Title

Hamiltonian Monte Carlo sampling to estimate past population dynamics using the skygrid coalescent model in a Bayesian phylogenetics framework

Authors
Baele, Guy
Gill, Mandev S
Lemey, Philippe
et al.

Publication Date
2020

DOI
10.12688/wellcomeopenres.15770.1

Peer reviewed

METHOD ARTICLE

Hamiltonian Monte Carlo sampling to estimate past population dynamics using the skygrid coalescent model in a Bayesian phylogenetics framework [version 1; peer review: 1 approved, 2 approved with reservations]

Guy Baele[1], Mandev S. Gill[1], Philippe Lemey[1], Marc A. Suchard[2]

[1]Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Herestraat 49, 3000, Leuven, Belgium
[2]Departments of Biostatistics, Biomathematics and Human Genetics, University of California, Los Angeles, 695 Charles E. Young Drive, Los Angeles, California, 90095-1766, USA

v1 **First published:** 30 Mar 2020, **5**:53
https://doi.org/10.12688/wellcomeopenres.15770.1

Latest published: 30 Mar 2020, **5**:53
https://doi.org/10.12688/wellcomeopenres.15770.1

Abstract

Nonparametric coalescent-based models are often employed to infer past population dynamics over time. Several of these models, such as the skyride and skygrid models, are equipped with a block-updating Markov chain Monte Carlo sampling scheme to efficiently estimate model parameters. The advent of powerful computational hardware along with the use of high-performance libraries for statistical phylogenetics has, however, made the development of alternative estimation methods feasible. We here present the implementation and performance assessment of a Hamiltonian Monte Carlo gradient-based sampler to infer the parameters of the skygrid model. The skygrid is a popular and flexible coalescent-based model for estimating population dynamics over time and is available in BEAST 1.10.5, a widely-used software package for Bayesian pylogenetic and phylodynamic analysis. Taking into account the increased computational cost of gradient evaluation, we report substantial increases in effective sample size per time unit compared to the established block-updating sampler. We expect gradient-based samplers to assume an increasingly important role for different classes of parameters typically estimated in Bayesian phylogenetic and phylodynamic analyses.

Keywords

Hamiltonian Monte Carlo, Markov chain Monte Carlo, Bayesian skygrid, phylogenetics, pathogen phylodynamics, BEAST, BEAGLE

Open Peer Review

Reviewer Status ✓ ? ?

	Invited Reviewers		
	1	**2**	**3**
version 1 30 Mar 2020	✓ report	? report	? report

1. **Mario dos Reis**, Queen Mary University of London, London, UK

2. **Shiwei Lan**, Arizona State University, Tempe, USA

3. **Nicola Müller**, Fred Hutchinson Cancer Research Center, Seattle, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Marc A. Suchard (msuchard@ucla.edu)

Author roles: Baele G: Formal Analysis, Investigation, Writing – Original Draft Preparation, Writing – Review & Editing; **Gill MS**: Investigation, Software, Writing – Review & Editing; **Lemey P**: Funding Acquisition, Writing – Review & Editing; **Suchard MA**: Conceptualization, Methodology, Software, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Wellcome Trust through funding to the Artic Network [206298]. GB acknowledges support from the Interne Fondsen KU Leuven / Internal Funds KU Leuven under grant agreement C14/18/094, and from the Research Foundation - Flanders (`Fonds voor Wetenschappelijk Onderzoek - Vlaanderen', G0E1420N). The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 725422 - ReservoirDOCS). PL acknowledges support by the Research Foundation -- Flanders (`Fonds voor Wetenschappelijk Onderzoek - Vlaanderen', G066215N, G0D5117N and G0B9317N). MAS is partly supported by NSF grant DMS 1264153 and NIH grants R01 HG006139, R01 AI107034 and U19 AI135995.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Baele G, Gill MS, Lemey P and Suchard MA. **Hamiltonian Monte Carlo sampling to estimate past population dynamics using the skygrid coalescent model in a Bayesian phylogenetics framework [version 1; peer review: 1 approved, 2 approved with reservations]** Wellcome Open Research 2020, **5**:53 https://doi.org/10.12688/wellcomeopenres.15770.1

First published: 30 Mar 2020, **5**:53 https://doi.org/10.12688/wellcomeopenres.15770.1

Introduction

Inference of effective population size over time from a sample of molecular sequences is a key aspect of many phylodynamics studies. Inference methods typically employ coalescent models that connect population dynamics to the shape of a genealogy relating such a sample. A wide range of coalescent models has been developed over the past decades, gradually extending the original coalescent theory of Kingman[1]. In particular, flexible nonparametric coalescent models have become widely used[2–8]. These models typically posit that the effective population size as a function of time (also referred to as the "demographic function") assumes a piecewise constant form, thereby avoiding restrictive *a priori* assumptions about the specific parametric form of the demographic function. In a Bayesian framework, coalescent models function as prior distributions for phylogenetic trees and, in conjunction with observed sequence data likelihoods based on continuous-time Markov models for molecular character evolution on trees[9], they enable the estimation of effective population size directly from molecular sequence data.

Among such nonparametric models, the Bayesian skygrid[7] has emerged as a popular choice for a number of reasons. Unlike most competing models, the skygrid can incorporate data from multiple unlinked genetic loci, which has proven to be invaluable for accurate reconstruction of past population dynamics[10]. Further, the skygrid has been extended to integrate external time-varying covariates[11], enabling the assessment of the relationship between effective population size and ecological and epidemiological indices, and also potentially yielding improved estimates of effective population size trajectories and genealogies[12]. Like the skyride model[6], the skygrid aims to construct a smooth population size trajectory over time through a Gaussian Markov random field (GMRF) smoothing prior. Finally, the skygrid is implemented in the widely used BEAST 1.10 software package[13,14], where it can be combined with a wide range of models for evolutionary heterogeneity, phylogeography, and phenotypic trait evolution to build sophisticated phylodynamic models. This enables the efficient analysis of large data sets using a combination of complex models, in large part owing to BEAST's integration with BEAGLE, a high-performance library for efficient phylogenetic likelihood calculation[15].

Inference in Bayesian phylogenetics relies on Markov chain Monte Carlo (MCMC) methods to sample from the posterior distribution[16,17]. MCMC methods simulate a correlated sample that converges in distribution to the posterior. The efficiency of a given MCMC algorithm depends on the transition kernel, which proposes a new simulated value based upon the current simulated value. Standard random walk transition kernels propose new values in a relatively blind fashion and generally update only one component of the multi-dimensional parameter space at a time. This can lead to a very slow, inefficient exploration of the posterior distribution where the MCMC algorithm would have to run for a relatively large number of iterations in order to simulate a suitable sample.

Fortunately, sophisticated, tailor-made transition kernels can often sample from the posterior much more efficiently. Made possible by the specific structure of the model, the skygrid adapts a highly efficient block-updating MCMC (BUMCMC) sampling scheme[18] that simultaneously proposes new values for the GMRF precision parameter and the effective population size values that correspond to the different levels of the piecewise constant demographic function. The GMRF smoothing prior along with the nonparametric coalescent likelihood gives the full conditional density of the effective population size in the form of a hidden Markov random field, allowing us to efficiently sample from its Gaussian approximation[19].

Hamiltonian Monte Carlo (HMC)[20,21] is an MCMC sampling scheme that bears some similarities to the BUMCMC sampler in that it aims to efficiently explore high probability regions of the posterior distribution and update all dimensions of the model parameter space simultaneously. HMC proceeds by introducing fictitious auxiliary "momentum" variables and reduces simulating from the posterior distribution to a matter of tracing Hamiltonian dynamics. While HMC's theoretical basis in differential geometry initially hindered its adoption, it has emerged in recent years as a widely-used method in statistical computing[22]. While adapting HMC to optimize the search through tree space is currently not possible, Dinh *et al.*[23] have developed an approach to sample from distributions on spaces with intricate combinatorial structure (such as for phylogenetic tree space). Applications of HMC in the field of Bayesian phylogenetics have started to emerge that focus on efficiently estimating classes of model parameters. Recently, Ji *et al.*[24] developed a linear-time algorithm for $\mathcal{O}(N)$-dimensional gradient evaluation – where N is the number of sampled molecular sequences – and showed HMC to greatly improve inference efficiency of branch-specific evolutionary rates.

Here, we present the implementation of an HMC transition kernel for the nonparametric skygrid coalescent model and compare its performance to the BUMCMC sampler. We compare the performance on three real viral data sets and find that in all cases HMC more efficiently explores the posterior distribution of skygrid model parameters. In some instances, the improvement afforded by HMC is especially striking, generating effectively independent posterior samples over five times as fast as BUMCMC.

Methods

The skygrid nonparametric coalescent model

The skygrid posits that demographic function $N_e(t)$ is a piecewise constant function that can change values only at pre-specified points in time known as "grid points." As in Gill *et al.*[7], let $x_1, \ldots, x_M$ denote the temporal grid points, where $x_1 \leq x_2 \leq \cdots \leq x_{M-1} \leq x_M$. The M grid points divide the demographic history timeline into $M + 1$ intervals so that the demographic function is fully specified by a vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{M+1})$ of values that it assumes on those intervals. Here, $N_e(t) = \theta_k$ for $x_{k-1} \leq t < x_k$, $k = 1, \ldots, M$, where it is understood that $x_0 = 0$. Also, $N_e(t) = \theta_{M+1}$ for $t \geq x_M$. Note that x_M is the time furthest back into the past at which the effective population size can change. The values of the grid points as well as the number M of total grid points are specified beforehand by the user. A typical way to select the grid points is to decide on a resolution M, let x_M assume the value furthest back in time for which the data are expected to be informative, and space the remaining grid points evenly between $x_0 = 0$ and x_M. Alternatively, as discussed in the next section, grid points can be selected to align with covariate sampling times in order to facilitate the modeling of associations between the log effective population size and external covariates.

Suppose we have m known genealogies $g_1, \ldots, g_m$ representing the ancestries of samples from m separate genetic loci with the same effective population size $N_e(t)$. We assume *a priori* that the genealogies are independent given $N_e(t)$. This assumption implies that the genealogies are unlinked which commonly occurs when researchers select loci from whole genome sequences or when recombination is very likely, such as between genes in retroviruses. The likelihood of the vector $\mathbf{g} = (g_1, \ldots, g_m)$ of genealogies can then be expressed as the product of likelihoods of individual genealogies:

$$P(\mathbf{g} \mid \boldsymbol{\theta}) = \prod_{i=1}^{m} P(g_i \mid \boldsymbol{\theta}). \tag{1}$$

To construct the likelihood of genealogy g_i, let t_{0_i} be the most recent sampling time of sequences contributing to genealogy i and t_{MRCA_i} be the time of the MRCA for locus i. Let x_{α_i} denote the minimal grid point greater than at least one sampling time in the genealogy, and x_{β_i} the greatest grid point less than at least one coalescent time. Let $u_{ik} = [x_{k-1}, x_k]$, $k = \alpha_i + 1, \ldots, \beta_i$, $u_{i\alpha_i} = [t_{0_i}, x_{\alpha_i}]$, and $u_{i(\beta_i+1)} = [x_{\beta_i}, t_{\mathrm{MRCA}_i}]$. For each u_{ik} we let t_{kj}, $j = 1, \ldots, r_k$, denote the ordered times of the grid points and sampling and coalescent events in the interval. With each t_{kj} we associate an indicator $\boldsymbol{\phi}_{kj}$ which takes a value of 1 in the case of a coalescent event and 0 otherwise. Finally, let v_{kj} denote the number of lineages present in the genealogy in the interval $[t_{kj}, t_{k(j+1)}]$. Following Griffiths and Tavaré[25], the likelihood of observing an interval is

$$P(u_{ik} \mid \theta_k) = \prod_{1 \le j < r_k : \phi_{kj}=1} \frac{v_{kj}(v_{kj}-1)}{2\theta_k} \times \prod_{j=1}^{r_k-1} \exp\left[-\frac{v_{kj}(v_{kj}-1)(t_{k(j+1)}-t_{kj})}{2\theta_k}\right] \tag{2}$$

for $k = \alpha_i, \ldots, \beta_i + 1$.

The product of interval likelihoods (2) yields the likelihood of coalescent times given the sampling times with genealogy g_i. To obtain the likelihood of the genealogy, however, we must account for the specific lineages that merge and result in coalescent events. Let $P_*(u_{ik} \mid \theta_k)$ denote $P(u_{ik} \mid \theta_k)$ except with factors of the form $\frac{v_{kj}(v_{kj}-1)}{2\theta_k}$ replaced by $\frac{2(2-1)}{2\theta_k} = \frac{1}{\theta_k}$. Then

$$P(g_i \mid \boldsymbol{\theta}) = \prod_{k=\alpha_i}^{\beta_i+1} P_*(u_{ik} \mid \theta_k). \tag{3}$$

We introduce some notation that will facilitate the derivation of a Gaussian approximation used to construct an MCMC transition kernel. If c_{ik} denotes the number of coalescent events which occur during interval u_{ik}, we can write

$$P(g_i \mid \boldsymbol{\theta}) = \prod_{k=\alpha_i}^{\beta_i+1} \left(\frac{1}{\theta_k}\right)^{c_{ik}} \exp\left[-\frac{S_{ik}}{\theta_k}\right], \tag{4}$$

where the S_{ik} are sufficient statistics from the genealogy. Rewriting this expression in terms of $\gamma_k = \log(\theta_k)$, we arrive at

$$P(g_i \mid \boldsymbol{\gamma}) = \prod_{k=\alpha_i}^{\beta_i+1} e^{-\gamma_k c_{ik}} \exp\left[-S_{ik} e^{-\gamma_k}\right] \tag{5}$$

$$= \prod_{k=\alpha_i}^{\beta_i+1} \exp\left[-\gamma_k c_{ik} - S_{ik} e^{-\gamma_k}\right]. \tag{6}$$

Invoking conditional independence of genealogies, the likelihood of the vector $\mathbf{g}$ of genealogies is

$$P(\mathbf{g} \mid \boldsymbol{\gamma}) = \prod_{i=1}^{m} P(g_i \mid \boldsymbol{\gamma}) \tag{7}$$

$$= \prod_{i=1}^{m} \prod_{k=\alpha_i}^{\beta_i+1} \exp\left[-\gamma_k c_{ik} - S_{ik} e^{-\gamma_k}\right] \tag{8}$$

$$= \exp\left[\sum_{k=1}^{M+1} \left[-\gamma_k c_k - S_k e^{-\gamma_k}\right]\right] \tag{9}$$

where $c_k = \sum_{i=1}^{m} c_{ik}$ and $S_k = \sum_{i=1}^{m} S_{ik}$; here, $c_{ik} = S_{ik} = 0$ if $k \notin [\alpha_i, \beta_i + 1]$.

The skygrid incorporates the prior assumption that effective population size changes continuously over time by placing a GMRF prior on $\boldsymbol{\gamma}$:

$$P(\boldsymbol{\gamma} \mid \tau) \propto \tau^{M/2} \exp\left[-\frac{\tau}{2}\sum_{i=1}^{M}(\gamma_{i+1}-\gamma_i)^2\right]. \tag{10}$$

This prior does not inform the overall level of the effective population size, just the smoothness of the trajectory. One can think of the prior as a first-order unbiased random walk with normal increments. The precision parameter τ determines how much differences between adjacent log effective population size values are penalized. We assign τ a gamma prior:

$$P(\tau) \propto \tau^{a-1} e^{-b\tau}. \tag{11}$$

In absence of prior knowledge about the smoothness of the effective population size trajectory, we choose $a = b = 0.001$ so that it is relatively uninformative. Conditioning on the vector of genealogies, we obtain the posterior distribution

$$P(\boldsymbol{\gamma}, \tau \mid \mathbf{g}) \propto P(\mathbf{g} \mid \boldsymbol{\gamma}) P(\boldsymbol{\gamma} \mid \tau) P(\tau). \tag{12}$$

Block-updating Markov chain Monte Carlo sampling

The original implementation of the skygrid adapts a BUMCMC sampling scheme for GMRFs[18] to sample $\boldsymbol{\gamma}$ and $\boldsymbol{\tau}$ when approximating the posterior (12). First, consider the full conditional density

$$P(\gamma|\mathrm{g},\tau) \propto P(\mathrm{g}|\gamma)P(\gamma|\tau) \\ \propto \tau^{M/2}\exp\left[-\frac{\tau}{2}\gamma'\mathrm{Q}\gamma-\sum_{k=1}^{M+1}\left(\gamma_k c_k+S_k e^{-\gamma_k}\right)\right]. \tag{13}$$

Let $h_k(\gamma_k) = (\gamma_k c_k + S_k e^{-\gamma_k})$. We can approximate each term $h_k(\gamma_k)$ by a second-order Taylor expansion about, say, $\hat{\gamma}_k$:

$$\begin{aligned} h_k(\gamma_k) \approx\ & h_k(\hat{\gamma}_k)+h_k'(\hat{\gamma}_k)(\gamma_k-\hat{\gamma}_k)+\frac{1}{2}h_k''(\hat{\gamma}_k)(\gamma_k-\hat{\gamma}_k)^2 \\ =\ & S_k e^{-\hat{\gamma}_k}\left(\frac{1}{2}\hat{\gamma}_k^{\,2}+\hat{\gamma}_k+1\right) \\ & +\left[c_k-S_k e^{-\hat{\gamma}_k}-S_k e^{-\hat{\gamma}_k}\,\hat{\gamma}_k\right]\gamma_k \\ & +\left[\frac{1}{2}S_k e^{-\hat{\gamma}_k}\right]\gamma_k^2. \end{aligned} \tag{14}$$

We center the Taylor expansion about a point $\hat{\gamma} = (\hat{\gamma}_1,\ldots,\hat{\gamma}_{M+1})$ obtained iteratively by the Newton-Raphson method:

$$\gamma_{(n+1)} = \gamma_{(n)} - [d^2 f(\gamma_{(n)})]^{-1}(df(\gamma_{(n)}))' \tag{15}$$

with $\boldsymbol{\gamma}_{(0)} = \boldsymbol{\gamma}^{(n)}$ the current value of $\boldsymbol{\gamma}$. Here,

$$f(\boldsymbol{\gamma}) = -\frac{1}{2}\boldsymbol{\gamma}'\tau\mathbf{Q}\boldsymbol{\gamma}-\sum_{k=1}^{M+1}\left(\gamma_k c_k+S_k e^{-\gamma_k}\right). \tag{16}$$

Replacing the terms $h_k(\gamma_k)$ with their Taylor expansions yields the following second-order Gaussian approximation to the full conditional density $P(\boldsymbol{\gamma}|\mathbf{g}, \tau)$:

$$\begin{aligned} P(\boldsymbol{\gamma}\,|\,\mathbf{g},\tau)\ \approx\ & \tau^{M/2}\exp\left[-\frac{1}{2}\boldsymbol{\gamma}'\left[\tau\mathbf{Q}+\mathrm{Diag}\left(S_k e^{-\hat{\gamma}_k}\right)\right]\right] \\ & -\exp\left[\sum_{k=1}^{M+1}\left(c_k-S_k e^{-\hat{\gamma}_k}-S_k e^{-\hat{\gamma}_k}\hat{\gamma}_k\right)\gamma_k\right], \end{aligned} \tag{17}$$

where Diag(·) is a diagonal matrix.

Starting from current parameter values $(\gamma^{(n)}, \tau^{(n)})$, we first generate a candidate value for the precision, $\tau^* = \tau^{(n)} f$, where f is drawn from a symmetric proposal distribution with density $P(f) \propto f + \frac{1}{f}$ defined on $[1/F, F]$. The tuning constant F controls the distance between the proposed and current values of the precision. Next, conditional on τ^*, we propose a new state γ^* using the Gaussian approximation (17) to the full conditional density $P(\gamma|\mathbf{g}, \tau^*)$. In the final step, the candidate state (τ^*, γ^*) is accepted or rejected according to the Metropolis-Hastings ratio[16,17].

Hamiltonian Monte Carlo sampling

HMC can be applied to most problems with continuous parameter spaces and produces distant proposals for the Metropolis algorithm[16] that nevertheless enjoy a high probability of acceptance. This enables efficient MCMC sampling by avoiding the slow exploration of the state space that accompanies simple random-walk proposals. Consider a d-dimensional *position* vector $\boldsymbol{\phi}$. This is the parameter whose posterior distribution we wish to sample from, and in the case of the skygrid, we have $\boldsymbol{\phi} = (\gamma, \tau, \boldsymbol{\beta})$. HMC proceeds by introducing a d-dimensional vector of auxiliary *momentum* variables $\mathbf{p}$ and sampling from the product distribution $\pi(\boldsymbol{\phi}, \mathbf{p}) = \pi(\boldsymbol{\phi})\pi(\mathbf{p})$ by simulating Hamiltonian dynamics. The Hamiltonian function is defined as

$$H(\boldsymbol{\phi},\mathrm{p}) = U(\boldsymbol{\phi}) + K(\mathrm{p}), \tag{18}$$

where $U(\boldsymbol{\phi})$, the *potential energy*, is defined as the negative log density of the position vector $\boldsymbol{\phi}$ and $K(\mathbf{p})$, the *kinetic energy* of the momentum variable $\mathbf{p}$ is defined as $K(\mathbf{p}) = \mathbf{p}^T \mathbf{M}^{-1}\mathbf{p}/2$, where $\mathbf{M}$ is a symmetric, positive-definite (usually diagonal) matrix known as the "mass matrix." For $\mathbf{p}$, we make the common assumption that it follows a multivariate normal distribution $\mathbf{p} \sim N(\mathbf{0}, \mathbf{M})$. It has become standard in basic HMC implementations to set $\mathbf{M} = \mathbf{I}$, but we will discuss a more informed choice later.

HMC generates a Metropolis proposal from the current state $(\boldsymbol{\phi}_0, \mathbf{p}_0)$ in the space $(\boldsymbol{\phi}, \mathbf{p})$ that evolves according to Hamilton's equations:

$$\begin{aligned} \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} &= -\nabla U(\boldsymbol{\phi}) = \nabla\log\pi(\boldsymbol{\phi}) \\ \frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}t} &= \nabla K(\mathbf{p}) = \mathrm{M}^{-1}\mathbf{p}. \end{aligned} \tag{19}$$

The *leapfrog* method to approximate a solution to Equation 19 performs the following updates for each of n leapfrog steps of size ϵ:

$$\begin{aligned} \mathbf{p}_{t+\epsilon/2} &= \mathbf{p}_t + \frac{\epsilon}{2}\nabla\log\pi(\boldsymbol{\phi}_t) \\ \boldsymbol{\phi}_{t+\epsilon} &= \boldsymbol{\phi}_t + \epsilon\mathbf{M}^{-1}\mathbf{p}_{t+\epsilon/2} \\ \mathbf{p}_{t+\epsilon} &= \mathbf{p}_{t+\epsilon/2} + \frac{\epsilon}{2}\nabla\log\pi(\boldsymbol{\phi}_{t+\epsilon}). \end{aligned} \tag{20}$$

The use of HMC therefore requires the user to specify these two parameters, i.e. the step size ϵ and the number of steps n. In addition, we assume a standard HMC transition kernel by employing an identity matrix $\mathbf{I}$ for the mass matrix $\mathbf{M}$. Through its internal auto-tuning capabilities, BEAST 1.10 enables tuning ϵ during an ongoing analysis, but n still needs to be provided by the user.

Data

We compare the performance of the BUMCMC and HMC transition kernels for the skygrid on three real data sets. First, we analyse the population dynamics of the rabies epizootic among raccoons in the northeastern United States starting in the late 1970s[26]. The sequence data consist of 47 sequences sampled from rabid raccoons between 1982 and 2004 and encompass the complete rabies nucleoprotein (N) genes as well as large portions of the glycoprotein (G) genes. Based on a previous analysis[11], we set a cutoff for the skygrid of 40 years during which we estimate the log population size for 50 time intervals.

We assume a single HKY nucleotide substitution model[27] while accounting for among-site rate variation[28], and assume an uncorrelated relaxed molecular clock with an underlying lognormal distribution[29].

The second data set consists of 196 Ebola virus (EBOV) sequences originating from Sierra Leone, previously analysed in Hill and Baele[14]. Based on information obtained from a large-scale study of the West African Ebola virus outbreak in 2013–2016[30], we set the cutoff for the skygrid to one year, and we estimate the log population size for each week for a total of 52 log population size estimates. We partition the coding part of the data set according to codon position, and create a fourth data partition containing the aggregated intergenic region data[30]. For each of the four resulting partitions, we assume an HKY nucleotide substitution model[27] while accounting for among-site rate variation[28]. We assume an uncorrelated relaxed molecular clock with an underlying lognormal distribution, which is shared across all partitions[29].

The third and final data set consists of 300 coat protein gene sequences of rice yellow mottle virus (RYMV), sampled from East to West Africa over a 46-year period from 1966 to 2012[31]. We set a skygrid cutoff value of 200 years, and we estimate the log population size for 100 time intervals. We assume an HKY nucleotide substitution model[27] for the first and second codon positions combined, and another for the third codon position, and we assume among-site rate variation[28] on each of these two partitions. Finally, we again assume an uncorrelated relaxed molecular clock with an underlying lognormal distribution[29].

Analysis

All analyses were performed using BEAST 1.10.5[13], along with the high-performance BEAGLE 3.2.0 library[15] for computational efficiency. Our central processing unit (CPU) analyses were performed on a single 18-core Intel Xeon 6140 Skylake processor running at a clock speed of 2.3 GHz. The Ebola data set, however, necessitates the use of a powerful graphical processing unit (GPU), and these analyses were hence performed on an NVIDIA Tesla P100 SMX2 graphics card designed for scientific computing. The rabies and RYMV data sets were run for 50 million iterations, logging every 2,000 iterations, whereas the Ebola data set was run for 100 million iterations, also logging every 2,000 iterations. The posterior samples were used to construct a maximum clade credibility (MCC) tree for each data set using TreeAnnotator, discarding 10% of the samples except for the RYMV data set for which we discarded 20% of the samples.

To directly compare the performance of the different transition kernels on estimating the parameters of the Bayesian skygrid model, we performed an initial analysis that focused solely on estimating the log population size and precision parameters. To this end, we fixed the phylogeny to the MCC tree and the non-skygrid parameters to their mean posterior estimates. All data sets were analysed in BEAST for 20,000 iterations, logging every two iterations.

We evaluated the performance of the different transition kernels by computing the effective sample size (ESS) for each parameter of interest using the coda package[32] in CRAN R[33]. The ESS estimates the number of in-dependent draws from the posterior distribution that an MCMC sample is equivalent to by accounting for the autocorrelation in the sample[34]. Thus the ESS per unit time provides a measure of how efficiently a given transition kernel is sampling from the posterior distribution. We report the difference in ESS per unit time of the skygrid's precision parameter, as well as the minimum and median ESS values of the log population size parameters.

Results

Inference on a fixed phylogeny

We first compare the performance of the different transition kernels when solely estimating the skygrid parameters, fixing the phylogeny to the MCC tree and all other parameters to their posterior mean estimates. For the rabies data set, Figure 1 shows a pronounced improvement in performance of HMC over BUMCMC. HMC generates a 3.91-fold and 5.35-fold improvement in median and minimum ESS per second over BUMCMC for the log population sizes, and a 1.90-fold improvement in ESS per second for the precision.

Figure 2 and Figure 3 show the performance improvements brought about by HMC over BUMCMC for the EBOV and RYMV data sets, respectively. For the Ebola data set, HMC yields a 1.41-fold performance increase in median ESS per second for the log population sizes and a 5.47-fold increase for the precision over BUCMC, but the latter offers a 1.08-fold improvement in minimum ESS per second over the former. Finally, for the RYMV data set, compared to BUMCMC the HMC transition kernel yields a 2.05-fold and 2.77-fold relative speedup in ESS per second in terms of the median and minimum ESS per second of the log population sizes respectively, while generating a 3.67-fold relative speedup in ESS per second for the precision. In conclusion, when focusing solely on estimation of the skygrid's parameters on a fixed phylogeny, HMC consistently reports substantial performance increases in estimating these parameters, with the magnitude of these improvements being dependent on the specific data set and the balance between the number of population sizes and the number of taxa available in the data set.

Joint inference

The improvements under HMC that we observe in analyses that solely infer skygrid parameters are mirrored in more comprehensive analyses that jointly infer the phylogeny and all other model parameters, which constitutes the most common use case for this model. Figure 4 shows a substantial performance increase in ESS per minute of HMC over BUMCMC for the rabies data set. Employing HMC results in a 3.38-fold and 1.51-fold relative speedup in the median and minimum ESS per minute, respectively – over BUCMC for the log population sizes and a 3.99-fold speedup for the precision.

For the EBOV data set, Figure 5 also shows a clear performance benefit of the HMC transition kernel over BUCMC, reporting a 1.48-fold and 1.35-fold speedup in median and minimum ESS per minute for the log population sizes and a 1.56-fold speedup for the precision. For the RYMV data set, the performance improvements of HMC over BUMCMC are

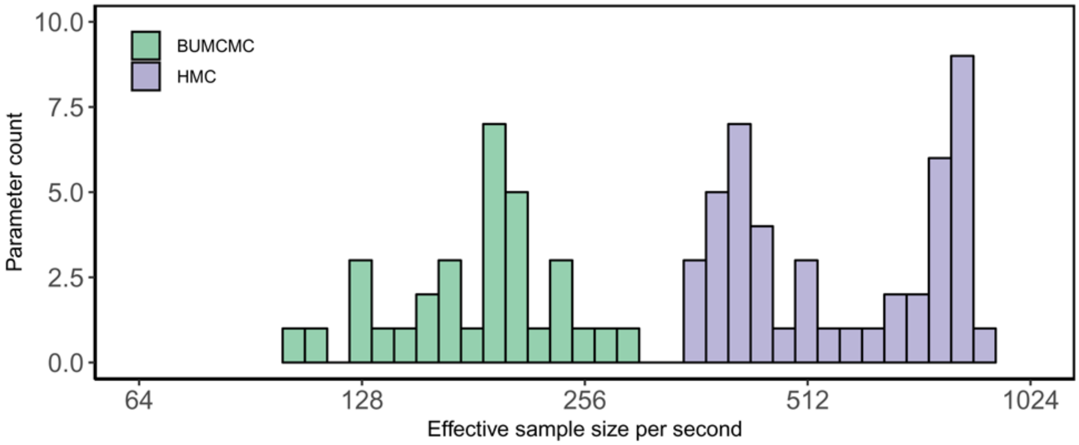

Figure 1. Rabies data set - fixed tree analysis. Bars correspond to the estimated effective sample size (ESS) per second averaged across five independent replicates for all log population size parameters, using the block-updating Markov chain Monte Carlo (BUMCMC) and Hamiltonian Monte Carlo (HMC) transition kernels. The height of each bar indicates the number of parameters that achieve the given ESS per second value. The HMC transition kernel improves upon the performance of the BUMCMC transition kernel by factors of 5.35 and 3.91 for the minimum and median ESS per second across all log population sizes.

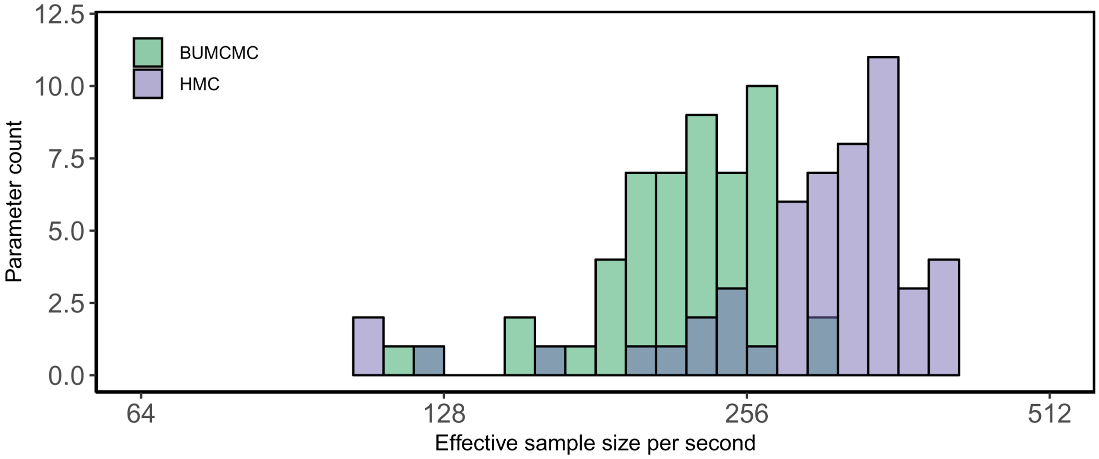

Figure 2. Ebola virus data set - fixed tree analysis. Bars correspond to the estimated effective sample size (ESS) per second averaged across five independent replicates for all log population size parameters, using the blockupdating Markov chain Monte Carlo (BUMCMC) and Hamiltonian Monte Carlo (HMC) transition kernels. The height of each bar indicates the number of parameters that achieve the given ESS per second value. The HMC transition kernel improves upon the performance of the BUMCMC transition kernel by a factor of 1.41 for the median ESS per second but BUMCMC yields a 1.08-fold improvement over HMC for the minimum ESS per second.

more modest, with Figure 6 showing near-identical performance between HMC and BUMCMC. HMC yields a 1.07-fold speedup in terms of minimum ESS per hour over BUMCMC for the log population sizes, whereas BUMCMC in turn yields a 1.08-fold improvement in median ESS per hour over HMC. Estimation efficiency of the skygrid's precision is however 1.45-fold faster using HMC compared to BUMCMC.

Discussion

Coalescent-based models that relate population dynamics to genealogical shapes are central to phylogenetic and phylodynamic inference. The increasing availability of large molecular sequence data sets is testing the limits of current Bayesian phylogenetic inference software, and estimation procedures that can scale efficiently are critically important. In statistics, HMC has emerged as one of the most powerful approaches in MCMC sampling, opening the door to more efficient exploration of high-dimensional distributions through accounting for the distribution's geometric structure. Here, we have evaluated the utility of HMC for posterior inference for the skygrid coalescent model.

In analyses of rabies, Ebola virus, and RYMV data sets, we observe that HMC consistently outperforms the standard skygrid BUMCMC transition kernel in terms of more efficiently generating effectively independent samples of skygrid model parameters.

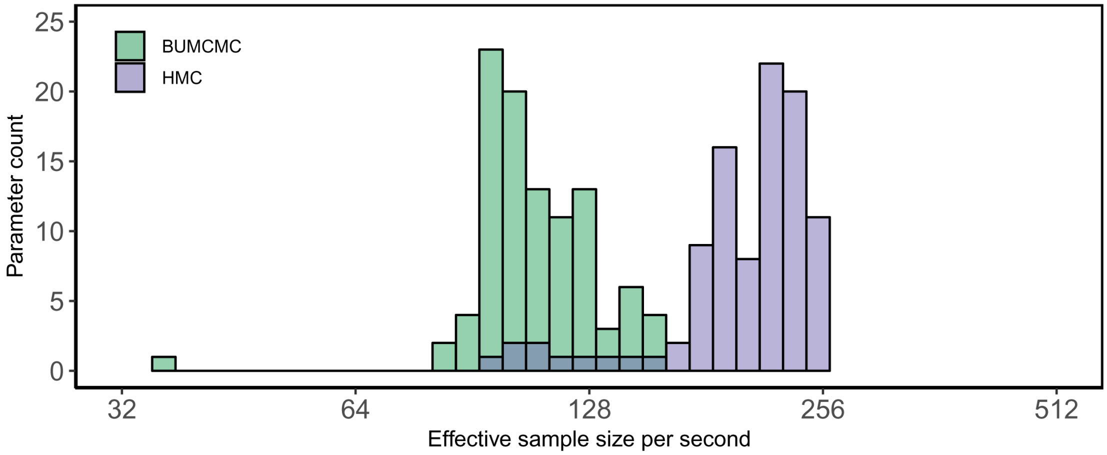

Figure 3. Rice yellow mottle virus data set - fixed tree analysis. Bars correspond to the estimated effective sample size (ESS) per second averaged across five independent replicates for all log population size parameters, using the block-updating Markov chain Monte Carlo (BUMCMC) and Hamiltonian Monte Carlo (HMC) transition kernels. The height of each bar indicates the number of log population size parameters that achieve the given ESS per second value. The HMC transition kernel improves upon the performance of the BUMCMC transition kernel by factors of 2.05 and 2.77 for the median and minimum ESS per second, respectively.

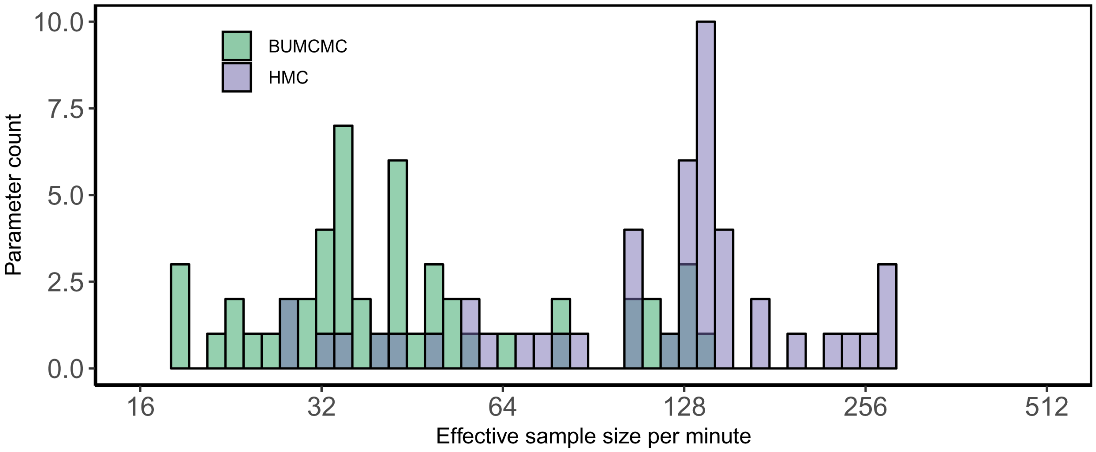

Figure 4. Rabies data set analysis using the skygrid coalescent model on a central processing unit (CPU). Bars correspond to the estimated effective sample size (ESS) per second averaged across five independent replicates for all log population size parameters and the precision parameter, using the block-updating Markov chain Monte Carlo (BUMCMC) and Hamiltonian Monte Carlo (HMC) transition kernels. The height of each bar indicates the number of skygrid parameters that achieve the given ESS per minute value. The HMC transition kernel improves upon the performance of the BUMCMC transition kernel by factors of 3.38 and 1.51 for the median and minimum ESS per minute, respectively, while a 3.99-fold improvement for the precision was generated.

For some data sets and model parameters, HMC is over five times as efficient as BUMCMC. Such gains in efficiency are especially valuable considering the increasingly important role that phylodynamic inference methods have assumed in real-time analysis of outbreak dynamics[35]. Advances in portable genomic sequencing capabilities have enabled sequencing during outbreaks in close to real-time[36], and phylodynamic inferences can potentially inform outbreak response efforts by public health apparatuses, provided that such inferences are made available in a timely manner.

Further performance improvements for the proposed HMC transition kernel may be achieved by replacing the standard choice of $\mathbf{I}$ with a more informative matrix $\mathbf{M}$, which is equivalent to preconditioning the posterior distribution by transforming the parameters ϕ. Girolami and Calderhead[37] suggest the negative Hessian as an alternative that better accounts for the target distribution's underlying geometry. However, such an approach is computationally prohibitive, necessitating numerical integrators that require several iterations of calculating and inverting the mass matrix at each step. Recently, however, Ji *et al.*[24]

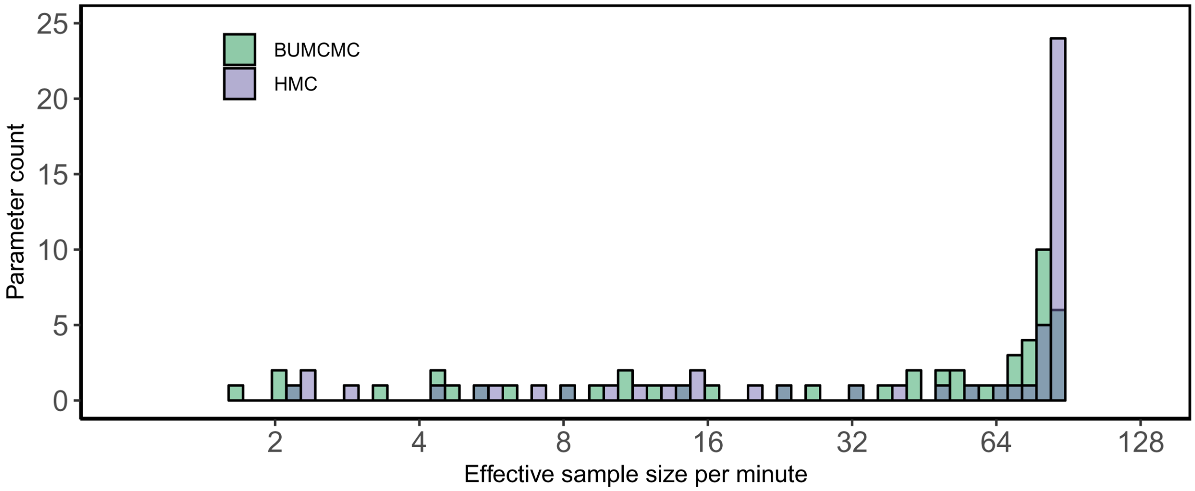

Figure 5. Ebola virus data set analysis using the skygrid coalescent model on a graphical processing unit (GPU). Bars correspond to the estimated effective sample size (ESS) per second averaged across five independent replicates for all log population size parameters and the precision parameter, using the block-updating Markov chain Monte Carlo (BUMCMC) and Hamiltonian Monte Carlo (HMC) transition kernels. The height of each bar indicates the number of skygrid parameters that achieve the given ESS per minute value. The HMC transition kernel improves upon the performance of the BUMCMC transition kernel by factors of 1.48 and 1.35 for the median and minimum ESS per minute, and a factor of 1.56 for the precision.

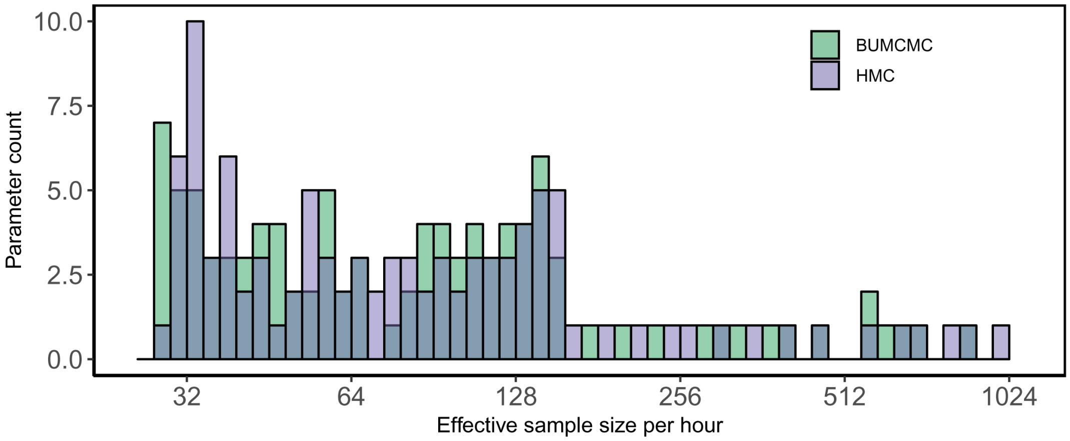

Figure 6. Rice yellow mottle virus data set analysis using the skygrid coalescent model on a central processing unit (CPU). Bars correspond to the estimated effective sample size (ESS) per hour averaged across five independent replicates for all log population size parameters and the precision parameter, using the block-updating Markov chain Monte Carlo (BUMCMC) and Hamiltonian Monte Carlo (HMC) transition kernels. The height of each bar indicates the number of skygrid parameters that achieve the given ESS per hour value. The HMC transition kernel improves upon the performance of the BUMCMC transition kernel by factors of 1.07 and 1.45 for the minimum ESS per hour of the log population sizes and the precision, respectively. In turn, BUMCMC yield a 1.08-fold improvement over HMC for the median ESS per hour of the log population sizes.

put forth a method to adaptively tune the mass matrix to estimate the expected Hessian averaged over the posterior distribution and avoid excessive computational burden. Extending the HMC approach for the skygrid model by tuning the mass matrix is the subject of future work.

The performance improvements that we see under HMC are also very encouraging in the context of further development and use of HMC methods in Bayesian phylogenetics and phylodynamics. It is particularly notable that a standard HMC implementation outperforms the BUMCMC transition kernel that was specifically designed for GMRF models and relied on and exploited many aspects of the skygrid model structure. In that regard, the performance improvements reported here are not directly comparable to those in the work of Ji *et al.*[24], who reported massive performance gains when comparing HMC transition kernels to simple univariate transition kernels. This illustrates the power of HMC and its potential for allowing statisticians to avoid developing estimation procedures that, while efficient, may only apply to a narrow range of models. Extensions to standard HMC that seek to improve sampling efficiency by better accounting for the posterior distribution's geometric structure[21,37,38] and optimizing path lengths for numerical solutions of Hamiltonian dynamics[39,40] offer further improvements and illustrate the need for continued development.

Data availability

Underlying data

Zenodo: GuyBaele/skygrid_hmc_data: First release of BEAST XML files for skygrid+HMC. https://doi.org/10.5281/zenodo.3715408[41]

This project contains the following underlying data:

- Rabies dataset BEAST 1.10.5 XML files using both BUMCMC and HMC transition kernels
- Ebola dataset BEAST 1.10.5 XML files using both BUMCMC and HMC transition kernels
- RYMV dataset BEAST 1.10.5 XML files using both BUMCMC and HMC transition kernels

Data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

The rabies virus sequences were originally published by Biek *et al.*[26], DOI: https://doi.org/10.1073/pnas.0700741104. The Ebola virus sequences were originally published by Dudas *et al.*[30], DOI: https://doi.org/10.1038/nature22040, and the subset analysed here was created in Hill and Baele[14], DOI: https://doi.org/10.1093/molbev/msz172. The rice yellow mottle virus sequences were originally published by Trovão *et al.*[31], DOI: https://doi.org/10.1093/ve/vev016.

References

1. Kingman JFC: **On the genealogy of large populations.** *J Appl Probab.* 1982; **19**(A): 27–43.
Publisher Full Text

2. Pybus OG, Rambaut A, Harvey PH: **An integrated framework for the inference of viral population history from reconstructed genealogies.** *Genetics.* 2000; **155**(3): 1429–1437.
PubMed Abstract | Free Full Text

3. Strimmer K, Pybus OG: **Exploring the demographic history of DNA sequences using the generalized skyline plot.** *Mol Biol Evol.* 2001; **18**(12): 2298–2305.
PubMed Abstract | Publisher Full Text

4. Drummond AJ, Rambaut A, Shapiro B, *et al.*: **Bayesian coalescent inference of past population dynamics from molecular sequences.** *Mol Biol Evol.* 2005; **22**(5): 1185–1192.
PubMed Abstract | Publisher Full Text

5. Opgen-Rhein R, Fahrmeir L, Strimmer K: **Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo.** *BMC Evol Biol.* 2005; **5**: 6.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Minin VM, Bloomquist EW, Suchard MA: **Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics.** *Mol Biol Evol.* 2008; **25**(7): 1459–1471.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Gill MS, Lemey P, Faria NR, *et al.*: **Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci.** *Mol Biol Evol.* 2013; **30**(3): 713–724.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Palacios JA, Minin VN: **Gaussian process-based Bayesian nonparametric inference of population size trajectories from gene genealogies.** *Biometrics.* 2013; **69**(1): 8–18.
PubMed Abstract | Publisher Full Text

9. Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach.** *J Mol Evol.* 1981; **17**(6): 368–376.
PubMed Abstract | Publisher Full Text

10. Felsenstein J: **Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci?** *Mol Biol Evol.* 2006; **23**(3): 691–700.
PubMed Abstract | Publisher Full Text

11. Gill MS, Lemey P, Bennett SN, *et al.*: **Understanding past population dynamics: Bayesian coalescent-based modeling with covariates.** *Syst Biol.* 2016; **65**(6): 1041–1056.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Faria NR, Kraemer MUG, Hill SC, *et al.*: **Genomic and epidemiological monitoring of yellow fever virus transmission potential.** *Science.* 2018; **361**(6405): 894–899.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Suchard MA, Lemey P, Baele G, *et al.*: **Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10.** *Virus Evol.* 2018; **4**(1): vey016.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Hill V, Baele G: **Bayesian estimation of past population dynamics in BEAST 1.10 using the Skygrid coalescent model.** *Mol Biol Evol.* 2019; **36**(11): 2620–2628.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Ayres DL, Cummings MP, Baele G, *et al.*: **BEAGLE 3: Improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics.** *Syst Biol.* 2019; **68**(6): 1052–1061.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Metropolis N, Rosenbluth AN, Rosenbluth MN, *et al.*: **Equation of state calculations by fast computing machines.** *J Chem Phys.* 1953; **21**: 1087–1092.
Publisher Full Text

17. Hastings WK: **Monte Carlo sampling methods using Markov chains and their applications.** *Biometrika.* 1970; **57**(1): 97–109.
Publisher Full Text

18. Knorr-Held L, Rue H: **On block updating in Markov random field models for desease mapping.** *Scand J Stat.* 2002; **29**(4): 597–614.
Publisher Full Text

19. Rue H: **Fast sampling of Gaussian Markov random fields.** *J R Stat Soc B.* 2001; **63**(2): 325–338.
Publisher Full Text

20. Duane S, Kennedy AD, Pendleton BJ, *et al.*: **Hybrid Monte Carlo.** *Phys Lett B.* 1987; **195**(2): 216–222.
Publisher Full Text

21. Neal RM: **MCMC using Hamiltonian dynamics.** *Handbook of Markov Chain Monte Carlo.* 2010; **54**: 113–162.
Reference Source

22. Betancourt M: **A conceptual introduction to Hamiltonian Monte Carlo.** 2017.
Reference Source

23. Dinh V, Bilge A, Zhang C, *et al.*: **Probabilistic path Hamiltonian Monte Carlo.** In *Proceedings of the 34th International Conference on Machine Learning.* 2017; **70**: 1009–1018.
Reference Source

24. Ji X, Zhang Z, Holbrook A, *et al.*: **Gradients do grow on trees: a linear-time O(N)dimensional gradient for statistical phylogenetics**. 2019.
Reference Source

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

26. Biek R, Henderson JC, Waller LA, *et al.*: **A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus.** *Proc Natl Acad Sci U S A.* 2007; **104**(19): 7993–7998.
PubMed Abstract | Publisher Full Text | Free Full Text

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

28. Yang Z: **Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods.** *J Mol Evol.* 1994; **39**(11): 306–314.
PubMed Abstract | **Publisher Full Text**

29. Drummond AJ, Ho SY, Phillips MJ, *et al.*: **Relaxed phylogenetics and dating with confidence.** *PLoS Biol.* 2006; **4**(5): e88.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

30. Dudas G, Carvalho LM, Bedford T, *et al.*: **Virus genomes reveal factors that spread and sustained the ebola epidemic.** *Nature.* 2017; **544**(7650): 309–315.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

31. Trovão NS, Baele G, Vrancken B, *et al.*: **Host ecology determines the dispersal patterns of a plant virus.** *Virus Evol.* 2015; **1**(1): vev016.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

32. Plummer M, Best N, Cowles K, *et al.*: **CODA: Convergence Diagnosis and Output Analysis for MCMC.** *R News.* 2006; **6**(1): 7–11.
Reference Source

33. R Core Team: **R: A Language and Environment for Statistical Computing.** R Foundation for Statistical Computing, Vienna, Austria, 2018.
Reference Source

34. Kass RE, Carlin BP, Gelman A, *et al.*: **Markov Chain Monte Carlo in practice: a roundtable discussion.** *Am Stat.* 1998; **52**(2): 93–100.
Publisher Full Text

35. Gardy J, Loman NJ, Rambaut A: **Real-time digital pathogen surveillance the time is now.** *Genome Biol.* 2015; **16**(1): 155.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

36. Quick J, Loman NJ, Durrafour S, *et al.*: **Real-time, portable genome sequencing for Ebola surveillance.** *Nature.* 2016; **530**(7589): 228–232.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

37. Girolami M, Calderhead B: **Riemann manifold Langevin and Hamiltonian Monte Carlo methods.** *J R Stat Soc B.* 2011; **73**: 123–214.
Publisher Full Text

38. Nishimura A, Dunson D: **Geometrically tempered Hamiltonian Monte Carlo**. 2016.
Reference Source

39. Hoffman MD, Gelman A: **The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo.** *J Mach Learn Res.* 2014; **15**: 1593–1623.
Reference Source

40. Wu C, Stoehr J, Robert CP: **Faster Hamiltonian Monte Carlo by learning leapfrog scale.** 2019.
Reference Source

41. Baele G: **GuyBaele/skygrid_hmc_data: First release of BEAST XML files for skygrid+HMC.** 2020.
http://www.doi.org/10.5281/zenodo.3715408

Open Peer Review

Current Peer Review Status: ✓ ? ?

Version 1

Reviewer Report 01 September 2020

https://doi.org/10.21956/wellcomeopenres.17296.r39859

?

Nicola Müller
Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Baele *et al*. introduce a Hamiltonian Markov Chain Operator to more efficiently explore non-parametric effective population size dynamics in BEAST. While not completely novel, this presents a useful addition to users of BEAST. The paper is overall well written and particularly the comparison of ESS per second on fixed trees nicely shows the advantages of using an HMC operator. The comparison when jointly inferring the trees alongside the Ne dynamics are a bit less convincing. I would suggest to add some test about what is limiting convergence in this scenario and to compare the performance of the HMC operator to the adaptive multivariate gaussian operator implemented by some of the same authors. Also, the convergence criteria should probably be changed (at least for the joint inferences).

Major comments:

- 5 runs are too few to compare ESS values across runs. Especially, to put an actual number on the fold increase of ESS values between the different operators. I would also argue that the ESS for a log Ne is not the most important ESS and would instead use the ESS of the posterior probabilities, in particular for the joint inferences. I would assume that the last few time intervals of the skygrid are often essentially sampling from the prior. So, by comparing ESS values of individual log Ne's, at least partly, the comparison is in how good the two operators are in sampling from the prior. There should also be some comparison of variance of ESS between runs to compare how reliable the fold increases are.

- The major limitation of joint analyses with the phylogenetic trees are typically the tree inference themselves. Operations on the trees are often also substantially more computationally expensive, while operations on the Ne's are fairly cheap (recomputing tree likelihoods is more expensive than recomputing the coalescent probability). This, in turn, would suggest that the weights of inefficient operators on the Ne trajectories could simply be increased without leading to a lot of extra computation time. I would suggest adding an experiment where the weights of the Ne operators are increased and ESS per unit of time are compared again.

- I've used the adaptive multivariate gaussian operator by Beale *et al.* for skyline type analyses. This one additionally provides the advantage of "learning" the correlation structure with other parameters (e.g. the clock rate). Is there any advantage of using the HMC approach instead of using the adaptive multivariate gaussian operator?

Minor comments:

- "In a Bayesian framework, coalescent models function ...." (Long and heavy sentence).

- "the skygrid can incorporate data from multiple unlinked genetic loci" same is true for EBSP.

- "For the rabies data set, Figure 1 shows a pronounced improvement" would replace "pronounced".

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bayesian Phylogenetics and Phylodynamics.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 13 August 2020

https://doi.org/10.21956/wellcomeopenres.17296.r39664

Shiwei Lan
Arizona State University, Tempe, AZ, USA

The submitted manuscript, *Hamiltonian Monte Carlo sampling to estimate past population dynamics using the skygrid coalescent model in a Bayesian phylogenetics framework*, applies Hamiltonian Monte Carlo (HMC) to estimate population dynamics for coalescent model in Bayesian phylogenetics.
While it is an important research question in the field and the paper provides interesting applications, it overlaps existing work thus affects its originality.
Gaussian approximation to HMC is not something new and has been extensively explored by, e.g. the following papers:

- Split Hamiltonian Monte Carlo (2014)[1]
- Semi-Separable Hamiltonian Monte Carlo for Inference in Bayesian Hierarchical Models (2014)[2]
- An efficient Bayesian inference framework for coalescent-based nonparametric phylodynamics (2015)[3]

Though not mentioned by the submitted paper, it shares substantial similarity with [3], which details and optimizes the splitting strategy to the almost same model. The only thing different is that the submitted work includes genealogies in the inference which is not present in [3].
[3] does a thorough study in this topic by comparing split HMC with many popular MCMC algorithms including MALA, adaptive MALA, elliptic slice sampler, and Bayesian Skyride as well.

The paper has the following issues.
('L' stands for line, and 'P' stands for page, negative number -XX means 'the last but XX'.)

Major issues:

- For all numeric experiments, how do their posterior estimates compare? It is not all about ESS per unit time. If the computationally light-weighted algorithm introduces more bias, or converges slowly, then we need to think of the trade off between efficiency and accuracy.
- Can you compare your method to other algorithms than BUMCMC?
- What does the blue-grey color stand for in the figures? It is not in the legend or explained.
- Did you repeat your experiment to reduce the sampling error?
- I am curious why collecting millions of samples? I understand that the estimate needs sufficient samples to reduce variance, but what is the computational cost?

Minor issues:

- P14L22 (below equation (1)): What is MRCA? Similarly HKY? Please consider giving the full name when you mention them for the first time.
- P5L-12 (below equation (17)): Could you please change the symbol f in $\tau^*=\tau^{(n)} f$ -- it is very easy to confuse with the function f as in equation (16) right above.
- P5L7: What is β in $\phi=(\gamma,\tau,\beta)$? Do you mean g?

References

1. Shahbaba B, Lan S, Johnson W, Neal R: Split Hamiltonian Monte Carlo. *Statistics and Computing*. 2014; **24** (3): 339-349 Publisher Full Text

2. Zhang Y, Sutton C: Semi-Separable Hamiltonian Monte Carlo for Inference in Bayesian Hierarchical Models. *Advances in Neural Information Processing Systems*. 2014. Reference Source

3. Lan S, Palacios JA, Karcher M, Minin VN, et al.: An efficient Bayesian inference framework for coalescent-based nonparametric phylodynamics.*Bioinformatics*. 2015; **31** (20): 3282-9 PubMed Abstract | Publisher Full Text

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Partly

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

No

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Statistical computing; Bayesian modeling; Uncertainty quantification

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 04 May 2020

https://doi.org/10.21956/wellcomeopenres.17296.r38402

✓ **Mario dos Reis**

School of Biological and Chemical Sciences, Queen Mary University of London, London, UK

The majority of Bayesian inference problems in phylogenetics are not analytically tractable. This means we must use the MCMC stochastic algorithm to obtain samples from the posterior distribution. The most widely used flavour of MCMC in phylogenetics is the Metropolis-Hastings sampler, which shows random walk behaviour and produces MCMC chains with high autocorrelation and high rejection rates, specially in problems with large number of parameters. The Hamiltonian Monte Carlo (HMC) sampler uses physics simulations to generate efficient MCMC proposals. In the HMC method, a particle is shot across the transformed posterior surface and the laws of physics are used to calculate the trajectory of the particle. The posterior is transformed in such a way that the particle becomes trapped in regions of high probability. Furthermore, by pushing the particle far enough, we ensure the particle lands far from its starting position. Thus, the HMC sampler tends to generate MCMC chains with high acceptance rates and low autocorrelation, making the sampler very efficient.

In phylogenetics, HMC has been used to estimate amino acid substitution matrices (Zhao *et al*. 2016)[1], to explore tree space (Dinh *et al*. 2017)[2] and to explore proposal efficiency in a two-species problem (Thawornwattana *et al*. 2018)[3]. Recently, preprints have appeared suggesting further applications (Ji *et al*. 2019; Bastide *et al*. 2020)[4,5]. Here, Baele *et al*. extend the application of HMC to sampling population parameters from the coalescent process in phylogenies.

This is a well thought out, novel application of HMC to phylogenetics. Baele *et al*. show that the new HMC sampler outperforms their previous BUCMC approach by roughly 2-5x when tested on several virus datasets. That is, HMC will generate effective sample sizes (ESS) for the coalescent parameters in up to less than one-fifth of the time required by the previous sampler (depending on the dataset). These are very impressive improvements, particularly given that phylogenomic analysis on large datasets can take several days to compute. But even more importantly, it appears that further improvements to the new HMC sampler can still be achieved: The authors used the identity matrix as the mass matrix of the particle. As they acknowledge, different choices for the mass matrix should be explored. In my experience, a well fine-tuned mass matrix can provide dramatic improvements on the efficiency of HMC samplers.

References

1. Zhao T, Wang Z, Cumberworth A, Gsponer J, et al.: Bayesian Analysis of Continuous Time Markov Chains with Application to Phylogenetic Modelling. *Bayesian Analysis*. 2016; **11** (4): 1203-1237 Publisher Full Text
2. Dinh V, Bilge A, Zhang C, , et al.: Probabilistic path Hamiltonian Monte Carlo. *Proceedings of the 34th International Conference on Machine Learning*. 2017. 1009-1018
3. Thawornwattana Y, Dalquen D, Yang Z: Designing Simple and Efficient Markov Chain Monte Carlo Proposal Kernels. *Bayesian Analysis*. 2018; **13** (4): 1037-1063 Publisher Full Text
4. Ji X, Zhang Z, Holbrook A, *et al.*: Gradients do grow on trees: a linear-time O(N)dimensional gradient for statistical phylogenetics. *arXiv:1905.12146*. 2019.
5. Bastide P, *et al.*: Efficient Bayesian Inference of General Gaussian Models on Large Phylogenetic Trees. *arXiv:2003.10336*. 2020.

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bayesian phylogenetics, molecular clock, divergence time estimation.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.