions to this framework include incorporation of overdispersion in secondary infections (Li, Grassly and Fraser, 2017). Recently, Tang et al. (2019) proposed to bypass PMCMC and used a linear noise approximation. The authors approximated the SIR transition density with a Gaussian density and developed an MCMC algorithm for this approximate inference.

Current implementations of the stochastic CP–EPI have a few limitations, many of which stem from computational cost. This reduces their utility in SARS-CoV-2 analyses. First, most methods have adopted an epidemic model with one infection compartment and ignore further population structure, such as spatial distribution and age. Second, statistical dependency between sampling times and latent prevalence is ignored. If the sampling process is known, we could incorporate sampling model directly as in the preferential sampling (Karcher et al., 2016), for improving parameter estimation; see Section 6. Finally, to fully account for phylogenetic uncertainty, a computationally efficient method for directly fitting stochastic epidemic models to genetic sequences will be needed.

5.2.2 *Phylodynamic inference with birth-death processes.* There have been a few recent developments in joint modeling of molecular data and case count records under the birth-death population dynamics. Recently, Gupta et al. (2020) extended the BDSP model (Stadler, 2010) to include case count data. The authors derive the density of the phylogeny jointly with case count data in terms of the BDSP rates. This work was later extended to model prevalence (Manceau et al., 2021); building on Gupta et al. (2020), the authors derived the density of the prevalence trajectory conditioned on the phylogeny and case count data. Finally, Andréoletti et al. (2020) extended this work to allow for piecewise constant rates and used it to estimate R_e and prevalence of the SARS-CoV-2 Diamond Princess epidemic that occurred in Jan–Feb 2020.

Vaughan et al. (2019) recently proposed a method that differs from the BDSP discussed in the previous paragraph. The authors propose to estimate the posterior distribution of the full epidemic trajectory, together with the phylogeny and model parameters from molecular sequence data and count data. Here, the authors express the density of the phylogeny jointly with case counts, conditionally on the full epidemic trajectory that consists of the sequence of events (infection, sampling and recovery) and their corresponding event times. The posterior distribution is estimated with PMCMC.

5.3 Phylodynamic Inference with Approximate Bayesian Computation and Deep Learning

As SARS-CoV-2 continues to spread, the virus is subject to strong host and *anthropogenic selective* pressures as has already been exemplified by the emergence of the new variants exhibiting adaptive *antigenic evolution* (Zhou et al., 2021, Lopez Bernal et al., 2021). As discussed in Section 3, however, the likelihood-based Bayesian computation methods are computationally expensive, prohibiting the application of more complex and realistic phylodynamic models such as those involving structured populations, natural selection and recombination. To overcome this obstacle, likelihood-free rejection sampling methods based on approximate Bayesian computation (ABC) (Beaumont, Zhang and Balding, 2002) have been developed for phylodynamic studies. The phylodynamic ABC methods (Ratmann et al., 2012, Poon, 2015, Saulnier, Gascuel and Alizon, 2017) first simulate a large number of phylogenies under complex epidemiological models with different parameter values, then quantify the discrepancy between simulated and "observed" phylogenies and accept the ones close to the target to construct an approximate posterior distribution of the model parameters. Here, the "true" phylogeny is unknown and an estimated phylogeny from sequence data is used as the "observed" phylogeny. The phylogenetic dissimilarity measure can be either a function of summary statistics (Saulnier, Gascuel and Alizon, 2017), where each extracts a specific feature of the phylogeny, or a metric defined directly on the space of phylogenies (Robinson and Foulds, 1981, Billera, Holmes and Vogtmann, 2001, Colijn and Plazzotta, 2017, Kim, Rosenberg and Palacios, 2020). To improve computational efficiency, Ratmann et al. (2012) and Poon (2015) used ABC-MCMC (Marjoram et al., 2003), while Saulnier, Gascuel and Alizon (2017) employed regression-based ABC (Blum and François, 2010).

The ABC-based methods, however, are known to be sensitive to the choice of summary statistics, similarity measures, and match tolerance (Lintusaari et al., 2016). As an alternative, Voznica et al. (2021) proposed a rejection-free approach for estimating epidemiological parameters and for model selection based on deep learning: feed-forward neural network (FFNN) with a large set of summary statistics that were curated for phylodynamic regression-ABC (Saulnier, Gascuel and Alizon, 2017) and convolutional neural network (CNN). A key component in their method is their proposed bijective encoding of (unlabeled) phylogenetic trees as vectors, amenable to standard deep learning methods. As the framework assumes a known phylogeny as an input, phylogenetic uncertainties are not accounted for. While the computational burden lies in simulating the training data and training the network, once trained, the parameter estimation is very efficient without having to retrain the model with new data. They show comparable accuracy under the basic BDSP model and better accuracy under more complex models, which incorporate factors such as superspreader events, when compared to the current popular likelihood-based methods. As the number of SARS-CoV-2 sequences grow exponentially and its disease dynamics varies across regions, the deep learning framework can offer a fast alternative for monitoring the epidemic.

6. PREFERENTIAL SAMPLING

In the standard CP, the temporal sampling process of sequences is assumed to be fixed and uninformative of model parameters. However, the sampling process that determines when sequences are collected can depend on model parameters such as the EPS in some situations. In spatial statistics, preferential sampling arises when the process that determines the data locations and the process under study are stochastically dependent (Diggle, Menezes and Su, 2010). In coalescent-based inference, this can be incorporated by modeling the sampling process as an inhomogeneous Poisson process (iPP) with a rate $\lambda := (\lambda(t))_{t\geq 0}$ that depends on N_e. If the model is correct, the sampling times can provide additional information about the EPS N_e. The statistical challenge is that when the model is misspecified, incorrectly accounting for preferential sampling can bias the estimation of the EPS. The same situation occurs in the BDSP in which the sampling process depends on the death rate (Stadler, 2010, Volz and Frost, 2014, Cappello and Palacios, 2021).

6.0.1 *Recent advances*. Table 1 lists different models and implementations that account for preferential sampling in phylodynamics. Among the parametric approaches, Volz and Frost (2014) model the EPS N_e as an exponentially growing function and λ is linearly dependent on the EPS. Karcher et al. (2016) assume that N_e is a continuous function modeled nonparametrically with Gaussian process priors, and at any time point t $\lambda(t) = \exp(\beta_0)N_e(t)^{\beta_1}$, for $\beta_0, \beta_1 \geq 0$, that is, the dependence between the sampling process and the effective sample size is described by a parametric model. While this model is computationally appealing, the strict parametric relationship between the sampling and coalescent rates can induce a bias if the sampling model is misspecified (see simulations in Cappello and Palacios, 2021).

Parag, du Plessis and Pybus (2020) propose an estimator called Epoch skyline plot, that allows the dependence between the rate of the sampling process and N_e to vary over time. In Parag, du Plessis and Pybus (2020), λ is a linear function of N_e within a given time interval, but the linear coefficient changes across time intervals. This framework allows practitioners to incorporate heterogeneity in the sampling design over time. Cappello and Palacios (2021) extends this approach, modeling both

TABLE 1

Implementations of phylodynamic methods and their applications to SARS-CoV-2 studies (where available). The details of methods are discussed in relevant sections, and their software availability can be found at Table 1 in the Supplementary Material (Cappello et al., 2022)

Method	Implementation	Author	COVID Application	Section
SMC	annealedSMC	Wang, Wang and Bouchard-Côté (2020)		3
Online Bayesian phylodynamics	BEAST 1.10	Gill et al. (2020)	Lemey et al. (2021); Thornlow et al. (2021)	
Variational	`bito`	Zhang and Matsen IV (2019)		
BDSP	`BDSKY`	Stadler (2010); Stadler et al. (2013)	Seemann et al. (2020);	5, 6
	in BEAST2		Featherstone et al. (2021); Hodcroft et al. (2021)	5
Structured BDSP	`bdmm` in BEAST2	Kühnert et al. (2016); Barido-Sottani, Vaughan and Stadler (2020)	Nadeau et al. (2021)	
Coalescent	`PhyDyn` in BEAST2	Volz and Siveroni (2018)	Miller et al. (2020); Geidelberg et al. (2021); Ragonnet-Cronin et al. (2021); Volz et al. (2021a)	
Coalescent	`LNAphyloDyn`	Tang et al. (2019)		
Parametric N_e and λ	*NA*	Volz and Frost (2014)		
Nonparametric N_e and $\lambda = \exp(\beta_0) N_e^{\beta_1}$	`phylodyn` in R	Karcher et al. (2016)	Cappello and Palacios (2021)	6
Epoch Skyline plot (Nonparametric N_e and λ)	`BESP` in BEAST2	Parag, du Plessis and Pybus (2020)		
AdaPref (Nonparametric N_e and $\lambda = \beta N_e$)	`adaPref` in R	Cappello and Palacios (2021)	Cappello and Palacios (2021)	
Nonparametric N_e and $\lambda = \exp(\beta_0) N_e^{\beta_1} + (\boldsymbol{\beta}' X(t))_{t \geq 0}$	BEAST	Karcher et al. (2020)		

N_e and λ nonparametrically, employing Markov random field priors on both N_e and a time-varying coefficient $\beta := (\beta(t))_{t \geq 0}$. Here, the dependence between the two processes is modeled through $\lambda(t) = \beta(t) N_e(t)$ for all $t \geq 0$. Cappello and Palacios (2021) show through simulations that the more flexible dependence between the sampling and the coalescent processes the less the risk of biasing the N_e estimate because of model misspecification while still retaining the advantages of the parametric approaches (narrower credible regions). Karcher et al. (2020) assume that $\lambda(t) = \exp(\beta_0) N_e^{\beta_1} + \boldsymbol{\beta}' X(t)$ for $t \geq 0$, where X is a vector of covariates and $\boldsymbol{\beta}'$ the corresponding linear coefficients. Here, a covariate can be a dummy variable indicating the implementation of lockdown measures. The covariate-dependent preferential sampling requires the availability of information related to the sampling design. Finally, the methodologies of Karcher et al. (2016) and Cappello and Palacios (2021) rely on a known phylogeny, while Stadler (2010), Parag, du Plessis and Pybus (2020), and Karcher et al. (2020) account for uncertainty in the phylogeny.

Application to SARS-CoV-2 sequences in Washington state. We continue the analysis of the Washington molecular sequences introduced in Section 4.1.2. We infer the EPS of the two groups (sequences with 614G and sequences with 614D) from the phylogenies inferred with BEAST2 and plotted in Figure 2. We compare three different models: one that ignores preferential sampling (Palacios and Minin, 2012), the parametric preferential sampling model of Karcher et al. (2016), and the adaptive preferential sampling of Cappello and Palacios (2021). All three models share a GMRF prior on N_e.

Figure 6 depicts the EPS posterior distributions obtained with the three methods applied to the two genealogies. At the bottom of each panel, heat maps represent the sampling times (intensity of the black color is proportional to the number of samples collected). The estimates of the two models accounting for preferential sampling are pretty similar, while not modeling the sampling process leads to a slightly different population size trajectory.

As expected, including a sampling process reduces the credible region width: the mean width of the 95% credible region is much wider for the model that ignores preferential sampling in the sequences with 614G with respect to any of the models accounting for preferential sampling (approximately 6 times large in the first two, and 3.5 times in the second row).

Under the preferential sampling assumption, the more sequences are collected, the higher the EPS is. The effect

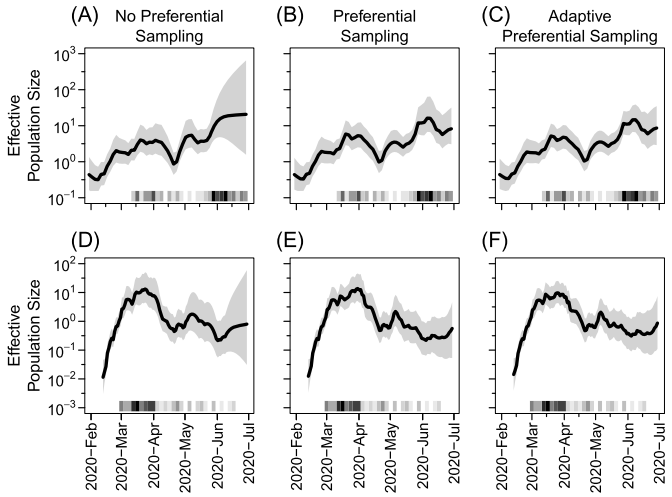

FIG. 6. *N_e estimated from SARS-CoV-2 phylogenies of sequences from Washington state. The first row (panels (A)-(C)) depicts EPS of the G type in codon position* 614 *and the second row (panels (D)-(E)) depicts EPS of the D type in codon position* 614. *The first column estimates are obtained with model of* Palacios and Minin (2012) *that ignores preferential sampling, the second column with the model of* Karcher et al. (2016) *that parametrically models preferential sampling, and the third column models with the adaptive preferential sampling model of* Cappello and Palacios (2021). *In each panel, black lines depict posterior medians and the gray areas the* 95% *credible regions of N_e. Sampling times are depicted by the heat maps at the bottom of each panel: the squares along the time axis depict the sampling time, while the intensity of the black color depicts the number of samples.*

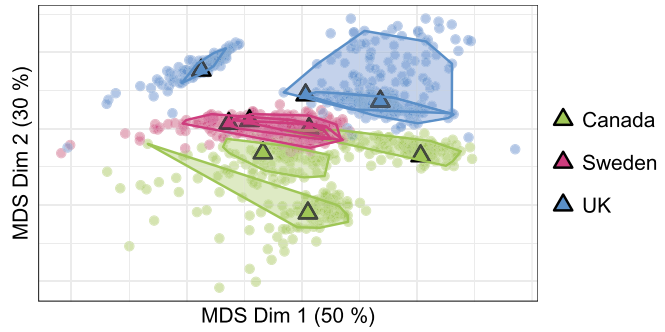

FIG. 7. *MDS plot of phylogenetic posterior distributions from Canada, Sweden, and UK. Each dot represents a phylogeny from one of the nine posterior distributions (three distributions per country). Each posterior (cluster of phylogenies) is estimated from* 100 *samples randomly chosen from November* 2020 *to February* 2021. *Metric on tree spaces given by* Kim, Rosenberg and Palacios (2020). *The triangles indicate the medoids of each distribution and the shaded regions corresponds to* 50% *credible convex hulls around the medoids.*

is evident in Figure 6. For example, in the month of June, we see that the EPS grows both for the sequences with 614G and 614D if we ignore sampling information. If we account for preferential sampling, the EPS of sequences with 614G "dips" because no sequences in the last week were part of our dataset.

This application offers a cautionary tale on this class of models. Modeling the sampling process not only reduces the credible region width, it can also affect the estimates heavily. This behavior signals that ignoring the sampling process leads to a bias if the preferential model is correctly specified. The opposite is also true: we could be biasing our estimates if the sampling model is incorrect by including sampling times information.

7. DISCUSSION

Statistical methods in molecular epidemiology offer powerful tools to help us understand and monitor a pandemic as it unfolds. Our paper has outlined some of the statistical models used for tracking SARS-CoV-2 and identified a few areas where state-of-the-art phylodynamic approaches fell short of delivering their full potential.

The lack of scalable inference methods that can analyze the unprecedented amount of molecular sequences available is a common theme among all SARS-CoV-2 analyses discussed here. Popular strategies include subsampling, inferring a fixed phylogeny and using a fixed phylogeny for partitioning the data. It is generally missing how stable the results are to these choices. We chose to evaluate the stability of phylogenetic posterior from subsampling.

Rajanala and Palacios (2021) proposed a visual inspection of several phylogenetic posteriors obtained from different samples, to investigate phylogenetic stability. If the distributions overlap, then there is indication of phylogenetic stability. We followed the proposed methodology to investigate the stability of the phylogenetic posteriors of SARS-CoV-2 obtained from Canada, Sweden and the UK. We took three samples, each containing 100 sequences chosen at random from each country available in GISAID, from November 1, 2020 to February 1, 2021. The nine posterior distributions are projected in two dimensions and depicted in the MDS plot of Figure 7. Here, Sweden is the only country that shows phylogenetic stability. We recommend performing stability analyses when choosing a small subset of samples.

We centered our scalability discussion on Bayesian algorithms that either aim to replace MCMC by sequential Monte Carlo or variational inference, or to potentially improve the convergence of MCMC. We focused on approaches that have already found an application in phylodynamics (or closely related fields like phylogenetics). Although in principle these methods can be extended to infer phylodynamic parameters such as EPS, it is not clear how much efficiency can be gained with these methods in comparison to current implementations that rely on Metropolis-Hastings steps. Bayesian computational statistics for large data sets is a very active area of research. New approaches, such as nonreversible MCMC schemes (Bierkens, 2016, Bouchard-Côté, Vollmer and Doucet, 2018b), are appealing, but to our knowledge,

have not yet found an application in this field. Other approaches include the online strategy, which updates of the posterior distribution as sequences become available sequentially, as well as the divide-and-conquer strategy, which divides the data into smaller subsets. We anticipate several advancements in this area in the future.

Apart from Bayesian computation, an important direction for improving scalability includes more efficient phylogenetic modeling. As the number of samples increases, the probability of observing sequences with identical genotypes also increases. Phylogenies with permuted labels of samples with identical genotypes have equal likelihood. States of the CP can then be lumped together in a situation like this. These lower resolutions of the coalescent have smaller cardinality and can potentially be more efficient (Sainudiin, Stadler and Véber, 2015, Palacios et al., 2019, Cappello, Veber and Palacios, 2020).

Another common theme has been a trade-off between interpretability of model parameters and model complexity. In order to make the CP and the BDSP amenable to infer relevant quantities such as prevelance, one needs to both impose more modeling assumptions and incorporate more data. Recall that under complex epidemiological models, the model becomes unidentifiable unless we pre-specify some of the parameter values or incorporate independent sources of information. However, incorporating other sources of information and their corresponding sampling models can also create biases when the models are misspecified. We envision future research that incorporates robust models against model misspecification such as the adaptive preferential sampling. As it is common in many other areas of science, the strive for a balance between realism on one side, and simplicity and interpretability on the other, is going to be an essential focus of future work.

Our discussion has omitted other phylodynamic models that have been used to track the evolution of the pandemic such as phylogeography models (Lemey et al., 2020), structured coalescent models (Müller, Rasmussen and Stadler, 2017), coalescent with recombination (Müller, Kistler and Bedford, 2022) and models of within-host-evolution (Jones et al., 2018). We have also omitted model selection from our discussion and refer the reader to Lewis et al. (2014). We do not discuss other data-quality associated statistical challenges such as sequencing errors (Turakhia et al., 2020, Morel et al., 2021) and underreporting of case count data (Wu et al., 2020).

APPENDIX A: DATA ANALYSES

A.1 Phylodynamic Analysis in California

Case counts for Figure 1 panel (A) were obtained from the New York Times repository (https://github.com/nytimes/covid-19-data). Molecular sequences for Figure 1 panel (B) were obtained from the GISAID repository (accession codes of the sequences used can be retrieved at https://github.com/JuliaPalacios/phylodyn/blob/master/data/California_statscience_ack.txt).

Given the molecular sequences, a viral phylogeny was obtained from a genetic distance-based method called serial UPGMA (Drummond and Rodrigo, 2000). Conditionally on the viral phylogeny, the EPS posterior was inferred with a Bayesian nonparametric method described in Palacios and Minin (2012). The posterior approximation is based on INLA.

The posterior EPS (Figure 5) was also inferred with Tang et al. (2019) by fitting the stochastic CP–EPI model assuming SIR dynamics to the fixed genealogy. The model assumed a fixed known total population size $N_{\text{pop}} = S(t) + I(t) + R(t)$, a constant removal rate γ, and a time-varying infection rate $\beta(t)$, which is then reparametrized with a time-varying basic reproduction number $R_0(t) = [\beta(t)N_{\text{pop}}]/\gamma$. We used $N_{\text{pop}} = 39.5 \times 10^6$ based on the California census population size in 2020 and a log-normal prior with parameters $(3.6, 0.2)$ in years based on the recovery period of 7-14 days. We ran the MCMC algorithm of Tang et al. (2019) for 100,000 iterations, with 10% of burn-in, and thinned to obtain a total of 1,000 samples.

A.2 Analysis of SARS-CoV-2 Sequences from Washington State

The two sets of 100 molecular sequences were analyzed independently with BEAST2 with the same model and MCMC parameters. We ran the chains for 20×10^6 iterations, thinning every 1000 and with a burnin of 10×10^6 iterations. We placed the Extended Bayesian Skyline prior on $N_e(t)$ (Heled and Drummond, 2008), the HKY mutation model with empirically estimated base frequencies (Hasegawa, Kishino and Yano, 1985), and the mutation rate fixed to 9×10^{-4} substitutions per site per year. The two phylogenies obtained are the maximum clade credibility trees of the posterior distributions obtained with TreeAnnotator (Bouckaert et al., 2019). Accession codes of the sequences can be retrieved at https://github.com/JuliaPalacios/phylodyn/blob/master/data/Washington_statscience_ack.txt. Details on the analyses done are included in the main text.

A.3 MDS Analysis

Molecular sequences used to generate Figure 7 were obtained from GISAID. We analyzed 9 samples of 100 sequences (3 samples of 100 sequences per country) independently with BEAST2 (Bouckaert et al., 2019) with the same model and MCMC parameters. We ran the chains for 50×10^6 iterations. We placed the Bayesian Skyline prior on $N_e(t)$, the HKY mutation model, and the mutation rate fixed to 9×10^{-4} substitutions per site per

year. Posterior samples were thinned to 100 samples. Pairwise distances were obtained using Kim, Rosenberg and Palacios (2020). Accession codes of the sequences can be retrieved at https://github.com/JuliaPalacios/phylodyn/blob/master/data/CanSweUk_statscience_ack.txt

APPENDIX B: GLOSSARY OF TERMS

Anthropogenic selection. A process where human-induced environmental changes, such as use of antiviral drugs, alter the direction and magnitude of selection.

Antigenic evolution. An evolutionary response of pathogens to host's antibody-mediated immunity selective pressures.

Consensus sequence. A consensus individual's sequence consists of those nucleotides with highest frequency at each position in an alignment (aligned to a reference genome) of multiple reads (Grubaugh et al., 2019); usually only those nucleotides with high frequency and high coverage (multiple reads per nucleotide) are used in the analyses.

Endemic equilibrium. A state at which the disease dynamics is in a steady state so the disease persists in the population.

Locus. The physical location of a specific gene on a chromosome. Here we assume there is no recombination within the locus.

Mutation. An alteration in a genetic sequence such as substitution, insertions, deletions, etc.

Neutral evolution. A theory that postulates that most evolutionary changes at the molecular level do not affect reproductive success (fitness), and can be described by random genetic drift of mutations that are selectively neutral.

Recombination. The exchange of genetic material between parental genomes by the breakage and rejoining of chromosomes, producing offspring genomes that carry genetic information distinct from its parental genomes.

Selection. A nonrandom difference in reproduction among individuals, often due to differential survival to specific environments, ensuring the transmission of beneficial traits to succeeding generations.

Substitutions. A type of mutation where a single nucleotide ("chemical letter") is replaced with a different nucleotide. There are two types of substitutions: transition and transversion.

Transition. A transition is a type of substitution mutations that occurs within each structural class of DNA: a purine nucleotide is substituted with another purine ($A \leftrightarrow G$) or a pyrimidine nucleotide is substituted with another pyrimidine ($C \leftrightarrow T$).

Transversion. A transversion is a type of substitution mutations that occurs across different structural classes of DNA: a purine nucleotide is substituted with a pyrimidine nucleotide or vice versa ($A \leftrightarrow C$, $A \leftrightarrow T$, $G \leftrightarrow C$, $G \leftrightarrow T$).

Wright–Fisher model. The model describes the sampling of alleles in a population with no selection, no mutation, no migration, nonoverlapping generation times and random mating. It is a Markov chain which samples with replacement a new generation from the previous one.

ACKNOWLEDGMENTS

We thank the Editor, the guest editors, and the anonymous reviewer for constructive comments. Computing resources were provided by Stanford University research computing center (Sherlock cluster).

FUNDING

J.A.P. acknowledges support from National Institutes of Health Grant R01-GM-131404.

SUPPLEMENTARY MATERIAL

Supplement to "Statistical challenges in tracking the evolution of SARS-CoV-2" (DOI: 10.1214/22-STS853SUPP; .pdf). Supplementary information.

REFERENCES

ALTEKAR, G., DWARKADAS, S., HUELSENBECK, J. P. and RONQUIST, F. (2004). Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* **20** 407–415.

AL KHATIB, H. A., BENSLIMANE, F. M., ELBASHIR, I. E., COYLE, P. V., AL MASLAMANI, M. A., AL-KHAL, A., AL THANI, A. A. and YASSINE, H. M. (2020). Within-host diversity of Sars-CoV-2 in Covid-19 patients with variable disease severities. *Front. Cell. Infect. Microbiol.* **10** 534. https://doi.org/10.3389/fcimb.2020.575613

ANDERSEN, K. G., RAMBAUT, A., LIPKIN, W. I., HOLMES, E. C. and GARRY, R. F. (2020). The proximal origin of Sars-CoV-2. *Nat. Med.* **26** 450–452. https://doi.org/10.1038/s41591-020-0820-9

ANDRÉOLETTI, J., ZWAANS, A., WARNOCK, R. C. M., AGUIRRE-FERNÁNDEZ, G., BARIDO-SOTTANI, J., GUPTA, A., STADLER, T. and MANCEAU, M. (2020). A skyline birth-death process for inferring the population size from a reconstructed tree with occurrences. *BioRxiv* 2020.10.27.356758. https://doi.org/10.1101/2020.10.27.356758

AYRES, D. L., DARLING, A., ZWICKL, D. J., BEERLI, P., HOLDER, M. T., LEWIS, P. O., HUELSENBECK, J. P., RONQUIST, F., SWOFFORD, D. L. et al. (2012). BEAGLE: An application programming interface and high-performance computing library for statistical phylogenetics. *Syst. Biol.* **61** 170–173.

BARIDO-SOTTANI, J., VAUGHAN, T. G. and STADLER, T. (2020). A multitype birth–death model for Bayesian inference of lineage-specific birth and death rates. *Syst. Biol.* **69** 973–986. https://doi.org/10.1093/sysbio/syaa016

BEAUMONT, M. A., ZHANG, W. and BALDING, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162** 2025–2035. https://doi.org/10.1093/genetics/162.4.2025

BEHR, M., ANSARI, M. A., MUNK, A. and HOLMES, C. (2020). Testing for dependence on tree structures. *Proc. Natl. Acad. Sci. USA* **117** 9787–9792. MR4236178 https://doi.org/10.1073/pnas.1912957117

BERESTYCKI, N. (2009). *Recent Progress in Coalescent Theory*. *Ensaios Matemáticos* [*Mathematical Surveys*] **16**. Sociedade Brasileira de Matemática, Rio de Janeiro. MR2574323

BIERKENS, J. (2016). Non-reversible Metropolis-Hastings. *Stat. Comput.* **26** 1213–1228. MR3538633 https://doi.org/10.1007/s11222-015-9598-x

BILLERA, L. J., HOLMES, S. P. and VOGTMANN, K. (2001). Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.* **27** 733–767. MR1867931 https://doi.org/10.1006/aama.2001.0759

BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. MR3671776 https://doi.org/10.1080/01621459.2017.1285773

BLUM, M. G. B. and FRANÇOIS, O. (2010). Non-linear regression models for approximate Bayesian computation. *Stat. Comput.* **20** 63–73. MR2578077 https://doi.org/10.1007/s11222-009-9116-0

BONI, M. F., LEMEY, P., JIANG, X., LAM, T. T.-Y., PERRY, B. W., CASTOE, T. A., RAMBAUT, A. and ROBERTSON, D. L. (2020). Evolutionary origins of the Sars-CoV-2 sarbecovirus lineage responsible for the Covid-19 pandemic. *Nat. Microbiol.* **5** 1408–1417.

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** 1–18. https://doi.org/10.1371/journal.pcbi.1003913

BOUCHARD-CÔTÉ, A., SANKARARAMAN, S. and JORDAN, M. I. (2012). Phylogenetic inference via sequential Monte Carlo. *Syst. Biol.* **61** 579–593.

BOUCHARD-CÔTÉ, A., VOLLMER, S. J. and DOUCET, A. (2018a). The bouncy particle sampler: A nonreversible rejection-free Markov chain Monte Carlo method. *J. Amer. Statist. Assoc.* **113** 855–867. MR3832232 https://doi.org/10.1080/01621459.2017.1294075

BOUCHARD-CÔTÉ, A., VOLLMER, S. J. and DOUCET, A. (2018b). The bouncy particle sampler: A nonreversible rejection-free Markov chain Monte Carlo method. *J. Amer. Statist. Assoc.* **113** 855–867. MR3832232 https://doi.org/10.1080/01621459.2017.1294075

BOUCKAERT, R., VAUGHAN, T. G., BARIDO-SOTTANI, J., DUCHÊNE, S., FOURMENT, M., GAVRYUSHKINA, A., HELED, J., JONES, G., KÜHNERT, D. et al. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **15** e1006650.

BRITTON, T. (2010). Stochastic epidemic models: A survey. *Math. Biosci.* **225** 24–35. MR2642269 https://doi.org/10.1016/j.mbs.2010.01.006

CAPPELLO, L. and PALACIOS, J. A. (2021). Adaptive preferential sampling in phylodynamics with an application to Sars-CoV-2. *J. Comput. Graph. Statist.* 1–12. https://doi.org/10.1080/10618600.2021.1987256

CAPPELLO, L., VEBER, A. and PALACIOS, J. A. (2020). The Tajima heterochronous n-coalescent: Inference from heterochronously sampled molecular data. Preprint. Available at arXiv:2004.06826.

CAPPELLO, L., KIM, J., LIU, S. and PALACIOS, J. A. (2022). Supplement to "Statistical Challenges in Tracking the Evolution of SARS-CoV-2." https://doi.org/10.1214/22-STS853SUPP

CHOI, S. C. (2020). A phylodynamic analysis of epidemiological situation of East Asia due to the coronavirus disease of 2019. *The Microbiological Society of Korea* **56** 241–253.

CHOPIN, N. and PAPASPILIOPOULOS, O. (2020). *An Introduction to Sequential Monte Carlo*. *Springer Series in Statistics*. Springer, Cham. MR4215639 https://doi.org/10.1007/978-3-030-47845-2

COLIJN, C. and PLAZZOTTA, G. (2017). A metric on phylogenetic tree shapes. *Syst. Biol.* **67** 113–126. https://doi.org/10.1093/sysbio/syx046

CYBIS, G. B., SINSHEIMER, J. S., BEDFORD, T., MATHER, A. E., LEMEY, P. and SUCHARD, M. A. (2015). Assessing phenotypic correlation through the multivariate phylogenetic latent liability model. *Ann. Appl. Stat.* **9** 969–991. MR3371344 https://doi.org/10.1214/15-AOAS821

DAVIES, N. G., ABBOTT, S., BARNARD, R. C., JARVIS, C. I., KUCHARSKI, A. J., MUNDAY, J. D., PEARSON, C. A., RUSSELL, T. W., TULLY, D. C. et al. (2021). Estimated transmissibility and impact of Sars-CoV-2 lineage B. 1.1. 7 in England. *Science* **372**.

DEARLOVE, B. and WILSON, D. J. (2013). Coalescent inference for infectious disease: Meta-analysis of hepatitis C. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **368** 20120314. https://doi.org/10.1098/rstb.2012.0314

DELLICOUR, S., DURKIN, K., HONG, S. L., VANMECHELEN, B., MARTÍ-CARRERAS, J., GILL, M. S., MEEX, C., BONTEMS, S., ANDRÉ, E. et al. (2021). A phylodynamic workflow to rapidly gain insights into the dispersal history and dynamics of Sars-CoV-2 lineages. *Mol. Biol. Evol.* **38** 1608–1613.

DENG, X., GU, W., FEDERMAN, S., DU PLESSIS, L., PYBUS, O. G., FARIA, N. R., WANG, C., YU, G., BUSHNELL, B. et al. (2020). Genomic surveillance reveals multiple introductions of Sars-CoV-2 into northern California. *Science* **369** 582–587.

DIGGLE, P. J., MENEZES, R. and SU, T. (2010). Geostatistical inference under preferential sampling. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **59** 191–232. MR2744471 https://doi.org/10.1111/j.1467-9876.2009.00701.x

DINH, V., DARLING, A. E. and IV, F. A. M. (2018). Online Bayesian phylogenetic inference: Theoretical foundations via sequential Monte Carlo. *Syst. Biol.* **67** 503–517. https://doi.org/10.1093/sysbio/syx087

DRUMMOND, A. and RODRIGO, A. G. (2000). Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA. *Mol. Biol. Evol.* **17** 1807–1815.

DU PLESSIS, L., MCCRONE, J. T., ZAREBSKI, A. E., HILL, V., RUIS, C., GUTIERREZ, B., RAGHWANI, J., ASHWORTH, J., COLQUHOUN, R. et al. (2021). Establishment and lineage dynamics of the Sars-CoV-2 epidemic in the UK. *Science* **371** 708–712.

EFRON, B., HALLORAN, E. and HOLMES, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. USA* **93** 7085–7090.

FAULKNER, J. R., MAGEE, A. F., SHAPIRO, B. and MININ, V. N. (2020). Horseshoe-based Bayesian nonparametric estimation of effective population size trajectories. *Biometrics* (in press).

FEATHERSTONE, L. A., DI GIALLONARDO, F., HOLMES, E. C., VAUGHAN, T. G. and DUCHÊNE, S. (2021). Infectious disease phylodynamics with occurrence data. *Methods Ecol. Evol.* **12** 1498–1507. https://doi.org/10.1111/2041-210X.13620

FELSENSTEIN, J. (1985). Phylogenies and the comparative method. *Amer. Nat.* **125** 1–15.

FELSENSTEIN, J. (2004). *Inferring Phylogenies* **2**. Sinauer associates Sunderland, MA.

FELSENSTEIN, J. (2005). Using the quantitative genetic threshold model for inferences between and within species. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **360** 1427–1434.

FELSENSTEIN, J. (2012). A comparative method for both discrete and continuous characters using the threshold model. *Amer. Nat.* **179** 145–156.

FELSENSTEIN, J. and RODRIGO, A. G. (1999). Coalescent approaches to HIV population genetics. In *The Evolution of HIV* 233–272. Johns Hopkins Univ. Press, Baltimore.

FITCH, W. M. (1971). Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Biol.* **20** 406–416.

FOURMENT, M., CLAYWELL, B. C., DINH, V., MCCOY, C., IV, F. A. M. and DARLING, A. E. (2018). Effective online Bayesian phylogenetics via sequential Monte Carlo with guided proposals. *Syst. Biol.* **67** 490–502. https://doi.org/10.1093/sysbio/syx090

FRICK, K., MUNK, A. and SIELING, H. (2014). Multiscale change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 495–580. MR3210728 https://doi.org/10.1111/rssb.12047

FROST, S. D. W., PYBUS, O. G., GOG, J. R., VIBOUD, C., BONHOEFFER, S. and BEDFORD, T. (2015). Eight challenges in phylodynamic inference. *Epidemics* **10** 88–92. https://doi.org/10.1016/j.epidem.2014.09.001

GEIDELBERG, L., BOYD, O., JORGENSEN, D., SIVERONI, I., NASCIMENTO, F. F., JOHNSON, R., RAGONNET-CRONIN, M., FU, H., WANG, H. et al. (2021). Genomic epidemiology of a densely sampled Covid-19 outbreak in China. *Virus Evolution* **7**. veaa102. https://doi.org/10.1093/ve/veaa102

GERNHARD, T. (2008). The conditioned reconstructed process. *J. Theoret. Biol.* **253** 769–778. MR2964590 https://doi.org/10.1016/j.jtbi.2008.04.005

GILL, M. S., LEMEY, P., SUCHARD, M. A., RAMBAUT, A. and BAELE, G. (2020). Online Bayesian phylodynamic inference in BEAST with application to epidemic reconstruction. *Mol. Biol. Evol.* **37** 1832–1842. https://doi.org/10.1093/molbev/msaa047

GRAFEN, A. (1989). The phylogenetic regression. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **326** 119–157.

GRENFELL, B. T., PYBUS, O. G., GOG, J. R., WOOD, J. L. N., DALY, J. M., MUMFORD, J. A. and HOLMES, E. C. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303** 327–332.

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

GRUBAUGH, N. D., GANGAVARAPU, K., QUICK, J., MATTESON, N. L., DE JESUS, J. G., MAIN, B. J., TAN, A. L., PAUL, L. M., BRACKNEY, D. E. et al. (2019). An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.* **20** 1–19.

GUPTA, A., MANCEAU, M., VAUGHAN, T., KHAMMASH, M. and STADLER, T. (2020). The probability distribution of the reconstructed phylogenetic tree with occurrence data. *J. Theoret. Biol.* **488** 110115, 10. MR4051870 https://doi.org/10.1016/j.jtbi.2019.110115

HADFIELD, J., MEGILL, C., BELL, S. M., HUDDLESTON, J., POTTER, B., CALLENDER, C., SAGULENKO, P., BEDFORD, T. and NEHER, R. A. (2018). Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics* **34** 4121–4123. https://doi.org/10.1093/bioinformatics/bty407

HASEGAWA, M., KISHINO, H. and YANO, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **2** 160–164.

HELED, J. and DRUMMOND, A. J. (2008). Bayesian inference of population size history from multiple loci. *BMC Evol. Biol.* **8** 289. https://doi.org/10.1186/1471-2148-8-289

HODCROFT, E. B., ZUBER, M. et al. (2021). Spread of a Sars-CoV-2 variant through Europe in the summer of 2020. *Nature* **595** 707–712. https://doi.org/10.1038/s41586-021-03677-y

HOFFMAN, M. D., BLEI, D. M., WANG, C. and PAISLEY, J. (2013). Stochastic variational inference. *J. Mach. Learn. Res.* **14** 1303–1347. MR3081926

HUANG, Z. and GELMAN, A. (2005). Sampling for Bayesian computation with large datasets. Available at SSRN 1010107.

HUELSENBECK, J. P. and RANNALA, B. (2003). Detecting correlation between characters in a comparative analysis with uncertain phylogeny. *Evolution* **57** 1237–1247.

JONES, B. R., KINLOCH, N. N., HORACSEK, J., GANASE, B., HARRIS, M., HARRIGAN, P. R., JONES, R. B., BROCKMAN, M. A., JOY, J. B. et al. (2018). Phylogenetic approach to recover integration dates of latent HIV sequences within-host. *Proc. Natl. Acad. Sci. USA* **115** E8958–E8967. https://doi.org/10.1073/pnas.1802028115

JORDAN, M. I., GHAHRAMANI, Z., JAAKKOLA, T. S. and SAUL, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.* **37** 183–233.

KARCHER, M. D., PALACIOS, J. A., BEDFORD, T., SUCHARD, M. A. and MININ, V. N. (2016). Quantifying and mitigating the effect of preferential sampling on phylodynamic inference. *PLoS Comput. Biol.* **12** e1004789.

KARCHER, M. D., SUCHARD, M. A., DUDAS, G. and MININ, V. N. (2020). Estimating effective population size changes from preferentially sampled genetic sequences. *PLoS Comput. Biol.* **in press**.

KERMACK, W. O., MCKENDRICK, A. G. and WALKER, G. T. (1927). A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. Ser. A, Contain. Pap. Math. Phys. Character* **115** 700–721. https://doi.org/10.1098/rspa.1927.0118

KIM, J., ROSENBERG, N. A. and PALACIOS, J. A. (2020). Distance metrics for ranked evolutionary trees. *Proc. Natl. Acad. Sci. USA* **117** 28876–28886. https://doi.org/10.1073/pnas.1922851117

KINGMAN, J. F. C. (1982a). The coalescent. *Stochastic Process. Appl.* **13** 235–248. MR0671034 https://doi.org/10.1016/0304-4149(82)90011-4

KINGMAN, J. F. C. (1982b). On the genealogy of large populations. *J. Appl. Probab.* **19A** 27–43. MR0633178

KOELLE, K. and RASMUSSEN, D. A. (2012). Rates of coalescence for common epidemiological models at equilibrium. *J. R. Soc. Interface* **9** 997–1007. https://doi.org/10.1098/rsif.2011.0495

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** 2102–2116. https://doi.org/10.1093/molbev/msw064

LAN, S., PALACIOS, J. A., KARCHER, M., MININ, V. N. and SHAHBABA, B. (2015). An efficient Bayesian inference framework for coalescent-based nonparametric phylodynamics. *Bioinformatics* **31** 3282–3289.

LEMEY, P., HONG, S. L., HILL, V., BAELE, G., POLETTO, C., COLIZZA, V., O'TOOLE, Á., MCCRONE, J. T., ANDERSEN, K. G. et al. (2020). Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of Sars-CoV-2. *Nat. Commun.* **11** 1–14.

LEMEY, P., RUKTANONCHAI, N., HONG, S. L., COLIZZA, V., POLETTO, C., DEN BROECK, F. V., GILL, M. S., JI, X., LEVASSEUR, A. et al. (2021). Untangling introductions and persistence in Covid-19 resurgence in Europe. *Nature* **595** 713–717. https://doi.org/10.1038/s41586-021-03754-2

LEUNG, K., SHUM, M. H., LEUNG, G. M., LAM, T. T. and WU, J. T. (2021). Early transmissibility assessment of the N501Y mutant strains of Sars-CoV-2 in the United Kingdom, October to November 2020. *Euro Surveill.* **26** 2002106.

LEWIS, P. O., XIE, W., CHEN, M.-H., FAN, Y. and KUO, L. (2014). Posterior predictive Bayesian phylogenetic model selection. *Syst. Biol.* **63** 309–321.

LI, L. M., GRASSLY, N. C. and FRASER, C. (2017). Quantifying transmission heterogeneity using both pathogen phylogenies and incidence time series. *Mol. Biol. Evol.* **34** 2982–2995. https://doi.org/10.1093/molbev/msx195

LINTUSAARI, J., GUTMANN, M. U., DUTTA, R., KASKI, S. and CORANDER, J. (2016). Fundamentals and recent developments in approximate Bayesian computation. *Syst. Biol.* **66** e66–e82. https://doi.org/10.1093/sysbio/syw077

LOPEZ BERNAL, J., ANDREWS, N., GOWER, C., GALLAGHER, E., SIMMONS, R., THELWALL, S., STOWE, J., TESSIER, E., GROVES, N. et al. (2021). Effectiveness of Covid-19 vaccines against the B.1.617.2 (delta) variant. *N. Engl. J. Med.* **385** 585–594. https://doi.org/10.1056/NEJMoa2108891

LOUCA, S., MCLAUGHLIN, A., MACPHERSON, A., JOY, J. B. and PENNELL, M. W. (2021). Fundamental identifiability limits in molecular epidemiology. *Mol. Biol. Evol.* **38** 4010–4024. https://doi.org/10.1093/molbev/msab149

MACCANNELL, T., BATSON, J., BONIN, B., KC, A., QUENELLE, R., STRONG, B., LIN, W., RUDMAN, S. L., DYNERMAN, D. et al. (2021). Genomic epidemiology and transmission dynamics of Sars-CoV-2 in congregate healthcare facilities in Santa Clara County, California. *Clin. Infect. Dis.*.

MACPHERSON, A., LOUCA, S., MCLAUGHLIN, A., JOY, J. B. and PENNELL, M. W. (2021). Unifying phylogenetic birth-death models in epidemiology and macroevolution. *Syst. Biol.*. syab049. https://doi.org/10.1093/sysbio/syab049

MANCEAU, M., GUPTA, A., VAUGHAN, T. and STADLER, T. (2021). The probability distribution of the ancestral population size conditioned on the reconstructed phylogenetic tree with occurrence data. *J. Theoret. Biol.* **509** Paper No. 110400, 18. MR4171763 https://doi.org/10.1016/j.jtbi.2020.110400

MARJORAM, P. and TAVARÉ, S. (2006). Modern computational approaches for analysing molecular genetic variation data. *Nat. Rev. Genet.* **7** 759–770.

MARJORAM, P., MOLITOR, J., PLAGNOL, V. and TAVARÉ, S. (2003). Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **100** 15324–15328. https://doi.org/10.1073/pnas.0306899100

MAURANO, M. T., RAMASWAMI, S., ZAPPILE, P., DIMARTINO, D., BOYTARD, L., RIBEIRO-DOS SANTOS, A. M., VULPESCU, N. A., WESTBY, G., SHEN, G. et al. (2020). Sequencing identifies multiple early introductions of Sars-CoV-2 to the New York City region. *Genome Res.* **30** 1781–1788.

MILLER, D., MARTIN, M. A., HAREL, N., TIROSH, O., KUSTIN, T., MEIR, M., SOREK, N., GEFEN-HALEVI, S., AMIT, S. et al. (2020). Full genome viral sequences inform patterns of Sars-CoV-2 spread into and within Israel. *Nat. Commun.* **11** 5518. https://doi.org/10.1038/s41467-020-19248-0

MINH, B. Q., SCHMIDT, H. A., CHERNOMOR, O., SCHREMPF, D., WOODHAMS, M. D., VON HAESELER, A. and LANFEAR, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37** 1530–1534.

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

MOREL, B., BARBERA, P., CZECH, L., BETTISWORTH, B., HÜBNER, L., LUTTEROPP, S., SERDARI, D., KOSTAKI, E.-G., MAMAIS, I. et al. (2021). Phylogenetic analysis of Sars-CoV-2 data is difficult. *Mol. Biol. Evol.* **38** 1777–1791.

MORENO, G. K., BRAUN, K. M., RIEMERSMA, K. K., MARTIN, M. A., HALFMANN, P. J., CROOKS, C. M., PRALL, T., BAKER, D., BACZENAS, J. J. et al. (2020). Revealing fine-scale spatiotemporal differences in Sars-CoV-2 introduction and spread. *Nat. Commun.* **11** 5558. https://doi.org/10.1038/s41467-020-19346-z

MÜLLER, N. F. and BOUCKAERT, R. R. (2020). Adaptive Metropolis-coupled MCMC for BEAST 2. *PeerJ* **8** e9473. https://doi.org/10.7717/peerj.9473

MÜLLER, N. F., KISTLER, K. E. and BEDFORD, T. (2022). Recombination patterns in coronaviruses. *BioRxiv*. https://doi.org/10.1101/2021.04.28.441806

MÜLLER, N. F., RASMUSSEN, D. A. and STADLER, T. (2017). The structured coalescent and its approximations. *Mol. Biol. Evol.* **34** 2970–2981. https://doi.org/10.1093/molbev/msx186

MÜLLER, N. F., WAGNER, C., FRAZAR, C. D., ROYCHOUDHURY, P., LEE, J., MONCLA, L. H., PELLE, B., RICHARDSON, M., RYKE, E. et al. (2021). Viral genomes reveal patterns of the Sars-CoV-2 outbreak in Washington state. *Sci. Transl. Med.* **13**.

NADEAU, S. A., VAUGHAN, T. G., SCIRE, J., HUISMAN, J. S. and STADLER, T. (2021). The origin and early spread of Sars-CoV-2 in Europe. *Proc. Natl. Acad. Sci. USA* **118**. https://doi.org/10.1073/pnas.2012008118

NEAL, R. M. (2001). Annealed importance sampling. *Stat. Comput.* **11** 125–139. MR1837132 https://doi.org/10.1023/A:1008923215028

NEE, S., MAY, R. M. and HARVEY, P. H. (1994). The reconstructed evolutionary process. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **344** 305–311.

NEISWANGER, W., WANG, C. and XING, E. (2013). Asymptotically exact, embarrassingly parallel MCMC. Preprint. Available at arXiv:1311.4780.

PAGEL, M. (1994). Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond., B Biol. Sci.* **255** 37–45.

PALACIOS, J. A. and MININ, V. N. (2012). Integrated nested Laplace approximation for Bayesian nonparametric phylodynamics. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence. UAI'*12 726–735. AUAI Press, Arlington, VA, United States.

PALACIOS, J. A. and MININ, V. N. (2013). Gaussian process-based Bayesian nonparametric inference of population size trajectories from gene genealogies. *Biometrics* **69** 8–18. MR3058047 https://doi.org/10.1111/biom.12003

PALACIOS, J. A., GILL, M. S., SUCHARD, M. A. and MININ, V. N. (2014). Bayesian nonparametric phylodynamics.

PALACIOS, J. A., VÉBER, A., CAPPELLO, L., WANG, Z., WAKELEY, J. and RAMACHANDRAN, S. (2019). Bayesian estimation of population size changes by sampling Tajima's trees. *Genetics* **213** 967–986.

PARAG, K. V., DU PLESSIS, L. and PYBUS, O. G. (2020). Jointly inferring the dynamics of population size and sampling intensity from molecular sequences. *Mol. Biol. Evol.* **37** 2414–2429.

PARIKH, V. N., IOANNIDIS, A., JIMENEZ-MORALES, D. et al. (2021). Multi-omic surveillance disambiguates social and biological determinants of COVID19 severity. *in Preparation*.

PARKER, J., RAMBAUT, A. and PYBUS, O. G. (2008). Correlating viral phenotypes with phylogeny: Accounting for phylogenetic uncertainty. *Infect. Genet. Evol.* **8** 239–246.

POON, A. F. Y. (2015). Phylodynamic inference with kernel ABC and its application to HIV epidemiology. *Mol. Biol. Evol.* **32** 2483–2495. https://doi.org/10.1093/molbev/msv123

POPINGA, A., VAUGHAN, T., STADLER, T. and DRUMMOND, A. J. (2015). Inferring epidemiological dynamics with Bayesian coalescent inference: The merits of deterministic and stochastic models. *Genetics* **199** 595–607. https://doi.org/10.1534/genetics.114.172791

RAGONNET-CRONIN, M., BOYD, O., GEIDELBERG, L., JORGENSEN, D., NASCIMENTO, F. F., SIVERONI, I., JOHNSON, R. A., BAGUELIN, M., CUCUNUBÁ, Z. M. et al. (2021). Genetic evidence for the association between Covid-19 epidemic severity and timing of non-pharmaceutical interventions. *Nat. Commun.* **12** 2188. https://doi.org/10.1038/s41467-021-22366-y

RAJANALA, S. and PALACIOS, J. A. (2021). Statistical summaries of unlabelled evolutionary trees and ranked hierarchical clustering trees. Preprint. Available at arXiv:2106.02724.

RASMUSSEN, D. A., RATMANN, O. and KOELLE, K. (2011). Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Comput. Biol.* **7** e1002136, 11. MR2845064 https://doi.org/10.1371/journal.pcbi.1002136

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 Comput. Biol.* **8** 1–14. https://doi.org/10.1371/journal.pcbi.1002835

ROBINSON, D. F. and FOULDS, L. R. (1981). Comparison of phylogenetic trees. *Math. Biosci.* **53** 131–147. MR0613619 https://doi.org/10.1016/0025-5564(81)90043-2

RODRIGO, A. G. and FELSENSTEIN, J. (1999). Coalescent approaches to HIV-1 population genetics. In *Molecular Evolution of HIV* (K. A. Crandell, ed.) Johns Hopkins Univ. Press, Baltimore, MD.

ROSENBERG, N. A. and NORDBORG, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat. Rev. Genet.* **3** 380–390.

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 319–392. MR2649602 https://doi.org/10.1111/j.1467-9868.2008.00700.x

SAGULENKO, P., PULLER, V. and NEHER, R. A. (2018). TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4** vex042. https://doi.org/10.1093/ve/vex042

SAINUDIIN, R., STADLER, T. and VÉBER, A. (2015). Finding the best resolution for the Kingman-Tajima coalescent: Theory and applications. *J. Math. Biol.* **70** 1207–1247. MR3323594 https://doi.org/10.1007/s00285-014-0796-5

SAN, J. E., NGCAPU, S., KANZI, A. M., TEGALLY, H., FONSECA, V., GIANDHARI, J., WILKINSON, E., NELSON, C. W., SMIDT, W. et al. (2021). Transmission dynamics of Sars-CoV-2 within-host diversity in two major hospital outbreaks in South Africa. *Virus Evolution* **7**. veab041. https://doi.org/10.1093/ve/veab041

SAULNIER, E., GASCUEL, O. and ALIZON, S. (2017). Inferring epidemiological parameters from phylogenies using regression-ABC: A comparative study. *PLoS Comput. Biol.* **13** 1–31. https://doi.org/10.1371/journal.pcbi.1005416

SCIRE, J., BARIDO-SOTTANI, J., KÜHNERT, D., VAUGHAN, T. G. and STADLER, T. (2020). Improved multi-type birth-death phylodynamic inference in BEAST 2. *BioRxiv*.

SEEMANN, T., LANE, C. R., SHERRY, N. L., DUCHENE, S., GONÇALVES DA SILVA, A., CALY, L., SAIT, M., BALLARD, S. A., HORAN, K. et al. (2020). Tracking the Covid-19 pandemic in Australia using genomics. *Nat. Commun.* **11** 4376. https://doi.org/10.1038/s41467-020-18314-x

SHU, Y. and MCCAULEY, J. (2017). GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22**. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494

SIMPER, M. and PALACIOS, J. A. (2020). An adjacent-swap Markov chain on coalescent trees. Preprint. Available at arXiv:2012.08030.

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

SRIVASTAVA, S., CEVHER, V., DINH, Q. and DUNSON, D. (2015). WASP: Scalable Bayes via barycenters of subset posteriors. In *Artificial Intelligence and Statistics* 912–920. PMLR.

STADLER, T. (2009). On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *J. Theoret. Biol.* **261** 58–66. MR2980272 https://doi.org/10.1016/j.jtbi.2009.07.018

STADLER, T. (2010). Sampling-through-time in birth-death trees. *J. Theoret. Biol.* **267** 396–404. MR2974417 https://doi.org/10.1016/j.jtbi.2010.09.010

STADLER, T. and BONHOEFFER, S. (2013). Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **368** 20120198. https://doi.org/10.1098/rstb.2012.0198

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). *Proc. Natl. Acad. Sci. USA* **110** 228–233. https://doi.org/10.1073/pnas.1207965110

SUCHARD, M. A., LEMEY, P., BAELE, G., AYRES, D. L., DRUMMOND, A. J. and RAMBAUT, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4** vey016. https://doi.org/10.1093/ve/vey016

TANG, M., DUDAS, G., BEDFORD, T. and MININ, V. N. (2019). Fitting stochastic epidemic models to gene genealogies using linear noise approximation. Preprint. Available at arXiv:1902.08877 [q-bio.PE].

TAVARÉ, S. (2004). *Ancestral Inference in Population Genetics. Lectures on Probability Theory and Statistics: Ecole D'Eté de Probabilités de Saint-Flour XXXI*-2001. Springer, Berlin.

THOMPSON, E. A. (1975). *Human Evolutionary Trees*. Cambridge Univ. Press, Cambridge.

THORNLOW, B., YE, C., DE MAIO, N., MCBROOME, J., HINRICHS, A. S., LANFEAR, R., TURAKHIA, Y. and CORBETT-DETIG, R. (2021). Online phylogenetics using parsimony produces slightly better trees and is dramatically more efficient for large SARS-CoV-2 phylogenies than de novo and maximum-likelihood approaches. BioRxiv.

TILLETT, R. L., SEVINSKY, J. R., HARTLEY, P. D., KERWIN, H., CRAWFORD, N., GORZALSKI, A., LAVERDURE, C., VERMA, S. C., ROSSETTO, C. C. et al. (2021). Genomic evidence for reinfection with Sars-CoV-2: A case study. *Lancet Infect. Dis.* **21** 52–58. https://doi.org/10.1016/S1473-3099(20)30764-7

TRUCCHI, E., GRATTON, P., MAFESSONI, F., MOTTA, S., CICCONARDI, F., MANCIA, F., BERTORELLE, G., D'ANNESSA, I. and DI MARINO, D. (2021). Population dynamics and structural effects at short and long range support the hypothesis of the selective advantage of the G614 Sars-CoV-2 spike variant. *Mol. Biol. Evol.* **38** 1966–1979.

TURAKHIA, Y., DE MAIO, N., THORNLOW, B., GOZASHTI, L., LANFEAR, R., WALKER, C. R., HINRICHS, A. S., FERNANDES, J. D., BORGES, R. et al. (2020). Stability of Sars-CoV-2 phylogenies. *PLoS Genet.* **16** e1009175.

VAN DORP, L., RICHARD, D., TAN, C. C., SHAW, L. P., ACMAN, M. and BALLOUX, F. (2020). No evidence for increased transmissibility from recurrent mutations in Sars-CoV-2. *Nat. Commun.* **11** 1–8.

VAUGHAN, T. G., LEVENTHAL, G. E., RASMUSSEN, D. A., DRUMMOND, A. J., WELCH, D. and STADLER, T. (2019). Estimating epidemic incidence and prevalence from genomic data. *Mol. Biol. Evol.* **36** 1804–1816. https://doi.org/10.1093/molbev/msz106

VOLZ, E. M. (2012). Complex population dynamics and the coalescent under neutrality. *Genetics* **190** 187–201. https://doi.org/10.1534/genetics.111.134627

VOLZ, E. M. and FROST, S. D. W. (2014). Sampling through time and phylodynamic inference with coalescent and birth & death models. *J. R. Soc. Interface* **11** 20140945. https://doi.org/10.1098/rsif.2014.0945

VOLZ, E. M., KOELLE, K. and BEDFORD, T. (2013). Viral phylodynamics. *PLoS Comput. Biol.* **9** e1002947, 12. MR3048921 https://doi.org/10.1371/journal.pcbi.1002947

VOLZ, E. M., ROMERO-SEVERSON, E. and LEITNER, T. (2017). Phylodynamic inference across epidemic scales. *Mol. Biol. Evol.* **34** 1276–1288. https://doi.org/10.1093/molbev/msx077

VOLZ, E. M. and SIVERONI, I. (2018). Bayesian phylodynamic inference with complex models. *PLoS Comput. Biol.* **14** 1–15. https://doi.org/10.1371/journal.pcbi.1006546

VOLZ, E. M., KOSAKOVSKY POND, S. L., WARD, M. J., LEIGH BROWN, A. J. and FROST, S. D. W. (2009). Phylodynamics of infectious disease epidemics. *Genetics* **183** 1421–1430. https://doi.org/10.1534/genetics.109.106021

VOLZ, E., BAGUELIN, M., BHATIA, S., BOONYASIRI, A., CORI, A., CUCUNUBÁ, Z., CUOMO-DANNENBURG, G., DONNELLY, C. A., DORIGATTI, I. et al. (2020). Phylogenetic analysis of SARS-CoV-2. Imperial College London (15-02-2020). https://doi.org/10.25561/77169

VOLZ, E., HILL, V., MCCRONE, J. T., PRICE, A., JORGENSEN, D., O'TOOLE, Á., SOUTHGATE, J., JOHNSON, R., JACKSON, B. et al. (2021a). Evaluating the effects of Sars-CoV-2 spike mutation D614G on transmissibility and pathogenicity. *Cell* **184** 64–75.

VOLZ, E., MISHRA, S., CHAND, M., BARRETT, J. C., JOHNSON, R., GEIDELBERG, L., HINSLEY, W. R., LAYDON, D. J., DABRERA, G. et al. (2021b). Assessing transmissibility of Sars-CoV-2 lineage B. 1.1. 7 in England. *Nature* **593** 266–269.

VOZNICA, J., ZHUKOVA, A., BOSKOVA, V., SAULNIER, E., LEMOINE, F., MOSLONKA-LEFEBVRE, M. and GASCUEL, O. (2021). Deep learning from phylogenies to uncover the transmission dynamics of epidemics. *BioRxiv*. https://doi.org/10.1101/2021.03.11.435006

WAKELEY, J. (2009). *Coalescent Theory*: *An Introduction*. Roberts and Co, Greenwood Village, CO.

WAKELEY, J. (2020). Developments in coalescent theory from single loci to chromosomes. *Theor. Popul. Biol.* **133** 56–64. https://doi.org/10.1016/j.tpb.2020.02.002

WAKELEY, J. and ALIACAR, N. (2001). Gene genealogies in a metapopulation *Genetics* **159** 893–905.

WAKELEY, J. and SARGSYAN, O. (2009). Extensions of the coalescent effective population size. *Genetics* **181** 341–345.

WANG, L., BOUCHARD-CÔTÉ, A. and DOUCET, A. (2015). Bayesian phylogenetic inference using a combinatorial sequential Monte Carlo method. *J. Amer. Statist. Assoc.* **110** 1362–1374. MR3449032 https://doi.org/10.1080/01621459.2015.1054487

WANG, L., WANG, S. and BOUCHARD-CÔTÉ, A. (2020). An annealed sequential Monte Carlo method for Bayesian phylogenetics. *Syst. Biol.* **69** 155–183.

WANG, T. H., DONALDSON, Y. K., BRETTLE, R. P., BELL, J. E. and SIMMONDS, P. (2001). Identification of shared populations of human immunodeficiency virus type 1 infecting microglia and tissue macrophages outside the central nervous system. *J. Virol.* **75** 11686–11699. https://doi.org/10.1128/JVI.75.23.11686-11699.2001

WHIDDEN, C. and MATSEN IV, F. A. (2015). Quantifying MCMC exploration of phylogenetic tree space. *Syst. Biol.* **64** 472–491.

WU, S. L., MERTENS, A. N., CRIDER, Y. S., NGUYEN, A., POKPONGKIAT, N. N., DJAJADI, S., SETH, A., HSIANG, M. S., COLFORD, J. M. et al. (2020). Substantial underestimation of Sars-CoV-2 infection in the United States. *Nat. Commun.* **11** 1–10.

YANG, Z. (2014). *Molecular Evolution*: *A Statistical Approach*. Oxford Univ. Press, London.

ZAREBSKI, A. E., DU PLESSIS, L., PARAG, K. V. and PYBUS, O. G. (2021). A computationally tractable birth-death model that combines phylogenetic and epidemiological data. *BioRxiv*. https://doi.org/10.1101/2020.10.21.349068

ZHANG, C. and MATSEN IV, F. A. (2019). Variational Bayesian phylogenetic inference. In *International Conference on Learning Representations*.

ZHANG, X., TAN, Y., LING, Y., LU, G., LIU, F., YI, Z., JIA, X., WU, M., SHI, B. et al. (2020). Viral and host factors related to the clinical outcome of Covid-19. *Nature* **583** 437–440.

ZHANG, Z., NISHIMURA, A., BASTIDE, P., JI, X., PAYNE, R. P., GOULDER, P., LEMEY, P. and SUCHARD, M. A. (2021). Large-scale inference of correlation among mixed-type biological traits with phylogenetic multivariate probit models. *Ann. Appl. Stat.* **15** 230–251. MR4255276 https://doi.org/10.1214/20-aoas1394

ZHOU, D., DEJNIRATTISAI, W., SUPASA, P., LIU, C., MENTZER, A. J., GINN, H. M., ZHAO, Y., DUYVESTEYN, H. M. E., TUEKPRAKHON, A. et al. (2021). Evidence of escape of Sars-CoV-2 variant B.1.351 from natural and vaccine-induced sera. *Cell* **184** 2348–2361.e6. https://doi.org/10.1016/j.cell.2021.02.037