HAMILTONIAN MONTE CARLO SAMPLING FOR THE SKYGRID COALESCENT-BASED MODEL WITH COVARIATES

by

MANINDER KAUR

(Under the Direction of Mandev Singh Gill)

ABSTRACT

Phylodynamic inference has emerged as a powerful tool for research in evolutionary biology. Coalescent-based models are widely used in phylodynamics, furnishing prior distributions for phylogenetic trees in Bayesian models and enabling inference of the effective population size, an abstract parameter that characterizes genetic diversity and is of fundamental importance in evolutionary biology. The Skygrid model enables the integration of external covariates into a coalescent-based model and provides a framework to study the relationship between past population dynamics and potential driving factors. However, the Skygrid's complexity makes posterior approximation challenging, and there is a need for algorithms that can allow it to scale efficiently to large genomic data sets. Here, we evaluate the effectiveness of a promising Markov chain Monte Carlo method, Hamiltonian Monte Carlo, for the Skygrid model with covariates. Through an analysis of three data sets, we show that Hamiltonian Monte Carlo generally outperforms earlier approaches.

INDEX WORDS: Coalescent, Phylodynamics, Markov chain Monte Carlo, Hamiltonian Monte Carlo.

HAMILTONIAN MONTE CARLO SAMPLING FOR THE SKYGRID COALESCENT-BASED MODEL WITH COVARIATES

by

MANINDER KAUR

B.S., Guru Nanak Dev University, India, 2012

M.S., Guru Nanak Dev University, India, 2014

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2023

HAMILTONIAN MONTE CARLO SAMPLING FOR THE SKYGRID COALESCENT-BASED MODEL WITH COVARIATES

by

MANINDER KAUR

Major Professor:	Mandev Singh Gill
Committee:	Liang Liu
	Jaxk Reeves

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
December 2023

ACKNOWLEDGMENTS

First and foremost, I extend my deepest gratitude to my advisor, Dr. Mandev Singh Gill, whose invaluable guidance was instrumental in the completion of this research. His generosity with his expertise and time has significantly shaped my research journey.

Additionally, my heartfelt thanks go to the members of my advisory committee, Dr. Jaxk Reeves and Dr. Liang Liu. Their insights and assistance have been crucial throughout the various stages of my degree, helping to navigate the challenges encountered along this academic path.

Beyond the realm of academic guidance, the journey through graduate school was made possible by a phenomenal circle of people whose inspiration, encouragement, and support were my pillars of strength. I am profoundly grateful for this community that believed in me, provided emotional sustenance, and cheered me on through every hurdle, making this journey possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

The term "phylodynamics" was introduced by Grenfell et al. (2004) and refers to the study of infectious disease dynamics that result from a combination of evolutionary, epidemiological and ecological processes. In particular, phylodynamic methods build on a foundation of phylogenetics, the study of evolutionary relationships among organisms, to analyze the genetic data of pathogens (such as viruses) in order to understand their spread, evolution, and population dynamics over time. Such an approach is possible for rapidly evolving pathogens because their evolutionary and epidemiological dynamics occur on the same timescale. Phylodynamic approaches have become increasingly popular in infectious disease epidemiology, thanks to advances in genomic sequencing technology, innovations in phylodynamic modeling, and increasing computational power (Pybus and Rambaut, 2009; Hill et al., 2021).

The evolutionary relationships in phylogenetics are characterized by a phylogenetic tree, a bifurcating graph representing the evolutionary history of different species, populations, or entities (Felsenstein, 2004; Yang, 2006). In the context of phylodynamics, the tips of a phylogenetic tree correspond to sampled pathogen sequences and internal nodes represent unobserved ancestral sequences (Pybus and Rambaut, 2009). Phylogenetic tree branch lengths correspond to genetic distances (Yang, 2006), and molecular clock models for relationships between genetic distance and time can be employed to estimate divergence dates of lineages (Pybus and Rambaut, 2009).

Probabilistic phylogenetic reconstruction methods assume a stochastic evolutionary model that acts along the branches of a given phylogenetic tree to produce an observed multiple sequence alignment corresponding to its tips. This enables computation of the observed data

likelihood for the tree (Felsenstein, 2004; Yang, 2006). Evolutionary models on trees typically make use of continuous-time Markov chain models for molecular character substitution that describe how molecular sequences mutate. The simplest models assume that the evolutionary process is the same for all phylogenetic tree branches and all molecular sequence alignment sites. More complex, realistic evolutionary models have been developed that build upon the substitution models to enable variation of the evolutionary process among lineages and sites (Yang, Ziheng, 1994; Drummond et al., 2006).

Such probabilistic modeling opened the door to maximum likelihood estimation of phylogenetic trees and evolutionary model parameters (Yang, 2006), and maximimum likelihood approaches remain widely used in phylogenetics. Beginning in the late 1990s, Bayesian modeling emerged as a popular alternative to maximum likelihood phylogenetic inference (Yang and Rannala, 1997; Drummond et al., 2002; Larget and Simon, 1999). Bayesian inference offers the flexibility to incorporate prior information, provides a natural framework to account for and quantify different sources of uncertainty, and allows for the development of more complex evolutionary models. Further, Bayesian phylogenetic inference have been shown to be more computationally efficient than maximum likelihood approaches that quantify uncertainty via bootstrap (Larget and Simon, 1999).

Phylogenetic reconstruction forms the core of phylodynamic inference, and time-measured phylogenetic trees can provide a great deal of valuable information about epidemic dynamics. For example, phylogenetic reconstruction can reveal transmission chains, identify and track viral mutations of interest, estimate the date at which an outbreak originated, and clarify the extent to which an outbreak is the result of multiple introductory events (Attwood et al., 2022; Hill et al., 2021). However, phylogenetic reconstruction is only the starting point for phylodynamic inference. Researchers have developed a wide array of phylodynamic models build upon a phylogenetic framework to model epidemiological processes in an evolutionary context. For example, phylodynamic models can estimate epidemiological parameters (Volz and Siveroni, 2018; Pybus et al., 2012), reconstruct the spatial and temporal spread of

pathogens (Lemey et al., 2010; De Maio et al., 2015) and elucidate the relationship between infectious disease dynamics and potential driving factors (Lemey et al., 2014).

One parameter of central importance in genomic epidemiology is the effective population size. The effective population size corresponds to the population size in an idealized Fisher-Wright model of reproduction (Wright, 1931) and as such does not usually correspond to the census population size. However, the effective population size is very valuable as a measure of a population's genetic diversity (Charlesworth, 2009) and, in the context of infectious diseases, serves as a proxy for viral circulation and provides valuable insights into how viruses mutate (Moya et al., 2004). For example, in viral populations with smaller effective population sizes, genetic variation may be lost more quickly due to genetic drift, making it harder for advantageous mutations to become fixed (Charlesworth, 2009). Notably, the effective population size is of fundamental interest in many different areas of evolutionary biology. In conservation biology, for instance, the effective population size helps assess the risk of genetic diversity loss in endangered species. Conservation efforts can be tailored based on the knowledge of the effective population size to prevent inbreeding and maintain healthy populations (Frankham, 2015; Frankham et al., 2010). As such, phylodynamic models that incorporate the effective population size can be applied beyond the realm of genomic epidemiology to study a wide variety of scientific problems.

Inference methods for the effective population size are often rooted in coalescent theory (Kingman, 1982a,b). The coalescent is a stochastic process that generates a genealogy that describes the ancestral relationships of a population arising from a classic Fisher-Wright model. Thus, the population size parameter in a coalescent process corresponds to the effective population size. The effective population size plays a large role in determining the genealogies generated by a coalescent process, and it is possible to exploit this link between the effective population size and genealogy to formulate coalescent-based methods for effective population size inference from a genealogy. Coalescent-based models can be seamlessly incorporated into Bayesian phylogenetic inference frameworks as prior distributions

for phylogenetic trees. Notably, this type of Bayesian modeling enables joint inference of the phylogenetic tree and population dynamics, and effective population size inferences will naturally take genealogical uncertainty into account. A number of widely used coalescent-based models for phylogenetic tree priors have been developed, including the Skyline (Drummond et al., 2005), Skyride (Minin et al., 2008) and Skygrid (Gill et al., 2013).

The development of flexible phylodynamic models that can incorporate different types of data is a necessary step in better understanding the interplay of evolutionary and epidemiological processes. However, it is essential for such model development to be accompanied by efficient computational algorithms that allow them to scale to large data sets. Phylogenetic inference is very computationally intensive: as the sample size increases, the number of possible phylogenetic trees that can describe their evolutionary history explodes (Suchard and Rambaut, 2009). Thus a major focus of Bayesian phylodynamics has been on Markov chain Monte Carlo (MCMC) algorithms that enable efficient exploration of this astronomically large parameter space. Given the close link between the phylogenetic tree and effective population size, effective sampling strategies for coalescent-based models is especially important. In recent years, Hamiltonian Monte Carlo (HMC) (Duane et al., 1987) has emerged as a promising sampling strategy. HMC employs ideas from Hamiltonian dynamics in an effort to reduce the correlation between successive samples and more efficiently explore the posterior distribution. To achieve this, HMC introduces an auxiliary "momentum" parameter, and at each iteration of the sampling scheme simulates Hamiltonian dynamics to arrive at new proposed values for the model parameter and momentum parameter. Through this strategy, HMC exploits the underlying geometry of the target posterior distribution to generate distant proposals to high probability regions of the posterior (Neal et al., 2011). HMC has proven to be especially effective at sampling from high-dimensional distributions whose complicated geometry confounds most standard MCMC algorithms (Betancourt, 2017).

HMC has been shown to be effective for a number of different phylodynamic models (Ji et al., 2020), including the Skygrid coalescent-based tree prior (Baele et al., 2020). However,

HMC has not been tested for an extension of the Skygrid model that incorporates external covariates (Gill et al., 2016). This model is especially important in phylodynamics because it provides a framework to test for relationships between the effective population size and potential driving factors, and it can leverage additional covariate data to yield improved estimates of past population dynamics. However, Skygrid models that include covariates can be especially challenging to sample from due to the introduction of covariate effect size parameters and their relationship with other model parameters. HMC may be especially well-suited to meet the challenge of sampling from such a complex model. In this thesis, we evaluate the effectiveness of HMC for Skygrid models that include covariates. We compare the performance of HMC with past MCMC sampling schemes for the model on three real data sets. We find that the HMC sampler generally performs better than the other approaches, providing further evidence for the promise of HMC in sampling for complex models, and potentially opening the door to easier, more widespread use of Skygrid analyses with covariates.

2 METHODS

2.1 Bayesian Inference

Bayes' theorem was formulated by Thomas Bayes in the 18th century. Given events A and B with $P(B) \neq 0$, the theorem states that

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}.$$

The application of Bayes' theorem in an inferential setting gave rise to Bayesian inference. Given observed data $\mathbf{Y}$ and model parameters $\boldsymbol{\theta}$, the goal is to compute the posterior distribution:

$$P(\boldsymbol{\theta}|\mathbf{Y}) = \frac{P(\boldsymbol{\theta})P(\mathbf{Y}|\boldsymbol{\theta})}{P(\mathbf{Y})}.$$

In a maximum likelihood inference framework, the parameter $\boldsymbol{\theta}$ is assumed to be fixed and unknown, and the likelihood $P(\mathbf{Y}|\boldsymbol{\theta})$ is maximized with respect to $\boldsymbol{\theta}$ to obtain a point estimate. In Bayesian inference, $\boldsymbol{\theta}$ is treated as a random variable, and prior knowledge or beliefs about $\boldsymbol{\theta}$ are expressed via the prior distribution $\boldsymbol{\theta}$. While the prior and likelihood are part of the Bayesian model specification, obtaining a closed form solution for the posterior requires computation of the marginal likelihood, or normalizing constant:

$$P(\mathbf{Y}) = \int P(\boldsymbol{\theta})P(\mathbf{Y}|\boldsymbol{\theta})d\boldsymbol{\theta}.$$

For most statistical models of interest, including phylodynamic models, the aforementioned integral is analytically intractable. This limitation (along with persistent philosophical debates about the use of a prior distribution) hampered the utility and adoption of Bayesian inference until advances in Bayesian computing. Markov chain Monte Carlo (MCMC) algorithms (Metropolis et al., 1953; Hastings, 1970) enable researchers to generate samples to empirically approximate previously intractable posterior distributions. While MCMC methods were originally developed in the 1950s by physicists to perform computer experiments, it was not until the landmark paper by Gelfand and Smith in 1990 (Gelfand and Smith, 1990) that the utility of MCMC in Bayesian inference was properly appreciated.

2.1.1 Markov Chain Monte Carlo (MCMC) Methods

MCMC methods aim to simulate a Markov chain $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$ that, under some mild regularity conditions, converges to the posterior distribution $P(\boldsymbol{\theta}|\mathbf{Y})$. Such a chain can be simulated via the Metropolis algorithm (Metropolis et al., 1953):

Metropolis Algorithm

1. **Initialization:**

    - Choose a starting value $\boldsymbol{\theta}^{(0)}$ for the Markov chain

- Set the length of the chain: N
- Specify a symmetric proposal distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$ to generate a new sample $\boldsymbol{\theta}^*$ given the current sample $\boldsymbol{\theta}$

2. **For** $i = 1$ **to** N**:**

   (a) **Proposal Step:**

   - Generate a candidate value $\boldsymbol{\theta}^*$ from the proposal distribution: $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)})$

   (b) **Acceptance Probability:**

   - Compute the following ratio:

$$\alpha = \min\left(1, \frac{P(\boldsymbol{\theta}^*|\mathbf{Y})}{P(\boldsymbol{\theta}^{(i-1)}|\mathbf{Y})}\right)$$

   (c) **Accept or Reject the Proposal:**

   - Generate a uniform random variable $U \sim \text{Uniform}(0, 1)$
   - If $U \leq \alpha$, accept the proposal and set $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^*$
   - If $U > \alpha$, reject the proposal and set $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)}$

3. **Output:**

   - Return the generated samples $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(N)}$

Because we can express the acceptance probability ratio as

$$\begin{aligned}
\frac{P(\boldsymbol{\theta}^*|\mathbf{Y})}{P(\boldsymbol{\theta}^{(i-1)}|\mathbf{Y})} &= \frac{\frac{P(\mathbf{Y}|\boldsymbol{\theta}^*)P(\boldsymbol{\theta}^*)}{P(\mathbf{Y})}}{\frac{P(\mathbf{Y}|\boldsymbol{\theta}^{(i-1)})P(\boldsymbol{\theta}^{(i-1)})}{P(\mathbf{Y})}} \\
&= \frac{P(\mathbf{Y}|\boldsymbol{\theta}^*)P(\boldsymbol{\theta}^*)}{P(\mathbf{Y}|\boldsymbol{\theta}^{(i-1)})P(\boldsymbol{\theta}^{(i-1)})},
\end{aligned}$$

the algorithm sidesteps the aforementioned hurdle of computing the marginal likelihood $P(\mathbf{Y})$. The Metropolis algorithm requires the *proposal distribution* (also known as the *tran-*

sition kernel) $q(\cdot|\cdot)$ to be symmetric: for any x and y we must have $q(x|y) = q(y|x)$. The algorithm was generalized to asymmetric proposals by Hastings (1970) to what is now known as the Metropolis-Hastings algorithm.

Metropolis-Hastings Algorithm

1. **Initialization:**
   - Choose a starting value $\boldsymbol{\theta}^{(0)}$ for the Markov chain
   - Set the length of the chain: N
   - Specify a proposal distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$ to generate a new sample $\boldsymbol{\theta}^*$ given the current sample $\boldsymbol{\theta}$
2. **For** $i = 1$ **to** N**:**
   (a) **Proposal Step:**
   - Generate a candidate value $\boldsymbol{\theta}^*$ from the proposal distribution: $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)})$

   (b) **Acceptance Probability:**
   - Compute the following ratio:

$$\alpha = \min\left(1, \frac{P(\boldsymbol{\theta}^*|\mathbf{Y})q(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{\theta}^*)}{P(\boldsymbol{\theta}^{(i-1)}|\mathbf{Y})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i-1)})}\right)$$

   (c) **Accept or Reject the Proposal:**
   - Generate a uniform random variable $U \sim \text{Uniform}(0, 1)$
   - If $U \leq \alpha$, accept the proposal and set $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^*$
   - If $U > \alpha$, reject the proposal and set $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)}$
3. **Output:**
   - Return the generated samples $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(N)}$

The key difference between the Metropolis algorithm and the Metropolis-Hastings algorithm is the contribution to the acceptance ratio made by the following factor:

$$\frac{q(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i-1)})}.$$

This factor is known as the Hastings ratio and can be viewed as a "correction factor" because it down-weights the acceptance probability for values that are more likely to be proposed by asymmetric transition kernels (Robert and Casella, 2004). Asymmetric transition kernels can enable more efficient sampling from certain distributions, such as skewed distributions or distributions with bounded parameter spaces.

MCMC methods have enabled the use of a wide range of Bayesian models that were previously intractable. Thus, MCMC methods have revolutionized Bayesian inference and brought it to the forefront of modern statistics. However, developing efficient MCMC algorithms for complex models designed to analyze large data sets remains challenging. In many cases, MCMC algorithms can take a long time to converge to the posterior distribution and, even after convergence has been achieved, be hampered by slow exploration of the different parts of the posterior (such as the different modes of a multimodal distribution). In particular, many standard Metropolis-Hastings transition kernels will propose distant "moves" that are in low probability regions of the posterior and are frequently rejected, or propose smaller, more conservative moves that are more likely to be accepted but remain in the same part of the posterior. In such instances, the high degree of correlation between the simulated values can render the MCMC sample unsuitable for posterior approximation, even after a very large number of iterations requiring a great deal of computational time. Researchers have sought to overcome such difficulties through the development of transition kernels that use the underlying geometry of the posterior to inform the direction in which to move at a given iteration. In this spirit, HMC aims to construct distant proposals with a high acceptance rate by simulating Hamiltonian dynamics (Betancourt, 2017).

2.1.2 Hamiltonian Monte Carlo

Hamiltonian dynamics, alternatively referred to as Hamiltonian mechanics or the Hamiltonian formalism, is a mathematical framework used to explain the motion of particles or systems within classical mechanics. Within this framework, the depiction of a mechanical system's state is accomplished through a collection of generalized coordinates, referred to as a *position* vector $\mathbf{q}$, along with a corresponding *momentum* vector, denoted as $\mathbf{p}$. Both $\mathbf{q}$ and $\mathbf{p}$ are posited to assume the same dimension d, so that the complete state space has a total dimension of $2d$ (Brooks et al., 2011). In Bayesian modeling, $\mathbf{q}$ corresponds to the model parameters whereas p are auxiliary variables that are introduced to facilitate posterior simulation that relies on Hamiltonian dynamics (Gelman et al., 2013).

The behavior of the aforementioned system is described by a mathematical function of $\mathbf{q}$ and $\mathbf{p}$ known as the *Hamiltonian*, which we denote by $H(\mathbf{q}, \mathbf{p})$. The evolution of the position and momentum vectors over time is determined by a system of differential equations known as *Hamilton's equations*:

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i},$$
$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i},$$

for $i = 1, \ldots, d$. If we let $\mathbf{z} = (\mathbf{q}, \mathbf{p})$ denote a concatenated vector of the position and momentum, Hamilton's equations can be expressed as :

$$\frac{d\mathbf{z}}{dt} = \mathbf{J}\nabla H(\mathbf{z}),$$

where ∇H is the gradient of H and

$$\mathbf{J} = \begin{bmatrix} \mathbf{0}_{dxd} & \mathbf{I}_{dxd} \\ -\mathbf{I}_{dxd} & \mathbf{0}_{dxd} \end{bmatrix}.$$

In the context of HMC, the Hamiltonian is typically decomposed as:

$$H(\mathbf{q}, \mathbf{p}) = U(\mathbf{q}) + K(\mathbf{p}),$$

where, $U(\mathbf{q})$ is known as the *potential energy*, and $K(\mathbf{p})$ is known as the *kinetic energy*. The potential energy is defined as the negative logarithm of the probability density function for $\mathbf{q}$. In Bayesian modeling, this corresponds to the negative logarithm of the posterior. The kinetic energy is typically defined as:

$$K(\mathbf{p}) = \mathbf{p}'\mathbf{M}^{-1}\mathbf{p}/2,$$

where $\mathbf{M}$ is a symmetric, positive definite matrix known as the *mass matrix*. $\mathbf{M}$ is typically chosen to be diagonal, in which case we can write

$$K(\mathbf{p}) = \sum_{i=1}^{d} \frac{p_i^2}{2m_i}.$$

Under this decomposition of the Hamiltonian, Hamilton's equations can be rewritten as follows:

$$\frac{dq_i}{dt} = [\mathbf{M}^{-1}\mathbf{p}]_i,$$

$$\frac{dp_i}{dt} = -\frac{\partial U}{\partial q_i},$$

for $i = 1, \ldots, d$ (Neal et al., 2011).

Except for simple cases, it is not possible to obtain analytic solutions to Hamilton's equations. Thus, we must resort to numerical approximation. This is accomplished by discretizing time by some small step size, ϵ, so that starting with the state at time $t = 0$, we iteratively approximate the state at times $t = \epsilon, 2\epsilon, 3\epsilon, \ldots$ up until the desired length. There are many numerical methods for solving systems of differential equations, but most solvers are plagued by error that accumulates in approximating relatively long trajectories. Such

error causes the approximated trajectory to drift progressively further away from the true trajectory. Fortunately, there is a class of methods for solving Hamilton's equations known as *symplectic integrators* which are robust to such error accumulation (Betancourt, 2017). A simple symplectic integrator that has been shown to work well in HMC is the *leapfrog* method, which simulates Hamiltonian dynamics according to the following procedure:

$$p_i(t + \epsilon/2) = p_i(t) - (\epsilon/2)\frac{\partial U}{\partial q_i}(\mathbf{q}(t)),$$

$$q_i(t + \epsilon) = q_i(t) + \epsilon\frac{p_i(t + \epsilon/2)}{m_i},$$

$$p_i(t + \epsilon) = p_i(t + \epsilon/2) - (\epsilon/2)\frac{\partial U}{\partial q_i}(\mathbf{q}(t + \epsilon)),$$

for $i = 1, \ldots, d$ (Neal et al., 2011).

HMC is a Metropolis algorithm that simulates Hamiltonian dynamics (using a method such as the leapfrog method) in order to generate proposals.

1. **Initialize Parameters:**
    - Choose a starting point for the Markov chain: x_0.
    - Set the number of leapfrog steps: L.
    - Set the step size: ϵ.
    - Set the total number of iterations: N.
2. **Iterate for i from 1 to N:**
    (a) **Generate New Values for Momentum:**
        - Sample a new momentum variable $\mathbf{p}' \sim N(0, \mathbf{M})$, where $\mathbf{M}$ is the mass matrix (usually a diagonal matrix)
    (b) **Leapfrog Integration:**

- Perform L leapfrog steps with step size ϵ to update the position and momentum, starting with the current state $(\mathbf{q}^{(i)}, \mathbf{p}')$

(c) **Negation of Momentum Variable:**

- Once the L leapfrog steps have been completed to update the position and momentum, negate the momentum variable to arrive at a proposed state $(\mathbf{q}^*, \mathbf{p}^*)$. The negation of the momentum variable ensures that the proposal is symmetric.

(d) **Metropolis-Hastings Acceptance:**

- The proposal $(\mathbf{q}^*, \mathbf{p}^*)$ is accepted or rejected with probability:

$$\min\left(1, \exp(-H(\mathbf{q}^*, \mathbf{p}^*) + H(\mathbf{q}^{(i)}, \mathbf{p}'))\right)$$

3. **Output the Samples:**

- Return the generated samples $(\mathbf{q}^{(1)}, \mathbf{p}^{(1)}), (\mathbf{q}^{(2)}, \mathbf{p}^{(2)}), \ldots, (\mathbf{q}^{(N)}, \mathbf{p}^{(N)})$.

2.2 Bayesian Phylodynamics

The core of a Bayesian phylodynamic inference framework is reconstruction of a phylogenetic tree from observed sequence data. Consider an observed data set consisting of an alignment of molecular sequence characters. Each molecular sequence is a string of discrete molecular character states. The characters may be nucleotide characters taking on one of four possible states: A, G, C, or T, which correspond, respectively, to DNA bases adenine, guanine, cytosine, and thymine. The character could also correspond to amino acids (featuring 20 possible states) or codons (61 possible states) (Yang, 2006).

Each observed sequence is thought of as corresponding to the tips of an unobserved phylogenetic tree that characterizes the evolutionary relationships of the sequences. Molecular sequence evolution is typically modeled by positing a continuous-time Markov chain (CTMC)

model for molecular character substitution that starts at the root of the tree and proceeds down its branches to produced the observed data. These CTMC models are characterized by instantaneous rate matrices that specify the rates of substitution from one molecular character to another. In nucleotide substitution, for example, a simple early approach was to assume equal nucleotide base frequencies and that each nucleotide has the same rate of mutating into any other nucleotide (Jukes and Cantor, 1969). The only parameter to estimate in this case is the overall substitution rate. Since then, a number of more flexible models for nucleotide substitution have been put forth. For example, the HKY model (Hasegawa et al., 1985) estimates base frequencies for the different nucelotide bases as well as a transition/transversion rate ratio κ. The bases A and G are known as purines while the bases C and T are pyrimidines. Substitutions form purine-to-purine or pyrimidine-to-pyrimidine are known as *transitions* whereas other kinds of substitutions are called *transversions*. Through the specification of κ, the HKY model allows for one substitution rate for transitions and another for transversions. Another popular substitution model is the general time-reversible (GTR) model (Tavaré, 1986; Waterman, 1986). The GTR model allows for nine free parameters and is the most general substitution model that allows the CTMC to be time-reversible.

Early approaches to evolutionary modeling assumed that such CTMC substitution models acted independently and identically across multiple sequence alignment sites and phylogenetic tree lineages. However, evolutionary processes are highly variable, and more realistic models that allow for such variation have become commonplace. For example, researchers often model variation across sites in the overall substitution rate (Yang, 1996) and specific character exchange rates (Pagel and Meade, 2004). It has also become common to model variation in evolutionary rates among different phylogenetic tree branches via "molecular clock" models (Drummond et al., 2006).

We can formulate a basic Bayesian phylogenetic model as follows. Let $\mathbf{Y}$ denote the observed molecular sequence data, let $\mathbf{\Lambda}$ denote a vector of parameters characterizing the mutation process (such as CTMC transition rate parameters, as well as parameters for across

site variation and molecular clocks), and let g denote the unobserved phylogenetic tree (or genealogy). The likelihood of the observed data given the mutation model parameters is then

$$P(\mathbf{Y}|g, \mathbf{\Lambda}).$$

For a Bayesian model, we need prior distributions for $\mathbf{\Lambda}$ and g. For the parameters constituting $\mathbf{\Lambda}$, different parametric priors are chosen for the different parameters, depending upon the support of the parameters. For example, Dirichlet distributions are common for equilibrium base frequencies whereas gamma, log normal and exponential distributions are common for rate parameters that are necessarily positive. The parameter values of the prior distributions can be determined based on prior biological information to formulate informative priors. Alternatively, they can be chosen to specify diffuse, relatively uninformative priors (Yang, 2006). Specifying a prior distribution for the phylogenetic tree g is more complicated. Some popular choices include Yule branching processes (Edwards, 1970) birth-death processes (Rannala and Yang, 1996), and coalescent processes (Kingman, 1982b). We will discuss coalescent-based priors in the next section. Suppose for now that the tree prior depends on some parameters $\boldsymbol{\gamma}$, giving us the prior $P(g|\boldsymbol{\gamma})$. Further, suppose that we can also assign to $\boldsymbol{\gamma}$ a suitable prior distribution $P(\boldsymbol{\gamma})$. With the likelihood and priors in hand, we wish to jointly infer the posterior distribution of phylogenetic trees, mutation parameters, and phylogenetic tree prior parameters:

$$P(g, \mathbf{\Lambda}, \boldsymbol{\gamma}|\mathbf{Y}) \propto P(\mathbf{Y}|g, \mathbf{\Lambda})P(g|\boldsymbol{\gamma})P(\boldsymbol{\gamma})P(\mathbf{\Lambda}).$$

This basic Bayesian phylogenetic model can be modified or extended in numerous ways, depending on the data at hand and the evolutionary and epidemiological phenomena we wish to model. For example, along with each observed molecular sequence, we may observe related information, such as a phenotypic trait or the sampling location of the sequence. We may then wish to simultaneously reconstruct the spatial and temporal history associated with

the sequence data along with its evolutionary history (for example, to study the spread of an outbreak in the context of viral sequence data). Or we may wish to study the evolutionary relationships between phenotypic traits. In such a context, we can model the dynamics of the phenotpyic or geographical data as stochastic processes on a phylogenetic tree (Felsenstein, 1985; Lemey et al., 2009). Specifically, suppose $\mathbf{X}$ is the observed phenotypic or geographical data and suppose we model its evolution as a stochastic process on g that depends on parameters $\boldsymbol{\Theta}$. Then our model could be extended to

$$P(g, \boldsymbol{\Lambda}, \boldsymbol{\Theta}, \boldsymbol{\gamma} | \mathbf{Y}, \mathbf{X}) \propto P(\mathbf{Y}|g, \boldsymbol{\Lambda}) P(\mathbf{X}|g, \boldsymbol{\Theta}) P(g|\boldsymbol{\gamma}) P(\boldsymbol{\gamma}) P(\boldsymbol{\Lambda}) P(\boldsymbol{\Theta}).$$

Of course, this is by no means the only way to extend a basic phylogenetic model or to reconstruct spatial and temporal histories in a phylogenetic context. In this thesis, we will work with a model that extends the basic Bayesian phylogenetic framework to model an evolutionary process for observed sequenced data from multiple unlinked genetic loci, and incorporate external covariates into the phylogenetic tree prior (Gill et al., 2016).

2.3 Coalescent-based Tree Priors

The coalescent (Kingman, 1982b) is a stochastic process that generates a genealogy relating a sample of "individuals" arising from an idealized Fisher-Wright reproductive model (Wright, 1931). The classic Fisher-Wright model makes a number of simplifying assumptions about the population: there are no overlapping generations, there is random mating, and there is no selection or migration, and the population size remains constant over time. Because of these simplifying assumptions, the population size parameter in the model (denoted N_e) typically does not correspond to the census population size of the population from which the individuals of interest arise. However, N_e characterizes the genetic diversity of the population of interest, as it describes the size of a Fisher-Wright population that gains and loses genetic diversity at the same rate as the population being studied. N_e is thus known as the *effective*

population size. Thus, in a Bayesian phylodynamic inference framework, the coalescent not only furnishes a prior distribution for the phylogenetic tree, but it also enables inference of the effective population size of the population from which the molecular sequence data are sampled.

The coalescent process begins at the sampling time $t = 0$ and proceeds backwards in time, merging lineages until all lineages have merged and we arrive at the most recent common ancestor of the sample, or root of the genealogy. Here, the time t is understood to represent the time prior to the sampling time and increases as we go further back in time. When two lineages merge, it is known as a *coalescent event*. If t_k represents the time of the $(n-k)$-th coalescent event for $k = 1, \ldots, n-1$, then under the coalescent, the waiting time $w_k = t_{k-1} - t_k$ follows an exponential distribution with rate $\frac{k(k-1)}{2N_e}$.

The coalescent has been extended from its original formulation to relax many of its restrictive assumptions. In our development, we make use of two extensions. First, we assume data that may be sampled at different times (Rodrigo and Felsenstein, 1999). Second, we assume that the effective population size may vary over time (Griffiths and Tavare, 1994).

In this case, we denote the effective population size by $N_e(t)$, and we also refer to it as the *demographic function*. The conditional distribution of waiting times w_k can then be expressed as:

$$P(w_k|t_k) = \frac{k(k-1)}{2N_e(w_k + t_k)} \exp\left[-\int_{t_k}^{w_k+t_k} \frac{k(k-1)}{2N_e(t)} dt\right].$$

Taking the product of such densities can give a joint distribution for waiting times between coalescent events. To arrive at a distribution for a specific genealogy, we must account for the fact that distinct genealogies can have identical coalescent times and thus identical waiting times between coalescent events. Suppose immediately preceding a coalescent time t' there are k' distinct lineages. Then there are $\frac{k'(k'-1)}{2}$ distinct pairs of lineages that can merge at time t' and result in a coalescent event. A different pair of lineages merging corresponds to a different genealogy. To obtain the likelihood of a specific genealogy, we consider only

the one specific pair lineages which merge at time t' in the specific genealogy. Thus, when taking the product of waiting time densities to obtain the likelihood of a specific genealogy, we replace $\frac{k(k-1)}{2}$ in each factor $P(w_k|t_k)$ by 1.

Many coalescent-based tree priors assume a simple parametric function for $N_e(t)$, which is suitable for modeling a constant effective population size, or an effective population size that exhibits exponential or logistic growth or decay. While such parametric functions are appealing in their simplicity, they are very restrictive. Assuming the wrong parametric form can lead to erroneous inferences, and testing a number of different parametric functions can be time consuming. Ultimately, the true dynamics of the demographic function $N_e(t)$ may not be adequately characterized by a simple parametric function. To overcome such difficulties, a number of coalescent-based prior distributions have been developed that approximate $N_e(t)$ as a piece wise constant function that can change at a number of time points (Drummond et al., 2005; Opgen-Rhein et al., 2005; Minin et al., 2008).

One such model is the Skygrid (Gill et al., 2013), which allows researchers to pre-specify the number and times of change points. This flexibility in specification of change points allows the change points to align with measurement times for external covariates, and this fact has been exploited to extend the Skygrid to a more general formulation that integrates covariates (Gill et al., 2016). The Skygrid model with covariates allows researchers to test for associations between the effective population size and epidemiological and ecological factors which may be hypothesized to be related to the effective populatioon size or, more specifically, be thought of as driving factors of population dynamics.

2.4 The Skygrid Model with Covariates

Here, we describe the details of the Skygrid model with covariates. As previously mentioned, the demographic function $N_e(t)$ is piecewise constant, with changes occurring only at specific temporal change points $x_1 < x_2 < \cdots < x_M$. Again, time is understood to increase into the past from the most recent sampling time $x_0 = 0$. These change points divide the population

history timeline into $M+1$ intervals, and the demographic function $N_e(t)$ is fully described by the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{M+1})$ of values that it assumes in the different intervals. In particular, $N_e(t) = \theta_k$ for $x_{k-1} \leq t < x_k$ for $k = 1, \ldots, M$, and $N_e(t) = \theta_{M+1}$ for $t \geq x_M$. Here, x_M is the earliest time in the past at which a change in the effective population size can occur, and it is recommended to choose x_M conservatively so that is earlier than we expect the time of the most recent common ancestor to be. After $\boldsymbol{\theta}$ has been inferred, it is customary to discard effective population size inferences for times that predate the bounds of the 95% Bayesian credibility interval for the time of the most recent common ancestor, since the data do not contain substantial information about the population dynamics for such times.

The Skygrid has been developed to accommodate data from multiple unlinked genetic loci that share the same demographic history but may have different evolutionary histories characterized by different genealogies. Genetic loci are (effectively) unlinked when recombination is very likely, as is the case between genes in retroviruses. In this scenario, we assume a data set the features m different loci and let $\mathbf{g} = (g_1, \ldots, g_m)$ denote the vector of corresponding genealogies. The genealogies are assumed to be conditionally independent, given the effective population size:

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

We can construct the likelihood of genealogy g_i as follows. Let t_{0_i} denote the most recent sampling time, and t_{MRCA_i} the time of the most recent common ancestor (or time of the root node) for g_i. Further, let x_{α_i} denote the smallest change point greater than at least one sampling time, and let x_{β_i} denote the largest change point less than at least one coalescent time. Let $u_{ik} = [x_{k-1}, x_k]$ for $k = \alpha_i + 1, \ldots, \beta_i$, $u_{i\alpha_i} = [t_{0_i}, x_{\alpha_i}]$, and $u_{i(\beta_i+1)} = [x_{\beta_i}, t_{MRCA_i}]$. For each u_{ik}, let t_{kj} for $j = 1, \ldots, r_k$ denote the ordered times of the change points and sampling and coalescent events that occur in the interval. Further, let c_{ik} denote the number

of coalescent events that occur in u_{ik}. Finally, let v_{kj} represent the number of lineages present in the genealogy interval $[t_{kj}, t_{k(j+1)}]$. We can express the likelihood of observing the interval u_{ik} in genealogy g_i as

$$P(u_{ik}|\theta_k) = \left(\frac{1}{\theta_k}\right)^{c_{ik}} \times \prod_{j=1}^{r_k-1} \exp\left[\frac{\nu_{kj}(\nu_{kj}-1)(t_{k(j+1)}-t_{kj})}{2\theta}\right],$$

for $k = \alpha_i, \ldots, \beta_i + 1$. We can then write the likelihood of g_i as

$$\begin{aligned} P(g_i|\boldsymbol{\theta}) &= \prod_{k=\alpha_i}^{\beta_i+1} P(u_{ik}|\theta_k) \\ &= \prod_{k=\alpha_i}^{\beta_i+1} \left(\frac{1}{\theta_k}\right)^{c_{ik}} \exp\left[-\frac{SS_{ik}}{\theta_k}\right], \end{aligned}$$

where the SS_{ik} are appropriate constants.

We will momentarily describe a prior distribution for the effective population size values, and the facilitate the specification of a prior, we will work with the log transformation of the effective population size. Let $\gamma_k = \log \theta_k$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{M+1})$. Using this new notation, we can write

$$\begin{aligned} P(g_i|\boldsymbol{\gamma}) &= \prod_{k=\alpha_i}^{\beta_i+1} P(u_{ik}|\gamma_k) \\ &= \prod_{k=\alpha_i}^{\beta_i+1} e^{-\gamma_k c_{ik}} \exp\left[SS_{ik} e^{-\gamma_k}\right] \\ &= \prod_{k=\alpha_i}^{\beta_i+1} \exp\left[-\gamma_k c_{ik} SS_{ik} e^{-\gamma_k}\right]. \end{aligned}$$

Finally, we can express the likelihood of all genealogies as

$$\begin{aligned} P(\mathbf{g}|\boldsymbol{\gamma}) &= \prod_{i=1}^{m} P(g_i|\boldsymbol{\gamma}) \\ &= \prod_{i=1}^{m} \prod_{k=\alpha_i}^{\beta_i+1} \exp\left[-\gamma_k C_{ik} - S_{ik} e^{-\gamma_k} \right] \\ &= \exp\left[\sum_{i=1}^{M+1} -\gamma_k C_{ik} - SS_{ik} e^{-\gamma_k} \right], \end{aligned}$$

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

Having specified the likelihood for $\mathbf{g}$, we now turn to the task of specifying a prior distribution for the vector $\boldsymbol{\gamma}$ of log effective population size values. It is through this prior that the Skygrid allows us to express our belief of a potential relationship between the effective population size and covariates. Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_P$ denote a collection of P covariates that we believe may be related to the effective population size. Suppose that for each covariate $\mathbf{Z}_j$, we have covariate values $Z_{1j}, \ldots, Z_{M+1,j}$ that are measured or observed at times that match up with the $M + 1$ intervals defined by the Skygrid change points. In practice, the Skygrid change points will be specified to match up with covariate measurement times. One goal of the Skygrid is to model the effective population size for a given interval as a log-linear function of covariates:

$$\gamma_k = \log\theta_k = \beta_1 Z_{k1} + ... + \beta_p Z_{kP} + w_k.$$

Here, β_j is an *effect size coefficient* which quantifies the relationship between the log effective population size and covarite $\mathbf{Z}_j$. Another goal of the Skygrid is to impose temporal dependence between the effective population size at adjacent intervals through appropriate modeling of the error term $w = (w_1, \ldots, w_{M+1})$. To fulfill these goals, the Skygrid employs

a Gaussian Markov random field (GMRF) prior distribution on $\boldsymbol{\gamma}$:

$$P(\boldsymbol{\gamma}|\mathbf{Z}, \boldsymbol{\beta}, \tau) \propto \tau^{\frac{M}{2}} \exp\left[-\frac{\tau}{2}(\boldsymbol{\gamma} - \boldsymbol{Z\beta})'\boldsymbol{Q}(\boldsymbol{\gamma} - \boldsymbol{Z\beta}) \right].$$

Here, $\mathbf{Z}$ is a matrix of covariate values with dimensions $(M+1) \times P$, and $\boldsymbol{\beta}$ is a vector with dimensions $P \times 1$. The precision matrix $\mathbf{Q}$ has dimensions $(M+1) \times (M+1)$ and is tridiagonal, with off-diagonal elements set to -1, $Q_{11} = Q_{M+1,M+1} = 1$, and $Q_{ii} = 2$ for $i = 2, \ldots, M$.

We can better understand the impact of the GMRF prior by examining the full conditional distributions for its coordinates. Let $\boldsymbol{\gamma}_{-i}$ denote the vector formed by excluding only the i-th component from the $\boldsymbol{\gamma}$ vector. Denote Z_i as the i-th row of the covariate matrix $\mathbf{Z}$. Then we have:

$$\gamma_1|\boldsymbol{\gamma}_{-1} \sim N\left(\mathbf{Z}_1'\boldsymbol{\beta} - \mathbf{Z}_2'\boldsymbol{\beta} + \gamma_2, \frac{1}{\tau} \right),$$

$$\gamma_1|\boldsymbol{\gamma}_{-i} \sim N\left(\boldsymbol{Z_i'\beta} + \frac{\gamma_{i-1} + \gamma_{i+1} - \boldsymbol{Z_{i-1}'\beta} - \boldsymbol{Z_{i+1}'\beta}}{2}, \frac{1}{2\tau} \right),$$

for $i = 2, \ldots, M$, and

$$\gamma_{M+1}|\boldsymbol{\gamma_{-(M+1)}} \sim N\left(\boldsymbol{Z_{M+1}'\beta} - \boldsymbol{Z_M'\beta} + \gamma_M, \frac{1}{\tau} \right).$$

These conditional distributions show that, conditional on $\boldsymbol{\gamma}_{-i}$, the i-th component γ_i is influenced solely by its immediate neighboring components. Notice that the precision parameter τ informs the smoothness of the effective population size trajectory.

To complete the Bayesian model specification, we assign diffuse prior distributions to the precision parameter τ and the effect size coefficients $\boldsymbol{\beta}$. To τ, we assign a gamma prior with shape and rate parameters equal to 0.001. We give $\boldsymbol{\beta}$ a multivariate normal prior with mean zero and covariance matrix $\boldsymbol{\Sigma}$ equal to a diagonal matrix with diagonal entries equal to 100.

We can now incorporate the coalescent prior for genealogies $\mathbf{g}$, $P(\mathbf{g}|\boldsymbol{\gamma})$, along with hyper-

priors $P(\boldsymbol{\gamma}|\mathbf{Z},\tau)$, $P(\boldsymbol{\beta})$, $P(\tau)$ into a Bayesian phylodynamic framework for inference directly from sequence data. For our setting of multilocus data, suppose observe aligned molecular sequence data $\mathbf{Y}_1,\ldots,\mathbf{Y}_m$ corresponding to the m genetic loci. For each a given locus i, we assume $\mathbf{Y}_i$ are generated by a CTMC characterized by mutation parameters $\boldsymbol{\Lambda}_i$ that acts along the unobserved genealogy g_i. We assume that $\boldsymbol{\Lambda}_1,\ldots,\boldsymbol{\Lambda}_m$ are *a priori* independent. We can then jointly infer the genealogies, mutation parameters, covariate effect size coefficients, precision, and vector of log effective population sizes through their posterior distribution:

$$P(\boldsymbol{g},\boldsymbol{\Lambda},\boldsymbol{\beta},\tau,\boldsymbol{\gamma}|\mathbf{Y},\mathbf{Z}) \propto \left[\prod_{i=1}^{m} P(\mathbf{Y}_i|g_i,\boldsymbol{\Lambda}_i)\right] \times P(\boldsymbol{\Lambda})P(\mathbf{g}|\boldsymbol{\gamma})P(\boldsymbol{\gamma}|\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\tau})P(\boldsymbol{\beta})P(\tau).$$

Notably, the inferences in this model are based on covariate data as well as sequence data. (Gill et al., 2016) show that incorporating covariate data into a coalescent prior can be valuable in a number of ways. First, it enables the inference of effect size coefficients to quantify and test the significance of relationships between the effective population size and covariates. Second, in the case of as significant associate between a covariate and the effective population size, inclusion of the covariate data in the model can lead to improved estimates of effective population size trajectories.

2.5 Earlier MCMC Sampling Schemes for the Skygrid Model with Covariates

We wish to compare the performance of HMC algorithms for sampling the Skygrid model parameters $\boldsymbol{\gamma},\tau,\boldsymbol{\beta}$ to previously used MCMC sampling schemes. In developing a sampling scheme for the Skygrid parameters, we can condition on the genealogies $\mathbf{g}$ and consider them fixed and known. An MCMC algorithm for the Skygrid parameters can then be combined with transition kernels for the other model parameters to jointly infer the posterior $P(\boldsymbol{g},\boldsymbol{\Lambda},\boldsymbol{\beta},\tau,\boldsymbol{\gamma}|\boldsymbol{Y},\boldsymbol{Z})$.

A simple MCMC algorithm to sample the Skygrid parameters relies on a random walk transition kernel to update the log effective population size parameters and coefficient effect size parameters. In particular, the transition kernel proposes new values for parameters one at a time by drawing a new value uniformly at random within some specified window size and adding that value to the current parameter value. For the precision parameter, new values are proposed by rescaling the current parameter value by some factor. In BEAST 1.10 (Suchard et al., 2018) the aforementioned window size and rescaling factor are automatically tuned to obtain a desired acceptance rate. The sampling scheme we have described is referred to in our data analysis section as the RW scheme (for the random walk that it employs for most model parameters).

An alternative to the simple RW scheme exploits the Skygrid's structure to adapt a fast block-updating sampler for highly structured Gaussian models (Knorr-Held and Rue, 2002). The full conditional distribution of the log effective population size vector $\boldsymbol{\gamma}$ can be approximated via a Gaussian distribution. A Metropolis-Hastings proposal for the parameters $\boldsymbol{\gamma}$ and τ is generated jointly as a block: first, a candidate value for τ is generated by rescaling the current value, then, conditional on the candidate value for τ, a candidate value for $\boldsymbol{\gamma}$ is generated via the aforementioned Gaussian approximation. The effect size coefficients $\boldsymbol{\beta}$ are updated separately via a standard random walk transition kernel, as in the RW sampler. We refer to this sampling scheme as the GA scheme (for the Gaussian approximation that forms the heart of it).

2.6 HMC for the Skygrid Model with Covariates

The main details of the HMC sampling scheme are explained in our earlier discussion of HMC. Here, we detail the gradient computations necessary to implement HMC for the Skygrid model with covariates. Adopting the HMC notation used earlier, we let $\mathbf{q}$ denote

$(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\tau})$. Because $\mathbf{g}$ is fixed in this context, the target density of interest is

$$P(\boldsymbol{\gamma}, \boldsymbol{\beta}, \tau | \mathbf{g}, \boldsymbol{Z}) \propto P(\mathbf{g}|\boldsymbol{\gamma}) P(\boldsymbol{\gamma}|\mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\tau}) P(\boldsymbol{\beta}) P(\tau). \tag{1}$$

We describe the computational details using general gamma and normal priors for τ and $\boldsymbol{\beta}$, respectively. In particular,

$$P(\tau) \propto \tau^{a-1} e^{-b\tau}, \tag{2}$$

and

$$P(\boldsymbol{\beta}) \propto \exp\left[-\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\right],$$

where $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$ for some variance value σ^2.

Let C denote a constant with respect to $\mathbf{q}$. We have:

$$\begin{aligned}
U(\mathbf{q}) &= -\log P(\boldsymbol{\gamma}, \boldsymbol{\beta}, \tau | \mathbf{g}, \mathbf{Z}) \\
&= -\log\left[P(\mathbf{g}|\boldsymbol{\gamma}) P(\boldsymbol{\gamma}|\mathbf{Z}, \boldsymbol{\beta}, \tau) P(\boldsymbol{\beta}) P(\tau)\right] + C \\
&= -\log\left[\exp\left[-\sum_{k=1}^{M+1}(\gamma_k c_k + S S_k e^{-\gamma_k})\right]\right] \\
&\quad -\log\left[\tau^{M/2}\exp\left[-\frac{\tau}{2}(\boldsymbol{\gamma} - \mathbf{Z}\boldsymbol{\beta})'\mathbf{Q}(\boldsymbol{\gamma} - \mathbf{Z}\boldsymbol{\beta})\right]\right] \\
&\quad -\log\left[e^{-\boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}/2}\right] - \log\left[\tau^{a-1}e^{-b\tau}\right] + C.
\end{aligned}$$

We can simplify this further to obtain

$$\begin{aligned}
U(\mathbf{q}) &= -\log \tau^{M/2}\exp\left[-\frac{\tau}{2}\boldsymbol{\gamma}'\mathbf{Q}\boldsymbol{\gamma} + \tau(\mathbf{Z}\boldsymbol{\beta})'\mathbf{Q}\boldsymbol{\gamma} - \frac{\tau}{2}(\mathbf{Z}\boldsymbol{\beta})'\mathbf{Q}\mathbf{Z}\boldsymbol{\beta}\right] \\
&\quad -\log\exp\left[-\sum_{k=1}^{M+1}(\gamma_k c_k + S S_k e^{-\gamma_k})\right] \\
&\quad -\log\left[e^{-\boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}/2}\right] - \log\left[\tau^{a-1}e^{-b\tau}\right] + C.
\end{aligned}$$

We can continue to simplify to get:

$$\begin{aligned} U(\mathbf{q}) &= -\frac{M}{2}\log\tau + \frac{1}{2}\tau\sum_{i=1}^{M}(\gamma_{i+1}-\gamma_i)^2 \\ &\quad -\tau(\gamma_1-\gamma_2)\sum_{k=1}^{P}Z_{1k}\beta_k - \tau(\gamma_{M+1}-\gamma_M)\sum_{k=1}^{P}Z_{(M+1),k}\beta_k \\ &\quad -\tau\sum_{i=2}^{M}\left[(-\gamma_{i-1}+2\gamma_i-\gamma_{i+1})\sum_{k=1}^{P}Z_{ik}\beta_k\right] \\ &\quad +\frac{1}{2}\tau\sum_{i=1}^{M}\left[\sum_{j=1}^{P}(Z_{i+1,j}-Z_{i,j})\,\beta_j\right]^2 \\ &\quad +\sum_{k=1}^{M+1}(\gamma_k c_k + SS_k e^{-\gamma_k}) + \frac{1}{2\sigma^2}\sum_{k=1}^{P}\beta_k^2 - (a-1)\log\tau + b\tau + C. \end{aligned}$$

For clarification, we detail some of the computations performed in these simplifications. First, note that

$$\mathbf{Q} = \begin{bmatrix} 1 & -1 & 0 & \dots & & \\ -1 & 2 & -1 & 0 & \dots & \\ 0 & -1 & 2 & -1 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix},$$

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots \\ 0 & \sigma^2 & 0 & \dots \\ \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix},$$

$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_{M+1} \end{bmatrix},$$

$$\mathbf{Z} = \begin{bmatrix} Z_{11} & Z_{12} & \dots & Z_{1P} \\ Z_{21} & Z_{22} & \dots & Z_{2P} \\ \vdots & \vdots & \vdots & \vdots \\ Z_{M+11} & Z_{M+12} & \dots & \dots Z_{M+1P} \end{bmatrix},$$

and

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}.$$

Then we can write:

$$\boldsymbol{\gamma}'\mathbf{Q}\boldsymbol{\gamma} = \sum_{i=1}^{M} (\gamma_{i+1} - \gamma_i)^2,$$

$$\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} = \frac{1}{\sigma^2} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix},$$

$$\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} = \frac{1}{2\sigma^2} \sum_{i=1}^{P} \beta_i^2,$$

$$\mathbf{Q}\boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 - \gamma_2 \\ -\gamma_1 + 2\gamma_2 - \gamma_3 \\ -\gamma_2 + 2\gamma_3 - \gamma_4 \\ \vdots \\ -\gamma_M + \gamma_{M+1} \end{bmatrix},$$

$$\mathbf{Z}\boldsymbol{\beta} = \begin{bmatrix} \sum_{i=1}^{P} Z_{1i}\beta_i \\ \sum_{i=1}^{P} Z_{2i}\beta_i \\ \vdots \\ \sum_{i=1}^{P} Z_{(M+1)i}\beta_i \end{bmatrix},$$

$$(\mathbf{Z}\boldsymbol{\beta})' = \begin{bmatrix} \sum_{i=1}^{P} Z_{1i}\beta_i & \sum_{i=1}^{P} Z_{2i}\beta_i & \dots & \sum_{i=1}^{P} Z_{(M+1)i}\beta_i \end{bmatrix},$$

$$(\mathbf{Z}\boldsymbol{\beta})'\mathbf{Q}\boldsymbol{\gamma} = \Bigg[(\gamma_1 - \gamma_2)\sum_{i=1}^{P} Z_{1i}\beta_i + (-\gamma_M + \gamma_{M+1})\sum_{i=1}^{P} Z_{(M+1)i}\beta_i + \sum_{j=2}^{M}\left[(-\gamma_{j-1} + 2\gamma_j - \gamma_{j+1})\sum_{i=1}^{P} Z_1 i\beta_i\right]\Bigg],$$

$$(\mathbf{Z}\boldsymbol{\beta})'\mathbf{Q}\mathbf{Z}\boldsymbol{\beta} = \sum_{j=1}^{M}\left[\sum_{i=1}^{P}(Z_{j+1,i} - Z_{j,i})\beta_i\right]^2.$$

Now we specify the gradient calculations required to implement the leapfrog algorithm. First, note that

$$\begin{aligned} \frac{\partial U}{\partial \tau} &= -\frac{M}{2\tau} + \frac{1}{2}\sum_{i=1}^{M}(\gamma_{i+1} - \gamma_i)^2 \\ &\quad -(\gamma_1 - \gamma_2)\sum_{k=1}^{P} Z_{1k}\beta_k - (\gamma_{M+1} - \gamma_M)\sum_{k=1}^{P} Z_{(M+1),k}\beta_k \\ &\quad -\sum_{i=2}^{M}\left[(-\gamma_{i-1} + 2\gamma_i - \gamma_{i+1})\sum_{k=1}^{P} Z_{ik}\beta_k\right] \\ &\quad +\frac{1}{2}\sum_{i=1}^{M}\left[\sum_{j=1}^{p}(Z_{i+1,j} - Z_{i,j})\,\beta_j\right]^2 \\ &\quad -(a-1)/\tau + b. \end{aligned}$$

Next, for $k = 1, \ldots, P$ we have

$$\begin{aligned}
\frac{\partial U}{\partial \beta_k} &= -\tau(\gamma_1 - \gamma_2)Z_{1k} - \tau(\gamma_{M+1} - \gamma_M)Z_{(M+1),k} \\
&\quad -\tau \sum_{i=2}^{M} \left[(-\gamma_{i-1} + 2\gamma_i - \gamma_{i+1})Z_{ik}\right] \\
&\quad +\tau \sum_{i=1}^{M} \left[\sum_{j=1}^{p} \left(Z_{i+1,j} - Z_{i,j}\right)\beta_j\right] \left(Z_{i+1,k} - Z_{i,k}\right) \\
&\quad +\frac{\beta_k}{\sigma^2}.
\end{aligned}$$

For $i = 2, \ldots, M$ we have

$$\begin{aligned}
\frac{\partial U}{\partial \gamma_i} &= \tau(-\gamma_{i-1} + 2\gamma_i - \gamma_{i+1}) \\
&\quad +\tau \sum_{k=1}^{P} Z_{(i+1),k}\beta_k - 2\tau \sum_{k=1}^{P} Z_{ik}\beta_k + \tau \sum_{k=1}^{P} Z_{(i-1),k}\beta_k \\
&\quad +c_i - SS_i e^{-\gamma_i}.
\end{aligned}$$

Finally, note that

$$\begin{aligned}
\frac{\partial U}{\partial \gamma_1} &= \tau(\gamma_1 - \gamma_2) \\
&\quad -\tau \sum_{k=1}^{P} Z_{1k}\beta_k \\
&\quad +\tau \sum_{k=1}^{P} Z_{2k}\beta_k \\
&\quad +c_1 - SS_1 e^{-\gamma_1},
\end{aligned}$$

and

$$\begin{aligned}\frac{\partial U}{\partial \gamma_{M+1}} &= \tau(\gamma_{M+1} - \gamma_M) \\ &\quad -\tau \sum_{k=1}^{P} Z_{(M+1),k}\beta_k \\ &\quad +\tau \sum_{k=1}^{P} Z_{Mk}\beta_k \\ &\quad +c_{M+1} - SS_{M+1}e^{-\gamma_{M+1}}.\end{aligned}$$

3 DATA AND ANALYSIS

Our goal is to gain a better understanding of the performance of HMC for the Skygrid model with covariates by analyzing three real data sets. The performance of HMC algorithms can depend considerably on the choice of tuning parameters: the step size and number of steps in the leapfrog algorithm (Betancourt, 2017; Neal et al., 2011). For each of the three data sets, we compare the performance of HMC under different combinations of tuning parameters. We also want to compare the performance of HMC to that of earlier MCMC sampling schemes that have been employed for the Skygrid model with covariates. These earlier approaches, which we refer to as the RW and GA sampling schemes, are described in detail in Section 2.5. We analyze each of the three real data sets using the RW, GA and HMC sampling schemes.

The first data set consists of 75 sequences of DENV-4, which were compiled by (Bennett et al., 2003) through the sequencing of randomly chosen DENV-4 isolates from Puerto Rico, sourced from the US Centers for Disease Control and Prevention (CDC) sample bank . DENV-4 is one of four closely related but distinct serotypes of the dengue virus (DENV). DENV is an RNA virus that causes the dengue viral infection. Dengue causes a severe flu-like illness, occasionally leading to potentially lethal syndromes (WHO: World Health Organization, 2015a). Each sequence in the data set encompasses approximately 40% of

the viral genome, including all structural genes (capsid: C; membrane: M; and envelope: E), as well as a subset of nonstructural genes (NS1, NS2A, and NS4B), in addition to the noncoding 3' NTR region. The sequences were sampled in various years, including 1982 (14 sequences), 1986/1987 (19 sequences), 1992 (15 sequences), 1994 (14 sequences), and 1998 (13 sequences). As a covariate, we use transformed case counts of the number of DENV-4 isolates recorded during every six-month period between 1981 and 1998. The case counts are transformed by the map $x \mapsto \log(x + 1)$ (the addition of one inside the logarithm is to allow for the transformation of isolate counts of zero).

The second data set comprises 47 rabies virus sequences sampled from rabid raccoons between 1982 and 2004 (Biek et al., 2007). These sequences encompass the complete rabies nucleoprotein (N) genes as well as substantial segments of the glycoprotein (G) genes. The rabies virus causes rabies, a severe zoonotic disease that results in over 50,000 human deaths annually and remains a major public health concern (WHO: World Health Organization, 2015b). For a covariate, we use the log transform of the cumulative area (in square kilometers) of counties affected by raccoon rabies between the years 1977 and 1999. The area of a county is added to the cumulative total for the month during which rabies is first reported in that county. There are 175 months for which the cumulative affected area changes, and we specify the Skygrid change points to coincide with those months.

Finally, our third data consists of ancient DNA sequences from musk ox (Campos et al., 2010). The sequences comprise 682 base pairs of the mitochondrial control region, obtained from 149 radiocarbon-dated specimens. The specimens encompass a considerable time range, from the present to 56,900 radiocarbon years before present (YBP), and come from a number of different geographic locations that represent the demographic range of ancient musk ox. The locations include the Taimyr Peninsula (54 sequences), the Urals (26 sequences), Northeast Siberia (12 sequences), North America (14 sequences), and Greenland (43 sequences). Musk ox were once widely distributed in the Holarctic ecozone but are now restricted to Greeland and the Arctic Archipelago. Campos et al. (2010) observe that time intervals dur-

ing which musk ox populations increase tend to be periods of global climatic cooling, while musk ox populations exhibit decline during warmer and climatically unstable periods. Environmental change thus appears to be a driving force behind musk ox population dynamics. To text this hypothesis, we make use of ice core δ^{18}O data sourced from the Greenland Ice Core Project (GRIP) (Dansgaard et al., 1989, 1993; GRIP Members, 1993; Grootes et al., 1993; Johnsen et al., 1997). δ^{18}O represents the oxygen isotope composition, where lower δ^{18}O values correspond to colder polar temperatures. As a covariate in a Skygrid analysis, we calculate the mean δ^{18}O value by averaging the δ^{18}O values corresponding to each 3000-year interval.

We analyze all data sets using a Bayesian phylodynamic model with an HKY nucleotide substitution model (Hasegawa et al., 1985). Further, we model branch-specific variation of evolutionary rates though an uncorrelated relaxed molecular clock, characterized by an underlying lognormal distribution (Drummond et al., 2006). All analyses are conducted using BEAST 1.10 (Suchard et al., 2018), and results are analzed via Tracer version 1.7.2 (Rambaut et al., 2018). For each analysis, we simulate and MCMC chain of 10 million states, discard the initial 1 million iterations as burn-in, and log every 2000 iterations. For each data set and each specific sampling scheme (including different HMC tuning parameter settings), we run three independent replicates and average results. We consider three different HMC tuning parameter combinations: 1) a step size of 0.01 with 50 steps, 2) a step size of 0.02 with 25 steps, and 3) a step size of 0.05 with 10 steps. To assess the effectiveness of differrent sampling schemes, we compute the effective sample size (ESS) for each Skygrid parameter via the coda R package (Plummer et al., 2006; R Core Team, 2018). The ESS for a parameter corresponds to the number of independent samples from the posterior distribution that the MCMC sample is equivalent to (Kass et al., 1998). Because different sampling schemes have varying amounts of computational burden, we cannot simply compare ESS values to compare efficiency. Instead, we compute and compare the ESS per unit time for each Skygrid parameter.

4 RESULTS

Table 1: Comparison of HMC performance under different tuning parameter combinations in analyses of dengue, rabies and musk ox data sets. Effective sample size per minute, averaged over three independent replicates, is reported for precision parameter τ and covariate effect size coefficient β. For vector of log effective population size parameters $\boldsymbol{\gamma}$, the range of effective sample size per minute, averaged over three independent replicates, is reported.

Example	Step size	Steps	$\boldsymbol{\gamma}$	τ	β
Dengue	0.01	50	12.38-233.09	16.71	70.34
Dengue	0.02	25	14.82-277.57	16.07	50.32
Dengue	0.05	10	23.91-286.25	35.02	66.02
Rabies	0.01	50	16.58-113.48,	23.44	66.72
Rabies	0.02	25	15.65-118.34	16.89	108.72
Rabies	0.05	10	9.37-149.27	12.56	82.53
Musk Ox	0.01	50	8.33-160.59	13.38	44.15
Musk Ox	0.02	25	8.26-135.38	16.14	44.45
Musk Ox	0.05	10	8.67-115.13	15.95	39.66

Table 2: Comparison of performance of HMC, GA and RW sampling schemes in analyses of dengue, rabies and musk ox data sets. Effective sample size per minute, averaged over three independent replicates, is reported for precision parameter τ and covariate effect size coefficient β. For vector of log effective population size parameters $\boldsymbol{\gamma}$, the range of effective sample size per minute, averaged over three independent replicates, is reported.

Example	$\boldsymbol{\gamma}$	τ	β
$\text{Dengue}_{\text{GA}}$	13.71-117.61	19.64	13.15
$\text{Dengue}_{\text{HMC}}$	23.91-286.25	35.02	66.02
$\text{Dengue}_{\text{RW}}$	3.22-23.18	2.98	3.75
$\text{Rabies}_{\text{GA}}$	9.02-114.85	11.10	51.96
$\text{Rabies}_{\text{HMC}}$	9.37-149.27	12.56	82.53
$\text{Rabies}_{\text{RW}}$	0.27-2.46	0.62	3.02
$\text{Musk Ox}_{\text{GA}}$	8.33-113.38	19.78	35.84
$\text{Musk Ox}_{\text{HMC}}$	8.33-160.59	13.38	44.15
$\text{Musk Ox}_{\text{RW}}$	1.23-14.016	1.89	3.69

We first evaluate the performance of HMC under different tuning parameter combinations for the three data sets. Figures 1-3 report the ESS per minute for the Skygrid model parameters for each HMC tuning parameter combination and each data set. In the case of the multidimensional log effective population size parameter, the range of ESS per minute

values is reported. In order to more clearly compare the performance in the case of the log effective population size, Figures 1-3 depict the distributions of ESS per minute values for the log effective population size for the different data sets and different HMC tuning parameter combinations.

For the dengue data set, Table 1 shows that the setting with a step size of 0.05 and 10 steps performs twice as well as the others for τ and comes in at a close second for β. Figure 1 shows the performance of HMC under the different settings for log effective population size $\boldsymbol{\gamma}$. For a step size of 0.01 with 50 steps: most parameters achieve an ESS per minute around the 100-120 range, with a few parameters even reaching 200-220. There is a good distribution, suggesting a decent level of efficiency for this setting. For a step size of 0.02 with 25 steps: most of its parameters are centered around the 50-70 ESS per minute range. For a step size of 0.05 with 10 steps: parameters achieve higher ESS per minute values in the ranges around 40-60 and 110-130 with some parameters even reaching 270-286. This shows that a significant number of parameters are achieving decent to high ESS values with this setting. The setting (Step Size=0.05, Steps=10) appears to have a good number of parameters that achieve higher ESS per minute values, suggesting it's potentially the most efficient setting among the three. The setting (Step Size=0.01, Steps=50) also has a good distribution of parameters around the 100-120 ESS per minute range but lacks the same number of parameters in the higher ESS categories compared to the step size=0.05, steps=10 The setting (Step Size=0.02, Steps=25) seems to be the least efficient, with a concentration of its parameters at lower ESS per minute values.

For the rabies data set, Table 1 shows that the setting with a step size of 0.01 and 50 steps performs the best for τ while the setting with a step size of 0.02 and 25 steps performs the best for β. Figure 2 shows in detail the performance of the log effective population size $\boldsymbol{\gamma}$. For a step size of 0.01 with 50 steps: the majority of parameters achieve an ESS per minute value between 40 and 50. There's a left skew, indicating that fewer parameters reach higher ESS per minute values with this setting. For a step size of 0.02 with 25 steps:

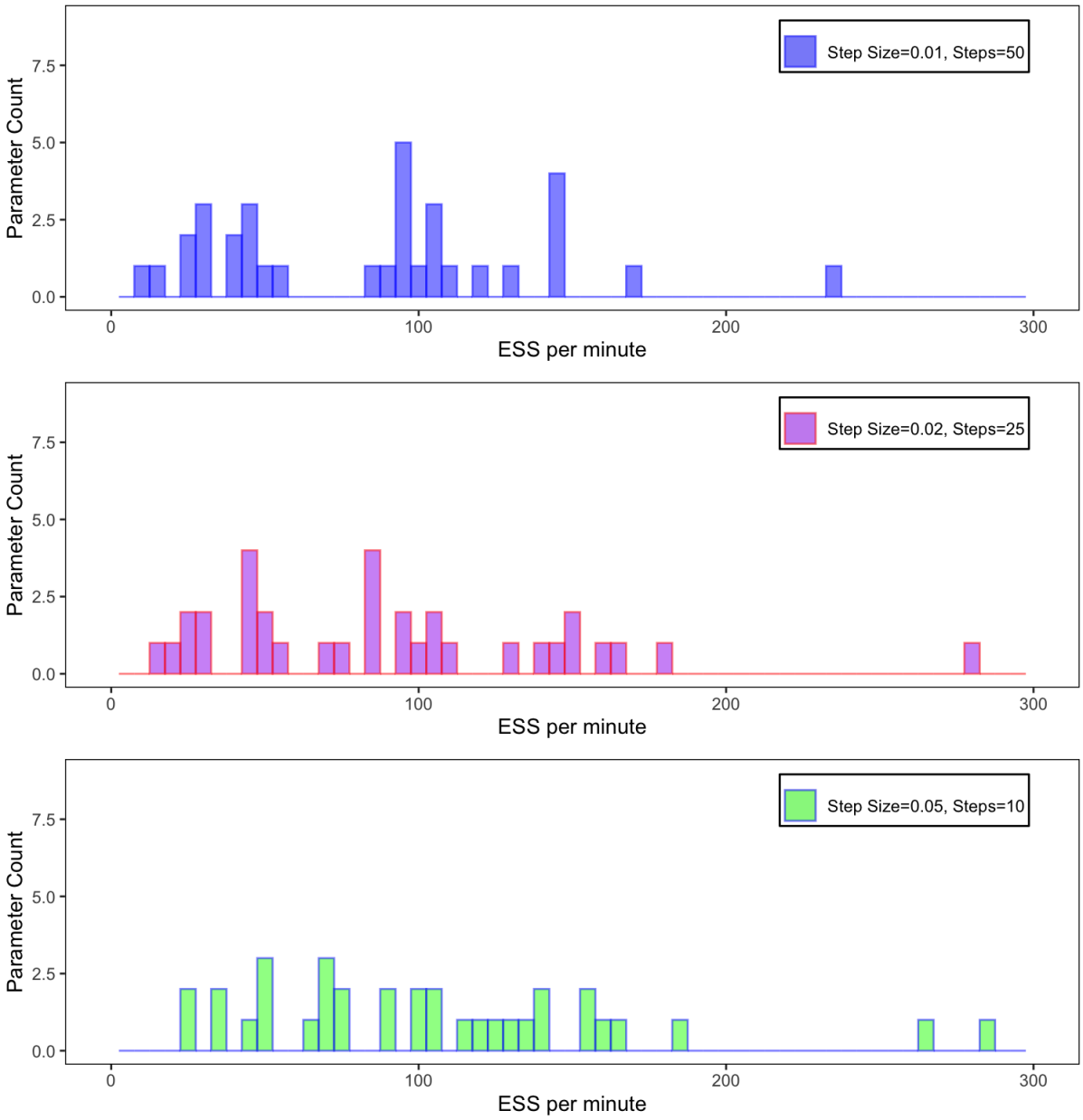

Figure 1: Comparison of HMC tuning parameter combinations for dengue data set. Bars correspond to the estimated effective sample size (ESS) per minute averaged across three independent replicates for all log effective population size parameters. The height of each bar indicates the number of parameters that achieve the given ESS per minute value. For the dengue data, the HMC transition kernel appears to performs best with 10 steps and step size = 0.05 .

The peak of this distribution is slightly leftward from that with a step size of 0.01 and is centered around the 20-30 ESS per minute range. This implies that this tuning parameter combination is somewhat less efficient than previous combination for a significant proportion of the parameters. However, some parameters still reach higher ESS per minute values, but not as many as in the blue histogram setting. For a step size of 0.05 with 10 steps: it has its highest peak around the 20-30 ESS per minute range. However, the distribution shows more variability, with noticeable peaks in the 40-50 and 60-70 ranges, suggesting that some parameters achieve higher efficiency, while others do not.

Table 1 shows that for the musk ox data set, a tuning parameter combination of a step size of 0.02 with 25 steps performs best for τ and β. However, the other tuning parameter combinations perform nearly as well. Figure 3 compares the performance in detail for the log effective population size $\boldsymbol{\gamma}$. For a step size of 0.01 with 50 steps: the distribution is left-skewed, with a peak around the 30-40 ESS per minute range. For a step size of 0.02 with 25 steps: this distribution showcases a peak slightly to the left of that of the previous combination, centered around 20-30 ESS per minute. This implies that, while these tuning parameters are efficient for a good number of parameters, they might be less efficient than the previous tuning parameter combination. There's also a few parameters that achieve higher ESS per minute, but they are fewer in number. For a step sizeo of 0.05 with 10 steps: the peak of the distribution is around the 20-30 ESS per minute range. This setting displays a clearer left-skew, indicating a consistent drop in the number of parameters as the ESS per minute value increases. The Step Size=0.01, Steps=50 combination appears to be the most efficient setting among the three, based on its peak at a relatively higher ESS per minute value.

In our comparison of HMC performance under different tuning parameter combinations, no combination emerged as the clear best choice. This illustrates the difficulty in choosing optimal values for tuning parameters, showing that for a specific model, the best choice can vary for different data sets and different parameters. The results are consistent with the

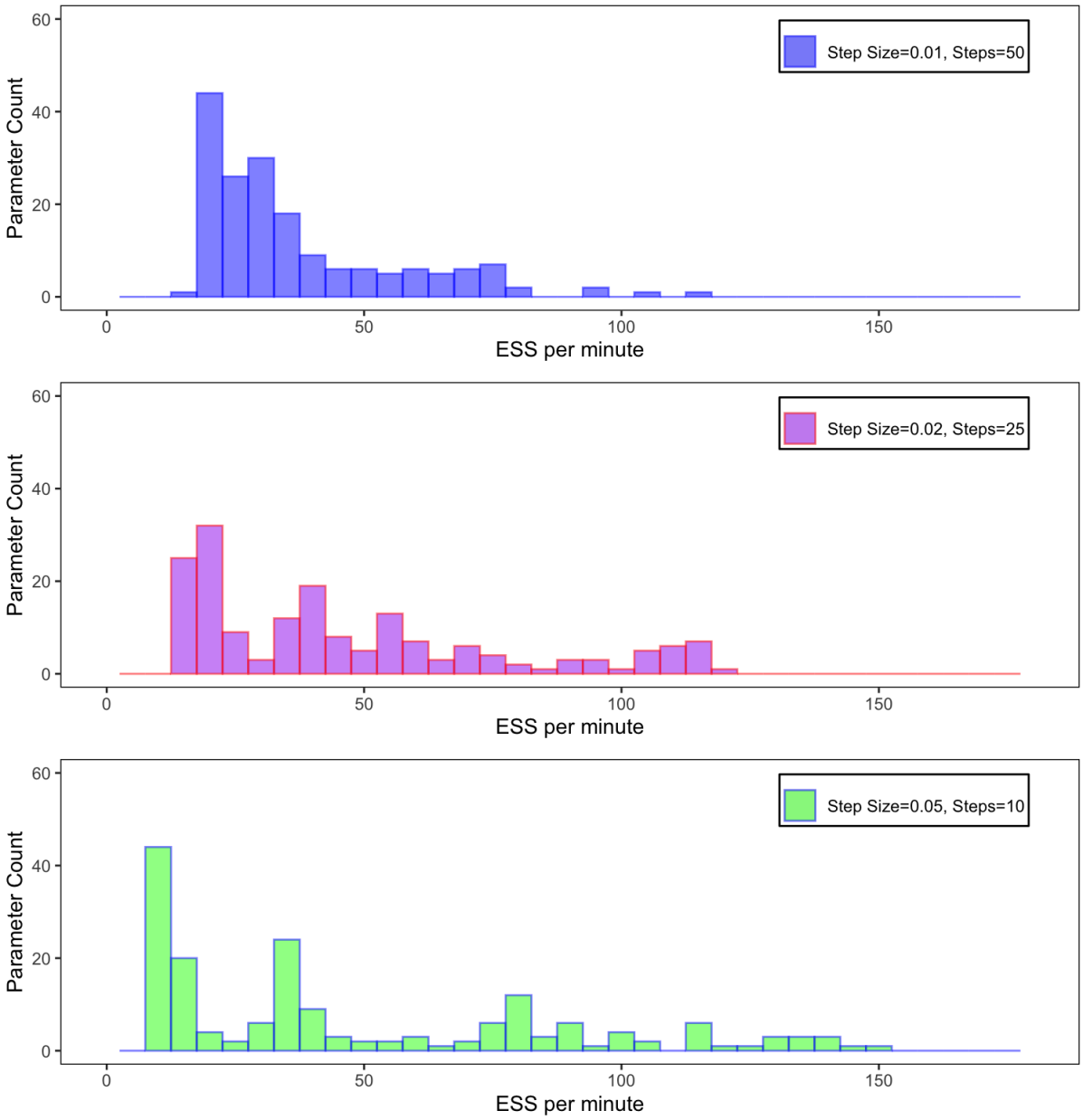

Figure 2: Comparison of HMC tuning parameter combinations for rabies data set. Bars correspond to the estimated effective sample size (ESS) per minute averaged across three independent replicates for all log effective population size parameters. The height of each bar indicates the number of parameters that achieve the given ESS per minute value. For the rabies data, the HMC transition kernel appears to performs best with 10 steps and step size = 0.05.

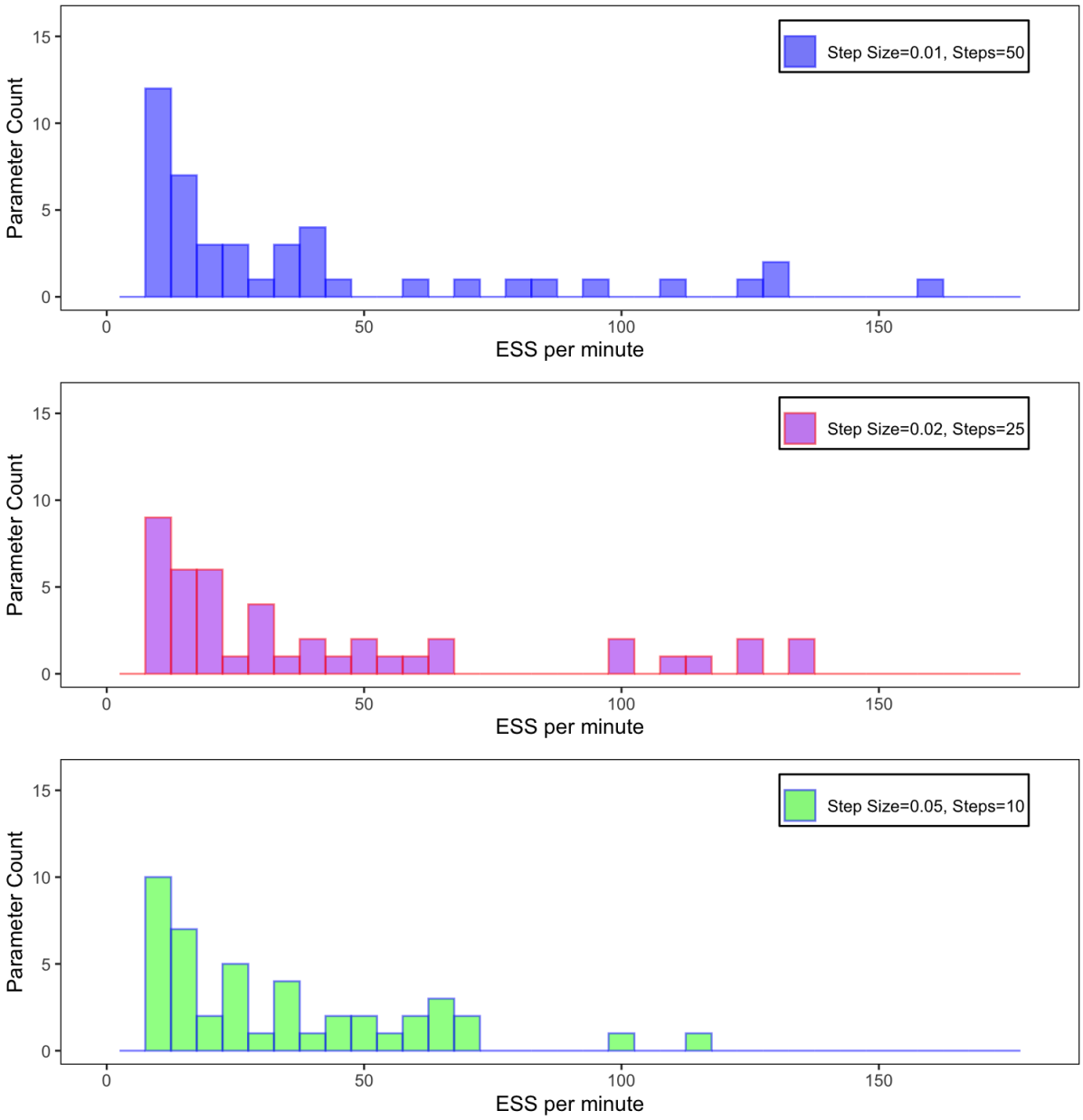

Figure 3: Comparison of HMC tuning parameter combinations for musk ox data set. Bars correspond to the estimated effective sample size (ESS) per minute averaged across three independent replicates for all log effective population size parameters. The height of each bar indicates the number of parameters that achieve the given ESS per minute value. For the musk ox data, the HMC kernel transition kernel appears to performs best with 50 steps and step size = 0.01

observations of other researchers that HMC performance can vary considerably depending on the chooice of tuning parameters (Betancourt, 2017). It is crucial to note that ESS per minute is just one metric. We did not deal with prolonged burn-in times in these analyses, but it is possible for some tuning parameters combinations to quick convergence and subsequent slow exploration of the posterior, or vice versa. Properly evaluating the performance of HMC in more challenging scenarios will require more care.

We now focus on comparing the performance of HMC (with a step size of 0.05 and 10 steps) with GA and RW sampling schemes for the three data sets. Table 2 shows that RW lags far behind GA and HMC for all Skygrid model parameters. In many instances, the RW sampling scheme fails to generate sufficient ESS for accurate posterior approximation after 10 million iterations. This illustrates how critically important an efficient MCMC sampling scheme is for complex, computationally intensive models that will be used to analyze large data sets. For the precision τ and covariate effect size coefficient β, Table 2 shows that HMC outperforms GA in 5 out of 6 scenarios. To compare the performance of HMC and GA for the log effective population $\boldsymbol{\gamma}$, we examine the distributions of ESS per minute values in Figures 4-6.

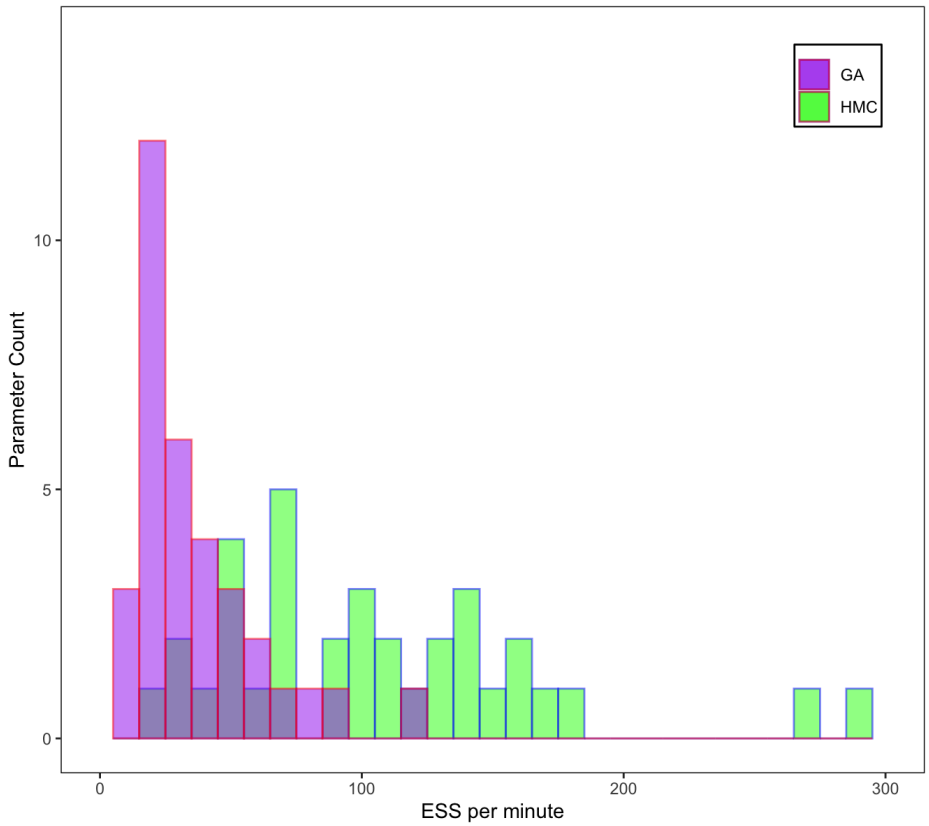

Figure 4: Comparison of HMC (with a step size of 0.05 and 10 steps) and GA for dengue data set. Bars correspond to the estimated effective sample size (ESS) per minute averaged across three independent replicates for all log effective population size parameters. The height of each bar indicates the number of parameters that achieve the given ESS per minute value.

Figure 4 compares the performance of HMC and GA in sampling $\boldsymbol{\gamma}$ for the dengue data set. The histogram representing GA is more concentrated towards the left side of the graph. The histogram representing HMC is spread out more evenly and extends much further to the right. HMC emerges as a clear winner in this comparison, achieving much higher ESS per minute values, in general. Figure 5 compares the performance of HMC and GA in sampling $\boldsymbol{\gamma}$ for the rabies data set. The histogram for the GA scheme has a prominent peak around the 20-30 ESS per minute mark, with a long right tail that has a smaller peak around 80. The histogram for HMC shows a wider spread with its highest peak at low ESS per minute values, but a long right tail with several smaller peaks at high ESS per minute values. It is difficult to pick a clear winner. The HMC histogram has a wider spread, but it has a

much longer right tail, and its best values are much better than the best values under GA, while the lower values under both methods are more similar. Finally, Figure 6 compares the performance of HMC and GA in sampling $\boldsymbol{\gamma}$ for the musk ox data set. While the histograms for both methods are concetrated towards the left of the plot, the HMC histogram shows more variation, with small peaks at relatively higher values and a longer right tail. While HMC does not perform overwhelmingly better than GA (as in the case of the dengue data set), it exhibits a small but clear edge.

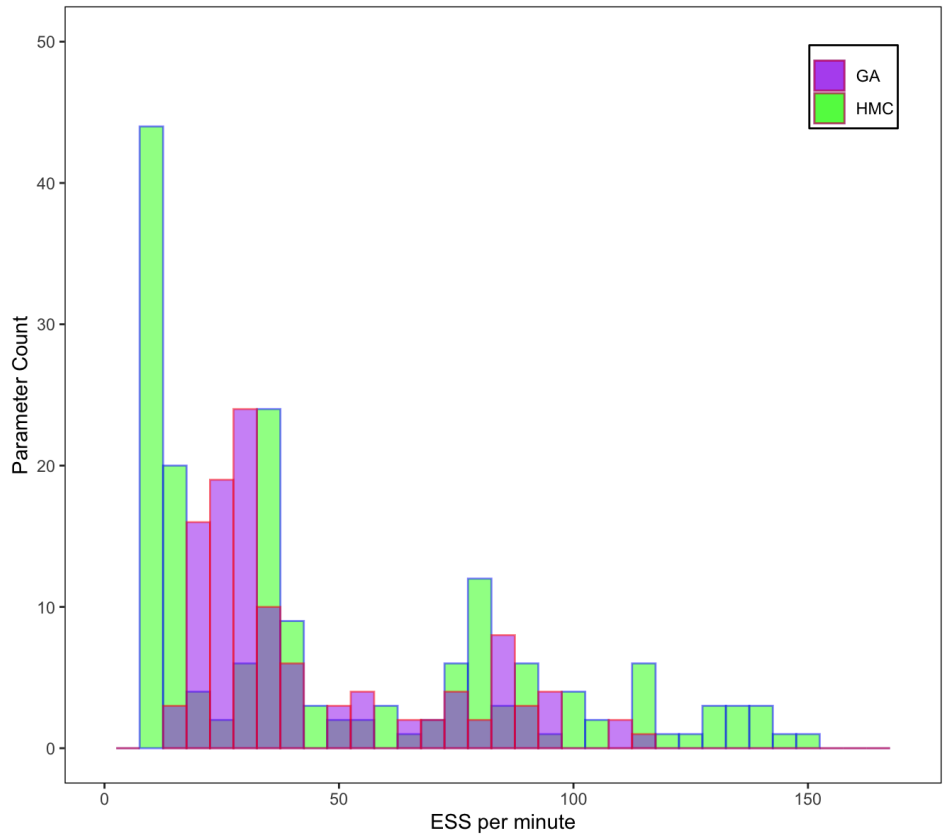

Figure 5: Comparison of HMC (with a step size of 0.05 and 10 steps) and GA for rabies data set. Bars correspond to the estimated effective sample size (ESS) per minute averaged across three independent replicates for all log effective population size parameters. The height of each bar indicates the number of parameters that achieve the given ESS per minute value.

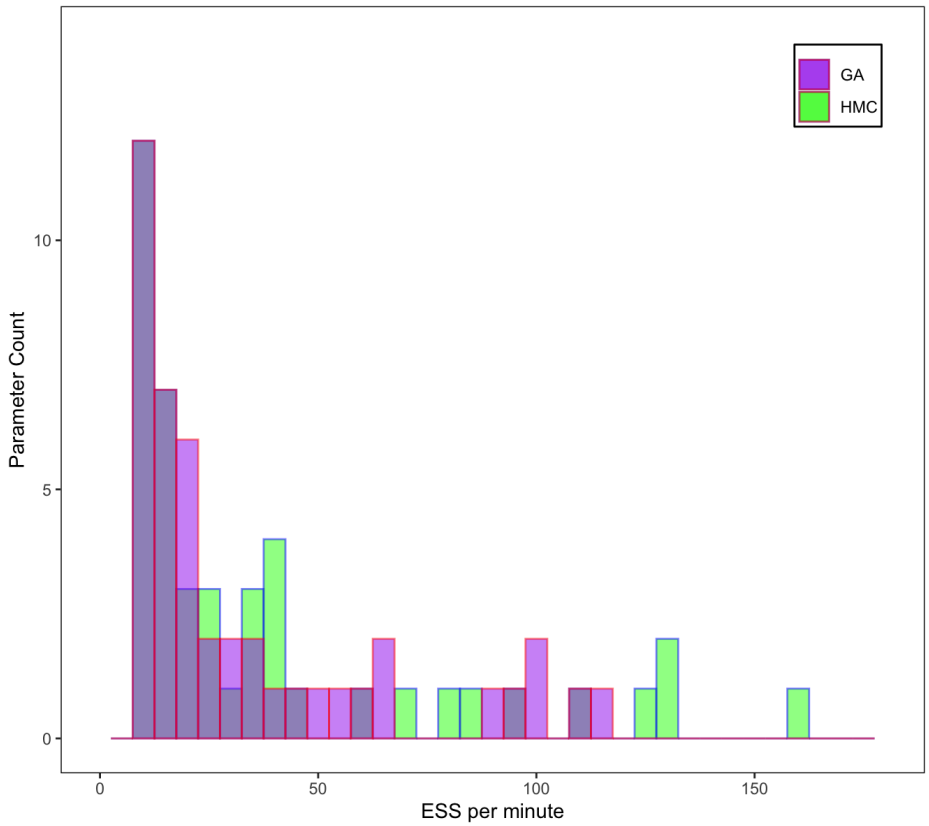

Figure 6: Comparison of HMC (with a step size of 0.05 and 10 steps) and GA for musk ox data set. Bars correspond to the estimated effective sample size (ESS) per minute averaged across three independent replicates for all log effective population size parameters. The height of each bar indicates the number of parameters that achieve the given ESS per minute value.

5 DISCUSSION

Phylodynamic inference is widely used throughout evolutionary biology and genomic epidemiology, but its potential impact is inhibited by its considerable computational burden. Continual development of computationally efficient algorithms for phylodynamic inference are therefore critically important. Coalescent-based models are ubiquitous in Bayesian phylodynamics as phylogenetic tree priors, and they also open the door for inference of the effective population size, which is of fundamental importance in population biology and epidemiology. Good coalescent tree priors can be vital for accurate phylogenetic inferences, but they will only enjoy widespread use if paired with efficient algorithms that allow them

to scale to large data sets that are commonplace thanks to advances in genomic sequencing technology. Here, we evaluate the performance of one of the most promising MCMC sampling techniques, HMC, for a novel coalescent-based model that links phylodynamic processes with other sources of epidemiological and ecological information. Compared to past MCMC sampling schemes that have been used for the Skygrid model with covarites, HMC consistently performs as well or (mostly) better. The fact that HMC outperforms the GA scheme in most scenarios, even though the GA sampler was designed specifically to exploit the Skygrid's special structure, is very noteworthy. If an algorithm such as HMC, which does not require very specific modifications to be employed for any given model, can outperform fast samplers designed specifically for certain types of models, it will free up researchers from the laborious and difficult task of designing highly tailored algorithms when implementing a new models.

6 FUTURE WORK

The general improvement of HMC over other samplers for the Skygrid model with covariates is encouraging, and it will be interesting to see if this success can carry over to even more complicated coalescent-based approaches that incorporate population structure and unite tree-generating and migration processes (De Maio et al., 2015). As we observed, the performance of HMC can vary depending on tuning parameters, and no combination of tuning parameter values is optimal for all data sets. There has been a great deal of focus on developing methods to adaptively optimize HMC tuning parameters (Hoffman and Gelman, 2014; Wu et al., 2018). There have also been adaptations of conventional HMC approaches that seek to gain further improvements in efficiency through greater exploitation of the geometric structure of the posterior distribution (Neal et al., 2011; Nishimura and Dunson, 2016). Building upon the work in this thesis, it would be interesting to evaluate the performance of such advanced methods for complex phylodynamic models.

REFERENCES

Attwood, S. W., Hill, S. C., Aanensen, D. M., Connor, T. R., and Pybus, O. G. (2022). Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic. *Nature Reviews Genetics*, 23(9):547–562.

Baele, G., Gill, M. S., Lemey, P., and Suchard, M. A. (2020). Hamiltonian Monte Carlo sampling to estimate past population dynamics using the skygrid coalescent model in a Bayesian phylogenetics framework. *Wellcome Open Research*, 5:53:1–17. [version 1; peer review: 1 approved, 2 approved with reservations].

Bennett, S., Holmes, E., Chirivella, M., Rodriguez, D., Beltran, M., Vorndam, V., Gubler, D., and McMillan, W. O. (2003). Selection-driven evolution of emergent dengue virus. *Molecular Biology and Evolution*, 20:1650–1658.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

Biek, R., Henderson, J. C., Waller, L. A., and et al. (2007). A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. *Proceedings of the National Academy of Sciences of the United States of America*, 104(19):7993–7998.

Brooks, S., Gelman, A., Jones, G. L., and Meng, X. L. (2011). *Handbook of Markov chain Monte Carlo*. CRC Press.

Campos, P., Willerslev, E., Sher, A., Orlando, L., Axelsson, E., Tikhonov, A., Aaris-Sorenson, K., Greenwood, A., Kahlke, R., Kosintsev, P., Krakhmalnaya, T., Kuznetsova,

T., Lemey, P., MacPhee, R., Norris, C., Shepherd, K., Suchard, M., Zazula, G., Shapiro, B., and Gilbert, M. (2010). Ancient DNA analyses exclude humans as the driving force behind late Pleistocene musk ox (*Ovibos moschatus*) population dynamics. *Proc. Natl Acad. Sci.*, 107:5675–5680.

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, 10(3):195–205.

Dansgaard, W., Johnsen, S., Clausen, H., Dahl-Jensen, D., Gundestrup, N., Hammer, C., Hvidberg, C., Steffensen, J., Sveinbjornsdottir, A., Jouzel, J., and Bond, G. (1993). Evidence for general instability of past climate from a 250 kyr ice-core record. *Nature*, 364:218–220.

Dansgaard, W., White, J., and Johnsen, S. (1989). The abrupt termination of the Younger Dryas climate event. *Nature*, 339:532–533.

De Maio, N., Wu, C.-H., O'Reilly, K. M., and Wilson, D. (2015). New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. *PLoS Genetics*, 11(8):e1005421.

Drummond, A. J. et al. (2002). Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data. *Genetics*, 161(3):1307–1320.

Drummond, A. J., Ho, S. Y., Phillips, M. J., and et al. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4(5):e88.

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Molecular Biology and Evolution*, 22(5):1185–1192.

Duane, S., Kennedy, A., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222.

Edwards, A. W. F. (1970). Estimation of the Branch Points of a Branching Diffusion Process. *Journal of the Royal Statistical Society. Series B (Methodological)*, 32(2):155–174.

Felsenstein, J. (1985). Phylogenies and the Comparative Method. *The American Naturalist*, 125(1):1–15.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates.

Frankham, R. (2015). Genetic rescue of small inbred populations: meta-analysis reveals large and consistent benefits of gene flow. *Molecular Ecology*, 24(11):2610–2618.

Frankham, R., Ballou, J. D., and Briscoe, D. A. (2010). *Introduction to conservation genetics*. Cambridge University Press.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press, 3rd edition.

Gill, M. S., Lemey, P., Faria, N. R., and et al. (2013). Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Molecular Biology and Evolution*, 30(3):713–724.

Gill, M. S., Lemey, P., Faria, N. R., and et al. (2016). Understanding Past Population Dynamics: Bayesian Coalescent-Based Modeling with Covariates. *Systematic Biology*, 65(6):1041–1056.

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L., Daly, J. M., Mumford, J. A., and Holmes, E. C. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science (New York, N.Y.)*, 303(5656):327–332.

Griffiths, R. C. and Tavare, S. (1994). Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1310):403–410.

GRIP Members (1993). Climate instability during the last interglacial period recorded in the GRIP ice core. *Nature*, 364:203–207.

Grootes, P., Stuiver, M., White, J., Johnsen, S., and Jouzel, J. (1993). Comparison of oxygen isotope records from the GISP2 and GRIP Greenland ice cores. *Nature*, 366:552–554.

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

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Hill, V., Ruis, C., Bajaj, S., Pybus, O. G., and Kraemer, M. U. (2021). Progress and challenges in virus genomic epidemiology. *Trends in Parasitology*, 37(12):1038–1049.

Hoffman, M. and Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J Mach Learn Res*, 15:1593–1623.

Ji, X., Zhang, Z., Holbrook, A., Nishimura, A., Baele, G., Rambaut, A., Lemey, P., and Suchard, M. A. (2020). Gradients do grow on trees: a linear-time O (N)-dimensional gradient for statistical phylogenetics. *Molecular Biology and Evolution*, 37(10):3047–3060.

Johnsen, S., Clausen, H., Dansgaard, W., Gundestrup, N., Hammer, C., Andersen, U., Andersen, K., Hvidberg, C., Dahl-Jensen, D., Steffensen, J., Shoji, H., Sveinbjornsdottir, A., White, J., Jouzel, J., and Fisher, D. (1997). The *d18O* record along the Greenland Ice Core Project deep ice core and the problem of possible Eemian climatic instability. *J. Geophys. Res.*, 102:26397–26410.

Jukes, T. and Cantor, C. (1969). Evolution of protein molecules. In 'Mammalian protein Metabolism'.(Ed. HN munro.) pp. 21–132. *Academic Press, New York)*, 1:504–511.

Kass, R. E., Carlin, B. P., Gelman, A., and et al. (1998). Markov chain Monte Carlo in practice: a roundtable discussion. *The American Statistician*, 52(2):93–100.

Kingman, J. F. (1982a). On the genealogy of large populations. *Journal of Applied Probability*, 19(A):27–43.

Kingman, J. F. C. (1982b). The coalescent. *Stochastic Processes and Their Applications*, 13(3):235–248.

Knorr-Held, L. and Rue, H. (2002). On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics*, 29(4):597–614.

Larget, B. and Simon, D. L. (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*, 16(6):750–759.

Lemey, P., Rambaut, A., Bedford, T., Faria, N., Bielejec, F., Baele, G., et al. (2014). Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2. *PLoS Pathogens*, 10(2):e1003932.

Lemey, P., Rambaut, A., Drummond, A. J., and Suchard, M. A. (2009). Bayesian Phylogeography finds its Roots. *PLoS Computational Biology*, 5(9):e1000520.

Lemey, P., Rambaut, A., Welch, J. J., and Suchard, M. A. (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology and Evolution*, 27(8):1877–1885.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Minin, V. N., Bloomquist, E. W., and Suchard, M. A. (2008). Smooth Skyride through a Rough Skyline: Bayesian Coalescent-Based Inference of Population Dynamics. *Molecular Biology and Evolution*, 25(7):1459–1471.

Moya, A., Holmes, E. C., and González-Candelas, F. (2004). The population genetics and evolutionary epidemiology of RNA viruses. *Nature Reviews Microbiology*, 2(4):279–288.

Neal, R. M. et al. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov chain Monte Carlo*, 2(11):2.

Nishimura, A. and Dunson, D. (2016). Geometrically tempered Hamiltonian Monte Carlo. *arXiv preprint arXiv:1604.00872*.

Opgen-Rhein, R., Fahrmeir, L., and Strimmer, K. (2005). Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo. *BMC Evolutionary Biology*, 5(6).

Pagel, M. and Meade, A. (2004). A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology*, 53(4):571–581.

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

Pybus, O. G. and Rambaut, A. (2009). Evolutionary analysis of the dynamics of viral infectious disease. *Nature Reviews. Genetics*, 10(8):540–550.

Pybus, O. G., Suchard, M. A., Lemey, P., Bernardin, F. J., Rambaut, A., Crawford, F. W., Gray, R. R., Arinaminpathy, N., Stramer, S. L., Busch, M. P., and Delwart, E. L. (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, 109(37):15066–15071.

R Core Team (2018). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology*, 67(5):901–904.

Rannala, B. and Yang, Z. (1996). Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution*, 43:304–311.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer Science & Business Media.

Rodrigo, A. G. and Felsenstein, J. (1999). Coalescent approaches to HIV population genetics. *The Evolution of HIV*, pages 233–272.

Suchard, M. A., Lemey, P., Baele, G., and et al. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4(1):vey016.

Suchard, M. A. and Rambaut, A. (2009). Many-core algorithms for statistical phylogenetics. *Bioinformatics*, 25(11):1370–1376.

Tavaré, S. (1986). Some probabilistic and statistical problems on the analysis of DNA sequence. *Lecture of Mathematics for Life Science*, 17:57.

Volz, E. M. and Siveroni, I. (2018). Bayesian phylodynamic inference with complex models. *PLoS Computational Biology*, 14(11):e1006546.

Waterman, M. S., editor (1986). *Some Mathematical Questions in Biology: DNA Sequence Analysis*. Lectures on Mathematics in the Life Sciences. American Mathematical Society, Providence, RI.

WHO: World Health Organization (2015a). Dengue. `http://www.who.int/topics/dengue/en/`.

WHO: World Health Organization (2015b). Rabies. `http://www.who.int/rabies/en/`.

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, 16(2):97–159.

Wu, C., Stoehr, J., and Robert, C. P. (2018). Faster Hamiltonian Monte Carlo by learning leapfrog scale. *arXiv preprint arXiv:1810.04449*.

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution*, 11(9):367–372.

Yang, Z. (2006). *Computational Molecular Evolution*. OUP Oxford.

Yang, Z. and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo Method. *Molecular Biology and Evolution*, 14(7):717–724.

Yang, Ziheng (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, 39(3):306–314.