UCLA
UCLA Previously Published Works

Title

Online Bayesian phylodynamic inference in BEAST with application to epidemic reconstruction

Permalink

https://escholarship.org/uc/item/7x12c60r

Journal

Molecular Biology and Evolution, 37(6)

ISSN

0737-4038

Authors

Gill, Mandev S
Lemey, Philippe
Suchard, Marc A
et al.

Publication Date

2020-06-01

DOI

10.1093/molbev/msaa047

Peer reviewed

Online Bayesian Phylodynamic Inference in BEAST with Application to Epidemic Reconstruction

Mandev S. Gill,[1] Philippe Lemey,[1] Marc A. Suchard,[2,3,4] Andrew Rambaut,[5,6] and Guy Baele ![ORCID][*,1]

[1]Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Leuven, Belgium
[2]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA
[3]Department of Biostatistics, School of Public Health, University of California, Los Angeles, CA
[4]Department of Biomathematics, David Geffen School of Medicine, University of California, Los Angeles, CA
[5]Institute of Evolutionary Biology, University of Edinburgh, United Kingdom
[6]Fogarty International Center, National Institutes of Health, Bethesda, MD

***Corresponding author:** E-mail: guy.baele@kuleuven.be.

Associate editor: Michael Rosenberg

Abstract

Reconstructing pathogen dynamics from genetic data as they become available during an outbreak or epidemic represents an important statistical scenario in which observations arrive sequentially in time and one is interested in performing inference in an "online" fashion. Widely used Bayesian phylogenetic inference packages are not set up for this purpose, generally requiring one to recompute trees and evolutionary model parameters de novo when new data arrive. To accommodate increasing data flow in a Bayesian phylogenetic framework, we introduce a methodology to efficiently update the posterior distribution with newly available genetic data. Our procedure is implemented in the BEAST 1.10 software package, and relies on a distance-based measure to insert new taxa into the current estimate of the phylogeny and imputes plausible values for new model parameters to accommodate growing dimensionality. This augmentation creates informed starting values and re-uses optimally tuned transition kernels for posterior exploration of growing data sets, reducing the time necessary to converge to target posterior distributions. We apply our framework to data from the recent West African Ebola virus epidemic and demonstrate a considerable reduction in time required to obtain posterior estimates at different time points of the outbreak. Beyond epidemic monitoring, this framework easily finds other applications within the phylogenetics community, where changes in the data—in terms of alignment changes, sequence addition or removal—present common scenarios that can benefit from online inference.

Key words: BEAST, Markov chain Monte Carlo, real-time analysis, Bayesian phylogenetics, pathogen phylodynamics, online inference.

Introduction

Changes in data during ongoing research commonly occur in many fields of research, including phylogenetics. These typically include the addition of new sequences as they become available—for example, during a large sequencing study or through data sharing—and updates of alignments of existing sequences, possibly as a result of correcting sequencing errors. Such changes usually lead to the discarding of results obtained prior to the revision of the data, and recommencing statistical analyses completely from scratch (de novo). Bayesian phylogenetic inference of large data sets can be very time-consuming, sometimes requiring weeks of computing time, even when using state-of-the-art hardware. A promising avenue to mitigate this problem is an online phylogenetic inference framework that can accommodate data changes in existing analyses and leverage intermediate results to shorten the run times of updated inferences.

Existing methods to update phylogenetic estimates in an online fashion are limited, but the initial concept dates back to seminal work by Felsenstein (1981), who proposed sequential addition of species to a topology as an effective search strategy in tree space. The stepwise addition approach inserts a new taxon on the branch of the tree that yields the highest likelihood (Felsenstein 1993), and was among the first heuristics to search for a maximum-likelihood tree topology. This concept has also been incorporated into the design of various tree transition kernels and estimation heuristics. For example, in searching for the optimal tree topology in a maximum-likelihood framework, Whelan (2007) proposed to first *pluck* a number of sequences from an existing tree and subsequently place each sequence onto the tree where it yields the highest likelihood value.

Initial developments to update phylogenies with new sequence data focused on methods for phylogenetic placement, where unknown query sequences—typically short reads obtained from next-generation sequencing—are placed onto a fixed tree precomputed from a reference alignment. Employing a likelihood-based approach, Matsen et al. (2010) proposed a two-stage search algorithm to accelerate

Open Access

placements for query sequences, where a quick first evaluation of the tree is followed by a more detailed search in high-scoring parts of the tree. An increasing body of work mainly targets such taxonomic identification methods, with recent developments confronting the increasing scalability issues associated with the high dimensions of modern data sets (Barbera et al. 2019; Czech et al. 2019).

Izquierdo-Carrasco et al. (2014) implemented an online framework to estimate phylogenetic trees using maximum-likelihood heuristics, which automatically extends an existing alignment when sufficiently new data have been generated and subsequently reconstructs extended phylogenetic trees by using previously inferred smaller trees as starting topologies. The authors compared their methodology to de novo phylogenetic reconstruction and found a slight but consistent improvement in computational performance and a similar topological accuracy.

Recent foundational work toward online Bayesian phylogenetic inference focuses on sequential Monte Carlo (SMC) methods to update the posterior distribution (Dinh et al. 2018; Fourment et al. 2018; Everitt et al. 2020). These methods approximate a posterior distribution using a set of particles that exist simultaneously, which are updated when new data arrive and are then resampled with weights determined by the unnormalized posterior density (Doucet et al. 2001). While SMC methods are not new to Bayesian phylogenetics, they have primarily been explored to increase computational efficiency in standard inference, for example, to infer rooted, ultrametric (Bouchard-Côté et al. 2012), and nonultrametric phylogenetic trees (Wang et al. 2015, 2020). Within an SMC framework, Everitt et al. (2020) introduced the use of deterministic transformations to move particles effectively between target distributions with different dimensions and applied this methodology to infer an ultrametric phylogeny of a bacterial population from DNA sequence data. A similar methodology was developed independently and almost simultaneously by Dinh et al. (2018), who also describe important theoretical results on the consistency and stability of SMC for online Bayesian phylogenetic inference. Building upon the work of Dinh et al. (2018), Fourment et al. (2018) showed that the total time to compute a series of unrooted phylogenetic trees as new sequence data arrive can be reduced significantly by proposing new phylogenies through *guided* proposals that attempt to match the proposal density to the posterior. All of these SMC approaches focus on the tree inference problem rather than the estimation of broader phylogenetic models where the goal is to marginalize these over plausible trees. They have also not yet led to implementations in widely used software packages for Bayesian phylogenetic inference.

The need for online phylogenetic inference is especially pressing in the growing field of phylodynamics (see, e.g., Baele et al. 2016, 2018 for an overview). Phylodynamic inference has emerged as an invaluable tool to understand outbreaks and epidemics (Pybus et al. 2012; Faria et al. 2014; Worobey et al. 2014; Nelson et al. 2015; Dudas et al. 2017; Metsky et al. 2017), and has the potential to inform effective control and intervention strategies (Al-Qahtani et al. 2017;

Dellicour et al. 2018). Importantly, phylodynamic analyses of pathogen genome sequences sampled over time reveal events and processes that shape epidemic dynamics that are unobserved and not obtainable through any other methods. The Bayesian Evolutionary Analysis by Sampling Trees (BEAST) version 1 software package (Suchard et al. 2018) has become a primary tool for Bayesian phylodynamic inference from genetic sequence data, offering a wide range of coalescent, trait evolution and molecular clock models to study the evolution and spread of pathogens, as well as potential predictors for these processes.

Recent advances in portable sequencing technology have led to a reduction in sequencing time and costs, enabling in-field sequencing and real-time genomic surveillance as an outbreak unfolds. This was demonstrated during the recent Ebola epidemic in West Africa (Arias et al. 2016; Quick et al. 2016), as well as the recent Zika outbreak in the Americas (Faria et al. 2017). Notably, Quick et al. (2016) were routinely able to sequence Ebola-positive samples within days of collection, and in some cases were able to obtain results within 24 h. Such a continuous stream of new sequence data creates the potential for phylodynamic inference to take up a more prominent role in the public health response by providing up-to-date, actionable epidemiological and evolutionary insights during the course of an ongoing outbreak. Bayesian modeling naturally accommodates uncertainty in the phylogeny and evolutionary model parameters, and therefore offers a coherent inference framework for relatively short outbreak time-scales for which the phylogeny may not be well-resolved.

However, the potential of phylodynamic methods in real-time epidemic response can only be fully realized if accurate up-to-date inferences are delivered in a timely manner. Fast maximum-likelihood-based methods, such as those adopted by Nextstrain (Hadfield et al. 2018), can provide rapid updates by relying on a pipeline of fast, but less rigorous heuristic methods (Sagulenko et al. 2018). Bayesian phylodynamic models rely on MCMC estimation procedures that can have very long run times, often requiring days or weeks to infer the posterior distribution for complex models. Having to restart these time-consuming procedures when new data become available thus represents a significant impediment to providing regular, updated phylodynamic inferences.

Here, we explore an approach that is conceptually simpler than SMC and consists of interrupting an ongoing MCMC analysis upon the arrival of new sequence data and after the current analysis has converged, placing the new sequences at plausible locations in the current tree estimate, and then resuming the analysis with the expanded data set. We apply this methodology to data from several time periods throughout the West African Ebola virus epidemic of 2013–2016 and show that resuming an interrupted analysis after inserting new sequences into the current tree estimate, as opposed to restarting from scratch, reduces the time necessary to converge to the posterior distribution. Specifically, our approach virtually eliminates the MCMC burn-in when computing updated inferences that incorporate new data sequenced during a subsequent epidemiological week (epi week, labeled 1–52). This improved efficiency will allow the analysis and

interpretation to more closely maintain a real-time relationship to the accumulation of data.

New Approaches

We present an online phylogenetic inference framework, implemented in the BEAST 1.10 software package, that allows incorporating new data into an ongoing analysis. Notably, this methodology efficiently updates the posterior distribution upon the arrival of new data by using previous inferences to minimize the burn-in time (the time necessary for the MCMC algorithm to converge to the posterior distribution) for analysis of the expanded data set that includes the new data (along with the previously available data). Additionally, our implementation includes a new feature for BEAST 1.10 that enables resuming an MCMC analysis from the iteration at which it was terminated (similar to the "stoppb" feature in the Bayesian phylogenetics package PhyloBayes; Lartillot et al. 2009).

When new sequence data become available and the current BEAST analysis has converged to the target distribution, the BEAST analysis is interrupted and a draw (featuring estimates of all model parameters) is taken from its posterior sample. We insert the new sequences into the phylogenetic tree estimate obtained from the draw in a stepwise fashion, where the location of each insertion is determined by computing the genetic distance between the new sequence and the taxa in the tree. Next, we impute plausible values for new model parameters that are necessitated by the increased dimensionality of the enlarged phylogenetic tree, such as branch-specific evolutionary rates. Parameter values for models unaffected by the increased data dimensionality are left unchanged. The BEAST analysis is then resumed with the simulation of an MCMC sample with starting parameter values that have been constructed from the aforementioned imputation and sequence insertion algorithm. Further, the resumed analysis employs a stored set of MCMC transition kernels that have been optimized for efficient sampling using BEAST's autotuning functionality.

To determine the performance of this framework, we carefully assess the reduction in time required to converge to the target posterior distribution by using both visual analyses of MCMC trace plots as well as a scripted sliding window approach to determine burn-in. The various steps of this approach are described in more detail in Materials and Methods. We provide BEAST XML input files for the analyses performed throughout this article as well as a tutorial on setting up these analyses at http://beast.community/online_inference.html. The tutorial also describes how to set up an MCMC analysis so that it can be resumed from the iteration at which it was terminated. This new feature in BEAST 1.10 will be useful in general (beyond an online inference setting), for example, in the case of a computer crash, or if an MCMC analysis needs to be run for longer to generate sufficient samples.

Results

We evaluate the performance of our BEAST 1.10 online inference framework by analyzing complete genome data from the West African Ebola virus epidemic of 2013–2016. The data comprise 1,610 whole genome sequences collected throughout the epidemic, from March 17, 2014 to October 24, 2015 (Dudas et al. 2017). Each sequence is associated with a particular epi week during which the sample was obtained, allowing us to recreate a detailed data flow of the actual epidemic. For the purpose of our performance comparisons, we assume that the genome data were made available immediately after the time of sampling, allowing us to assess potential efficiency gains in a scenario where a Bayesian phylodynamic reconstruction would be attempted once per epi week, incorporating the newly obtained genome data into the inference up to the previous epi week.

Although our previous study on these data was performed toward the end of the epidemic (Dudas et al. 2017), during this work we were still confronted with new genome sequences becoming available, requiring us to frequently restart our MCMC analyses de novo. Considering the size of the data set, this required tremendous computational effort to obtain updated results. Here, we evaluate our online procedure by computing updated inferences corresponding to increases in data during consecutive epi weeks at different time points during the epidemic. For each time point we consider two consecutive epi weeks, which we shall refer to as the *first* and *second* epi weeks in this context. We analyze the cumulative data available by the end of the second epi week using two methods: Our proposed online inference framework which augments a previous analysis with newly obtained data (see Materials and Methods), and a de novo analysis using a randomly generated starting tree and default starting values for the model parameters following a typical Bayesian phylogenetic analysis. We use a slightly different phylodynamic model setup than in our previous study (Dudas et al. 2017), that is, an exponential growth coalescent model as the prior density on trees (Griffiths and Tavaré 1994), and an HKY+$\Gamma 4$ substitution model (Hasegawa et al. 1985; Yang 1996) for each of the four nucleotide partitions (the three codon positions and the noncoding intergenic regions) with different relative rates across the partitions. Evolutionary rates were allowed to vary across branches according to an uncorrelated relaxed molecular clock model with an underlying log-normal distribution (Drummond et al. 2006). The overall evolutionary rate was given an uninformative continuous-time Markov chain (CTMC) reference prior (Ferreira and Suchard 2008), whereas the rate multipliers for each partition were given a joint Dirichlet prior. The BEAST 1.10 XML files used in our analyses are available at http://beast.community/online_inference.html.

We consider five different pairs of consecutive epi weeks from the 2013 to 2016 Ebola epidemic: Epi weeks 25 and 26 of 2014, epi weeks 30 and 31 of 2014, epi weeks 41 and 42 of 2014, epi weeks 1 and 2 of 2015, and the final epi weeks 41 and 42 of 2015. These sets of epi weeks constitute a relatively broad range of possible sequence addition scenarios, as they

Table 1. Reduced Burn-In (in millions of iterations) Achieved with Online Bayesian Phylodynamic Inference.

Data	Sequences		Standard Analysis		Online Analysis	
	Total	Added	Burn-In (G)	Burn-In (ESS)	Burn-In (G)	Burn-In (ESS)
2014, Epi week 26	158	13	0.2 (<0.1)	<0.1 (<0.1)	<0.1 (<0.1)	<0.1 (<0.1)
2014, Epi week 31	240	8	0.8 (0.3)	<0.1 (<0.1)	<0.1 (<0.1)	0.4 (0.9)
2014, Epi week 42	706	32	8.6 (2.1)	10.2 (10.3)	0.6 (0.9)	1.0 (1.0)
2015, Epi week 2	1,072	24	16.4 (7.3)	17.6 (7.1)	0.6 (0.5)	0.4 (0.5)
2015, Epi week 42	1,610	2	49.6 (20.6)	60.2 (15.4)	<0.1 (<0.1)	0.6 (1.3)

NOTE.—Comparison of burn-in for the log joint density sample resulting from two different analysis methods applied to Ebola virus data taken from the West African Ebola epidemic of 2013–2016. The standard de novo approach of analyzing the full data set from scratch is compared with the online inference approach that updates inferences from the previous epi week upon the arrival of new data. The length of burn-in (in millions of states) is determined through a graphical approach (G) that consists of analyzing posterior trace plots, as well as by computing the amount of discarded burn-in that maximizes the ESS. Results are averaged over five replicates for each analysis, with standard deviation in parentheses.
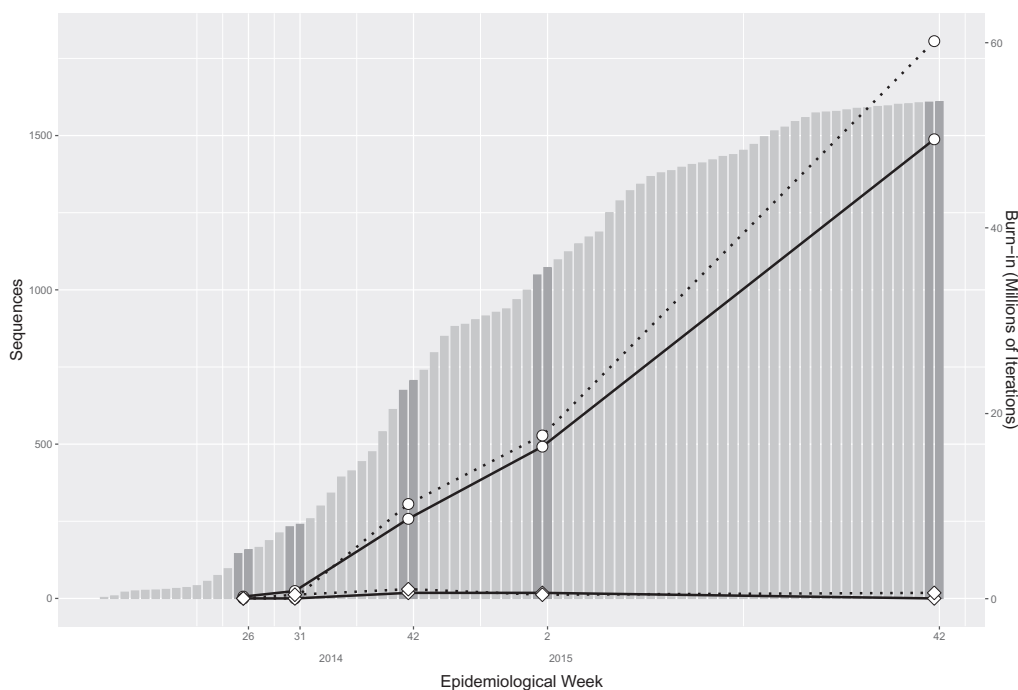

FIG. 1. Comparison of burn-in resulting from standard de novo analyses versus online Bayesian analyses to compute updated inferences from data taken from different time points of the West African Ebola virus epidemic. The data flow of the epidemic, in terms of total sequence available during each epi week, is recreated in the background of the plot in gray bars. Dark gray bars show the data corresponding to the five time points at which we compute updated inferences. The plots chart the burn-in required by de novo analyses, represented by circles, and online analyses, represented by diamonds. Solid lines correspond to burn-in estimates based on visual analyses of trace plots whereas dotted lines correspond to burn-in estimates based on maximizing ESS values.

occurred during the actual epidemic. We provide details on the number of sequences for these scenarios in table 1 and figure 1. As a Markov chain constitutes a stochastic process, for each time point, we perform five independent replicates of a standard de novo analysis of the data available by the end of first epi week, five independent replicates of a standard de novo analysis of the data available by the end of the second epi week, and five independent replicates of an online analysis of the data available by the end of the second epi week. Note that each online analysis proceeds by updating inferences from one of the de novo analyses of the data available by the end of the first epi week. We examine split frequencies for tree samples from independent replicates to compare replicates and ensure convergence to the same posterior

distribution (see supplementary material, Supplementary Material online). In particular, in all analyses we observe an average standard deviation of split frequencies (ASDSF) that meets the guideline of being <0.01 (see Materials and Methods). The replicates are independent in that the MCMC simulations start from different trees. In particular, standard de novo analyses use randomly generated starting trees, and online analyses feature starting trees that differ because they are constructed by augmenting different tree estimates from different de novo analyses of the data available by the end of the first epi week. For each time period, we determine a random order for the new sequences and insert them into the tree estimate in the same order for each of the five replicates.
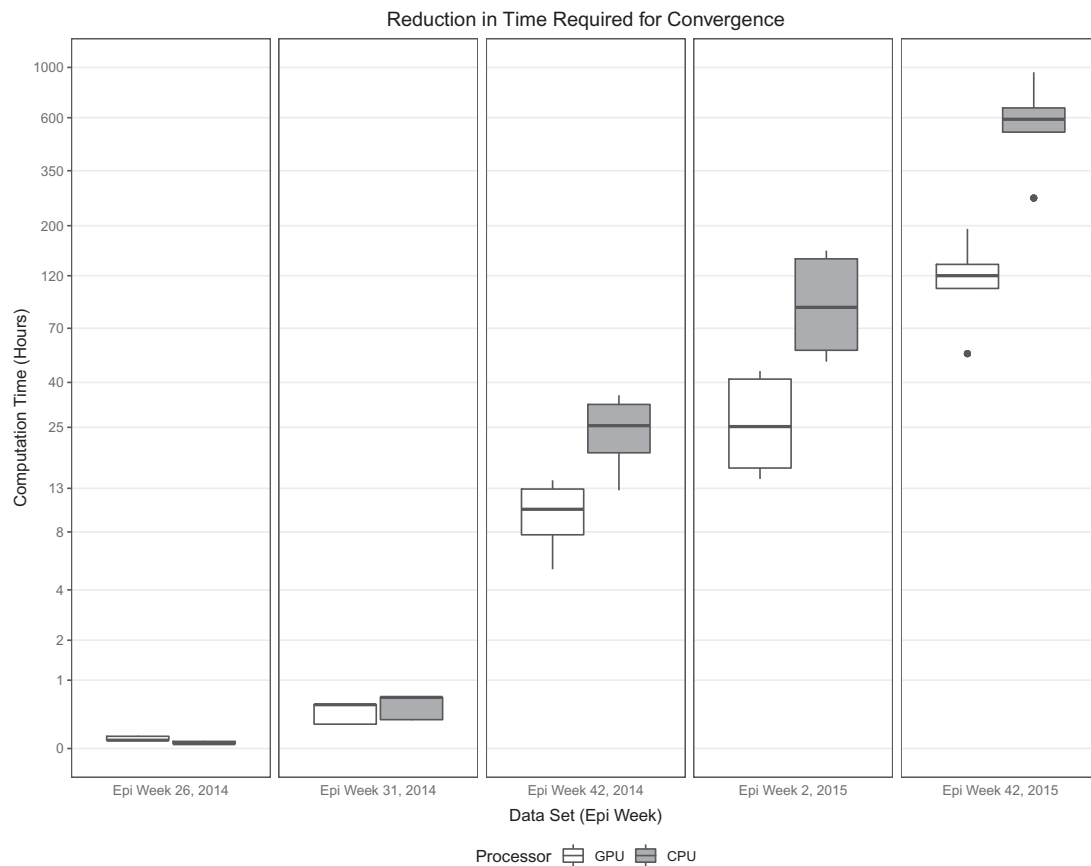
Fig. 2. Box plots show distribution of savings in computation time by using online inference as compared with standard de novo analyses to update inferences for data from different time points in the West African Ebola virus epidemic. White box plots correspond to analyses using a Tesla P100 graphics card for scientific computing and gray boxes correspond to analyses using a multi-core CPU. Irrespective of the actual hardware used, the time savings are substantial with up to 600 h on average saved using our online approach on CPU for our most demanding scenario. The axis corresponding to running time (in hours) is log-transformed to allow for greater visibility of plots for smaller data sets.

For each pair of consecutive epi weeks, we compare the burn-in for the sample of the log joint density (which is proportional to the posterior density) resulting from online and standard de novo analyses. Figure 1 and table 1 show the results, averaged over five replicates. The different methods of determining the burn-in (see Materials and Methods) yield very similar estimates. We assess the sensitivity of sequence insertion order by performing five additional replicates each for epi weeks 41 and 42 of 2014 and epi weeks 1 and 2 of 2015. Each of the additional replicates for a given time period augments the same inferences through a different, random sequence insertion order. We find that the estimated burn-in for each additional replicate is in line with the burn-in estimate for the corresponding time period in table 1, lying within two standard deviations of the mean.

The results show that our online inference framework can reduce burn-in by a significant amount (P-values are <0.01 for t-tests comparing burn-in from online and standard analyses for the latter three epi weeks). Although the burn-in for epi weeks 26 and 31 of 2014 is negligible in both online and standard analyses, the standard approach requires substantial burn-in in the latter three cases. By reducing the average burn-in to one million iterations or less for each of these three epi weeks, the online approach virtually eliminates the burn-

in in these analyses. The results for epi week 42 of 2015 data are particularly remarkable (see supplementary figs. S1–S3, Supplementary Material online for a comparison of posterior trace plots from five replicates of all test cases), showing average reductions of burn-in by 50–60 million iterations.

To put these efficiency gains into perspective, it is useful to translate the reduction of burn-in into actual saved computing time using a multi-core CPU (in our case, a 14-core 2.20 GHz Intel Xeon Gold 5120 CPU) as well as using a state-of-the-art hardware setup enhanced by a GPU (e.g., a Tesla P100 graphics card intended for scientific computing). We use BEAGLE 2.1.2 (Ayres et al. 2012) to enable such GPU computation within BEAST. Figure 2 depicts the savings in computation time by using online inference as compared with standard de novo analyses to update inferences for data from different time points in the West African Ebola virus epidemic. Dunn tests (Dunn 1961) indicate that the savings under online inference for each time point are significant ($P < 0.01$). We note that running time depends on burn-in length as well as data set size, with larger data sets requiring more time per iteration. Our online inference approach leads to higher computation time savings as the complexity of the data increases, with up to 600 h being saved on average on a modern multi-core processor. State-of-the-art

graphics cards targeting the scientific computing market are able to reduce this number to 120 h on average of savings, but such cards may not be readily available, especially in resource-limited settings.

Discussion

We present a framework for online Bayesian phylodynamic inference that accommodates a continuous data flow, as exemplified by an epidemic scenario where continued sampling efforts yield a series of genome sequences over time. This framework has been implemented in BEAST 1.10, a popular software package for Bayesian phylogenetic and phylodynamic inference. Through empirical examples taken from the 2013 to 2016 West African Ebola epidemic, we show that our online approach can significantly reduce burn-in and, consequently, the time necessary to generate sufficient samples from the posterior distribution of a phylodynamic model being applied to a growing data set. The savings in computation time can amount to days or even weeks, depending on the computational infrastructure, the complexity of the data and hence also the accompanying phylodynamic model.

The improvements in computational efficiency through minimizing burn-in that we observe are encouraging, but there is a need to continue improving efficiency in multiple directions. First, alternative sequence insertion and branch rate imputation procedures may yield better performance in certain situations. Desper and Gascuel (2002), for instance, employ a minimum evolution criterion for stepwise addition of taxa. As another example, an insertion procedure that allows new sequences to have insertion times that are deeper than the root of the current tree estimate may be more suitable in the case that new sequences are distantly related to the sequences that already exist in the tree. Under the current implementation, MCMC transition kernels enable the insertion point of a new sequence to eventually be repositioned deeper than the root of the starting tree. However, allowing a sequence to be directly inserted deeper than the root may save computational time.

Second, even if burn-in is minimized, generating sufficient samples from the Markov chain after it has converged to the posterior distribution can still be very time-consuming. A popular approach to generate samples more quickly is to run multiple independent chains, starting from different random locations in search space, in parallel and combine the posterior samples. However, the time saved through such a strategy depends on the burn-in phase, which must elapse for each chain before its samples can be used. From this perspective, the advances of our online framework are especially important. Another strategy for more efficient sampling is to evaluate past MCMC performance during pauses to incorporate new data and make informed adjustments prior to resuming the analysis. For instance, transition kernel weights can be modified to focus on parameters with low ESS values. Progress can also be made through advances in MCMC sampling that enable more efficient exploration of posterior distributions. Innovative sampling techniques that have already

shown promise in the context of phylogenetics and are ripe for further development include adaptive MCMC (Baele et al. 2017) and Hamiltonian Monte Carlo (Neal 2010; Lan et al. 2015; Ji et al. 2019). Finally, the computational performance will undoubtedly benefit from continued development of high-performance libraries for phylogenetic likelihood calculation (Ayres et al. 2019).

The implementation we present here differs from other recent work on online Bayesian phylogenetic inference, which relies on SMC to update phylogenies (Dinh et al. 2018; Fourment et al. 2018; Everitt et al. 2020). Although SMC represents a principled approach to infer a distribution of growing dimensions, the SMC-based methods for online Bayesian phylogenetics are limited to inferring phylogenetic trees. It would be beneficial to integrate SMC algorithms for updating phylogenies with MCMC methods to sample other evolutionary model parameters, and ultimately to implement a complementary online inference framework in BEAST. Such an implementation would enable direct comparison of the current online framework with SMC-based approaches, allowing researchers to assess the benefits and drawbacks of each approach and helping to streamline future development of online Bayesian phylogenetic inference.

Our development has been primarily motivated by epidemic scenarios that entail a continuous stream of new sequence data becoming available during the course of an outbreak. In our empirical assessment of the West African Ebola virus epidemic, we have assumed that the genome data were made available close to the time of sampling, which represents the ideal scenario in an outbreak response. In reality, during the epidemic, there was considerable variation in how rapidly virus genome data were available for analysis. There were many reasons for this, but even when genomes were being shared as rapidly as possible, the batch shipping of samples to high-throughput sequencing centers resulted in a minimum delay of many weeks (Gire et al. 2014; Park et al. 2015). This changed toward the end of the epidemic as new, portable, sequencing instruments were installed in Ebola treatment centers in Guinea and Sierra Leone (Arias et al. 2016; Quick et al. 2016), producing virus genome sequences from patients within days or hours of a sample being taken. We expect that the use of such instruments at the point of diagnosis will increase and the resulting stream of sequence data will mean that the computational analysis will become the bottleneck in using the data to inform the response. From this perspective, the reduction in time necessary to compute updated inferences on data from the Ebola virus epidemic through our online inference framework is promising, and continued efforts to further improve efficiency are crucial.

Beyond computational efficiency, additional development is needed in order to maximize the potential impact of our framework as a support tool during outbreaks. The current implementation must be extended to accommodate more sophisticated phylodynamic models, especially methods that integrate sequence data with other epidemiological data to elucidate different phylodynamic processes (Lemey et al. 2009; Gill et al. 2013; Lemey et al. 2014; Gill et al. 2016). For many of these models—for example, a phylogeographic

model for which a sequence from a previously unsampled location is being added—the addition of novel sequence data will increase their dimensionality, and methods that augment the models in an intelligent manner are essential. Adding sequence data may also require increasingly complex models to accurately describe the underlying evolutionary processes as the data set grows (e.g., transitioning from a strict to a relaxed clock model), a process that should ideally not require user interactions. This could potentially be addressed by developing nonparametric Bayesian models for evolutionary heterogeneity that can dynamically accommodate increasing model complexity. Finally, we have focused on evaluating the performance of updating phylogenetic inferences conditional on pre-aligned sequence data. However, a comprehensive system for real-time evolutionary analysis will need to include an alignment step when new sequence data become available.

Finally, while real-time monitoring of infectious disease outbreaks has motivated much of our development, we anticipate that our online inference framework will be more broadly useful, allowing researchers to save precious time in any context in which new data become available that extend a previously analyzed data set. Many large-scale sequencing efforts in a wide range of research fields generate a steady flow of genomic data sequences, which often involve a phylogenetic component, and as such online Bayesian phylogenetic inference will prove useful beyond the field of pathogen phylodynamics.

Materials and Methods

Online Bayesian Phylogenetic Inference

Our strategy to increase efficiency through an online inference framework in BEAST 1.10 builds on using estimates from a previous MCMC analysis in order to minimize time to convergence to the new posterior distribution. In MCMC simulation, this *burn-in* period corresponds to a transient phase of the Markov chain during which the simulated values reflect the influence of the starting values of the chain and are from low-probability regions of the target posterior distribution (Brooks and Roberts 1998). The burn-in period ends once the chain achieves stationary behavior and has converged to the posterior distribution. Including simulated values from the burn-in phase of the chain in approximations of the posterior distribution can lead to substantial bias and it has therefore become common practice to discard samples taken during the burn-in period. Burn-in phases for standard phylodynamic models on realistic data sets can be extremely long, and through minimizing burn-in, we can save a potentially large proportion of the computational time usually required to generate a good posterior sample.

Online inference can be viewed as a series of steps (or generations) with increasing amounts of data, with each step consisting of sampling from the posterior distribution for the model specified at the given step. The model must be adjusted when transitioning from one step to the next in order to accommodate the growth in data. Consider an ongoing (or completed) analysis at step i of a data set of N_i

sequences with a phylodynamic model that includes a choice of substitution model(s) (Jukes and Cantor 1969; Hasegawa et al. 1985; Tavaré 1986), a strict or uncorrelated relaxed molecular clock model (Drummond et al. 2006), and a parametric coalescent tree prior (Griffiths and Tavaré 1994). Assume that at step i, the analysis has achieved convergence and has generated samples from the posterior distribution. Upon the arrival of M_{i+1} new sequences, we interrupt the step i analysis (if it has not yet run to completion), augment the analysis with the new sequences, and proceed to step $i + 1$, during which we will analyze the expanded data set of $N_{i+1} = N_i + M_{i+1}$ sequences.

We take a random draw θ_i from the posterior sample (i.e., excluding the burn-in) generated in step i that consists of estimates of the phylogenetic tree and all other model parameters. Further, BEAST automatically optimizes transition kernel tuning parameters during an MCMC analysis in order to maximize sampling efficiency (Suchard et al. 2018), and we extract the optimized tuning parameter values from step i. We modify the elements of θ_i in order to obtain $\theta_{i+1}^{(0)}$, the starting model parameter values for the MCMC chain simulated in step $i + 1$. The aim in our construction of $\theta_{i+1}^{(0)}$ is to leverage the values of θ_i to obtain starting parameter values that are in, or relatively close to, a high-probability region of the target posterior in step $i + 1$, and thereby minimize the step $i + 1$ burn-in phase. This is in contrast to the typical approach of using default or randomly generated starting parameter values (including the phylogenetic tree) that can be very distant from high-probability regions of the posterior. Such suboptimal starting values are a major cause of long burn-in periods.

The algorithm to augment θ_i to $\theta_{i+1}^{(0)}$ starts with expanding the tree from θ_i by inserting a new sequence into it. The sequence insertion process is illustrated in figure 3. First we find the observed sequence already in the tree that is closest to the new sequence in terms of genetic distance, where genetic distance is based on a simple nucleotide substitution model (we refer to this sequence as the *closest* sequence). We compute the genetic distance in all analyses using a JC69 model (Jukes and Cantor 1969), but our implementation also offers an F84 model (Felsenstein and Churchill 1996).

We then insert a common ancestor node for the new sequence and its *closest* sequence. To determine the height at which to insert the new ancestor node, we first translate the genetic distance d between the two sequences to a distance d_t in units of time by dividing d by the evolutionary rate associated with the branch leading to the *closest* sequence. Further, let t_n denote the sampling time (in terms of time units prior to the present time) of the new sequence, t_c the sampling time of the *closest* sequence, and t_{insert} the time at which we will insert the new ancestor node. Assume, without loss of generality, that the new sequence has a more recent sampling time (so that $t_c > t_n$). Consider

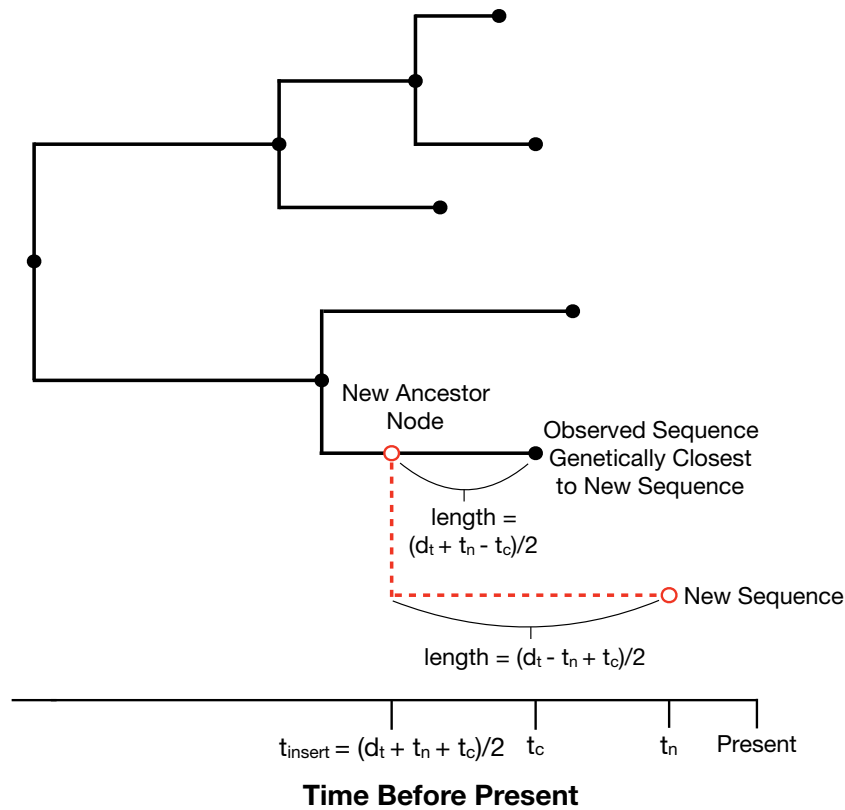$$t^* = t_c + \frac{d_t - (t_c - t_n)}{2} = \frac{d_t + t_c + t_n}{2}. \tag{1}$$

We set

FIG. 3. A new sequence is inserted into an existing phylogenetic tree by determining the closest observed sequence (in terms of genetic distance) already in the tree, and inserting a new ancestor node for the new sequence and its closest sequence. The genetic distance between the new sequence and its closest sequence is converted into a distance in units of time, d_t, by dividing by the evolutionary rate associated with the branch leading to the closest sequence. To determine the insertion time t_{insert} of the new ancestor node (in terms of time prior to the present time), we require $(t_{insert} - t_c) + (t_{insert} - t_n) = d_t$, where t_n is the sampling time of the new sequence, and t_c the sampling time of its closest sequence. This yields $t_{insert} = (d_t + t_n + t_c)/2$.

$$t_{insert} = t^*, \qquad (2)$$

(except in special cases, which we discuss shortly) because this ensures that the placement of the new ancestor node is consistent with d_t in that $(t_{insert} - t_c) + (t_{insert} - t_n) = d_t$. Notably, this method of determining the insertion height allows the new branch to emanate from an external branch or internal branch, with the latter case accommodating realistic insertion of divergent lineages. In certain cases, however, we use an alternative insertion time because setting $t_{insert} = t^*$ results in $t_{insert} < t_c$, or a new branch of length 0, or $t_c \geq t_{root}$ (where t_{root} is the root height of the tree). In these cases, we let ϵ denote a scalar in the interval (0, 1), let *split-child* refer to the child node of the branch that will be split by the insertion of the new ancestor node, let l_b denote the length of the aforementioned branch, and let t_{sc} denote the height of the *split-child*. We then set

$$t_{insert} = t_{sc} + \epsilon * l_b. \qquad (3)$$

Here, if $t^* < t_c$, the *split-child* is the *closest* sequence, and if t^* is equal to the height of an ancestral node of the *closest* sequence, then this ancestral node's child is the *split-child*. Finally, if $t^* \geq t_{root}$, the *split-child* is the child node of the root that is an ancestor of the *closest* sequence. See supplementary

algorithm S1, Supplementary Material online for further details.

Next, the growth of the tree after a sequence insertion requires branch-specific aspects of the evolutionary model to assume a greater dimension. In particular, our implementation allows for specification of either a strict or uncorrelated relaxed molecular clock model. Under the uncorrelated relaxed clock, each branch-specific clock rate is drawn independently from an underlying rate distribution (e.g., an exponential or log-normal distribution). The underlying rate distribution is discretized into a number of categories equal to the number of branches, and each branch receives a unique clock rate corresponding to its assigned category. We impute clock rates on the branches of the enlarged tree by assigning branches to rate categories according to a deterministic procedure described in detail in supplementary material, Supplementary Material online.

The algorithm continues in this fashion: The remaining new sequences are inserted into the growing phylogenetic tree one at a time, and uncorrelated relaxed clock rates associated with tree branches are updated after each insertion. The order of insertion can be specified by the user in the XML (in the Ebola virus example, a sensitivity analysis detailed in the Results section suggests that the performance does not

depend on insertion order). Aspects of the model that remain compatible with an increase in sequence data, such as substitution model specification, are left unaltered, and the parameters that characterize these aspects are identical in both θ_i and $\theta_{i+1}^{(0)}$.

The final part of step $i + 1$ is to simulate a Markov chain, with starting model parameter values $\theta_{i+1}^{(0)}$ and initial tuning parameter values taken, pre-optimized, from step i. We note that there is no hard-encoded stopping rule, and the termination of the simulated chain at step $i + 1$ is left to the user's discretion. The simulation should continue at least until the chain has achieved stationarity, and until either new data become available (and the simulation can be interrupted to incorporate the new data), or a sufficient posterior sample for inference has been produced. However, there is no need to completely terminate the chain at step $i + 1$ if it is interrupted to incorporate new data because the step $i + 1$ chain can be resumed after the interruption, and the step $i + 2$ simulation for the expanded data set can be started as an independent process. Indeed, if step $i + 1$ has yet to produce sufficient posterior samples it may be optimal to resume its simulation to obtain provisional inferences (that could go toward informing the response to an outbreak, for instance) while waiting for the step $i + 2$ chain to converge.

Performance

We assess burn-in using two different approaches. First, we use Tracer (Rambaut et al. 2018), a popular software package for posterior summarization in Bayesian phylogenetics, to visually examine trace plots of the posterior distribution. The earliest iteration after which the plot exhibits stationarity is taken to be the end of the burn-in period. Second, we use the R (R Core Team 2018) package coda (Plummer et al. 2006) to compute the effective sample size (ESS) of the log joint (likelihood × prior) density sample after discarding the first n samples, and we adopt the value of n that yields the maximal ESS as the burn-in. The ESS is a statistic that estimates the number of independent draws from the target distribution that an MCMC sample corresponds to by accounting for the autocorrelation in the sample (Kass et al. 1998), and the joint density is often, even by us, called the "posterior" in BEAST. This is inexact because the joint density is an unnormalized rescaling of the posterior. Discarding highly correlated burn-in iterates from the sample leads to a greater ESS and, in effect, a more informative sample.

We compare the frequencies of splits (or clades) across multiple independent Markov chains in order to ensure that the independent replicates for a given time point in the Ebola virus epidemic converge to the same stationary distribution. In particular, we compare chains generated by the same method (standard inference or online inference) and by different methods by considering all possible pairwise comparisons for chains corresponding to the same data set. For each pair of chains, we use the R We There Yet (RWTY) software package (Warren et al. 2017) to create a plot of split frequencies as well compute their correlation and the ASDSF (Lakner et al. 2008). As the different chains converge to the same stationary distribution, the ASDSF should approach 0. We adopt the guideline that an ASDSF <0.05 (ideally, <0.01) supports topological convergence (Ronquist et al. 2011).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

References

Al-Qahtani AA, Baele G, Khalaf N, Suchard MA, Al-Anazi MR, Abdo AA, Sanai FM, Al-Ashgar HI, Khan MQ, Al-Ahdal MN, et al. 2017. The epidemic dynamics of hepatitis C virus subtypes 4a and 4d in Saudi Arabia. *Sci Rep.* 7:44947.

Arias A, Watson SJ, Asogun D, Tobin EA, Lu J, Phan MVT, Jah U, Wadoum REG, Meredith L, Thorne L, et al. 2016. Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases. *Virus Evol.* 2(1):vew016.

Ayres DL, Cummings MP, Baele G, Darling AE, Lewis PO, Swofford DL, Huelsenbeck JP, Lemey P, Rambaut A, Suchard MA. 2019. BEAGLE 3: improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. *Syst Biol.* 68(6):1052–1061.

Ayres DL, Darling A, Zwickl DJ, Beerli P, Holder MT, Lewis PO, Huelsenbeck JP, Ronquist F, Swofford DL, Cummings MP, et al. 2012. BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Syst Biol.* 61(1):170–173.

Baele G, Dellicour S, Suchard MA, Lemey P, Vrancken B. 2018. Recent advances in computational phylodynamics. *Curr Opin Virol.* 31:24–32.

Baele G, Lemey P, Rambaut A, Suchard MA. 2017. Adaptive MCMC in Bayesian phylogenetics: an application to analyzing partitioned data in BEAST. *Bioinformatics* 33(12):1798–1805.

Baele G, Suchard MA, Rambaut A, Lemey P. 2017. Emerging concepts of data integration in pathogen phylodynamics. *Syst Biol.* 66(1):e47–e65.

Barbera P, Kozlov AM, Czech L, Morel B, Darriba D, Flouri T, Stamatakis A. 2019. EPA-ng: massively parallel evolutionary placement of genetic sequences. *Syst Biol.* 68(2):365–369.

Bouchard-Côté A, Sankararaman S, Jordan MI. 2012. Phylogenetic inference via sequential Monte Carlo. *Syst Biol.* 61(4):579–593.

Brooks S, Roberts G. 1998. Convergence assessment techniques for Markov chain Monte Carlo. *Stat Comput.* 8(4):319–335.

Czech L, Barbera P, Stamatakis A. 2019. Methods for automatic reference trees and multilevel phylogenetic placement. *Bioinformatics* 35(7):1151–1158.

Dellicour S, Baele G, Dudas G, Faria N, Pybus O, Suchard M, Rambaut A, Lemey P. 2018. Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak. *Nat Commun.* 9(1):2222.

Desper R, Gascuel O. 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J Comput Biol.* 9(5):687–705.

Dinh V, Darling A, Matsen F. 2018. Online Bayesian phylogenetic inference: theoretical foundations via sequential Monte Carlo. *Syst Biol.* 67(3):503–517.

Doucet A, de Freitas N, Gordon N, editors. 2001. Sequential Monte Carlo methods in practice. New York: Springer.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4(5):e88.

Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, Faria NR, Park DJ, Ladner JT, Arias A, Asogun D, et al. 2017. Virus genomes reveal factors that spread and sustained the ebola epidemic. *Nature* 544(7650):309–315

Dunn OJ. 1961. Multiple comparisons among means. *J Am Stat Assoc.* 56:54–64.

Everitt R, Culliford R, Medina-Aguayo F, Wilson D. 2020. Sequential Monte Carlo with transformations. *Stat Comput.* 30(3):663–676.

Faria NR, Quick J, Claro IM, Thézé J, de Jesus JG, Giovanetti M, Kraemer MUG, Hill SC, Black A, da Costa AC, et al. 2017. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature* 546(7658):406–410.

Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, Tatem AJ, Sousa JD, Arinaminpathy N, Pépin J, et al. 2014. 2014. The early spread and epidemic ignition of HIV-1 in human populations. *Science* 346(6205):56–61.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol.* 17(6):368–376.

Felsenstein J. 1993. PHYLIP: phylogenetic inference package. Version 3.5c. Distributed by the author. Seattle (WA): Department of Genome Sciences, University of Washington.

Felsenstein J, Churchill GA. 1996. A hidden Markov model approach to variation among sites in rate of evolution. *Mol Biol Evol.* 13(1):93–104.

Ferreira MAR, Suchard MA. 2008. Bayesian analysis of elapsed times in continuous-time Markov chains. *Can J Stat.* 26:355–368.

Fourment M, Claywell B, Dinh V, McCoy C, Matsen F, Darling A. 2018. Effective online Bayesian phylogenetics via sequential Monte Carlo with guided proposals. *Syst Biol.* 67(3):490–502.

Gill MS, Lemey P, Bennett SN, Biek R, Suchard MA. 2016. Understanding past population dynamics: Bayesian coalescent-based modeling with covariates. *Syst Biol.* 65(6):1041–1056.

Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. 2013. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol Biol Evol.* 30(3):713–724.

Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G, et al. 2014. Genomic surveillance elucidates ebola virus origin and transmission during the 2014 outbreak. *Science* 345(6202):1369–1372.

Griffiths R, Tavaré S. 1994. Sampling theory for neutral alleles in a varying environment. *Philos Trans R Soc Lond B Biol Sci.* 344(1310):403–410.

Hadfield J, Megill C, Bell S, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher R. 2018. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 34(23):4121–4123.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol.* 22(2):160–174.

Izquierdo-Carrasco F, Cazes J, Smith S, Stamatakis A. 2014. PUmPER: phylogenies updated perpetually. *Bioinformatics* 30(10):1476–1477.

Ji X, Zhang Z, Holbrook A, Nishimura A, Baele G, Rambaut A, Suchard MA. 2019. Gradients do grow on trees: a linear-time O(N)-dimensional gradient for statistical phylogenetics. Available from: Https://arxiv.org/pdf/1905.12146.pdf.

Jukes T, Cantor C. 1969. Evolution of protein molecules. Cambridge (MA): Academic Press. p. 21–132.

Kass R, Carlin B, Gelman A, Neal R. 1998. Markov Chain Monte Carlo in practice: a roundtable discussion. *Am Stat.* 52:93–100.

Lakner C, van der Mark P, Huelsenbeck JP, Larget B, Ronquist F. 2008. Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. *Syst Biol.* 57(1):86–103.

Lan S, Palacios JA, Karcher M, Minin VN, Shahbaba B. 2015. An efficient Bayesian inference framework for coalescent-based nonparametric phylodynamics. *Bioinformatics* 31(20):3282–3289.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25(17):2286–2288.

Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, Baele G, Russell CA, Smith DJ, Pybus OG, Brockmann D, et al. 2014. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS Pathog.* 10(2):e1003932.

Lemey P, Rambaut A, Drummond AJ, Suchard MA. 2009. Bayesian phylogeography finding its roots. *PLoS Comput Biol.* 5(9):e1000520.

Matsen F, Kodner R, Armbrust E. 2010. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* 11(1):538.

Metsky HC, Matranga CB, Wohl S, Schaffner SF, Freije CA, Winnicki SM, West K, Qu J, Baniecki ML, Gladden-Young A, et al. 2017. Zika virus evolution and spread in the Americas. *Nature* 546(7658):411–415.

Neal RM. 2010. MCMC using Hamiltonian dynamics. In: *Handbook of Markov chain Monte Carlo.* Vol. 54. Boca Raton (FL): Chapman & Hall/CRC. p. 113–162.

Nelson MI, Viboud C, Vincent AL, Culhane MR, Detmer SE, Wentworth DE, Rambaut A, Suchard MA, Holmes EC, Lemey P. 2015. Global migration of influenza A viruses in swine. *Nat Commun.* 6:6696.

Park DJ, Dudas G, Wohl S, Goba A, Whitmer SLM, Andersen KG, Sealfon RS, Ladner JT, Kugelman JR, Matranga CB, et al. 2015. Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. *Cell* 161(7):1516–1526.

Plummer M, Best N, Cowles K, Vines K. 2006. CODA: convergence diagnosis and output analysis for MCMC. *R News.* 6:7–11.

Pybus OG, Suchard MA, Lemey P, Bernardin FJ, Rambaut A, Crawford FW, Gray RR, Arinaminpathy N, Stramer SL, Busch MP, et al. 2012. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc Natl Acad Sci USA.* 109(37):15066–15071.

Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A, et al. 2016. Real-time, portable genome sequencing for Ebola surveillance. *Nature* 530(7589):228–232.

R Core Team 2018. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior summarization in Bayesian phylogenetics using tracer 1.7. *Syst Biol.* 67(5):901–904.

Ronquist F, Huelsenbeck JP, Teslenko M. 2011. Draft MrBayes version 3.2 manual: tutorials and model summaries. Available from: http://mrbayes.sourceforge.net/mb3.2_manual.pdf.

Sagulenko P, Puller V, Neher R. 2018. TreeTime: maximum-likelihood phylodynamic analysis. *Virus Evol.* 4(1):vex042.

Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* 4(1):vey016.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. In: Waterman MS, editor. Some mathematical questions in biology: DNA sequence analysis. Providence (RI): American Mathematical Society. p. 57–86.

Wang L, Bouchard-Côté A, Doucet A. 2015. Bayesian phylogenetic inference using a combinatorial sequential Monte Carlo method. *J Am Stat Assoc.* 110(512):1362–1374.

Wang L, Wang S, Bouchard-Côté A. 2020. An annealed sequential Monte Carlo method for Bayesian phylogenetics. *Syst Biol.* 69(1):155–183.

Warren DL, Geneva AJ, Lanfear R. 2017. RWTY: (R We There Yet): an R package for examining convergence of Bayesian phylogenetic analyses. *Mol Biol Evol.* 34:1016–1020.

Whelan S. 2007. New approaches to phylogenetic tree search and their application to large numbers of protein alignments. *Syst Biol.* 56(5):727–740.

Worobey M, Han GZ, Rambaut A. 2014. A synchronized global sweep of the internal genes of modern avian influenza virus. *Nature* 508(7495):254–257.

Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol.* 11(9):367–372.