University of California
Los Angeles

Bayesian Modeling of Viral Phylodynamics

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biostatistics

by

Mandev Singh Gill

2016

Abstract of the Dissertation

Bayesian Modeling of Viral Phylodynamics

by

Mandev Singh Gill

Doctor of Philosophy in Biostatistics

University of California, Los Angeles, 2016

Professor Marc Adam Suchard, Chair

Viral phylodynamics is the study of how immunodynamics, epidemiology, and evolutionary processes act and interact to shape viral phylogenies. We build upon the foundation of Bayesian phylogenetic inference to develop statistical tools to address phylodynamic problems. First, we present a flexible nonparametric Bayesian framework to infer the effective population size as a function of time directly from molecular sequence data. The effective population size is an abstract quantity that characterizes a population's genetic diversity, and it is of fundamental interest in population genetics, conservation biology, and infectious disease epidemiology. Our model is based on the coalescent, a stochastic process that relates phylogenies to population dynamics. We enforce temporal smoothing of inferred trajectories via a Gaussian Markov random field prior. Notably, our framework incorporates data from multiple genetic loci to achieve improved inference of population dynamics. Next, we turn to phylogenetic trait evolution. Modeling the processes giving rise to nonsequence traits associated with molecular sequence data is crucial in comparative studies of phenotypic traits as well as in phylogeographic analyses that reconstruct the spatiotemporal spread of viruses. A popular, yet restrictive approach to modeling such processes is Brownian diffusion along a phylogeny. We relax a major restriction by introducing a nontrivial estimable drift vector into the Brownian diffusion. Importantly, we implement a relaxed drift process that permits the drift vector to vary along the phylogeny. We showcase improved trait evolutionary inference in three viral examples. Finally, we return to effective population size inference and extend our framework to include covariates, enabling modeling of associations between past population dynamics and external factors. We apply our model to four examples. We reconstruct the demographic history of raccoon rabies in North America and find a significant association with the spatiotemporal spread of the outbreak. Next, we examine the effective population size trajectory of the DENV-4 virus in Puerto Rico along with viral isolate count data and find similar cyclic patterns. We compare the population history of the HIV-1 CRF02_AG clade in Cameroon with HIV incidence and prevalence data and find that the effective population size is more reflective of incidence rate. Finally, we explore the hypothesis that the population dynamics of musk ox during the Late Quaternary period were related to climate change. Incorporating covariates into the demographic inference framework enables the modeling of associations between the effective population size and covariates while accounting for uncertainty in population histories. Furthermore, it can lead to more precise estimates of population dynamics.

The dissertation of Mandev Singh Gill is approved.

James Oliver Lloyd-Smith

Janet Suzanne Sinsheimer

Robert Erin Weiss

Marc Adam Suchard, Committee Chair

University of California, Los Angeles

2016

To my family

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I want to thank my advisor Marc Suchard. It has been a privilege to work with and learn from Marc, and I can't imagine a better advisor. Everyone knows that Marc is an exceptional statistician, but he is also a great mentor who cares deeply about his students and is concerned with every aspect of their professional development. Most importantly, Marc stood by me when I struggled and never stopped believing in me. For that, I will always be grateful.

I also want to thank the other members of my doctoral committee, Janet Sinsheimer, Robert Weiss, and James Lloyd-Smith, for their very helpful comments and suggestions regarding my research. Dr. Sinsheimer and Dr. Weiss also played important roles during my first year of graduate school. It was in their courses that I was introduced to statistical methods in genetics and Bayesian inference, and what I learned helped set the stage for my research.

I have had the fortune of working with a wonderful set of collaborators: Philippe Lemey, Nuno Faria, Andrew Rambaut, Beth Shapiro, Lam Si Tung Ho, Guy Baele, Shannon Bennett, Roman Biek, Julia Palacios, and Vladimir Minin. I want to thank all of them, especially Philippe Lemey, who I've worked with on all of my projects.

I've had a great time at the UCLA Department of Biostatistics, and I want to thank all of the faculty who I've had an opportunity to learn from in courses and tea room discussions. I especially want to thank Professor William Cumberland, who supported me on his AIDS training grant.

Vita

2006	B.A., Mathematics University of California, Berkeley
2009	M.S., Mathematics University of California, Riverside
2007-2009	Teaching Assistant, Department of Mathematics, University of California, Riverside
2014	Teaching Assistant, Department of Biostatistics, University of California, Los Angeles

Publications and Presentations

Gill MS and Suchard MA. Phylogenetic trait evolution with drift. Presented at the Joint International Chinese Statistical Association Applied Statistics Symposium and Graybill Conference, Fort Collins, Colorado, June 2015.

Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, and Suchard MA. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Poster session presented at the UCLA Biomathematics Symposium, Los Angeles, California, March 2015.

Gill MS and Suchard MA. Diffusion models for phylogenetic trait evolution. Presented at the Joint Statistical Meetings, Boston, Massachusetts, August 2014.

Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, and Suchard MA. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Presented at the

International Indian Statistical Association Conference, Riverside, California, July 2014.

Palacios JA, Gill MS, Suchard MA, and Minin VN. *Bayesian Phylogenetics: Methods, Algorithms, and Applications,* chapter Bayesian Nonparameteric Phylodynamics. Chapman & Hall/CRC Mathematical and Computational Biology, 2014.

Gill MS and Suchard MA. Continuous phylogeographic diffusion with drift. Presented at the Joint Statistical Meetings, Montreal, Canada, August 2013.

Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, and Suchard MA. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Presented at the Western North American Region of the International Biometric Society Annual Meeting, Los Angeles, California, June 2013.

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

Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, and Suchard MA. Bayesian coalescent-based inference of population dynamics from multiple loci. Presented at the Joint Statistical Meetings, San Diego, California, August 2012.

CHAPTER 1

Introduction

The term "phylodynamics" was introduced in 2004 by Grenfell et al. (2004) to describe "the melding of immunodynamics, epidemiology, and evolutionary biology" required to "clarify how pathogen genetic variation, modulated by host immunity, transmission bottlenecks, and epidemic dynamics, determines the wide variety of pathogen phylogenies observed at scales that range from individual host to population." At the heart of this "melding" is the development of novel mathematical models and statistical inference frameworks. We contribute to this development by creating new statistical methods to enhance our understanding of infectious disease dynamics. In particular, we connect phylogenetic information to genetic diversity over time, the spatiotemporal spread of viruses, and phenotypic trait evolution.

All of our methodology is developed in the context of Bayesian phylogenetics. In Chapter 2, we review basic concepts in Bayesian phylogenetics in order to elucidate the foundations that we build upon. First, we review continuous-time Markov chain models for molecular character mutation on phylogenies. Then we discuss models for evolutionary rate variation among different DNA sequence sites, as well as different approaches to model evolutionary rate variation among lineages. We also outline the basic Bayesian modeling framework for phylogenetic inference.

Effective population size is fundamental in population genetics and characterizes genetic diversity. To infer past population dynamics from molecular sequence data, coalescent-based models have been developed for Bayesian nonparametric estimation of effective population size over time. Among the most successful is a Gaussian Markov random field (GMRF) model for a single gene locus. In Chapter 3, we present a generalization of the GMRF model that allows for the analysis of multilocus sequence data. Using simulated data, we demonstrate the improved performance of our method to recover true population trajectories and the time to the most recent common ancestor (TMRCA). We analyze a multilocus alignment of HIV-1 CRF02_AG gene sequences sampled from Cameroon. Our results are consistent with HIV prevalence data and uncover some aspects of the population history that go undetected in Bayesian parametric estimation. Finally, we recover an older and more reconcilable TMRCA for a classic ancient DNA data set. Chapter 3 is joint work with Philippe Lemey, Nuno Faria, Andrew Rambaut, Beth Shapiro, and Marc Suchard and has been published in *Molecular Biology and Evolution* (Gill et al., 2013).

Understanding the processes that give rise to quantitative characters associated with molecular sequence data remains an important issue in statistical phylogenetics. Examples of such characters include geographic coordinates in the context of phylogeography and phenotypic traits in the context of comparative studies. A popular approach is to model the evolution of continuously varying traits as a Brownian diffusion process acting on a phylogenetic tree. However, standard Brownian diffusion is quite restrictive and may not accurately characterize certain trait evolutionary processes. In Chapter 4, we relax one of the major restrictions of standard Brownian diffusion by incorporating a nontrivial estimable drift into the process. We introduce a relaxed drift diffusion model for the evolution of multivariate continuously varying traits along a phylogenetic tree via Brownian diffusion with drift. Notably, the relaxed drift model accommodates branch-specific variation of drift rates while preserving model identifiability. Furthermore, our development of a computationally efficient dynamic programming approach to compute the data likelihood enables scaling of our method to large data sets frequently encountered in viral evolution. We implement the relaxed drift model in a novel Bayesian inference framework to simultaneously reconstruct the evolutionary histories of molecular sequence data and associated multivariate continuous trait data, and provide tools to visualize evolutionary reconstructions. We illustrate the utility of our approach in three viral examples. In the first two, we examine the spatiotemporal spread of HIV-1 in central Africa and the West Nile virus in North America and show that a relaxed drift approach uncovers a clearer, more detailed picture of the dynamics of viral dispersal than standard Brownian diffusion. Finally, we study antigenic evolution in the context of

HIV-1 resistance to broadly neutralizing antibodies PG9, PG16 and VRC01. Our analysis reveals evidence of a continuous drift of HIV-1 at the population level towards enhanced resistance to VRC01 neutralization over the course of the epidemic. Chapter 4 is joint work with Philippe Lemey, Lam Si Tung Ho, Guy Baele, and Marc Suchard.

In Chapter 5, we return to effective population size inference and extend the nonparametric Bayesian framework from Chapter 3. Despite considerable progress, there remains a need for further development of statistical tools to understand past population dynamics. A major goal of demographic reconstruction is understanding the association between the effective population size and potential explanatory factors. We introduce a flexible model that incorporates external time-varying covariates into the inference framework. We demonstrate the utility of such an approach on four examples. First, we find striking similarities between the demographic and spatial expansion of raccoon rabies in North America. Second, we compare and contrast the cyclic epidemiological dynamics of dengue in Puerto Rico with patterns of viral diversity. Third, we examine the population history of the HIV-1 CRF02_AG clade in Cameroon and find that the effective population size is more reflective of HIV incidence than prevalence. Finally, we explore the relationship between musk ox population dynamics and climate change during the Late Quaternary period. Our extension to the Skygrid proves to be a useful framework for ascertaining the association between effective population size and external covariates while accounting for demographic uncertainty. Furthermore, we show that incorporating covariates into the demographic inference framework can improve estimates of effective population size trajectories, increasing precision and uncovering patterns in the population history that integrate the covariate data in addition to the sequence data. Chapter 5 is joint work with Philippe Lemey, Shannon Bennett, Roman Biek, and Marc Suchard.

Finally, we outline some future research directions in Chapter 6. In Chapter 3, we observe that combining data from several effectively unlinked genetic loci with the same demographic history enables improved inference of population trajectories. Similarly, combining data from multiple populations can improve inference. Demographic histories of different populations may exhibit similar, although not identical, trends over time. We outline a hierarchical modeling approach to share information about population dynamics among the different populations without strictly enforcing that they follow the same demographic history. We also discuss how hierarchical modeling can build upon the framework of Chapter 5 to pool information from several similar data sets to model associations with covariates.

CHAPTER 2

Bayesian Phylogenetic Inference

2.1 Introduction

We develop all of our methodology in the context of Bayesian phylogenetics. This section provides a brief overview of the standard Bayesian phylogenetic inference framework.

2.2 Phylogenetic Trees and Evolution

A phylogenetic tree τ is a directed graph that describes the evolutionary relationship between a group of, say, N taxa and their common ancestors. The taxa can represent species, genes, populations, or even individuals (Yang, 2006). The graph consists of edges and vertices, which are referred to as *branches* and *nodes*, respectively, in the context of a tree. Lengths of branches (edge weights) represent elapsed evolutionary time between the nodes they connect. The tree has N nodes of degree 1, known as *tips* or *leaves*, corresponding to the N taxa. The tips are assigned indices $i = 1, \ldots, N$. There are $N - 2$ nodes of degree 3 called *internal nodes*, with indices $i = N + 1, \ldots, 2N - 2$. Finally, there one node of degree 1 called the *root*, with index $2N - 1$. Internal nodes represent common ancestors of two or more taxa, while the root node represents the most recent common ancestor of all N taxa. Associated with each tip i of the phylogenetic tree is an observed nucleotide sequence of, say, length L: $\mathbf{S}_i = (S_{i1}, \ldots, S_{iL})$. Additionally, we posit unobserved sequences $\mathbf{S}_{N+1}, \ldots, \mathbf{S}_{2N-1}$ at the $N - 1$ root and internal nodes. The characters S_{ij} take on values in the set $\{A, G, C, T\}$, whose elements correspond to the DNA bases adenine (A), guanine (G), cytosine (C) and thymine (T). In RNA, uracil (U) is substituted for thymine.

The evolutionary process that generates the observed sequence data is typically modeled using continuous-time Markov chain (CTMC) models for nucleotide substitution. Many models make the following simplifying assumptions: all sites evolve according to the same tree, all sites evolve independently, all sites evolve according to the same stochastic process, and, conditional on the character at a given site of an internal node, evolution proceeds independently at the site along the two branches descending from the node (Lange, 2002). The character at a given site j mutates according to an instantaneous rate matrix $\mathbf{Q}$, and transition probabilities $\mathbf{P}(t)$ can be obtained through matrix exponentiation as $\mathbf{P}(t) = \exp(t\mathbf{Q})$. The matrix $\mathbf{P}(t)$ contains the probabilities of starting at a given nucleotide at time 0 and ending up at another nucleotide at time t. Equilibrium base frequencies are denoted $\pi_A, \pi_G, \pi_C, \pi_T$. Let b_{ik} denote a hypothetical branch connecting nodes i and k (where node k is a direct descendant of node i), let t_{ik} represent the length of a branch b_{ik}, and let $\mathcal{B}$ denote the set of all branches that exist in the phylogenetic tree τ. Then the likelihood of a particular set of sequences being observed on the nodes of the tree can be expressed as

$$P(\mathbf{S}_1, \ldots, \mathbf{S}_{2N-1} | \tau, \mathbf{Q}) = \prod_{b_{ik} \in \mathcal{B}} \prod_{j=1}^{L} \pi_{S_{(2N-1),j}} P_{S_{ij} S_{kj}}(t_{ik}). \tag{2.1}$$

The likelihood of observed sequences $\mathbf{S}_1, \ldots, \mathbf{S}_N$ can be obtained by summing over unknown ancestral sequences:

$$P(\mathbf{S}_1, \ldots, \mathbf{S}_N | \tau, \mathbf{Q}) = \sum_{\mathbf{S}_{N+1}} \cdots \sum_{\mathbf{S}_{2N-1}} \prod_{b_{ik} \in \mathcal{B}} \prod_{j=1}^{L} \pi_{S_{(2N-1),j}} P_{S_{ij} S_{kj}}(t_{ik}). \tag{2.2}$$

Computation of the likelihood (2.2) is made feasible through Felsenstein's tree-pruning algorithm (Felsenstein, 1981).

2.3 Nucleotide Substitution Models

There has been considerable development of different nucleotide substitution models, defined by different parameterizations of the CTMC instantaneous rate matrix $\mathbf{Q}$. The substitution models differ in biological assumptions about the mutation process and the number of free parameters that must be estimated. In our discussion of substitution models, the elements of **Q** are

$$\mathbf{Q} = \begin{pmatrix} - & Q_{AG} & Q_{AC} & Q_{AT} \\ Q_{GA} & - & Q_{GC} & Q_{GT} \\ Q_{CA} & Q_{CG} & - & Q_{CT} \\ Q_{TA} & Q_{TG} & Q_{TC} & - \end{pmatrix}, \tag{2.3}$$

where Q_{AG}, for instance, represents the instantaneous rate of substitution from nucleotide A to nucleotide G. The other nondiagonal elements of **Q** are defined analogously. Here, the character "-" in each row is determined by the requirement that the elements of each row must sum to 0.

2.3.1 The JC69 Model

The JC69 model (Jukes and Cantor, 1969) is one of the simplest models of nucleotide substitution. It assumes equal base frequencies $\pi_A = \pi_G = \pi_C = \pi_T = \frac{1}{4}$, and that every nucleotide has the same rate λ of mutating into any other nucleotide. Its substitution rate matrix is

$$\mathbf{Q} = \begin{pmatrix} - & \lambda & \lambda & \lambda \\ \lambda & - & \lambda & \lambda \\ \lambda & \lambda & - & \lambda \\ \lambda & \lambda & \lambda & - \end{pmatrix}, \tag{2.4}$$

yielding the transition probabilities

$$P_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}t} & \text{if } i = j \\ \frac{1}{4} - \frac{1}{4}e^{-\frac{4}{3}t} & \text{if } i \neq j. \end{cases} \tag{2.5}$$

2.3.2 The K80 Model

The DNA bases A and G are called purines, and the bases C and T are known as pyrimidines. Substitutions from purine-to-purine and pyrimidine-to-pyrimidine are called transitions, while substitutions from purine-to-pyrimidine and pyrimidine-to-purine are said to be transversions. Transitions typically occur more frequently than transversions in real data, and Kimura (1980) proposed a substitution model that allows for differing evolutionary rates of α for transitions and β for transversions. Like the JC69 model, however, the K80 model still assumes identical base frequencies $\pi_A = \pi_G = \pi_C = \pi_T = \frac{1}{4}$. Under these assumptions, we have

$$\mathbf{Q} = \begin{pmatrix} - & \alpha & \beta & \beta \\ \alpha & - & \beta & \beta \\ \beta & \beta & - & \alpha \\ \beta & \beta & \alpha & - \end{pmatrix}. \tag{2.6}$$

Alternatively, κ can denote the transition/transversion ratio, and the rate of transversions can be set to 1, giving us

$$\mathbf{Q} = \begin{pmatrix} - & \kappa & 1 & 1 \\ \kappa & - & 1 & 1 \\ 1 & 1 & - & \kappa \\ 1 & 1 & \kappa & - \end{pmatrix}. \tag{2.7}$$

The transition probabilities are

$$P_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{1}{4}e^{-4\beta t} + \frac{1}{2}e^{-2(\alpha+\beta)t} & \text{if } i = j \\ \frac{1}{4} + \frac{1}{4}e^{-4\beta t} - \frac{1}{2}e^{-2(\alpha+\beta)t} & \text{if substitution is transition} \\ \frac{1}{4} - \frac{1}{4}e^{-4\beta t} & \text{if substitution is transversion.} \end{cases} \tag{2.8}$$

2.3.3 The F81 Model

Assuming equal base frequencies is unrealistic for virtually every real data set (Yang, 2006). Felsenstein (1981) extended the JC69 model by allowing unequal base frequencies, subject to the constraints that each base frequency is nonnegative, and the frequencies sum to 1. The rate matrix of the F81 model is

$$\mathbf{Q} = \begin{pmatrix} - & \pi_G & \pi_C & \pi_T \\ \pi_A & - & \pi_C & \pi_T \\ \pi_A & \pi_G & - & \pi_T \\ \pi_A & \pi_G & \pi_C & - \end{pmatrix}. \tag{2.9}$$

The transition probabilities are

$$P_{ij}(t) = \begin{cases} e^{-\beta t} + \pi_j(1 - e^{-\beta t}) & \text{if } i = j \\ \pi_j(1 - e^{-\beta t}) & \text{if } i \neq j, \end{cases} \tag{2.10}$$

where

$$\beta = \frac{1}{1 - \pi_A^2 - \pi_C^2 - \pi_G^2 - \pi_T^2}. \tag{2.11}$$

2.3.4 The HKY85 and TN93 Models

The HKY85 model (Hasegawa et al., 1985) combines the distinction between transitions and transversions of the K80 model with the allowance of unequal base frequencies found in the F81 model. The rate matrix is parameterized as follows:

$$\mathbf{Q} = \begin{pmatrix} - & \alpha\pi_G & \beta\pi_C & \beta\pi_T \\ \alpha\pi_A & - & \beta\pi_C & \beta\pi_T \\ \beta\pi_A & \beta\pi_G & - & \alpha\pi_T \\ \beta\pi_A & \beta\pi_G & \alpha\pi_C & - \end{pmatrix}. \tag{2.12}$$

The TN93 model (Tamura and Nei, 1993) goes one step further than the HKY85 model by allowing different rates for the two different types of transitions (A $\longleftrightarrow$ G and C $\longleftrightarrow$ T). All transversions are assumed to occur at the same rate. The TN93 rate matrix is

$$\mathbf{Q} = \begin{pmatrix} - & \alpha_1\pi_G & \beta\pi_C & \beta\pi_T \\ \alpha_1\pi_A & - & \beta\pi_C & \beta\pi_T \\ \beta\pi_A & \beta\pi_G & - & \alpha_2\pi_T \\ \beta\pi_A & \beta\pi_G & \alpha_2\pi_C & - \end{pmatrix}. \tag{2.13}$$

Notably, setting $\alpha_1 = \alpha_2 = \alpha$ yields the HKY85 model (and additionally fixing $\pi_A = \pi_G = \pi_C = \pi_T = \frac{1}{4}$ gives us the K80 model). Setting $\alpha_1 = \alpha_2 = \alpha = \beta$ yields the F81 model.

2.3.5 The GTR Model

A CTMC is defined to be time-reversible if

$$\pi_i q_{ij} = \pi_j q_{ji}, \text{ for all } i \neq j. \tag{2.14}$$

This is known as the *detailed balance* condition and means that the "flow" from state i to state j, $\pi_i q_{ij}$, is equal to the flow in the opposite direction. There is no biological reason to expect the substitution process to satisfy time-reversibility, so it is simply a mathematical convenience (Yang, 2006). All of the nucleotide substitution models we have discussed are time-reversible, and it is natural to seek the most general such model. An equivalent condition for time-reversibility is that the rate matrix can be expressed as the product of a symmetric matrix and a diagonal matrix. Under such a parameterization, the diagonals in the diagonal matrix will specify the base equilibrium frequencies. Thus, the most general time-reversible substitution model is defined by the rate matrix

$$\mathbf{Q} = \begin{pmatrix} - & a & b & c \\ a & - & d & e \\ b & d & - & f \\ c & e & f & - \end{pmatrix} \begin{pmatrix} \pi_A & & & \\ & \pi_G & & \\ & & \pi_C & \\ & & & \pi_T \end{pmatrix} \tag{2.15}$$

$$= \begin{pmatrix} - & a\pi_G & b\pi_C & c\pi_T \\ a\pi_A & - & d\pi_C & e\pi_T \\ b\pi_A & d\pi_G & - & f\pi_T \\ c\pi_A & e\pi_G & f\pi_C & - \end{pmatrix}. \tag{2.16}$$

The resulting substitution model is known as the general time-reversible (GTR) model (Tavare, 1986).

2.4 Variation of Substitution Rates Across Sites

The mutation models described in Section 2.3 all assume that different sites in a sequence evolve according to the same process and at the same rate. This is an unrealistic assumption, and failure to account for rate variation can greatly impact a phylogenetic analysis (Yang, 1996; Sullivan and Swofford, 2001). Adopting a unique rate for every site leads to overparameterization, and a more sensible approach is to model rate variation with a statistical distribution (Yang, 2006).

The discrete-rates model of Yang (1995) posits that the sites are a finite mixture of, say, K classes. The rate at any site takes a value r_k with probability p_k, $k = 1, \ldots, K$, and the r_k and p_k are estimated via maximum likelihood from the data. Hasegawa et al. (1985) adopt a special case of this model assuming two classes: a class of invariable sites with rate $r_0 = 0$ and another class of sites with constant rate r_1.

A popular approach using a continuous distribution is to assume that the rate r for any site is a random variable drawn from a gamma distribution. However, algorithms for computing the likelihood under a continous-rates model can be slow (Yang, 1993). Another strategy is to use the discrete-rates model (Yang, 1995) as an approximation to the continous gamma-distributed model, and is called the discrete-gamma model (Yang, 1994).

2.5 Molecular Clocks

The molecular clock hypothesis posits that the rate of DNA sequence evolution is constant over time and among evolutionary lineages. The molecular clock provides a simple way of dating evolutionary events, with the evolutionary distance between sequences increasing linearly with the time of divergence. (Here, the evolutionary distance is defined as the expected number of nucleotide substitutions per site.) However, while the molecular clock assumption can hold for very closely related species, it is usually inappropriate in more distant comparisons (Yang, 2006).

A number of models have been developed to accommodate variation in evolutionary rates from lineage to lineage. Local clock models (Hasegawa et al., 1989; Rambaut and Bromham, 1998; Yoder and Yang, 2000) permit different user-specified regions of a phylogenetic tree to have different evolutionary rates. However, assigning branches to different rate groups can be difficult without strong *a priori* knowledge. Furthermore, they do not account for uncertainty in positions of rate changes between local clock regions, and they do not account for uncertainty in the phylogenetic tree topology. Drummond and Suchard (2010) overcome such difficulties by developing a random local clock model that permits multiple evolutionary rates on a tree, but estimates the number of different rates and positions of rate changes from the data. Furthermore, their random local clock accounts for phylogenetic tree uncertainty. Importantly, the random local clock model favors a small number of data-driven rate changes as opposed to numerous small or smoothly changing events.

An alternative to local clock models are "relaxed clock" models (Thorne et al., 1998; Aris-Brosou and Yang, 2002; Drummond et al., 2006). Thorne et al. (1998) and Aris-Brosou and Yang (2002) permit evolutionary rate variation among lineages in an autocorrelated manner, with the rate on each branch being drawn *a priori* from a distribution whose mean is a function of the evolutionary rate on the parent branch. Drummond et al. (2006) adopt a hierarchical approach under which the rate on each branch of the tree is drawn independently and identically from an underlying rate distribution.

2.6 Bayesian Inference

Let $\boldsymbol{\theta}$ denote the phylogenetic parameters (which, in the simplest case, include the phylogenetic tree τ and mutation process rate matrix $\mathbf{Q}$), and let $\mathbf{S} = (\mathbf{S}_1, \ldots, \mathbf{S}_N)$ denote the observed sequence data. In Section 2.2, we observed the construction of the likelihood $P(\mathbf{S}|\tau, \mathbf{Q})$. Maximum likelihood techniques are popular and exploit efficient algorithms to obtain point estimates (Felsenstein, 1981). However, parameter confidence interval estimates require computationally expensive bootstrapping techniques (Felsenstein, 1985a). Bayesian approaches offer an alternative, enabling direct measures of parameters of interest on a probability scale that is easy to interpret. Furthermore, Bayesian modeling allows prior biological knowledge to be incorporated into the inference framework. The Bayesian framework also makes it possible to quantify the impact of prior information on parameter estimates and their uncertainties (Drummond et al., 2002). On the other hand, Bayesian approaches can also be computationally onerous (though somewhat less so than bootstrapping under likelihood (Yang, 2006)), and the elicitation of appropriate prior distributions can be difficult.

The posterior distribution of phylogenetic parameters $\boldsymbol{\theta}$ can be computed through Bayes' theorem:

$$P(\boldsymbol{\theta}|\mathbf{S}) = \frac{P(\mathbf{S}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathbf{S})}. \tag{2.17}$$

Here, $P(\mathbf{S}|\boldsymbol{\theta})$ is the data likelihood developed in Section 2.2. $P(\boldsymbol{\theta})$ is the prior distribution of the phylogenetic parameters, quantifying our *a priori* beliefs about $\boldsymbol{\theta}$, and $P(\mathbf{S})$ is the marginal likelihood of the sequence data. Prior choices for the parameters of $\mathbf{Q}$ vary depending on the exact substitution model, with gamma and Dirichlet distributions commonly employed. A popular prior distribution for the phylogenetic tree is the birth-death prior (Rannala and Yang, 1996). This prior models trees as a forward-time process with new lineages arising through splits that occur with a birth rate λ, and with lineages going extinct at a death rate μ. A different class of popular phylogenetic tree priors are based on coalescent processes (Kingman, 1982b,a) that reconstruct genealogies from contemporary samples by proceeding backwards in time and merging lineages, with waiting times between such mergers being exponentially distributed. Coalescent tree priors are discussed in detail in Chapters 3 and 5.

Computation of the posterior distribution $P(\boldsymbol{\theta}|\mathbf{S})$ requires the integration

$$P(\mathbf{S}) = \int P(\mathbf{S}|\boldsymbol{\theta})P(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{2.18}$$

which is analytically intractable. Fortunately, Markov chain Monte Carlo (MCMC) algorithms afford an alternative path to estimate the posterior, only requiring evaluation of $P(\mathbf{S}|\boldsymbol{\theta})P(\boldsymbol{\theta})$. MCMC methods generate a dependent sample $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ from the posterior distribution $P(\boldsymbol{\theta}|\mathbf{S})$. In particular, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ form an ergodic discrete-time Marko chain with equilibrium distribution $p(\boldsymbol{\theta}|\mathbf{S})$. Because the chain is guaranteed to converge to the equilibrium distribution, the states of the chain after a given number of steps can be used a sample from the posterior and used to compute summary statistics. The most common approach to construct such a chain is the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970). Given current state $\boldsymbol{\theta}^{(n)}$, a new state $\boldsymbol{\theta}^*$ is proposed according to a proposal distribution $Q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(n)})$. The new state is accepted $\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^*$ with probability

$$\alpha(\boldsymbol{\theta}^{(n)}, \boldsymbol{\theta}^*) = \min \left[1, \frac{Q(\boldsymbol{\theta}^{(n)}|\boldsymbol{\theta}^*)P(\boldsymbol{\theta}^*|\mathbf{S})}{Q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(n)})P(\boldsymbol{\theta}^{(n)}|\mathbf{S})} \right]. \tag{2.19}$$

If the proposed new state is not accepted, then $\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)}$. Finding a suitable proposal density that ensures sufficient mixing in the target distribution can be challenging and is of central importance in construction of MCMC algorithms.

CHAPTER 3

Improving Bayesian Population Dynamics Inference: a Coalescent-Based Model for Multiple Loci

3.1 Introduction

Coalescent theory has become a cornerstone of computational population genetics. First introduced by Kingman (1982b), the coalescent is a stochastic process that generates genealogies relating a random sample of individuals arising from a classic forward-time population model (such as the Fisher-Wright model). The basic assumptions on such an idealized population are a constant population size, no selection or migration, nonoverlapping generations, and an equal propensity among individuals to produce offspring.

Researchers have extended coalescent theory to accomodate a range of relaxed assumptions about the population of interest, including a variable population size (Griffiths and Tavaré, 1994; Donnelly and Tavaré, 1995), and serially-sampled data (Rodrigo and Felsenstein, 1999). Notably, coalescent-based inference methods enable estimation of population genetic parameters from a fixed genealogy and, because genealogical shapes leave their imprints in the genomes of sampled individuals, directly from molecular sequence data (Hein et al., 2005).

One parameter of great scientific interest is the effective population size over time (often called the demographic model or demographic function). The effective population size is an abstract quantity that corresponds to the population size under an idealized model of reproduction. The census population size can be recovered from the effective population size by appropriate scaling. The utility of the effective population size is that it provides a measure of genetic diversity and its fluctuations over time, and acts as a "common denominator", allowing researchers to compare populations arising from different reproductive models. As recent examples, Campos et al. (2010) reconstruct the demographic history of musk ox from ancient DNA sequences to examine the cause of the reduction in their mitochondrial diversity, Rambaut et al. (2008) uncover trends of genetic diversity of the influenza A virus and compare them to the seasonal occurence of influenza, and Faria et al. (2012) explore past population dynamics of HIV-1 CRF02_AG gene sequences sampled in Cameroon.

Computational biologists and statisticians have posited a number of coalescent-based models to infer population dynamics across time. Many of these models (Kuhner et al., 1998; Drummond et al., 2002) characterize the effective population size over time using simple parametric functions (examples of such scenarios include constant size, exponentially growing, or logistically growing populations). This approach is advantageous in that there are relatively few parameters to be estimated, and hypothesis testing is convenient. However, *a priori* parametric functions may not accurately characterize important aspects of the population history of a given sample, and finding an appropriate parametric model can be difficult and time consuming (Baele et al., 2012a). Accordingly, nonparametric approaches have become popular in recent years. These approaches typically center around approximating the population history with a piecewise constant or linear function. Some of the first nonparametric models (Pybus et al., 2000; Strimmer and Pybus, 2001) provide fast but noisy estimation of population trajectories from a fixed genealogy. More recent models (Drummond et al., 2005; Opgen-Rhein et al., 2005; Minin et al., 2008) estimate population trajectories jointly, along with genealogies and mutational parameters, directly from sequence data in a Bayesian framework. These models differ primarily by the priors they place on the effective population size, and the choice of prior influences not only the estimated effective population size trajectory but also the estimated genealogy (in particular, the time of the most recent common ancestor). Opgen-Rhein et al. (2005) and Drummond et al. (2005) use multiple change-point models to estimate past population dynamics. The latter model is called the Bayesian Skyline, and Heled and Drummond (2008) have developed an extension called the Extended Bayesian Skyline Plot (EBSP) model that incorporates data from multiple unlinked genetic loci. The model proposed by Minin et al. (2008), called the Skyride, exploits a Gaussian Markov random field (GMRF) prior to achieve temporal smoothing.

Here, we present a novel Bayesian nonparametric model, named the Skygrid, to estimate effective population size trajectories. Like the Skyride, we parameterize the effective population size as a piecewise constant function and employ a GMRF prior to smooth the trajectory. However, while the Skyride allows the estimated trajectory to change at coalescent times, our improved method does so at pre-specified fixed points in real time. This grants the user extra flexibility and provides a natural framework to extend the model in the future to incorporate covariate values. Furthermore, this distinction enables the Skygrid's GMRF prior to be independent of the genealogy, which has important implications for estimation of the time to the most recent common ancestor. Another departure from the Skyride, and a major advantage of our model, is the ability to base the estimation on data from multiple unlinked genetic loci. Data from effectively unlinked loci are rapidly becoming the norm in the era of next-generation sequencing. Through simulation, we demonstrate that increasing the number of loci improves estimation of past population dynamics in terms of both bias and precision. We also compare the performance of the Skygrid with the EBSP in two different simulation scenarios and find that the Skygrid compares favorably. The limited number of scenarios prevents us from a comprehensive comparison of the models, but we can still conclude that the Skygrid is a competitive alternative to the EBSP. We also show the improvement of our model over the existing Skyride and Bayesian Skyline models in terms of estimation of the time to the most recent common ancestor for single locus data sets arising from three different demographic models. We analyze a multilocus data set of CRF02_AG gene sequences sampled in Cameroon and demonstrate that our nonparametric approach is able to recover characteristics of the sample's population history that are undetected by existing parametric models.

3.2 New Approaches

The Skygrid is a Bayesian nonparametric model that estimates $N_e(t)$, the effective population size over time, directly from a sample of multilocus molecular sequence data. Here, $t = 0$ is the most recent sampling time and the time t increases into the past. Thus, $N_e(0)$ is the effective population size at the most recent sampling time and $N_e(t)$ is the effective population size t time units prior to that. We estimate the effective population size trajectory as a piecewise constant function that changes values at pre-specified times called grid points. The user is allowed to specify the number of grid points M and a cutoff value K. The grid points are typically equally spaced between times $t = 0$ and $t = K$. The estimated trajectory is constant between grid points and constant for all times further into the past than the cutoff value K, and the values it assumes come in the form of a vector of length $M+1$. To smooth the trajectory, we place a GMRF prior on the vector of effective population sizes. The effective population size is estimated jointly along with mutation parameters, a GMRF precision parameter, and genealogies representing the ancestries of samples at the different genetic loci. We highly recommend reading the Methods section for further details before proceeding.

3.3 Results

3.3.1 Simulation Studies

We assess the performance of our model in recovering population dynamics in a series of simulation studies. In all our analyses, we transform the effective population size by taking the natural logarithm. To generate a synthetic data set, we first simulate a genealogy assuming one of following demographic models:

1. Constant population size: $\log N_e(t) = 1$
2. Exponential growth: $\log N_e(t) = \log 150 - t$

3. Exponential growth followed by a crash:

$$\log N_e(t) = \begin{cases} \log 150 - t & \text{if } t > 1.5 \\ \log(7.4681) + t & \text{if } t \in [0, 1.5] \end{cases} \tag{3.1}$$

In these models, we measure time in expected mutations per site. The genealogy has 30 tips sampled at time $t = 0$. Next, we use a molecular sequence simulator available in BEAST to generate sequence data on the tips of the genealogy. We assume a molecular clock under the HKY85 CTMC model (Hasegawa et al., 1985) with a transition/transversion rate ratio fixed to 4.0. To simulate a data set with n unlinked loci, we repeat this process n times. We consider data sets with 1, 2, 5 and 10 loci.

We analyze all data sets using the Skygrid model with 29 grid points and a cutoff value of 10. This way, the vector of effective population sizes has length of 30, equal to the number of individuals sampled in the data set. Furthermore, the cutoff value is greater than the root heights of typical genealogies generated by the coalescent under the aforementioned demographic scenarios. This goes towards ensuring that we capture as much of the population trajectory as possible given the data at hand.

Figure 3.1 illustrates the results of estimating the effective population size trajectories of constant size populations. The bold lines in the plots correspond to posterior medians and 95% Bayesian credibility intervals (BCIs) are shown as gray shaded areas. The dashed lines represent the true population trajectories. The model does a reasonably good job of recovering the true effective population size trajectory. In each plot, the BCIs increase as we move from the present time to the past. This is representative of the fact that, for constant populations, coalescent events become increasingly rare as we move away from the tips of the genealogy and towards the root. In other words, there are typically fewer data points (coalescent events) to inform the estimation near the root of the tree. We also see that the width of the BCI region decreases as more loci are incorporated into the analysis. The shrinkage is most dramatic as we go further back in time where data are scarce. This is due to the fact that increasing the number of loci is a very effective way of providing precious extra information in that time frame, and it illustrates a major advantage of performing a multilocus analysis.

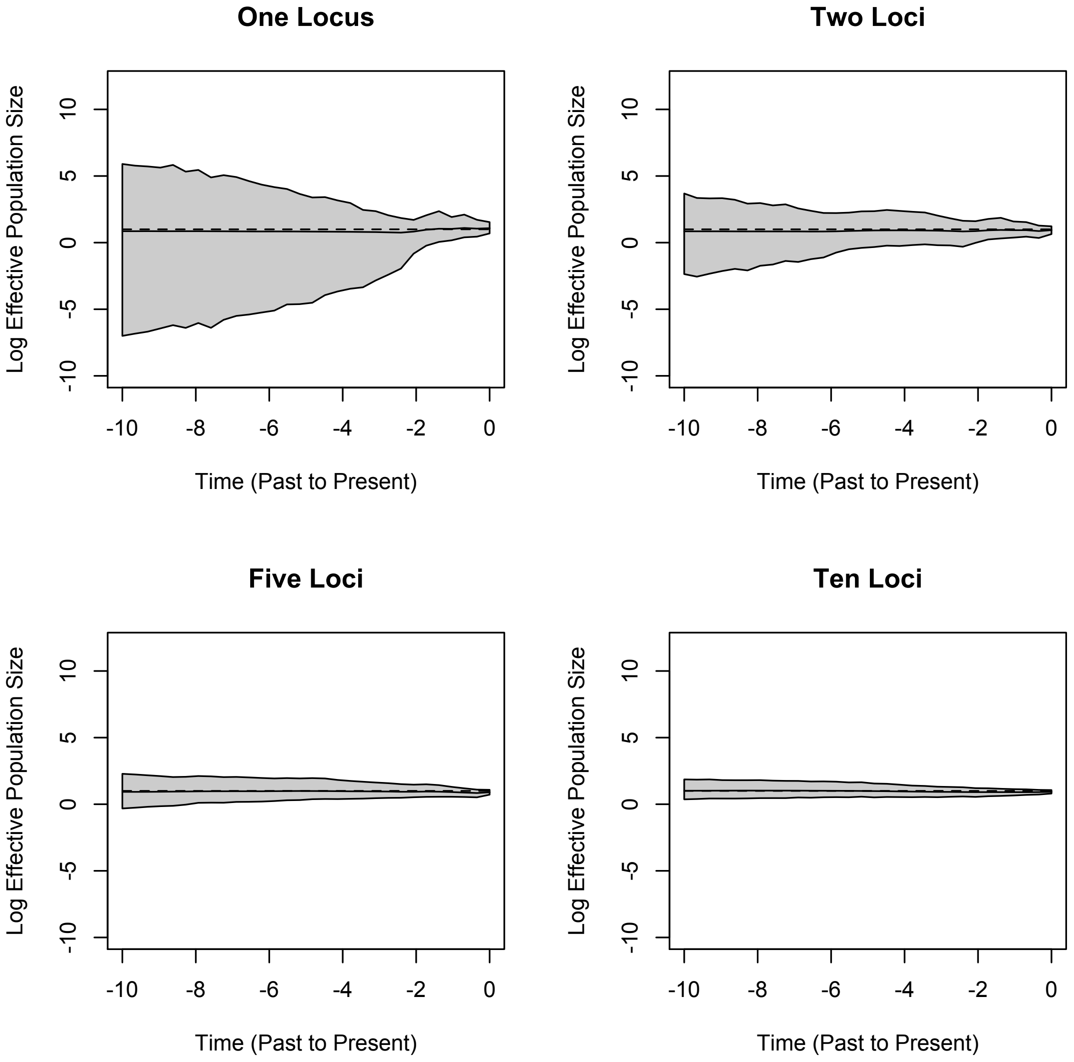

Figure 3.1: Constant population size simulation. We present plots of posterior medians (solid black lines) and 95% Bayesian credibility intervals (gray shading) of the effective population size $N_e(t)$ based on data sets with 1, 2, 5 and 10 loci. The true population size trajectories are depicted by dashed lines. Here and in all subsequent plots of effective population sizes, we use the log transformation of the population size axis.

Figure 3.2 shows the results under the exponential growth demographic model. As is the case with the constant demographic model, including data from additional loci leads to more precise esimation. Note that in each plot, following the trajectory from right to left (from present to past), the posterior median curve is very close to the true effective population size until it reaches a certain point, after which the curve follows a constant trajectory. In each plot, the posterior median becomes constant around the time of the greatest of the root heights of the coalescent trees which are used to generate the data. For instance, the greatest root height of the trees used to generate the 10-loci data set is 6.07. This flattening occurs because, beyond the greatest root height, the estimated effective population size is primarily informed by the prior rather than the non-existent data. It is important when drawing inferences to take note of the estimated root height to get an idea of where the trajectories are informative and where they are not.

Figure 3.3 depicts results in the case of populations that undergo a period of exponential growth followed by a period of exponential decline. As in the exponential growth case, the estimated trajectory is constant (and uninformative) during the time frame preceding the greatest root height of the trees used to generate the data. We do not accurately recover the overall trend of the demographic history in the one locus plot. While it does show a clear period of growth, the decline is rather mild and the time of transition from growth to decline is imprecise and occurs before the actual transition time. However, the remaining plots show that we infer with greater accuracy the transition time and the rates of growth and decline as we incorporate more loci into the analysis. These findings illustrate that important aspects of a population's demographic history may go undetected in a standard one locus analysis, but that increasing the number of loci can recover them.

Let $\hat{N}(t)$ denote the estimated posterior median effective population size, and $\hat{N}_{2.5}(t)$ and $\hat{N}_{97.5}(t)$ the 2.5 and 97.5 percent quantiles of the estimated posterior effective population size, respectively. To provide a comparative summary of the performance of our model for data sets with varying numbers of loci, we use the percent error and size, which are defined as follows:

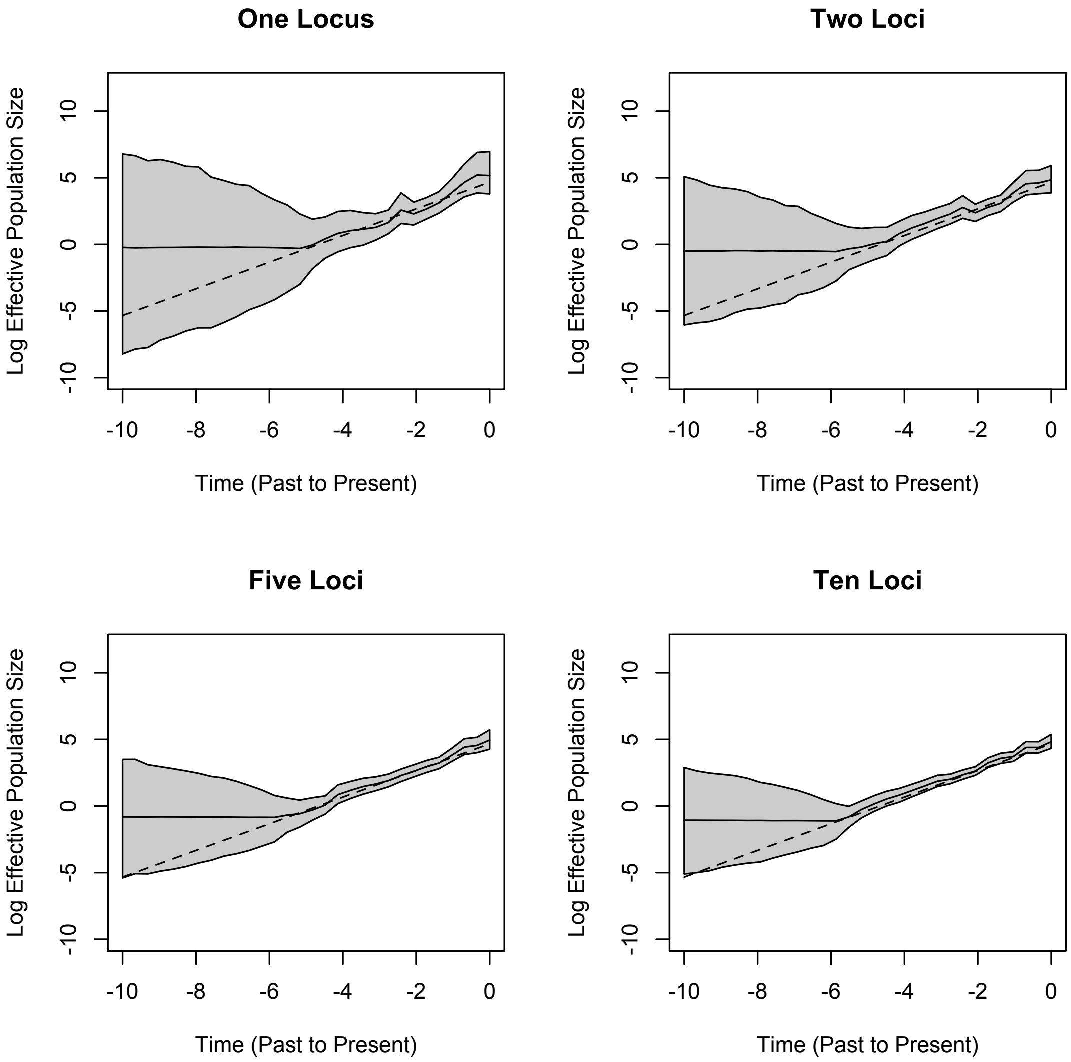

Figure 3.2: Exponential growth simulation. See Figure 3.1 for the legend explanation. The times of divergence between the estimated trajectories in solid black lines and the true trajectories depicted by dashed lines correspond approximately to the greatest root heights of the trees used to generate the data sets and illustrate the importance of the estimated root height in understanding the range over which the plots are informative.

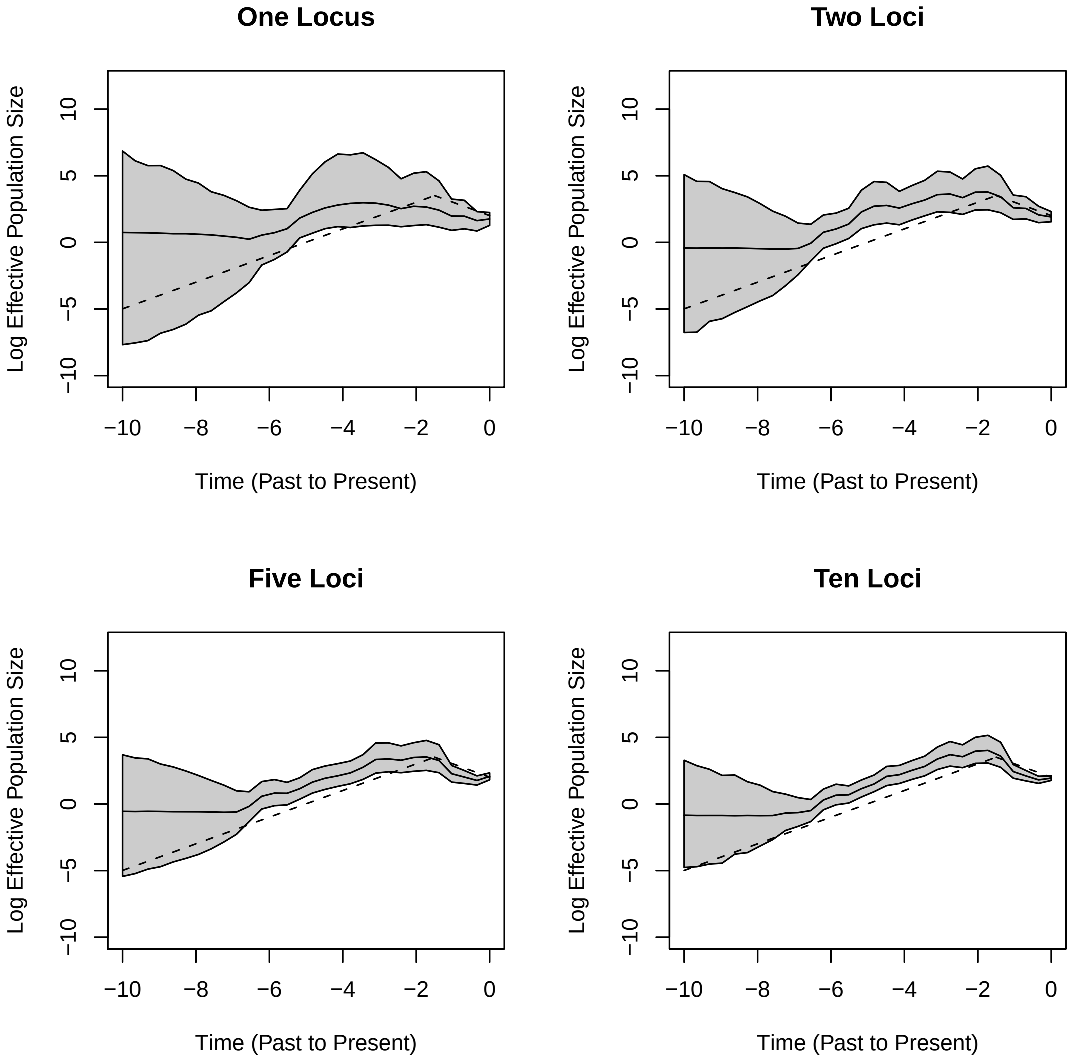

Figure 3.3: Simulation for a population that experiences exponential growth followed by a decline. See Figure 1 for the legend explanation. As in Figure 3.2, the trajectories are constant (and not informative) during time range $(-10, -7)$ which precedes the greatest root height of the trees used to generate the data sets. The plots illustrate the improvement in correctly recovering past population trends by incorporating data from additional loci.

$$\text{Percent Error} = 100 \times \frac{1}{R_{\max}} \int_0^R \frac{|\hat{N}(t) - N_e(t)|}{N_e(t)} \mathrm{d}t \tag{3.2}$$

and

$$\text{Size} = \frac{1}{R_{\max}} \int_0^R \frac{|\hat{N}_{97.5}(t) - \hat{N}_{2.5}(t)|}{N_e(t)} \mathrm{d}t. \tag{3.3}$$

Here, R is the maximum of the mean estimated root heights for a given data set and $R_{\max}$ is the root height of the tallest tree used to generate the data set. We use R as the upper limit in the integrals because the maximum root height provides an indication of how far back in time the data are informative. Dividing by $R_{\max}$ adjusts the metrics to ensure they provide measures of bias and variance that are not inflated for data sets that are informative for longer time spans.

Our results based on 100 simulated data sets are summarized in Table 3.1. We report relative percent error, which we obtain by dividing the mean percent error of 100 simulated data sets for a given number of loci by the mean percent error of 100 simulated one-locus data sets. We also report the relative size, which is defined analogously.

Under all three demographic models, the size and percent error decrease as the number of loci increases. In other words, multilocus data improve estimation in terms of both bias and precision.

To compare the Skygrid with the EBSP, we analyze the same simulated datasets generated for the Skygrid performance analysis using the EBSP. We compare these two models since they are, to our knowledge, the only coalescent-based nonparametric Bayesian models that infer population dynamics from multilocus data. In Table 3.2, we report the relative percent error and size, where we define the relative value of each metric as the mean value over 100 simulations based on EBSP analysis divided by the mean value over 100 simulations based on the Skygrid analysis. The Skygrid almost always outperforms the EBSP by a wide margin in terms of percent error. The EBSP analyses generally have smaller sizes, but in light of the much greater percent error, this extra “precision” is not especially meaningful. Indeed, an investigation of 95% BCI regions and the proportion of the true trajectory that

	Constant		Exponential		Crash	
	Relative		Relative		Relative	
Loci	Percent Error	Size	Percent Error	Size	Percent Error	Size
1	1.00	1.00	1.00	1.00	1.00	1.00
2	0.89	0.31	0.75	0.59	0.80	0.68
5	0.69	0.16	0.57	0.38	0.71	0.25
10	0.55	0.13	0.46	0.27	0.69	0.19

Table 3.1: Improvement of Skygrid performance with additional loci. Percent error and size, relative to estimates for one locus data sets, in simulations under the demographic scenarios of constant population size, exponential growth and exponential growth followed by a crash.

	Constant		Exponential		Crash	
	Relative		Relative		Relative	
Loci	Percent Error	Size	Percent Error	Size	Percent Error	Size
1	2.88	0.76	1.23	0.32	0.97	0.20
2	2.94	1.07	1.66	0.50	1.28	0.47
5	2.28	0.93	1.90	0.58	1.26	0.73
10	1.45	0.62	2.41	0.63	1.19	0.76

Table 3.2: Performance of Extended Bayesian Skyline Plot (EBSP) model relative to Skygrid. Percent error and size based on EBSP analyses, relative to estimates based on Skygrid analyses, in simulations under the demographic scenarios of constant population size, exponential growth and exponential growth followed by a crash.

each BCI region covers reveals that the Skygrid outperforms the EBSP three-fold for the exponentially growing populations and by 6-19% for the constant size populations. The Skygrid thus emerges as a better overall choice in common situations comparable to our simulation set-up.

3.3.2 Performance and Mixing

In all simulation studies, we simulate MCMC chains of length set to 20 million steps and sub-sample the chain every 1000 states, after discarding the first 10% as burn-in. To confirm sufficient mixing within the MCMC chain, we examine the effective sample size (ESS) scores of the model parameters and note that all ESS scores for effective population size parameters are over 1000.

The inclusion of data from additional loci adds to the complexity of the model and increases the run time necessary to achieve sufficient mixing. To investigate the computational cost of increasing the number of loci in a Skygrid analysis, we examine the ESS scores per unit time for effective population size parameters. We conduct all analyses on a 2.93 Ghz Intel Core 2 Duo processor with 4 GB of RAM. ESS per minute for data sets with 1, 2, 5 and 10 loci have respective ranges of $126.9 - 493.1$, $124.8 - 271.5$, $69.2 - 245.7$ and $41.3 - 137.4$ across all effective sample size parameters. These findings suggest that, while increasing the number of loci in a Skygrid analysis necessitates longer run times, the marginal cost is not especially high. For instance, a ten-fold increase in the number of loci does not require a ten-fold increase to achieve the same ESS. The feasibility of Skygrid analyses of data sets with large numbers of loci is encouraging in light of the increasing availability of multilocus data sets and the improvements they confer upon inference of past population dynamics.

3.3.3 Choice of Cutoff Values and Grid Points

Because it is up to the user to specify the cutoff value K and number of grid points M, it is natural to wonder how this choice will influence inference. A natural desired feature of the cutoff value is that it be sufficiently greater than the root height of the unobserved coalescent process, thereby allowing the analysis to capture as much information about the population dynamics as the data allow. An initial choice of a cutoff value can be informed by prior knowledge or a time frame of scientific interest. Otherwise, we recommend performing a preliminary analysis and examining the estimated root height to determine the need to possibly adjust the cutoff.

To investigate the sensitivity of the estimated root height to different choices of cutoff values, we consider a 100-taxa data set simulated under an exponential growth demographic model with demographic function $\log N_e(t) = \log 150 - \frac{t}{10}$. We assume a molecular clock under the HKY85 CTMC model (Hasegawa et al., 1985) with a transition/transversion rate ratio fixed to 4.0. The true root height of the coalescent tree used to generate the data is 27.1. We estimate the root height under our model using cutoff values of $2, 4, 6, \ldots, 28$. We adjust the number of grid points in each case so that the grid points remain 0.2 units apart. The results are summarized in Figure 3.4. The black dots represent posterior mean root height estimates and are connected by a dotted line, while the true root height is marked by a dashed line. The shaded gray rectangles show the coverage of the 95% BCI for each cutoff value.

As we see, increasing the cutoff value generally leads to more precise and less biased estimates of the root height. Since low cutoff values force the estimated population size to be constant for the bulk of the demographic history, this illustrates the advantage, when estimating the root height, of using a flexible model that allows the population trajectory to change over time. At the same time, the posterior mean estimated root heights and 95% BCIs using relatively low cutoff values are not especially far off from the estimates found using a cutoff value greater than the true root height. This is convenient because because it allows the user to make an informative adjustment of the cutoff value for a subsequent analysis if the current cutoff value turns out to be too low. In our example, low cutoff values lead to overestimates of the root height because constant size populations generally have greater root heights than exponentially growing populations.

In our analysis of cutoff values, we observe another nice feature of the Skygrid, in addition to the generally low root height estimate sensitivity: the choice of cutoff value does not have

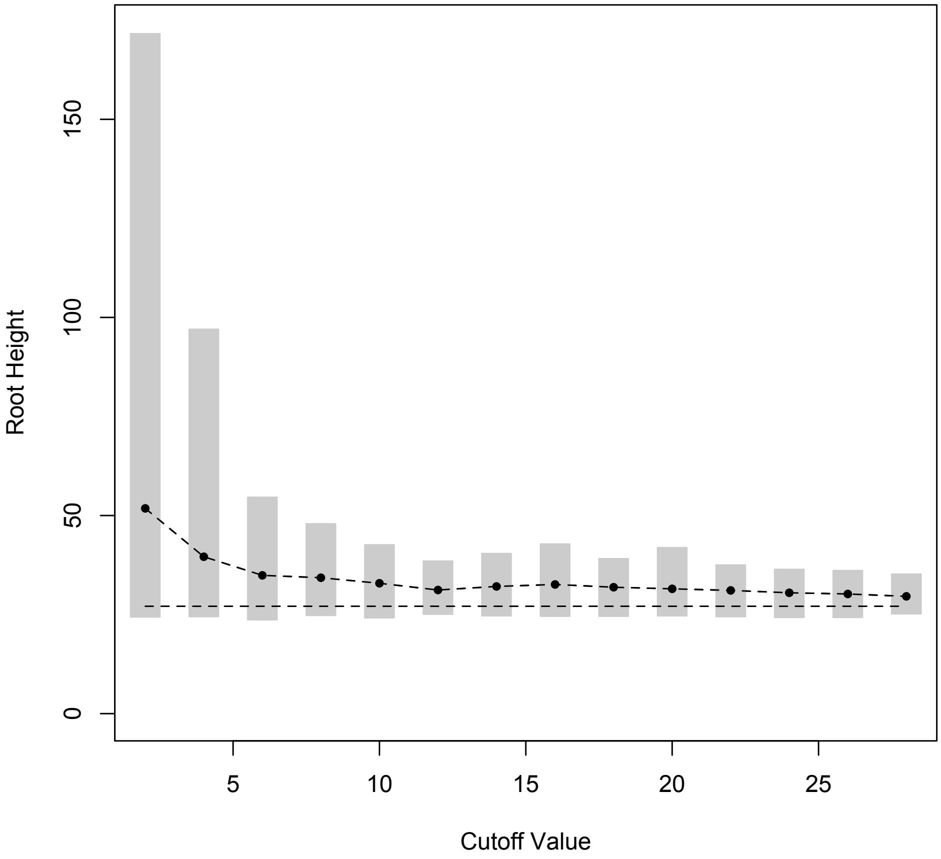

Figure 3.4: Root height estimate sensitivity. Here we compare the estimated root height for a simulated exponential growth data set using different cutoff values. The black dots represent posterior mean root height estimates, and the shaded gray rectangles represent the 95% Bayesian credibility intervals. The dashed line indicates the true root height of 27.1.

any notable impact on the trajectory of the estimated effective population size prior to the cutoff value.

With respect to the number of grid points, it is advisable to specify enough points at different times to capture any possible population trends. Our default suggestion is to specify one less grid point than the number of taxa (so that the length of the population size vector will be the same as the number of taxa) and space grid points evenly. This spacing gives us an equal opportunity to detect trends at different times, and the resolution coincides in a rough sense with the amount of available data. However, a major advantage of our model is the ability to set grid points at any desired time. The user has the flexibility, for instance, to concentrate grid points in time intervals in which the data are more informative or in regions in which prior beliefs suggest rapid changes.

3.3.4 Estimation of Time to Most Recent Common Ancestor

It is often of interest to estimate the time to the most recent common ancestor (TMRCA), also known as the root height, from genetic sequence data. While the Skygrid is in a sense a generalization of the Skyride, the Skyride's GMRF prior conditions on the genealogy whereas the Skygrid's does not, and this can affect TMRCA estimation. We wish to compare the performance of the Skygrid, Skyride and Bayesian Skyline models in estimating the TMRCA from one-locus data sets. We consider the one-locus case because the Skyride is not equipped analyze multilocus data sets. We conduct a series of simulations under three different demographic scenarios. First, a constant population with demographic function $\log N_e(t) = 1$, and second, an exponentially growing population with demographic function $\log N_e(t) = \log 150 - t$. The third demographic scenario is a four-epoch piecewise exponential model motivated by the Beringian bison data set discussed below.

To analyze a data set of 152 mtDNA control region sequences from ancient bison in Beringia (Siberia, Alaska, and north-western Canada) and central North America, Shapiro et al. (2004) implement a coalesecent-based two-epoch parametric demographic model in BEAST. The model is characterized by two phases of exponential growth at different rates,

				Frequentist
Model	Demographic	Percent Error	Size	Coverage
Skyride	Constant	3.99	0.77	89
Skyline	Constant	3.69	0.79	94
Skygrid	Constant	3.66	0.79	92
Skyride	Exponential	0.99	0.26	92
Skyline	Exponential	1.02	0.26	91
Skygrid	Exponential	0.99	0.26	94
Skyride	Beringian bison	69.79	24621.74	1
Skyline	Beringian bison	9.90	77742.94	96
Skygrid	Beringian bison	9.29	74634.97	96

Table 3.3: Estimation of time to most recent common ancestor. Size is measured in years for Ancient DNA demographic and in substitutions per site for other demographic models. Here, Skyline refers to the Bayesian Skyline.

and a transition time between the phases. Their analysis suggests an initial phase of exponential growth followed by a period of exponential decline, with a transition time around 32-43 ka BP (where 1 ka BP is 1000 years before present). The estimated TMRCA has a posterior mean of 136 ka BP with a 95% BCI of (111 ka BP, 164 ka BP).

We analyze the same data using the Skyride as well as our Skygrid model (with 150 grid points and a cutoff of 150 ka BP). Both analyses suggest a period of sustained population growth, peaking at about 35-45 ka BP, followed by a period of decline bottoming out around 10 ka BP, and then a post-bottleneck recovery. The post-bottleneck recovery, which is not identified by the two-epoch parametric model, is also observed in a nonparametric analysis by Drummond et al. (2005) using the Bayesian Skyline. While all of the nonparametric analyses uncover similar demographic histories, the same cannot be said for estimating the TMRCA. The Skyride gives us a posterior mean TMRCA of 101.45 ka BP with a 95% BCI of (87.12 ka BP, 117.5 ka BP), the Bayesian Skyline gives us a posterior mean TMRCA of 133.56 ka BP with a 95% BCI of (103.86 ka BP, 167.63 ka BP), and the Skygrid yields a posterior mean of 130.39 ka BP with a 95% BCI of (99.99 ka BP, 159.54 ka BP).

The Skygrid and Bayesian Skyline estimates are similar to those of Shapiro et al. (2004), while the Skyride analysis paints a substantially different picture. The Skyride results do not agree with the North American fossil record; bison are known to have been present in Alaska during the last interglacial interval (150-100 ka BP). To further investigate which estimates are closer to the truth, we test the Skygrid, Bayesian Skyline and Skyride on simulated data sets that are similar to the bison data set. We generate the data sets using evolutionary parameter values similar to the estimated values from the bison data set along with a four-epoch demographic model (which we refer to as the 'Ancient DNA' model) that grows and declines exponentially at approximately the same times and rates as the estimated trajectory using the Skygrid model on the bison data.

For each of the three demographic scenarios, we simulate 100 one-locus genetic sequence data sets and estimate the root heights using the three different models. To provide a comparative summary of the performance, we define the percent error as follows:

$$\text{Percent Error} = 100 \times \frac{|\text{Estimated Mean TMRCA - True TMRCA}|}{\text{True TMRCA}}. \tag{3.4}$$

Also, we define the size of each estimate as the length of the 95% BCI. Finally, we monitor the percentage of BCIs that contain the true root height as a measure of frequentist coverage, a useful property for inference tools that will be applied to many independent data sources. Ideally, estimated coverage should approach its nominal level; 0.95 in this case.

The simulation results are presented in Table 3.3. The three different models exhibit similar performance in the constant and exponential growth demographic scenarios. The Skygrid performs slightly better than the other two models in the case of exponentially growing populations. For the constant population simulations, each of the three performance metrics identifies a different model as the best, and none of the models dramatically outperforms the others in any way. In the Ancient DNA demographic scenario, the Skygrid outperforms both models. The contrast with the Skyride in terms of relative error and frequentist coverage of the true root height is especially dramatic. For each of the three demographic situations, the Skygrid model performs as good or better than the Skyride and Bayesian Skyline. Our simulation studies thus offer support for the Skygrid as the best of the three models for estimating the TMRCA from populations with a variety of demographic histories.

3.3.5 Population History of HIV-1 CRF02_AG Clade in Cameroon

Circulating recombinant forms (CRFs) are genomes that result from recombination of two or more different HIV-1 subtypes and that have been found in at least 3 epidemiologically unrelated individuals. CRF02_AG is globally responsible for 7.7% of HIV infections (Hemelaar et al., 2011) , but HIV/AIDS surveillance studies indicate that it accounts for approximately 60% of infections in Cameroon (Brennan et al., 2008).

Faria et al. (2012) investigate the population dynamics of the CRF02_AG lineage through a multilocus alignment of 336 *gag* (HXB2: 1255-1682), *pol* (HXB2: 4228-5093) and *env* (HXB2: 7890-8266) CRF02_AG gene sequences sampled between 1996 and 2004 from blood donors from Yaounde and Douala (Brennan et al., 2008). Given the high rate of recombination in HIV, it is common to assume these three genes are unlinked. Following this assumption, Faria et al. (2012) use BEAST to conduct a multilocus analysis employing a parametric piecewise constant-logistic demographic tree prior model. Their analysis suggests a period of exponential growth of the viral effective population size until the mid 1990s at which point the growth levels off. The estimated origins of the most recent common ancestors for the *env*, *gag* and *pol* sequences are, respectively, 1967.6 (1962.4, 1972.4, 95% CI), 1967.6 (1962.5, 1972.5, 95% CI) and 1968.1 (1962.8, 1972.8, 95% CI).

We perform a multilocus Skygrid analysis of the same data with 50 grid points and a cutoff value of 50 years. Figure 3.5 depicts the resulting estimated posterior median log effective population size along with estimated HIV prevalence counts in Cameroon from 1990-2004 (UNAIDS/WHO, 2008). Like the parametric multilocus analysis, the Skygrid analysis points to a period of exponential growth in effective population size from 10-30 years prior to the most recent sampling time. It also yields similar results regarding the origin of the HIV-1 CRF02_AG clade. The most recent common ancestors for the *env*, *gag* and *pol* sequences have estimated origins of 1965.2 (1959.6, 1970.1, 95% BCI), 1967.3 (1962.8, 1971.3, 95% BCI) and 1969.3 (1963.1, 1974.1, 95% BCI), respectively. However, in contrast to the parametric multilocus analysis, the Skygrid analysis suggests a dip in effective population size over the 5 years prior to the most recent sampling time. This finding is supported by the drop in HIV-1 prevalence in Cameroon from 2000-2004, but is not detected by the parametric multilocus analysis due to the *a priori* constraints on the shape of the effective population size trajectory imposed by the logistic-constant demographic model prior. It should be noted that the CRF02_AG population in Cameroon will have some gene flow with the worldwide population of CRF02_AG, and this is not modeled in our Skygrid analysis or the earlier parametric analysis. This may account for some of the discordance between the inferred population sizes and the Cameroon prevalence counts.

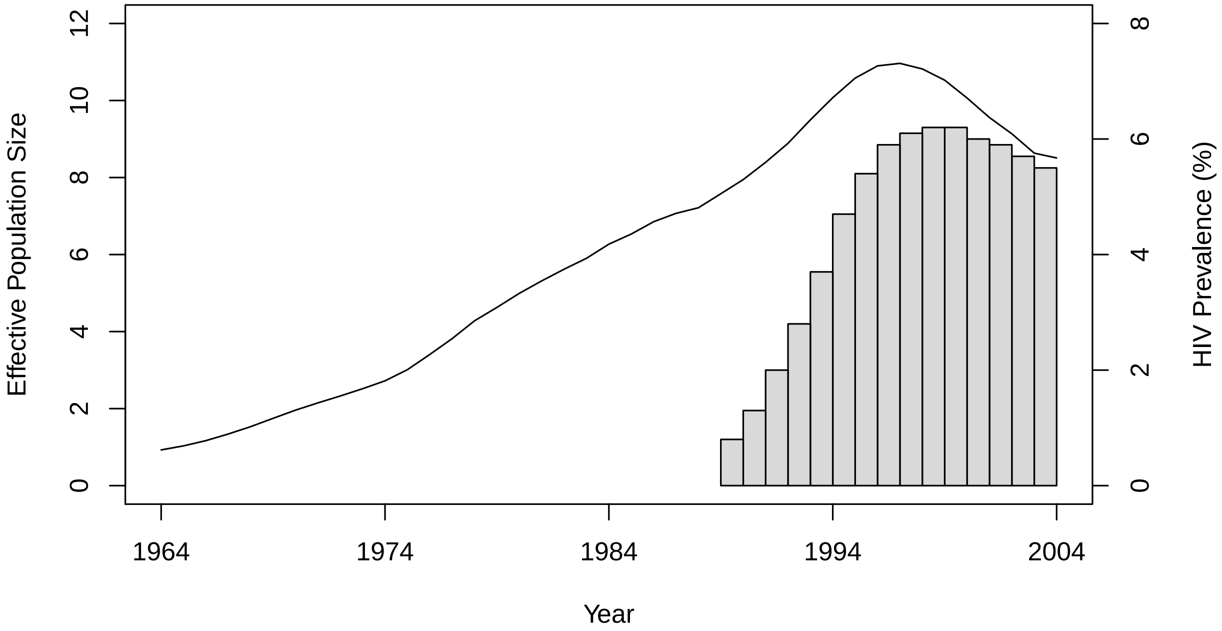

Figure 3.5: Population history of HIV-1 CRF02_AG clade in Cameroon. The curve represents the estimated median log effective population size estimated from a multilocus alignment of 336 *gag*, *pol* and *env* sequences sampled between 1996 and 2004. The bars represent estimated HIV prevalence counts in Cameroon.

3.3.6 Prior Sensitivity

The GMRF smoothing prior we place on the vector γ of log effective population sizes informs our model about the smoothness of the trajectory. The precision parameter τ governs the level of smoothness. There is usually little *a priori* knowledge regarding the smoothness of the effective population size trajectory, and in all of our examples we assign τ a relatively uninformative gamma prior. To investigate the sensitivity of our results to different hyperprior parameter values, we follow the suggestion of Minin et al. (2008) and analyze the Beringian bison data set with five different values of a: 0.001, 0.002, 0.005, 0.01, and 0.1, leaving b unchanged. These choices correspond to increasing prior means of 1, 2, 5, 10 and 100, respectively. Table 3.4 presents the estimated posterior means and 95% BCIs of τ. The results demonstrate that the posterior distribution of τ is robust to alterations of the hyperprior parameter a. Moreover, they suggest that the data contain sufficient information to estimate τ.

Prior	Posterior	
Mean	Mean	95% BCI
1	5.27	(0.58, 11.82)
2	5.37	(0.63, 11.60)
5	5.19	(0.50, 11.42)
10	5.12	(0.59, 11.68)
100	5.40	(0.76, 12.50)

Table 3.4: GMRF precision sensitivity to prior. Posterior estimates of precision parameter τ corresponding to different choices of prior mean. We use the Beringian bison data.

3.4 Discussion

The Skygrid is a powerful, flexible new model for nonparametric coalescent-based inference of past population dynamics from molecular sequence data. It incorporates a Gaussian

Markov random field smoothing scheme similar to that of the Skyride, and provides smooth and realistic estimates of demographic histories. Like the Skyride, the Skygrid model does a fairly good job of recovering essential features of simulated data based on standard parametric coalescent models.

However, the Skygrid is an improvement over the Skyride in a number of important ways. It allows for estimation based on multilocus data, yields improved TMRCA estimation, and it gives the user additional flexibility.

Molecular sequence data sets from effectively unlinked loci are becoming increasingly common thanks to lower DNA sequencing costs. Accordingly, there is a need for multilocus statistical approaches to reap the benefits. The Skygrid provides estimates of effective population size trajectories based on samples from several different genetic loci with the same demographic histories. One of the primary difficulties in coalescent-based approaches is that most of the coalescent events in the reconstructed genealogy usually occur in a short time span. During the long periods of time in which few coalescent events occur, there is not much data to infer the population dynamics. This problem is mitigated to a certain extent by increasing the sample size, but the additional coalescent events tend to occur in a small stretch of time. Increasing the number of loci more effectively provides extra information during the long stretches of time with few coalesecent events (Felsenstein, 2006). We demonstrate through a series of simulations that incorporating data from additional loci yields more precise and less biased estimates of past population dynamics. We also note that multilocus data are especially helpful in improving estimation during time periods for which single locus data are not very informative.

We compare our Skygrid model to existing multilocus approaches. As seen in the analysis of HIV-1 CRF02_AG gene sequences sampled from Camaroon, our nonparametric approach enables detection of a decline in the effective population size that is supported by HIV-1 prevalence data. This aspect of the population history went unnoticed in a multilocus analysis employing a parametric constant-logistic demographic tree model prior. The only other currently available nonparametric Bayesian model that enables estimation of past population dynamics from multilocus data is the EBSP. We analyze simulated data sets with the Skygrid and the EBSP and find that the Skygrid performs more favorably.

Bayesian nonparametric models for inference of population histories typically estimate genealogies and mutation parameters jointly along with effective population size trajectories. The different priors placed on the effective population size that distinguish these models can affect estimation of quantities other than the population history, notably the TMRCA. In simulation studies to explore TMRCA estimation, we consider data sets generated from a variety of different parametric demographic scenarios. These include typical constant and exponential growth demographic models, as well as a more complicated piecewise-exponential model motivated by a data set of ancient DNA from Beringian bison. Considered along with the Skyride and Bayesian Skyline models, the Skygrid emerges as the best overall choice for TMRCA estimation in these examples.

Unlike the Skyride, the Skygrid allows the user to specify the spacing of points where the effective population size of the estimated trajectory can change. This flexibility can be especially convenient for future extensions of the model which incorporate covariate values which must, necessarily, be measured at fixed times. We anticipate that such extensions will lead to further improvement in estimation of the effective population size over time and, for instance, enable statistical testing of environmental effects on population histories.

3.5 Methods

3.5.1 Coalescent Background

Coalescent theory was first developed by Kingman (1982b). Considering a random population sample of n individuals arising from a classic Fisher-Wright population model of constant size N_e, Kingman developed a stochastic process called the coalescent to generate genealogies relating the sample. The process begins at a sampling time $t = 0$ and proceeds backward in time as t increases, successively merging lineages until all lineages have merged. The merging of lineages is called a coalescent event and there are $n - 1$ coalescent events in all. Let t_k denote the time of the $(n - k) - th$ coalescent event for $k = 1, \ldots, n - 1$ and

$t_n = 0$ denote the sampling time. Then for $k = 2, \ldots, n$ the waiting time $w_k = t_{k-1} - t_k$ is exponentially distributed with rate $\frac{k(k-1)}{2N_e}$.

Griffiths and Tavaré (1994) provide a generalization of the coalescent that allows for the effective population size $N_e = N_e(t)$ to change over time. Here, $N_e(0)$ is the effective population size at the sampling time, and $N_e(t)$ is the effective population size t time units before the sampling time. In this case, the waiting time w_k is given by the conditional density

$$P(w_k|t_k) = \frac{k(k-1)}{2N_e(w_k + t_k)} \exp\left[-\int_{t_k}^{w_k+t_k} \frac{k(k-1)}{2N_e(t)} dt\right]. \tag{3.5}$$

Taking the product of such densities yields the joint density of intercoalescent waiting times, and this fact can be exploited to obtain the probability of observing a particular genealogy given a demographic function. Here, we consider a piecewise constant demographic function that changes values at pre-specified times.

3.5.2 Piecewise Constant Demographic Model

We start by assuming there are m known genealogies. Let $g = (g_1, g_2, \ldots, g_m)$ be a vector of genealogies representing the ancestry of populations with the same effective population size $N_e(t)$, where the time t increases into the past. We assume *a priori* that the genealogies are independent given $N_e(t)$. This assumption implies that the genealogies are unlinked which commonly occurs when researchers select loci from whole genome sequences or when recombination is very likely, such as between genes in retroviruses. Let M denote the number of points we desire for a fixed-time grid, and let K be a positive real cutoff value. Then the temporal grid points $x_1, ..., x_M$ are $x_1 = \frac{K}{M}, x_2 = 2 \times \frac{K}{M}, \ldots, x_M = K$. Here, we assume the grid points are equally spaced, but the model easily extends to arbitrarily spaced grid points.

We estimate the effective population size as a piecewise constant function that changes values only at grid points. The cutoff value is the time furthest back into the past at which the effective population size changes. Notice that for all times $t \geq K$ further into the past than the cutoff value, $N_e(t) = N_e(K)$. Let $\theta = (\theta_1, \ldots, \theta_{M+1})$ be the vector of effective population sizes. Here, $N_e(t) = \theta_k$ for $x_{k-1} \leq t < x_k$, $k = 1, \ldots, M$ where it is understood that $x_0 = 0$. Also, $N_e(t) = \theta_{M+1}$ for $t \geq x_M$.

To construct the likelihood of genealogy i, let t_{0_i} be the most recent sampling time and t_{MRCA_i} the time of the most recent common ancestor (also referred to as the root height of genealogy i). Let x_{α_i} denote the smallest grid point greater than at least one sampling time in the genealogy, and x_{β_i} the greatest grid point less than at least one coalescent time. Let $u_{ik} = [x_{k-1}, x_k]$, $k = \alpha_i + 1, \ldots, \beta_i$, $u_{i\alpha_i} = [t_{0_i}, x_{\alpha_i}]$, and $u_{i(\beta_i+1)} = [x_{\beta_i}, t_{\text{MRCA}_i}]$. For each u_{ik} we let t_{kj}, $j = 1, \ldots, r_k$, denote the ordered times of the grid points and sampling and coalescent events in the interval. With each t_{kj} we associate an indicator ϕ_{kj} which takes a value of 1 in the case of a coalescent event and 0 otherwise. Also, let v_{kj} denote the number of lineages present in the genealogy in the interval $[t_{kj}, t_{k(j+1)}]$. Following Griffiths and Tavaré (1994), the likelihood of observing an interval is

$$P(u_{ik}|\theta_k) \quad = \prod_{1 \leq j < r_k : \phi_{kj} = 1} \frac{v_{kj}(v_{kj}-1)}{2\theta_k} \prod_{j=1}^{r_k - 1} \exp\left[-\frac{v_{kj}(v_{kj}-1)(t_{k(j+1)} - t_{kj})}{2\theta_k}\right] \tag{3.6}$$

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

Let $P_*(u_{ik}|\theta_k)$ denote $P(u_{ik}|\theta_k)$ except with any factors of the form $\frac{v_{kj}(v_{kj}-1)}{2\theta_k}$ replaced by $\frac{2(2-1)}{2\theta_k} = \frac{1}{\theta_k}$; this is for the purpose of computing the probability of a genealogy, where the specific branches of a tree which coalesce matters. Then

$$P(g_i|\theta) = \prod_{k=\alpha_i}^{\beta_i+1} P_*(u_{ik}|\theta_k). \tag{3.7}$$

We introduce some notation that will facilitate the derivation of the Gaussian approximation in the next section. If c_{ik} denotes the number of coalescent events which occur during interval u_{ik}, we can write

$$P(g_i|\theta) = \prod_{k=\alpha_i}^{\beta_i+1} \left(\frac{1}{\theta_k}\right)^{c_{ik}} \exp\left[-\frac{SS_{ik}}{\theta_k}\right], \tag{3.8}$$

where the SS_{ik} are appropriate constants. Rewriting this expression in terms of $\gamma_k = \log(\theta_k)$, we arrive at

$$P(g_i|\gamma) = \prod_{k=\alpha_i}^{\beta_i+1} e^{-\gamma_k c_{ik}} \exp[-SS_{ik}e^{-\gamma_k}] = \prod_{k=\alpha_i}^{\beta_i+1} \exp[-\gamma_k c_{ik} - SS_{ik}e^{-\gamma_k}]. \tag{3.9}$$

Assuming conditional independence of genealogies, the likelihood of the vector g of genealogies is

$$\begin{aligned} P(g|\gamma) &= \prod_{i=1}^{m} P(g_i|\gamma) && (3.10) \\ &= \prod_{i=1}^{m} \prod_{k=\alpha_i}^{\beta_i+1} \exp[-\gamma_k c_{ik} - SS_{ik} e^{-\gamma_k}] && (3.11) \\ &= \exp\left[\sum_{k=1}^{M+1} \left[-\gamma_k c_k - SS_k e^{-\gamma_k}\right]\right] && (3.12) \end{aligned}$$

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

To incorporate the prior assumption that effective population size changes continuously over time we put the following Gaussian Markov random field prior on γ:

$$P(\gamma|\tau) \propto \tau^{M/2} \exp\left[-\frac{\tau}{2} \sum_{i=1}^{M} (\gamma_{i+1} - \gamma_i)^2\right]. \qquad (3.13)$$

This prior posits that differences between adjacent elements of γ are normally distributed with mean 0 and estimable precision τ, drawing motivation from a Brownian diffusion process. Let Q be a square matrix of dimension $M + 1$ with entries $Q_{ij} = -1$ for $j = i + 1$ and $j = i - 1$, $Q_{ii} = 2$ for $i = 2, \ldots, M$ and $Q_{ii} = 1$ for $i = 1, M + 1$. Then we can write

$$P(\gamma|\tau) \propto \tau^{M/2} \exp\left[-\frac{\tau}{2} \gamma' Q \gamma\right]. \qquad (3.14)$$

Finally, we assign τ a gamma prior:

$$P(\tau) \propto \tau^{a-1} e^{-b\tau}. \qquad (3.15)$$

This yields the following posterior distribution:

$$P(\gamma, \tau|g) \propto P(g|\gamma) P(\gamma|\tau) P(\tau). \qquad (3.16)$$

It should be noted that the GMRF prior does not inform the overall level of the estimated effective population size, just the smoothness of the trajectory. The degree of smoothness is determined by the precision τ. Researchers typically do not have any prior knowledge about the smoothness of the effective population size trajectory, and in such cases it is appropriate to use relatively uninformative priors. Accordingly, we choose $a = b = 0.001$ in our examples, giving τ a prior mean of 1 and variance of 1,000.

3.5.3 Markov Chain Monte Carlo Sampling Scheme

We use a block-updating Markov chain Monte Carlo sampling scheme (Knorr-Held and Rue, 2002) to approximate the posterior given in Equation (3.16). First, consider the full conditional density

$$\begin{aligned}
P(\gamma|\tau, g) &\propto P(g|\gamma)P(\gamma|\tau) \\
&\propto \exp\left[\sum_{k=1}^{M+1}(-\gamma_k c_k - SS_k e^{-\gamma_k})\right]\tau^{M/2}\exp\left[-\frac{1}{2}\gamma' Q\gamma\right] \\
&= \tau^{M/2}\exp\left[-\frac{1}{2}\gamma' Q\gamma - \sum_{k=1}^{M+1}(\gamma_k c_k + SS_k e^{-\gamma_k})\right]. \qquad (3.17)
\end{aligned}$$

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

$$\begin{aligned}
h_k(\gamma_k) &\approx h_k(\hat{\gamma}_k) + h_k'(\hat{\gamma}_k)(\gamma_k - \hat{\gamma}_k) + \frac{1}{2}h_k''(\hat{\gamma}_k)(\gamma_k - \hat{\gamma}_k)^2 \\
&= SS_k e^{-\hat{\gamma}_k}\left(\frac{1}{2}\hat{\gamma}_k{}^2 + \hat{\gamma}_k + 1\right) \\
&\quad + \left[c_k - SS_k e^{-\hat{\gamma}_k} - SS_k e^{-\hat{\gamma}_k}\hat{\gamma}_k\right]\gamma_k \\
&\quad + \left[\frac{1}{2}SS_k e^{-\hat{\gamma}_k}\right]\gamma_k^2. \qquad (3.18)
\end{aligned}$$

This yields the following second-order Gaussian approximation:

$$\begin{aligned}
P(\gamma|\tau, g) \propto \tau^{M/2}\exp\Bigg[&-\frac{1}{2}\gamma'[Q + \mathrm{Diag}(SS_k e^{-\hat{\gamma}_k})]\gamma \\
&- \sum_{k=1}^{M+1}(c_k - SS_k e^{-\hat{\gamma}_k} - SS_k e^{-\hat{\gamma}_k}\hat{\gamma}_k)\gamma_k\Bigg], \qquad (3.19)
\end{aligned}$$

where Diag$(\cdot)$ is a diagonal matrix.

Now suppose we have current parameter values $(\tau^{(n)}, \gamma^{(n)})$. First, we generate a candidate value for the precision, $\tau^* = \tau^{(n)}f$, where f is drawn from a symmetric proposal distribution with density $P(f) \propto f + \frac{1}{f}$ defined on $[1/F, F]$. The tuning constant F controls the distance between the proposed and current values of the precision. Next, conditional on τ^*, we propose a new state γ^* using the aforementioned Gaussian approximation to the full conditional density $P(\gamma^{(n)}|\tau^*, g)$. In the Gaussian approximation, we center the Taylor expansion about a point $\hat{\gamma}$ obtained iteratively by the Newton-Raphson method:

$$\gamma_{(n+1)} = \gamma_{(n)} - [d^2 f(\gamma_{(n)})]^{-1}(df(\gamma_{(n)}))' \tag{3.20}$$

with $\gamma_{(0)} = \gamma^{(n)}$. Here,

$$f(\gamma) = -\frac{1}{2}\gamma' Q \gamma - \sum_{k=1}^{M+1}(\gamma_k c_k + SS_k e^{-\gamma_k}) \tag{3.21}$$

and

$$df(\gamma) = -\gamma' Q - [c_1 - SS_1 e^{-\gamma_1}, \ldots, c_{M+1} - SS_{M+1} e^{-\gamma_{M+1}}] \tag{3.22}$$

and

$$d^2 f(\gamma) = -Q - \text{diag}[SS_k e^{-\gamma_k}]. \tag{3.23}$$

Finally, the candidate state (τ^*, γ^*) is accepted or rejected in a Metropolis-Hastings step.

3.5.4 Incorporation of Genealogical Uncertainty

In our development thus far, we have assumed the genealogies $g_1, \ldots, g_m$ are known and fixed. However, in reality we observe sequence data rather than genealogies. We can think of the aligned sequence data $Y = (Y_1, \ldots, Y_m)$ as arising from continuous-time Markov chain (CTMC) models for molecular character substitution that act along the hidden genealogies. Each CTMC depends on a vector of mutational parameters Q_i, that include, for example, an overall rate multiplier, relative exchange rates among characters and across-site variation specifications. We let $Q = (Q_1, \ldots, Q_m)$. We then jointly estimate the genealogies, mutational parameters, precision, and vector of effective population sizes through their posterior distribution

$$P(g, Q, \tau, \gamma | Y) \propto \left[\prod_{i=1}^{m} P(Y_i | g_i, Q_i)\right] P(Q) P(g|\gamma) P(\gamma|\tau) P(\tau). \tag{3.24}$$

Here, the coalescent acts as a prior for the genealogies, and we assume that Q and g are *a priori* independent of each other. Hierarchical models are however available to share information about Q among loci without strictly enforcing that they follow the same evolutionary process (Edo-Matas et al., 2011).

We achieve joint estimation by integrating the block-updating MCMC scheme for the fixed-trees case into the software package BEAST (Drummond et al., 2012). We plan to provide a user-friendly interface to this joint model in the next public release of BEAUti (Drummond et al., 2012), a graphical user interface application for generating BEAST model and data description files. In the meantime, we welcome users to exploit this multilocus model in the development branch of the BEAST source code repository (http://beast-mcmc.googlecode.com/svn/trunk). Examples of XML specification for the model are available at http://beast.bio.ed.ac.uk.

Acknowledgments

We would like to thank Vladimir Minin and Alexei Drummond for helpful discussions. This work is supported by NIH grants R01 GM086887, R01 HG006139 and 5T32AI007370-22, and NESCent through its BEAST working group. N.R.F. is supported by Fundação para a Ciência e a Tecnologia (Grant No. SFRH/BD/64530/2009). The research leading to these results has received funding from the European Union Seventh Framework Programme [FP7/2007-2013] under Grant Agreement no. 278433-PREDEMICS and ERC Grant agreement no. 260864.

CHAPTER 4

A Relaxed Drift Diffusion Model for Phylogenetic Trait Evolution

4.1 Introduction

Phylogenetic inference has emerged as an important tool for understanding patterns of molecular sequence variation over time. Along with the increasing availability of molecular sequence data, there has been a growth of associated nonsequence data, underscoring the need for integrated models of phylogenetic evolution of sequences and traits.

Much of the development of trait evolution models has been motivated by phylogenetic comparative approaches focusing on phenotypic and ecological traits. A proper understanding of patterns of correlation between traits can be achieved only by accounting for their shared evolutionary history (Felsenstein, 1985b; Harvey and Pagel, 1991), and comparative methods focus on relating observed phenotype information to an evolutionary history.

Trait evolution has been tackled from another angle in phylogeographic approaches focusing on geographic locations rather than phenotypic traits. Evolutionary change is better understood when accounting for its geographic context, and phylogeographic inference methods aim to connect the evolutionary and spatial history of a population (Bloomquist et al., 2010). Phylogeographic techniques have allowed researchers to better understand the origin, spread, and dynamics of emerging infectious diseases. Examples include the human influenza A virus (Rambaut et al., 2008; Smith et al., 2009; Lemey et al., 2009b), rabies viruses (Biek et al., 2007; Seetahal et al., 2013), dengue virus (Bennett et al., 2010; Allicock et al., 2012) and hepatitis B virus (e.g. Mello et al. (2013)).

While methods for phenotypic and phylogeographic analyses are developed with different data in mind, they address similar situations and it is appropriate to speak more generally of trait evolution. Two key components required for modeling phylogenetic trait evolution are a method for incorporating phylogenetic information and a model of an evolutionary process on a phylogeny giving rise to the observed traits. Many popular approaches first reconstruct a phylogenetic tree and condition inferences pertaining to the trait evolution process on this fixed tree. However, computational advances, particularly in Markov chain Monte Carlo (MCMC) sampling techniques, have made it possible to control for phylogenetic uncertainty (as well as uncertainty in other important model parameters) through integrated models that jointly estimate parameters of interest (Huelsenbeck and Rannala, 2003; Lemey et al., 2010).

The evolution of discrete traits has typically been modeled using continuous-time Markov chains (Felsenstein, 1981; Pagel, 1999; Lemey et al., 2009a), analogous to substitution models for molecular sequence characters. However, phenotypic and geographic traits are often continuously distributed, and while meaningful inferences may still be made partitioning the state space into finite parts, stochastic processes with continuous state spaces represent a more natural approach. A popular choice to model continous trait evolution along the lineages of a phylogenetic tree is Brownian diffusion (Felsenstein, 1985b). Lemey et al. (2010) and Pybus et al. (2012) have recently developed a computationally efficient Brownian diffusion model for evolution of multivariate traits in a Bayesian framework that integrates it with models for phylogenetic reconstruction and molecular evolution. Notably, their full probabilistic approach accounts for uncertainty in the phylogeny, demographic history and evolutionary parameters. Trait evolution is modeled as a multivariate time-scaled mixture of Brownian diffusion processes with a zero-mean displacement (or, in other words, neutral drift) along each branch of the possibly unknown phylogeny.

While adopting a mixture of Brownian diffusion processes is a popular and useful approach, it may not appropriately describe the evolutionary process in certain situations. Such scenarios are more realistically modeled by more sophisticated diffusion processes. There may, for example, be selection toward an optimal trait value. To address this phenomenon, there has been considerable development of mean-reverting Ornstein-Uhlenbeck process models for trait evolution, featuring a stochastic Brownian component along with a deterministic component (Hansen, 1997; Butler and King, 2004; Bartoszek et al., 2012).

Another trait evolutionary process inadequately modeled by standard Brownian diffusion is one characterized by directional trends. The need for relaxing the assumption of neutral drift is highlighted by a number of evolutionary scenarios in which there are apparent trends in the direction of variations, including antigenic drift in influenza (Bedford et al., 2014), the evolution of body mass in carnivores (Lartillot and Poujol, 2011), and dispersal patterns of viral outbreaks (Pybus et al., 2012). To this end, we extend the Bayesian multivariate Brownian diffusion framework of Lemey et al. (2010) to allow for an unknown estimable nonzero drift vector for the mean displacement in a computationally efficient manner. While inclusion of a nontrivial drift represents a promising first step, a constant drift rate may not hold over an entire evolutionary history. We address this issue by presenting a flexible relaxed drift model that permits multiple drift rates on a phylogenetic tree. Importantly, we equip the model with machinery to infer the number of different drift rates supported by the data as well as the locations of rate changes.

We apply our relaxed drift diffusion methodology to three viral examples of clinical importance. In the first two examples, we illustrate our approach in a phylogeographic setting by investigating the spatial diffusion of HIV-1 in central Africa and the West Nile virus in North America. For the third example, we explore antigenic evolution in the context of enhanced resistance of HIV-1 to broadly neutralizing antibodies over the course of the epidemic. We employ model selection techniques to compare the nested drift-neutral, constant drift and relaxed drift Brownian diffusion models. We demonstrate a better fit by relaxing the restrictive drift-neutral assumption, and an improved ability to uncover and quantify key aspects of trait evolution dynamics.

4.2 Methods

We start by assuming we have a dataset of N aligned molecular sequences $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)$ along with N associated M-dimensional, continuously varying traits $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_N)$. The sequence and trait data correspond to the tips of an unknown yet estimable phylogenetic tree τ. Later we will discuss accounting for phylogenetic uncertainty, modeling the molecular evolution process giving rise to $\mathbf{X}$ and integrating it with a model for trait evolution. But first, we explore trait evolution on a fixed phylogeny via a diffusion process acting conditionally independently along its branches.

The N-tipped bifurcating phylogenetic tree τ is a graph with a set of vertices $\mathcal{V} = (\mathcal{V}_1, \ldots, \mathcal{V}_{2N-1})$ and edge weights $\mathcal{T} = (t_1, \ldots, t_{2N-2})$. The vertices correspond to nodes of the tree and, as the length of the tree τ is measured in units of time, $\mathcal{T}$ consists of times corresponding to branch lengths. Each external node $\mathcal{V}_i$ for $i = 1, \ldots, N$ is of degree 1, with one parent node $\mathcal{V}_{pa(i)}$ from within the internal or root nodes. Each internal node $\mathcal{V}_i$ for $i = N+1, \ldots, 2N-2$ is of degree 3 and the root node $\mathcal{V}_{2N-1}$ is of degree 2. An edge with weight t_i connects $\mathcal{V}_i$ to $\mathcal{V}_{pa(i)}$, and we refer to this edge as branch i. In addition to the observed traits $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ at the external nodes, we posit for mathematical convenience unobserved traits $\mathbf{Y}_{N+1}, \ldots, \mathbf{Y}_{2N-1}$ at the internal nodes and root.

Brownian diffusion (also known as a Wiener process) is a continuous-time stochastic process originally developed to model the random motion of a physical particle (Brown, 1828; Wiener, 1958). Formally, a standard multivariate Brownian diffusion process $\mathbf{W}(t)$ is characterized by the following properties: $\mathbf{W}(0) = \mathbf{0}$, the map $t \mapsto \mathbf{W}(t)$ is almost surely continuous, $\mathbf{W}(t)$ has independent increments and, for $0 \leq s \leq t$, $\mathbf{W}(t) - \mathbf{W}(s)$ has a multivariate normal distribution with mean $\mathbf{0}$ and variance matrix $(t-s)\mathbf{I}$.

Recent phylogenetic comparative methods (Vrancken et al., 2014; Cybis et al., 2015) aim to model the correlated evolution between multiple traits and, to this end, employ a correlated multivariate Brownian diffusion with displacement variance $(t-s)\mathbf{P}^{-1}$. Here, $\mathbf{P}$ is an $M \times M$ infinitesimal precision matrix. The mean of $\mathbf{0}$ posits a neutral drift so that the traits do not evolve according to any systematic directional trend. Matrix $\mathbf{P}$ determines the intensity and correlation of the trait diffusion after controlling for shared evolutionary history.

The Brownian diffusion process along a phylogeny produces the observed traits by starting at the root node and proceeding down the branches of τ. The displacement $\mathbf{Y}_i - \mathbf{Y}_{pa(i)}$ along a branch is multivariate normally distributed, centered at $\mathbf{0}$ with variance $t_i\mathbf{P}^{-1}$ proportional to the length of the branch. Therefore, conditioning on the trait value $\mathbf{Y}_{pa(i)}$ at the parent node, we have

$$\mathbf{Y}_i|\mathbf{Y}_{pa(i)} \sim N\left(\mathbf{Y}_{pa(i)}, t_i\mathbf{P}^{-1}\right). \tag{4.1}$$

An extension that introduces branch-specific mixing parameters ϕ_i into the process that rescale $t_i \mapsto \phi_i t_i$ yields a mixture of Brownian processes and remains popular in phylogeography (Lemey et al., 2010).

4.2.1 Drift

To incorporate a directional trend, we adopt a multivariate correlated Brownian diffusion process with a non-neutral drift. In our drift diffusion process we replace the zero mean of the increment $\mathbf{W}(t) - \mathbf{W}(s)$ with the time-scaled mean vector $(t-s)\boldsymbol{\mu}$. The expected difference between the trait values of a descendant and its ancestor is determined by the overall drift vector $\boldsymbol{\mu}$ and the time elapsed between descendant and ancestor. This yields what we will call the constant drift model:

$$\mathbf{Y}_i|\mathbf{Y}_{pa(i)} \sim N\left(\mathbf{Y}_{pa(i)} + t_i\boldsymbol{\mu}, t_i\mathbf{P}^{-1}\right). \tag{4.2}$$

While this approach is useful for modeling general directional trends, it is quite restrictive in that the drift $\boldsymbol{\mu}$ is fixed over the entire phylogeny. We can relax this assumption by introducing branch-specific drift vectors $\boldsymbol{\mu}_i$:

$$\mathbf{Y}_i|\mathbf{Y}_{pa(i)} \sim N\left(\mathbf{Y}_{pa(i)} + t_i\boldsymbol{\mu}_i, t_i\mathbf{P}^{-1}\right) \tag{4.3}$$

for $i = 1, \ldots, 2N-2$. We assign the root a conjugate prior

$$\mathbf{Y}_{2N-1} \sim N\left(\boldsymbol{\mu}^*, (\phi\mathbf{P})^{-1}\right), \tag{4.4}$$

that is relatively uninformative for small values of ϕ.

Conditioning on the trait value $\mathbf{Y}_{2N-1}$ at the root of τ, the joint distribution of observed traits $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ can be expressed as

$$\text{vec}\left[\mathbf{Y}\right] |(\mathbf{Y}_{2N-1}, \mathbf{P}, \mathbf{V}_\tau, \boldsymbol{\mu}_\tau) \sim N\left(\mathbf{Y}_{\text{root}} + (\mathbf{T} \otimes \mathbf{I}_M)\,\boldsymbol{\mu}_\tau, \mathbf{P}^{-1} \otimes \mathbf{V}_\tau\right), \tag{4.5}$$

building on a similar construction for drift-neutral Brownian diffusion (Felsenstein, 1973; Freckleton et al., 2002). Here, vec$[\mathbf{Y}]$ is the vectorization of the column vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$, while $\mathbf{I}_M$ is an $M \times M$ identity matrix, and $\otimes$ is the Kronecker product. $\mathbf{Y}_{\text{root}}$ is the $NM \times 1$ vector $(\mathbf{Y}^t_{2N-1}, \ldots, \mathbf{Y}^t_{2N-1})^t$, and $\boldsymbol{\mu}_\tau$ is the $(2N-2)M \times 1$ drift rate vector $(\boldsymbol{\mu}^t_1, \ldots, \boldsymbol{\mu}^t_{2N-2})^t$. The $N \times N$ variance matrix $\mathbf{V}_\tau$ is a deterministic function of τ and represents the contribution of the phylogenetic tree to the covariance structure. Its diagonal entries V_{ii} are equal to the distance in time between the tip $\mathcal{V}_i$ and the root node $\mathcal{V}_{2N-1}$, and off-diagonal entries V_{ij} correspond to the distance in time between the root node $\mathcal{V}_{2N-1}$ and the most recent common ancestor of tips $\mathcal{V}_i$ and $\mathcal{V}_j$. Finally, the $N \times (2N-2)$ matrix $\mathbf{T}$ is defined as follows: $T_{ij} = t_j$, the length of branch j, if branch j is part of the path from the external node i to the root, and $T_{ij} = 0$ otherwise.

Our development thus far clarifies some important issues. First, while it is tempting to model a unique drift rate on each branch, not all $\boldsymbol{\mu}_i$ are uniquely identifiable in the likelihood (4.5). Care must be taken to impose necessary restrictions to ensure identifiability while still permitting sufficient drift rate variation, and we discuss an approach to achieve this in section 4.2.3. Second, the variance matrix $\mathbf{P}^{-1} \otimes \mathbf{V}_\tau$ in (4.5) suggests a computational order of $\mathcal{O}(N^3M^3)$ to evaluate the density. Repeated evaluation of (4.5) is necessary for numerical integration in Bayesian modeling, and viral data sets may encompass thousands of sequences. Fortunately, Pybus et al. (2012) demonstrate that phylogenetic Brownian diffusion likelihoods can be evaluated in computational order $\mathcal{O}(NM^2)$ by modeling in terms of the precision matrix $\mathbf{P}$ (as opposed to the variance) and adopting a dynamic programming approach. In section 4.2.2, we present an adaptation of their algorithm for our drift diffusion likelihood.

4.2.2 Multivariate Trait Peeling

Under our Brownian drift diffusion process, the joint distribution of all traits is straightforwardly expressed as the product

$$P(\mathbf{Y}_1, \ldots, \mathbf{Y}_{2N-1}|\tau, \mathbf{P}, \boldsymbol{\mu}, \phi) = \left(\prod_{i=1}^{2N-2} P(\mathbf{Y}_i|\mathbf{Y}_{pa(i)}, \mathbf{P}, t_i, \boldsymbol{\mu}_i) \right) P(\mathbf{Y}_{2N-1}|\mathbf{P}, \boldsymbol{\mu}^*, \phi), \qquad (4.6)$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{2N-2}, \boldsymbol{\mu}^*)$. The density of the observed traits can then be obtained by integrating over all possible realizations of the unobserved traits at the root and internal nodes. We adopt a dynamic programming approach that is analogous to Felsenstein's pruning method (Felsenstein, 1981) and has been employed for drift-neutral Brownian diffusion likelihoods (Pybus et al., 2012; Vrancken et al., 2014; Cybis et al., 2015) .

We wish to compute the density

$$\begin{aligned} P(\mathbf{Y}_1, \ldots, \mathbf{Y}_N) &= \int \cdots \int P(\mathbf{Y}_1, \ldots, \mathbf{Y}_{2N-1}) d\mathbf{Y}_{N+1} \ldots d\mathbf{Y}_{2N-1} & (4.7) \\ &= \int \cdots \int \left(\prod_{i=1}^{2N-2} P(\mathbf{Y}_i|\mathbf{Y}_{pa(i)}) \right) P(\mathbf{Y}_{2N-1}) d\mathbf{Y}_{N+1} \ldots d\mathbf{Y}_{2N-1}. & (4.8) \end{aligned}$$

We have omitted dependence on the parameters $\tau, \mathbf{P}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{2N-2}, \boldsymbol{\mu}*$ and ϕ from the notation for the sake of clarity. The integration proceeds in a postorder traversal integrating out one internal node trait at a time. Let $\{\mathbf{Y}_i\}$ denote the set of observed trait values descendant from and including the node V_i, and suppose $pa(i) = pa(j) = k$. Our traversal requires computing integrals of the form

$$P(\{\mathbf{Y}_k\}|\mathbf{Y}_{pa(k)}) = \int P(\{\mathbf{Y}_i\}|\mathbf{Y}_k) P(\{\mathbf{Y}_j\}|\mathbf{Y}_k) P(\mathbf{Y}_k|\mathbf{Y}_{pa(k)}) d\mathbf{Y}_k. \qquad (4.9)$$

Because the integrand is proportional to a multivariate normal density, it suffices to keep track of partial mean vectors $\mathbf{m}_k$, partial precision scalars p_k and normalizing constants ρ_k.

Let MVN(.;$\boldsymbol{\kappa}$,$\boldsymbol{\Lambda}$) denote a multivariate normal probability density function with mean $\boldsymbol{\kappa}$ and precision $\boldsymbol{\Lambda}$. We can rewrite conditional densities to facilitate integration with respect to the trait at the parent node:

$$\begin{aligned} P(\mathbf{Y}_i|\mathbf{Y}_k) &= \text{MVN}\left(\mathbf{Y}_i; t_i \boldsymbol{\mu}_i + \mathbf{Y}_k, \frac{1}{t_i}\mathbf{P} \right) & (4.10) \\ &= \text{MVN}\left(\mathbf{Y}_i - t_i \boldsymbol{\mu}_i; \mathbf{Y}_k, \frac{1}{t_i}\mathbf{P} \right). & (4.11) \end{aligned}$$

For $i = 1, \ldots, N$, set normalizing constant $\rho_i = 1$, partial mean

$$\mathbf{m}_i = \mathbf{Y}_i - t_i \boldsymbol{\mu}_i \tag{4.12}$$

and partial precision

$$p_i = \frac{1}{t_i}. \tag{4.13}$$

Then

$$\begin{aligned} P(\{\mathbf{Y}_i\}|\mathbf{Y}_k)P(\{\mathbf{Y}_j\}|\mathbf{Y}_k) &= \rho_i\rho_j \times \mathrm{MVN}(\mathbf{m}_i; \mathbf{Y}_k, p_i\mathbf{P}) \times \\ &\quad \mathrm{MVN}(\mathbf{m}_j; \mathbf{Y}_k, p_j\mathbf{P}) \qquad (4.14) \\ &= \rho_k \times \mathrm{MVN}(\mathbf{Y}_k; \mathbf{m}_k^*, (p_i + p_j)\mathbf{P}) \qquad (4.15) \end{aligned}$$

where partial unshifted mean

$$\mathbf{m}_k^* = \frac{p_i\mathbf{m}_i + p_j\mathbf{m}_j}{p_i + p_j}, \tag{4.16}$$

and normalizing constant

$$\rho_k = \rho_i\rho_j \left(\frac{p_ip_j}{2\pi(p_i + p_j)}\right)^{d/2} |\mathbf{P}|^{1/2} \frac{\exp\left[-\frac{p_i}{2}\mathbf{m}_i'\mathbf{P}\mathbf{m}_i - \frac{p_j}{2}\mathbf{m}_j'\mathbf{P}\mathbf{m}_j\right]}{\exp\left[-\frac{p_i+p_j}{2}\mathbf{m}_k^{*\prime}\mathbf{P}\mathbf{m}_k^*\right]}. \tag{4.17}$$

Multiplying by $P(\mathbf{Y}_k|\mathbf{Y}_{pa(k)})$ and integrating with respect to $\mathbf{Y}_k$, we get

$$\begin{aligned} P(\{\mathbf{Y}_k\}|\mathbf{Y}_{pa(k)}) &= \int P(\{\mathbf{Y}_i\}|\mathbf{Y}_k)P(\{\mathbf{Y}_j\}|\mathbf{Y}_k)P(\mathbf{Y}_k|\mathbf{Y}_{pa(k)})d\mathbf{Y}_k \qquad (4.18) \\ &= \rho_k \times \mathrm{MVN}(\mathbf{Y}_{pa(k)}; \mathbf{m}_k, p_k\mathbf{P}), \qquad (4.19) \end{aligned}$$

where

$$\mathbf{m}_k = \mathbf{m}_k^* - t_k\boldsymbol{\mu}_k, \tag{4.20}$$

and

$$p_k = \frac{1}{t_k + \frac{1}{p_i+p_j}}. \tag{4.21}$$

Integrating out all internal node traits yields

$$P(\mathbf{Y}_1, \ldots, \mathbf{Y}_N|\mathbf{Y}_{2N-1}) = \rho_{2N-1} \times \mathrm{MVN}(\mathbf{Y}_{2N-1}; \mathbf{m}_{2N-1}^*, (p_{2N-2} + p_{2N-3})\mathbf{P}). \tag{4.22}$$

For the final step, we multiply by the conjugate root prior and integrate:

$$\begin{aligned} P(\mathbf{Y}_1, \ldots, \mathbf{Y}_N) &= \int P(\mathbf{Y}_1, \ldots, \mathbf{Y}_N|\mathbf{Y}_{2N-1})P(\mathbf{Y}_{2N-1})d\mathbf{Y}_{2N-1} \qquad (4.23) \\ &= \rho_{2N-1}\mathrm{MVN}(\mathbf{m}_{2N-1}^*; \boldsymbol{\mu}^*, p_{2N-1}\mathbf{P}), \qquad (4.24) \end{aligned}$$

where

$$p_{2N-1} = \frac{(p_{2N-2} + p_{2N-3})\phi}{p_{2N-2} + p_{2N-3} + \phi}. \tag{4.25}$$

In practice, the algorithm visits each phylogeny node once and computes partial unshifted means $\mathbf{m}_k^*$, partial means $\mathbf{m}_k$, partial precisions p_k, and normalizing constants ρ_k.

4.2.3 Identifiability and Relaxed Drift

Ideally, we would like to model a unique drift rate $\boldsymbol{\mu}_i$ on each branch i of the phylogenetic tree. However, such lax assumptions open the door to misleading inferences. Adopting the notation of section 4.2.1, there can exist distinct drift rate vectors $\boldsymbol{\mu}_\tau \neq \boldsymbol{\mu}_\tau^*$ such that

$$\left(\mathbf{T} \otimes \mathbf{I}_M\right) \boldsymbol{\mu}_\tau = \left(\mathbf{T} \otimes \mathbf{I}_M\right) \boldsymbol{\mu}_\tau^*, \tag{4.26}$$

yielding identical trait likelihoods (4.5). The lack of model identifiability presents an obstacle to uncovering the “true” values of the drift rates that characterize the trait evolution process.

We propose a relaxed drift model that allows for drift rate variation along a phylogenetic tree while maintaining model identifiability. This is achieved by having branches inherit drift rates from ancestral branches by default, but allowing a random number of certain types of rate changes to occur along the tree. We describe the model here and refer readers to the Appendix for a detailed argument establishing identifiability.

We begin at the unobserved branch leading to the root, or most recent common ancestor (MRCA), of the phylogenetic tree τ and associate with it the drift $\boldsymbol{\mu}_{\text{MRCA}}$. Then the two branches emanating from the root node either both inherit the drift rate $\boldsymbol{\mu}_{\text{MRCA}}$, or a rate change occurs and one branch receives a new rate while the other branch assumes the rate $\boldsymbol{\mu}_{\text{MRCA}}$. Similarly, whenever a branch splits into two anywhere in τ, either both child branches assume the same drift rate as the parent branch, or one child branch takes on a new value while the other inherits its drift from the parent branch. Both child branches taking on different drift rates than the parent branch is not permitted.

Importantly, rather than fix the type of drift rate transfer that occurs at a given node, we estimate it from the data. The benefits of this choice are twofold. First, a rate change is not forced when the data do not suggest a need for one. Unnecessarily imposing a large number of unique drift rates to be inferred from limited data can lead to high variance estimates. Second, in the event of a rate change occurring at a node, only one of the two child branches can assume a new drift rate. We let the data determine which of the child branches assumes the new rate. The data may support new rates on both child branches. While our model may seem too restrictive to accommodate such a scenario at first glance, we are able to infer the relative support for each child, and it is reflected in the posterior distribution in terms of the probabilities of the two types of changes. Thus summaries of the posterior distribution can capture the true nature of drift rate variation in spite of the identifiability restrictions.

It is important to handle the initial drift rate $\boldsymbol{\mu}_{\text{MRCA}}$ with care. One option is to estimate $\boldsymbol{\mu}_{\text{MRCA}}$ from the data just as with all other drift rates. However, such a choice may not be ideal for data sets that exhibit relatively long periods of divergence from the MRCA to the sampling times. There is generally more information about diffusion dynamics during time periods overlapping with or close to sampling times. Likewise, the further removed the MRCA is from sampling times, the less information there is about $\boldsymbol{\mu}_{\text{MRCA}}$ and other drift rates near the MRCA. Because of the interconnectedness of branch drift rates in the relaxed model, estimates of $\boldsymbol{\mu}_{\text{MRCA}}$ and neighboring drift rates under such circumstances may primarily reflect information about drift rates on branches near sampling times. To mitigate misleading inferences of drift near the MRCA, we can adopt an initial drift $\boldsymbol{\mu}_{\text{MRCA}} = 0$ and still interprete changes in drift rates across the tree.

To parameterize the model, we associate a ternary variable $\boldsymbol{\delta}_k$ with each internal node k specifying how it passes on its drift rate to its child nodes. Suppose node k has left child node i and right child node j. If $\boldsymbol{\delta}_k = -1$, then $\boldsymbol{\mu}_i = \boldsymbol{\mu}_k$ and node j assumes a new rate $\boldsymbol{\mu}_j = \boldsymbol{\mu}_k + \boldsymbol{\alpha}_j$. If $\boldsymbol{\delta}_k = 1$, then node i assumes a new rate $\boldsymbol{\mu}_i = \boldsymbol{\mu}_k + \boldsymbol{\alpha}_i$ while $\boldsymbol{\mu}_j = \boldsymbol{\mu}_k$. If $\boldsymbol{\delta}_k = 0$, then no drift rate changes occur and $\boldsymbol{\mu}_k = \boldsymbol{\mu}_i = \boldsymbol{\mu}_j$. To map the ternary $\boldsymbol{\delta}_k$ variables to binary indicators $\boldsymbol{\gamma}_i$ of rate changes for child branches, we define

$$\boldsymbol{\gamma}_i = \frac{1 + \boldsymbol{\delta}_k}{2} |\boldsymbol{\delta}_k|, \tag{4.27}$$

and

$$\boldsymbol{\gamma}_j = \frac{1-\boldsymbol{\delta}_k}{2}|\boldsymbol{\delta}_k|. \tag{4.28}$$

Thus

$$\boldsymbol{\mu}_i = \boldsymbol{\mu}_{pa(i)} + \boldsymbol{\gamma}_i \boldsymbol{\alpha}_i. \tag{4.29}$$

for $i = 1, \dots, 2N-2$. Working with the binary $\boldsymbol{\gamma}_i$ eases our understanding of the MCMC procedure to infer drift rate changes, discussed below. However, we parameterize the model in terms of the ternary $\boldsymbol{\delta}_k$ to facilitate enforcement of the model restrictions.

Of particular interest is the random number $K \in 0, \dots, N-1$ of rate changes that occur in τ. We can write K in terms of the $\boldsymbol{\delta}_i$,

$$K = \sum_{i=N+1}^{2N-1} |\boldsymbol{\delta}_i|, \tag{4.30}$$

and it provides us with a natural way to think of the vector $\boldsymbol{\delta} = (\boldsymbol{\delta}_{N+1}, \dots, \boldsymbol{\delta}_{2N-1})$. For example, we can express our prior beliefs about $\boldsymbol{\delta}$ in terms of K. A popular prior for count data is the Poisson distribution

$$K \sim \text{Poisson}(\lambda). \tag{4.31}$$

Here, λ is the expected number of rate changes in τ. In our analyses, we set $\lambda = \log(2)$, which places 50% prior probability on the hypothesis of no rate changes.

In order to infer the nature of the drift rate transitions that occur at the nodes of the phylogenetic tree, we borrow ideas from Bayesian stochastic search variable selection (BSSVS) (George and McCulloch, 1993; Kuo and Mallick, 1998; Chipman et al., 2001). BSSVS is typically applied to model selection problems in a linear regression setting. In this framework, we begin with a large number P of potential predictors $\mathbf{X}_1, \dots, \mathbf{X}_P$ and seek to determine which of them associate linearly with an N-dimensional outcome $\mathbf{Y}$. The full model with all predictors is

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \cdots + \mathbf{X}_P\boldsymbol{\beta}_P + \boldsymbol{\epsilon}, \tag{4.32}$$

where the $\boldsymbol{\beta}_i$ are regression coefficients and $\boldsymbol{\epsilon}$ is a vector of normally distributed error terms with mean $\mathbf{0}$. When a particular $\boldsymbol{\beta}_i$ is determined to differ significantly from 0, the corresponding $\mathbf{X}_i$ helps predict $\mathbf{Y}$. If not, $\mathbf{X}_i$ contributes little additional information and is fit to be removed from the model by forcing $\boldsymbol{\beta}_i = 0$. Predictors may be highly correlated, and deterministic model selection strategies tend not to find the optimal set of predictors without exploring all possible subsets. There exist 2^P such subsets, so exploring all of them is computationally unfeasible in general and fails completely for $P > N$.

BSSVS efficiently explores the possible subsets of model predictors by augmenting the model state space with a vector $\boldsymbol{\delta} = (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_P)$ of binary indicator variables that dictate which predictors to include. The indicators $\boldsymbol{\delta}_i$ impose a prior on the regression coefficients $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_P)$ with mean $\mathbf{0}$ and variance proportional to a $P \times P$ diagonal matrix with its diagonal equal to $\boldsymbol{\delta}$. If $\boldsymbol{\delta}_i = 0$, then the prior variance on $\boldsymbol{\beta}_i$ shrinks to 0 and forces $\boldsymbol{\beta}_i = 0$ in the posterior. The joint space $(\boldsymbol{\beta}, \boldsymbol{\delta})$ is explored simultaneously through MCMC.

We apply BSSVS in our relaxed drift setting to determine the types of drift rate transfers that occur. We achieve this by exploring the joint space $(\boldsymbol{\alpha}, \boldsymbol{\delta})$ of rate differences between parent and child branches, and ternary rate change indicators. The $\boldsymbol{\delta}_k$ map to binary indicators $\boldsymbol{\gamma}_i$, as shown in (4.27) and (4.28). We assume that drift rate differences $\boldsymbol{\alpha}_i = \boldsymbol{\mu}_i - \boldsymbol{\mu}_{pa(i)}$ are *a priori* independent and normally distributed,

$$\boldsymbol{\alpha}_i \sim N(\mathbf{0}, \boldsymbol{\gamma}_i \sigma^2 \mathbf{I}). \tag{4.33}$$

If $\boldsymbol{\gamma}_i = 0$, then the prior variance σ^2 on the components of $\boldsymbol{\alpha}_i$ shrinks to 0. This forces $\boldsymbol{\alpha}_i = \mathbf{0}$, and hence $\boldsymbol{\mu}_i = \boldsymbol{\mu}_{pa(i)}$, in the posterior.

We complete our drift diffusion model specification by assigning the precision matrix $\mathbf{P}$ a Wishart prior with, say, degrees of freedom v and scale matrix $\mathbf{V}$. Importantly, the Wishart distribution is conjugate to the observed trait likelihood. Indeed, invoking the notation for partial means and precisions from section 4.2.2, the posterior

$$P(\mathbf{P}|\mathbf{Y}_1, \ldots, \mathbf{Y}_N) \propto P(\mathbf{Y}_1, \ldots, \mathbf{Y}_N|\mathbf{P})P(\mathbf{P}) \tag{4.34}$$

has a Wishart distribution with $N + v$ degrees of freedom and scale matrix

$$\left(\mathbf{V}^{-1} + p_{2N-1}(\mathbf{m}^*_{2N-1} - \boldsymbol{\mu}^*)(\mathbf{m}^*_{2N-1} - \boldsymbol{\mu}^*)' + \sum_{k=N+1}^{2N-1} [p_i \mathbf{m}_i \mathbf{m}_i' + p_j \mathbf{m}_j \mathbf{m}_j' - (p_i + p_j)\mathbf{m}^*_k \mathbf{m}^{*'}_k]\right)^{-1}. \tag{4.35}$$

Lemey et al. (2010) exploit a similar conjugacy to construct an efficient Gibbs sampler for $\mathbf{P}$, and our adoption of the Wishart prior conveniently allows us to extend use of the sampler to our model that now includes a relaxed drift process.

4.2.4 Joint Modeling and Inference

A major strength of our Bayesian framework is that it jointly models sequence and trait evolution. Adopting a standard phylogenetic approach, we assume the sequence data $\mathbf{X}$ arise from a continuous-time Markov chain (CTMC) model for character evolution acting along the unobserved phylogenetic tree τ. The CTMC is characterized by a vector $\mathbf{Q}$ of mutation parameters that may include, for instance, relative exchange rates among characters, an overall rate multiplier and across-site variation specifications. The traits $\mathbf{Y}$ arise from a Brownian drift diffusion process acting on τ, governed by parameters $\mathbf{\Lambda}$. A crucial assumption is that the processes giving rise to the observed sequences and traits are conditionally independent given the phylogenetic tree τ:

$$P(\mathbf{X}, \mathbf{Y}|\tau, \mathbf{Q}, \mathbf{\Lambda}) = P(\mathbf{X}|\tau, \mathbf{Q})P(\mathbf{Y}|\tau, \mathbf{\Lambda}), \tag{4.36}$$

enabling us to write the joint model posterior distribution as

$$P(\tau, \mathbf{Q}, \mathbf{\Lambda}|\mathbf{X}, \mathbf{Y}) \propto P(\mathbf{X}|\tau, \mathbf{Q})P(\mathbf{Y}|\tau, \mathbf{\Lambda})P(\tau)P(\mathbf{Q})P(\mathbf{\Lambda}). \tag{4.37}$$

We implement the joint model by integrating our drift diffusion framework for trait evolution into the Bayesian Evolutionary Analysis Sampling Trees (BEAST) software package (Drummond et al., 2012). BEAST provides an array of efficient methods for Bayesian phylogenetic inference, particularly to estimate phylogenies and model molecular sequence evolution. For the phylogeny τ, we choose from flexible coalescent-based priors that do not make strong *a priori* assumptions about the population history (Minin et al., 2008; Gill et al., 2013). For sequence evolution, we have access to a range of classic substitution models (Kimura, 1980; Felsenstein, 1981; Hasegawa et al., 1985), gamma distributed rate heterogeneity among sites (Yang, 1994), and strict and relaxed molecular clock models for branch rates (Drummond et al., 2006).

Estimation of the full joint posterior (4.37) is achieved through MCMC sampling (Metropolis et al., 1953; Hastings, 1970). We employ standard Metropolis-Hastings transition kernels available in BEAST to integrate over the parameter spaces of $\mathbf{Q}$ and τ. To sample realizations of the drift diffusion precision matrix $\mathbf{P}$, we adapt a Gibbs sampler developed for drift-neutral Brownian diffusion (Lemey et al., 2010). For the relaxed drift model, we need transition kernels to explore the space $(\boldsymbol{\alpha}, \boldsymbol{\delta})$ of branch rate differences and ternary rate change indicators. We propose new rate differences $\boldsymbol{\alpha}_i^*$ component-wise through a random walk transition kernel that adds random values within a specified window size to the current $\boldsymbol{\alpha}_i$.

For $\boldsymbol{\delta}$, we implement a trit-flip transition kernel that chooses one of the $N-1$ ternary indicators $\boldsymbol{\delta}_k$ uniformly at random and proposes a new state $\boldsymbol{\delta}_k^*$ assuming one of the two possible values not equal to $\boldsymbol{\delta}_k$ with equal probability. For example, if $\boldsymbol{\delta}_k = 0$, then

$$\boldsymbol{\delta}_k^* = \begin{cases} -1 & \text{with probability } \frac{1}{2} \\ 1 & \text{with probability } \frac{1}{2}. \end{cases} \tag{4.38}$$

We have parameterized our prior on $\boldsymbol{\delta}$ in terms of the number K of rate changes, and this parameterization should be retained for the transition kernel in order to ensure the correct Metropolis-Hastings proposal ratio (Drummond and Suchard, 2010). A proposed increase in rate changes occurs when we choose a $\boldsymbol{\delta}_k$ with value 0, so

$$q(K^* = K+1|K) = \frac{N-1-K}{N-1}. \tag{4.39}$$

If we choose a $\boldsymbol{\delta}_k$ with a nonzero value, we propose the other nonzero value with probability 0.5 (corresponding to $K^* = K$), and we propose 0 with probability 0.5 (which means a decrease in rate changes, $K^* = K-1$). Therefore

$$q(K^* = K|K) = q(K^* = K-1|K) = \frac{1}{2}\frac{K}{N-1}. \tag{4.40}$$

These calculations yield the following proposal ratio for K:

$$\frac{q(K|K^*)}{q(K^*|K)} = \begin{cases} \frac{1}{2}\frac{K+1}{N-1-K} & \text{if } K^* = K+1 \\ 1 & \text{if } K^* = K \\ \frac{2(N-K)}{K} & \text{if } K^* = K-1. \end{cases} \tag{4.41}$$

In addition to the parameters characterizing the trait and sequence evolution processes, we may wish to make inferences about the posterior distribution of traits at the root and internal nodes, or at any arbitrary time point in the past. We equip BEAST with the ability to generate posterior trait realizations at these nodes by implementing a preorder, tree-traversal algorithm.

A natural choice to summarize the results of a Bayesian phylogenetic analysis is a maximum clade credibility (MCC) tree. To form an MCC tree, the posterior sample of trees is examined to determine posterior clade probabilities, and the tree with the maximum product of posterior clade probabilities is the MCC tree. The branches and nodes of MCC trees can be annotated with inferred drift rates and trait values, along with other quantities of interest. Annotated MCC trees can be summarised using TreeAnnotator, available as part of the BEAST distribution, and visualized using FigTree.

4.2.5 Model Selection

We can formally compare the constant and relaxed drift diffusion models through Bayes factors (BFs). A Bayes factor (Jeffreys, 1935, 1961) compares the fit of two models, M_1 and M_0, to observed data $(\mathbf{X}, \mathbf{Y})$ by taking the ratio of marginal likelihoods:

$$BF_{10} = \frac{P(\mathbf{X}, \mathbf{Y} | M_1)}{P(\mathbf{X}, \mathbf{Y} | M_0)} = \frac{P(M_1 | \mathbf{X}, \mathbf{Y})}{P(M_0 | \mathbf{X}, \mathbf{Y})} \bigg/ \frac{P(M_1)}{P(M_0)} \, . \tag{4.42}$$

BF_{10} quantifies the evidence in favor of model M_1 over M_0. Kass and Raftery (1995) provide guidelines for assessing the strength of the evidence against M_0: BFs between 1 and 3 are not worth more than a bare mention, while values between 3 and 20 are considered positive evidence against M_0. BFs in the ranges 20-150 and >150 are considered to be strong and very strong evidence against M_0, respectively.

Evaluation of Bayes factors has become a popular approach to model selection in Bayesian phylogenetics (Sinsheimer et al., 1996; Suchard et al., 2001, 2005). Marginal likelihood estimation can be quite difficult in a phylogenetic context, and stepping-stone sampling estimators have been implemented to address this (Baele et al., 2012a,b). Following the approach of Drummond and Suchard (2010), however, we are able to straightforwardly compute the

Bayes factor BF_C supporting the constant drift model M_C over the relaxed drift model M_R. The model M_C is nested within the more general model M_R and occurs when $K = 0$. This enables us to write

$$BF_C = \frac{P(\mathbf{X}, \mathbf{Y}|M_C)}{P(\mathbf{X}, \mathbf{Y}|M_R)} = \frac{P(M_C|\mathbf{X}, \mathbf{Y})}{P(M_R|\mathbf{X}, \mathbf{Y})} \Bigg/ \frac{P(M_C)}{P(M_R)} \tag{4.43}$$

$$= \frac{P(K = 0|\mathbf{X}, \mathbf{Y}, M_R)}{1 - P(K = 0|\mathbf{X}, \mathbf{Y}, M_R)} \Bigg/ \frac{P(K = 0|M_R)}{1 - P(K = 0|M_R)}\ , \tag{4.44}$$

requiring only our prior probability of no rate changes under the relaxed drift model, and the posterior probability of zero rate changes.

4.3 The Spread of HIV-1 in Central Africa

Faria et al. (2014) explore the early spatial expansion and epidemic dynamics of HIV-1 in central Africa by analyzing sequence data sampled from countries in the Congo River basin. The authors employ a discrete phylogeographic inference framework (Lemey et al., 2009a) and show that the pandemic likely originated in Kinshasa (in what is now the Democratic Republic of Congo) in the 1920s. Furthermore, viral spread to other population centers in sub-Saharan Africa was aided by a combination of factors, including strong railway networks, urban growth, and changes in sexual behavior.

We follow up the analysis of Faria et al. (2014) by applying our continuous drift diffusion approach to one of the data sets analysed in this study. The data set consists of HIV-1 sequences sampled between 1985-2004 from the Democratic Republic of Congo and the Republic of Congo and includes 96 sequences from Kinshasa (Kalish et al., 2004; Vidal et al., 2000, 2005; Yang et al., 2005), 96 sequences from Mbuji-Mayi (Vidal et al., 2000, 2005), 96 from Brazzaville (Bikandou et al., 2004; Niama et al., 2006), 76 from Lubumbashi (Vidal et al., 2005), 33 from Bwamanda (Vidal et al., 2000), 24 from Likasi (Kita et al., 2004), 23 from Kisangani (Vidal et al., 2005), and 22 sequences from Pointe-Noire (Bikandou et al., 2000). We reconstruct the spatial dynamics under drift-neutral, constant drift, and relaxed drift Brownian diffusion on a maximum clade credibility tree estimated from the sequences and their locations of sampling. The traits in this instance are bivariate longitude and

	No Drift		Constant Drift		Relaxed Drift	
Drift (Lat.)	-	-	-0.09	(-0.11, -0.06)	-0.03	(-0.05, -0.01)
Rate Changes (Lat.)	-	-	-	-	28.13	(27.0, 29.0)
Variance (Lat.)	0.25	(0.23, 0.27)	0.23	(0.21, 0.25)	0.13	(0.12, 0.14)
Drift (Long.)	-	-	0.30	(0.26, 0.33)	0.12	(0.08, 0.16)
Rate Changes (Long.)	-	-	-	-	1.48	(1.00, 3.00)
Variance (Long.)	0.59	(0.55, 0.64)	0.37	(0.34, 0.40)	0.43	(0.39, 0.45)
Correlation	-0.47	(-0.54, -0.40)	-0.40	(-0.47, -0.32)	-0.84	(-0.87, -0.82)

Table 4.1: Spatiotemporal dynamics of HIV-1 in central Africa. Model comparison of Brownian diffusion with no drift, constant drift Brownian diffusion, and relaxed drift Brownian diffusion. We report posterior mean estimates along with 95% Bayesian credibility intervals (BCIs). Drift rates for latitude and longitude coordinates are reported in units of degrees per year.

latitude coordinates, with observed traits corresponding to sampling locations. Table 4.1 reports posterior estimates of drift.

Under the constant drift model, we infer a significant longitudinal drift with posterior mean 0.30 degrees per year and a 95% Bayesian credibility interval (BCI) of (0.26, 0.33), as well as a significant latitudinal drift with posterior mean of -0.09 degrees per year and BCI (-0.11, -0.06). Furthermore, for each coordinate the Bayes factor in favor of a constant drift model over a drift-neutral model is greater than 1,000, indicating a substantially better fit for the constant drift model. These results imply general eastward and southward trends in the spread of HIV from the Kinshasa-Brazzaville-Pointe-Noire area to other population centers. They also reflect the composition of sampling locations: in terms of longitude, a majority are far to the east of the believed origin while the rest are relatively close to it. Similarly, nearly 90% of the sequences come from locations south of Kinshasa or from neighboring locations of similar latitude. On the other hand, the existence of samples from cities north of Kinshasa, Bwamanda and Kisangani, suggests that diffusion with a northward trend may more accurately characterize part of the spatial history. We explore this possibility with the relaxed drift model.

Under the relaxed drift model, there is significant evidence of multiple longitudinal drift rates. We estimate a posterior mean of 1.48 drift rate changes with BCI (1, 3), and the Bayes factor in favor of relaxed drift over constant drift is greater than 1,000. The posterior mean longitudinal drift rate across all branches is 0.12 with BCI (0.08, 0.16). Figure 4.1 shows the maximum clade credibility tree colored according to drift rates. The tree is essentially divided up into two clades: one clade with green-colored branches and another with brown-colored branches. We infer eastward drift rates of 0.18 degrees per year on green branches, and drift rates close to or equal to 0 on brown branches. Tree nodes in Figure 4.1 are depicted as circles of different sizes, with the size of each circle being determined by the longitude of the observed or inferred location corresponding to the node. Larger circles represent more eastward locations. The observed and inferred longitudes provide better understanding of the difference in drift rates between the two clades. During the period 1960-2004, there is generally greater eastward movement along the lineages of the green clade. Although

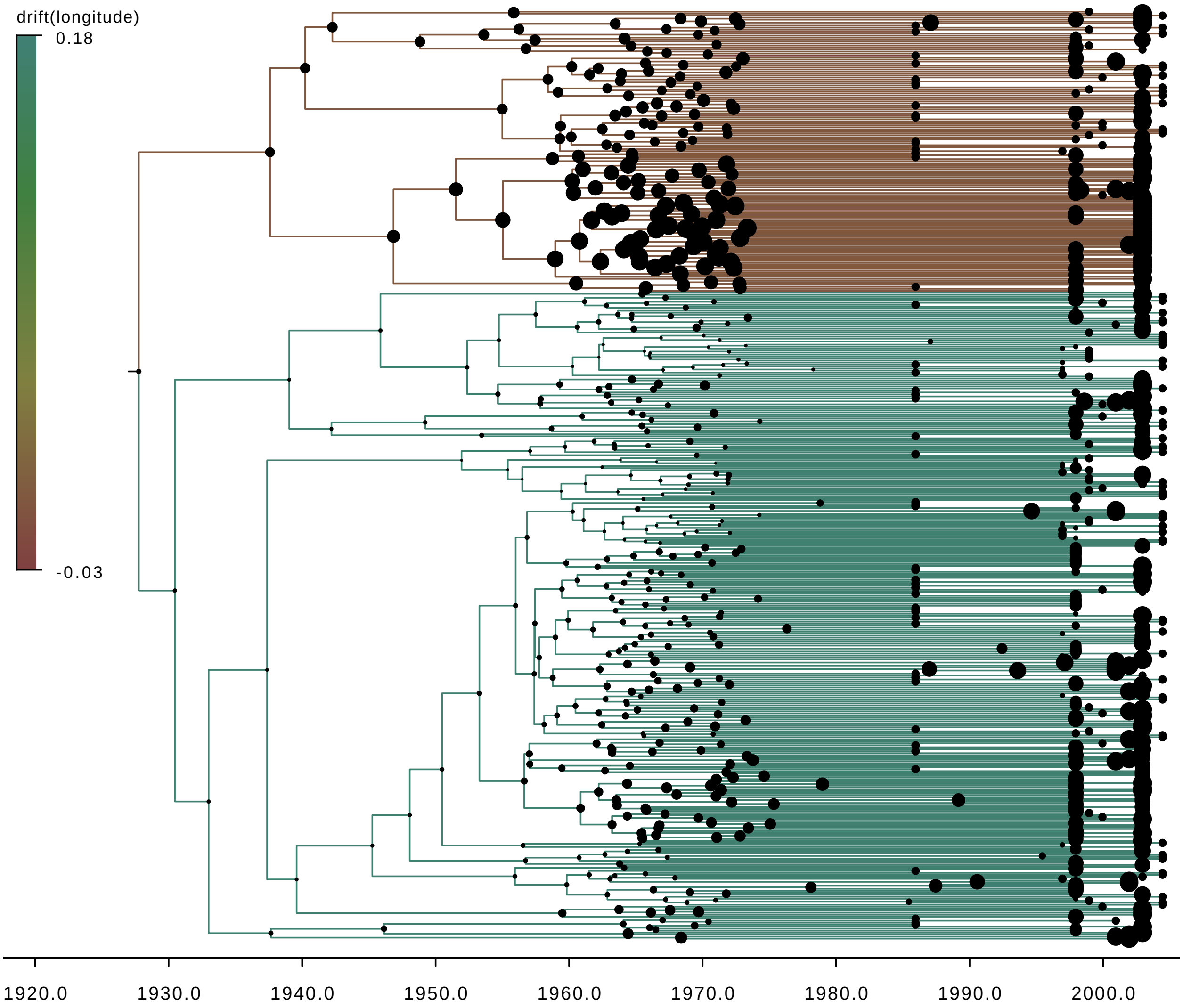

Figure 4.1: Maximum clade credibility tree for spread of HIV-1 in central Africa. The posterior mean longitudinal drift is depicted using a color gradient along the branches. Green indicates an eastward drift while brown signifies drift rates close to or equal to zero. Tree nodes are depicted as circles of different sizes. The size of each circle is determined by the longitude of the observed or inferred location corresponding to the node. Larger circles represent more eastward locations.

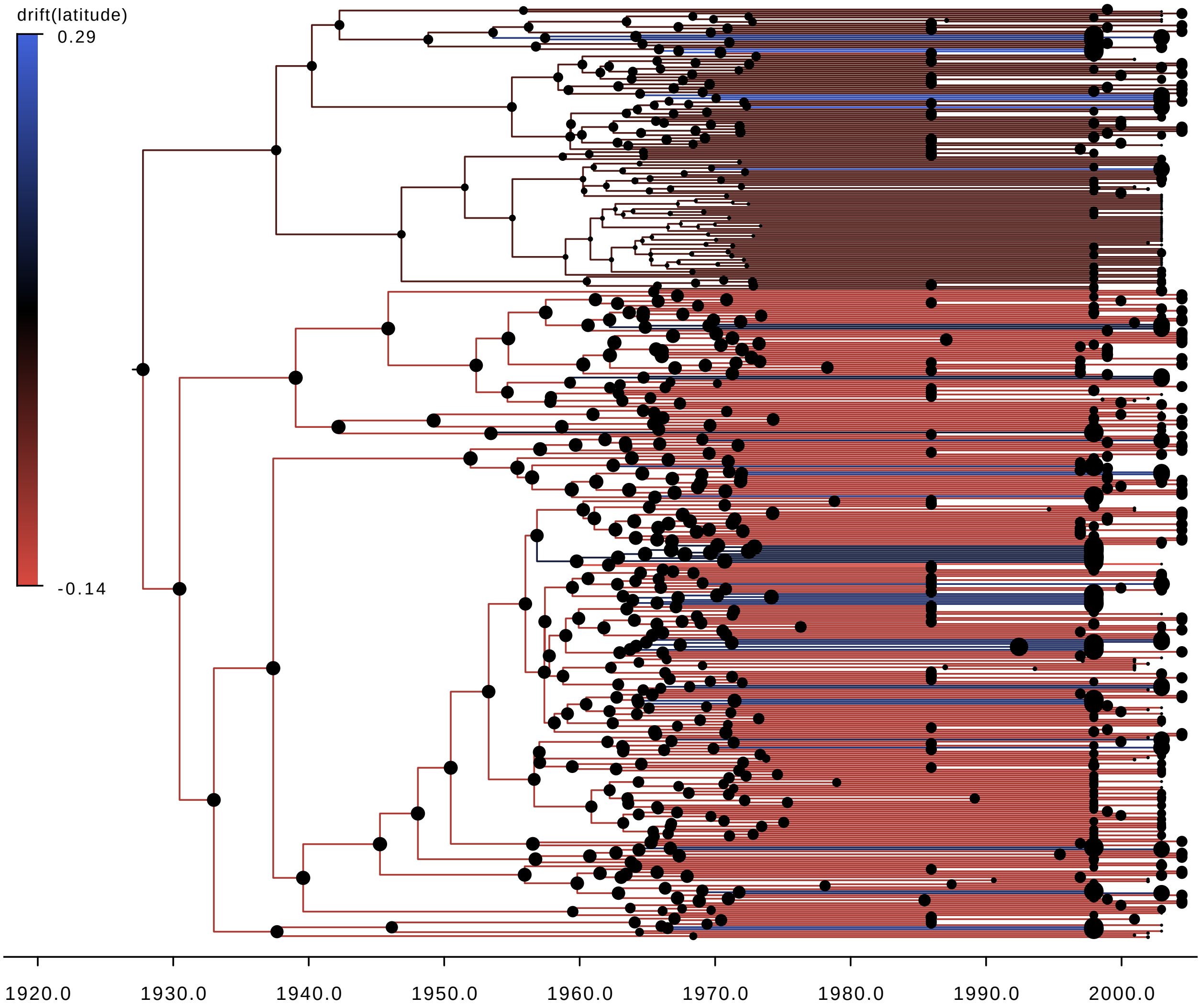

Figure 4.2: Maximum clade credibility tree for spread of HIV-1 in central Africa. The posterior mean latitudinal drift is depicted using a color gradient along the branches. The colors range between red and blue, with the former indicating a southward drift and the latter a northward drift. Tree nodes are depicted as circles of different sizes. The size of each circle is determined by the longitude of the observed or inferred location corresponding to the node. Larger circles represent more northward locations.

the lineages of the brown clade show greater eastward spread during the first half of the evolutionary history, the drift rates in the two clades are driven by the trends of the second half of the evolutionary history. The second half accounts for a much greater proportion of tree branches and, because it overlaps with all sampling times, contains more information about the spatial diffusion process.

For latitudinal drift, we estimate 28.13 rate changes with BCI (27, 29). Furthermore, relaxed drift is supported over constant drift with a Bayes factor greater than 1,000. The overall posterior mean drift rate across all branches is -0.03 with BCI (-0.05, -0.01). Figure 4.2 depicts an annotated maximum clade credibility tree colored according to latitudinal drift rates. The color gradient ranges from red to black to blue, with red indicating a southward drift, blue a northward drift, and black a neutral drift. As in Figure 4.1, tree nodes are depicted as circles of varying sizes, with larger circles representing greater observed or inferred latitude coordinates. Notably, external branches with positive drift rates lead to samples from locations north of the origin (Bwamanda and Kisangani), while external branches with nonpositive drift rates lead to samples from locations with latitudes south of or similar to the origin.

Importantly, adopting a zero-mean displacement distribution when the diffusion process is more accurately described with a nontrivial mean can result in inflated displacement variance rate estimates. (Recall that the displacement variance along a branch of length t_i is $t_i\mathbf{P}^{-1}$, so by "displacement variance rates," we mean the diagonal elements of the variance rate matrix $\mathbf{P}^{-1}$.) By incorporating drift into the model, we are able to disentangle the drift from the variance and uncover a clearer picture of the movement. The variance rate estimates in Table 4.1 illustrate this point. Including a constant drift reduces the variance rate of the displacement of the longitude coordinate from 0.59 to 0.37. For the latitude, on the other hand, the variance rate decreases modestly from 0.25 to 0.23 and the BCIs of (0.23, 0.27) and (0.21, 0.25) overlap. The lack of an appreciable reduction in variance rate may be explained by the fact that the apparent northward drift to Bwamanda and Kisangani remains unaccounted for by inclusion of a constant southward drift rate. Indeed, by accommodating drift rate changes under the relaxed drift model, the variance rate of the latitudinal displacement drops to 0.13 with BCI (0.12, 0.14).

4.4 West Nile Virus

West Nile virus (WNV) is the most important mosquito-transmitted arbovirus now native to the U.S. Birds are the most common host, although WNV has been known to infect other animals including humans. About 70-80% of infected humans do not develop any symptoms, and most of the remaining infections result in fever and other symptoms such as headaches, body aches, joint pains, vomiting, diarrhea or rash. Fewer than 1% of infected humans will develop a serious neurologic illness such as encephalitis or meningitis. However, recovery from neurologic infection can take several months, some neurologic effects can be permanent, and about 10% of such infections lead to death. Since 1999, WNV has been responsible for more than 1,500 deaths in the United States. (CDC, 2013).

WNV was first detected in the United States in New York City in August 1999, and is most closely related to a highly pathogenic WNV lineage isolated in Israel in 1998 (Lanciotti et al., 1999). Surveillance records of WNV incidence show a wave of infection that spread westward and reached the west coast by 2004 (CDC, 2013). Thus the spread of WNV in North America can be naturally modeled as a spatial diffusion process with a directional trend.

We analyze a data set of 104 WNV complete genome sequences sampled between 1999 and 2008 and isolated from a number of different host and vector species (Pybus et al., 2012). The sequences were sampled from a wide variety of locations in the United States and Mexico. We represent the sampling locations as bivariate traits consisting of latitude and longitude coordinates and conduct a phylogeographic analysis using our Brownian drift diffusion model. Results are presented in Table 4.2.

We begin by analyzing the data with the constant drift model. For the longitude, the posterior mean drift rate is -1.77 degrees per year with a 95% BCI of (-3.43, -0.27). This indicates a significant westward trend in the spread of WNV, consistent with WNV incidence data. Furthermore, a Bayes factor of 18.23 lends considerable support to the constant drift model over the drift-neutral model. For the latitude, we have an estimated posterior mean drift of -0.25 degrees per year, but the 95% BCI of (-1.05, 0.08) contains zero. Moreover, the Bayes factor in favor of constant drift over no drift is 0.71. Therefore there is not much evidence of a significant North-South drift. We check for the possibility of multiple drift rates under the relaxed drift model. However, the inferred number of rate changes in the latitudinal drift is 0.63 with BCI (0, 2), and the Bayes factor in favor of relaxed drift over constant drift is 0.87. Similarly, for the longitude coordinate, we estimate 0.71 rate changes with BCI (0, 2), and Bayes factor of 1.04. Thus the data do not support relaxed drift over constant drift.

While inclusion of drift leads to smaller displacement variance rates in our analysis of HIV-1 dispersal dynamics, Table 4.2 shows that not to be the case for the WNV data. Even for the longitude coordinate where there is significant evidence of a nontrivial drift, there is a great deal of BCI overlap in the estimated displacement variance rate for drift-neutral and constant drift diffusion. This difference can be explained by comparing the magnitude of the displacement drift relative to the displacement standard deviation in the WNV and HIV examples. Here, we work with displacement standard deviation rather than displacement variance in order to make comparisons on the same scale. Recall that the displacement mean and variance along a branch are both proportional to the branch length. To get an estimate of the average displacement mean and standard deviation for each example, we use the mean branch length. The mean branch length is 16.9 years for the HIV data set, and 2.4 years for the WNV data set. Under constant drift, we get an average longitudinal displacement mean of 5.07 for the HIV data, and -4.25 for the WNV data. The average longitudinal displacement standard deviation without drift is 3.25 for the HIV data, and 7.49 for the WNV data. In the HIV analysis, the average displacement mean with drift is 1.61 times as much as the average displacement standard deviation without drift. In the case of the WNV, the absolute value of the average displacement mean with drift is 0.57 times as much as the average displacement standard deviation without drift. So while the displacement standard deviation in the drift-neutral HIV analysis is inflated by the hidden drift, the latent drift represents a much smaller contribution to the drift-neutral displacement standard deviation in the WNV analysis.

	No Drift		Constant Drift	
Drift (Lat.)	-	-	-0.25	(-1.05, 0.08)
Variance (Lat.)	6.74	(4.83, 15.81)	6.38	(4.60, 14.61)
Drift (Long.)	-	-	-1.77	(-3.43, -0.27)
Variance (Long.)	23.37	(16.95, 54.04)	20.06	(14.67, 45.52)
Correlation	0.22	(0.03, 0.43)	0.18	(-0.03, 0.36)

Table 4.2: Spatiotemporal dynamics of West Nile virus in North America. Model comparison of Brownian diffusion with no drift and constant drift Brownian diffusion. We report posterior mean estimates along with 95% BCIs. Drift rates for latitude and longitude coordinates are reported in units of degrees per year.

4.5 HIV-1 Resistance to Broadly Neutralizing Antibodies

It is widely believed that a successful HIV-1 vaccine will require the elicitation of neutralizing antibodies (Johnston and Fauci, 2007; Barouch, 2008; Walker and Burton, 2008). Most neutralizing antibodies are strain-specific and therefore not so attractive for vaccine design (Weiss et al., 1985; Mascola and Montefiori, 2010). It is important to identify and characterize antibody specificities that are effective against a large number of currently circulating HIV-1 variants (Burton, 2002, 2004). Several broadly neutralizing monoclonal antibodies have been recently isolated, including PG9 and PG16 (Walker et al., 2009), and VRC01 (Zhou et al., 2010).

Studies comparing viruses isolated from individuals who seroconverted early in the HIV-1 epidemic to viruses from individuals who seroconverted in recent years have shown that HIV-1 has become increasingly resistant to antibody neutralization over the course of the epidemic (Bunnik et al., 2010; Euler et al., 2011; Bouvin-Pley et al., 2013). Bunnik et al. (2010) demonstrate a decreased sensitivity to polyclonal antibodies and to monoclonal antibody

Antibody	Constant Drift		Relaxed Drift	
PG9	0.08	(-0.05, 0.19)	0.07	(-0.04, 0.18)
PG16	0.10	(-0.02, 0.20)	0.11	(-0.01, 0.24)
VRC01	0.15	(0.06, 0.24)	0.15	(0.09, 0.21)

Table 4.3: HIV-1 resistance to broadly neutralizing antibodies. Mean drift rates under constant and relaxed drift models for $\log(\mathrm{IC}_{50})$ measurements corresponding to monoclonal neutralizing antibodies PG9, PG16, and VRC01. Higher $\log(\mathrm{IC}_{50})$ values represent lower sensitivity to antibody neutralization, and positive drift rates indicate a trend over time toward greater resistance. We report posterior means along with 95% BCIs.

	Displacement Variance					
Antibody	No Drift		Constant Drift		Relaxed Drift	
PG9	0.28	(0.18, 0.57)	0.26	(0.16, 0.53)	0.26	(0.17, 0.55)
PG16	0.26	(0.17, 0.51)	0.23	(0.15, 0.49)	0.23	(0.15, 0.51)
VRC01	0.19	(0.11, 0.43)	0.13	(0.08, 0.27)	0.06	(0.04, 0.14)

Table 4.4: HIV-1 resistance to broadly neutralizing antibodies. Displacement variance rate under drift-neutral, constant drift and relaxed drift models for $\log(\mathrm{IC}_{50})$ measurements corresponding to monoclonal neutralizing antibodies PG9, PG16, and VRC01. We report posterior means along with 95% BCIs.

b12. Euler et al. (2011) extend those findings by investigating whether HIV-1 has adapted to the neutralization activity of PG9, PG16, and VRC01. Their results show that HIV-1 has become significantly more resistant to neutralization by VRC01 and also provide some support for increased resistance to neutralization by PG16.

These studies typically do not account for phylogenetic dependence among the sampled viruses. Vrancken et al. (2014) examine the data set of Euler et al. (2011) with a Brownian diffusion trait evolution model that simultaneously infers phylogenetic signal, the degree to which resemblance in traits reflects phylogenetic relatedness. They find moderate phylogenetic signal and, through ancestral trait value reconstruction, more evidence of decreased sensitivity of HIV-1 to VRC01 and PG16 neutralization at the population level.

We follow up on the analysis of Vrancken et al. (2014) by incorporating drift into the Brownian trait evolution. The data set is comprised of clonal HIV-1 variants from “historic” and “contemporary” seroconverters with an acute or early subtype B HIV-1 infection. The 14 historic seroconverters have a known seroconversion date between 1985 and 1989, and the 21 contemporary seroconverters have a seroconversion date between 2003 and 2006. The percent neutralization is determined by calculating the reduction in p24 production in the presence of the neutralizing agent compared to the p24 levels in the cultures with virus only. The trait values of interest are 50% inhibitory concentration (IC_{50}) assay values that summarize the percent neutralization by antibodies PG9, PG16 and VRC01, measured in units of μg/ml. We take the log-transform of IC_{50} values in order to ensure that concentration values are strictly positive under the diffusion process. Higher $\log(\mathrm{IC}_{50})$ values correspond to greater resistance to antibody neutralization. For viruses with $\log(\mathrm{IC}_{50})$ values that fall outside the tested antibody concentration range, we integrate out the concentration over a plausible IC_{50} interval.

First, we analyze the data with the constant drift model (see Table 4.3). The results are essentially consistent with the findings of Euler et al. (2011). For VRC01, we estimate a posterior mean drift of 0.15 with 95% BCI (0.06, 0.24), signaling a significant drift toward higher resistance to VRC01 neutralization. Furthermore, a Bayes factor of 32.33 lends strong support to a constant drift over no drift. There is not as much evidence of a trend for PG16.

On one hand, the posterior probability that the drift rate is positive is 0.953, providing some corroboration for a decreased sensitivity to PG16 neutralization. However, we infer a mean drift rate of 0.10 with a 95% BCI of (-0.02, 0.20) that contains zero. Furthermore, the Bayes factor in favor of constant drift over no drift is just 1.32, showing little support for inclusion of a drift term. We do not detect a significant drift in the case of PG9: the posterior mean is 0.08 with 95% BCI (-0.05, 0.19) and the Bayes factor is 1.17.

To take a closer look, we fit the data to the relaxed drift model. Along with branch specific drift rates, we examine the posterior mean rates over the entire evolutionary history (see Table 4.3). First, we consider the results for PG9 and PG16. The posterior mean drift rates are similar to the inferred drift rates under the constant drift model, and their 95% BCIs contain 0. There is little support for any drift rate changes occurring along the phylogeny. The mean estimated number of rate changes are 0.17 and 0.2 for PG9 and PG16, respectively, and the Bayes factors in favor of relaxed drift over constant drift are 0.19 and 0.20. Hence there is not much evidence of localized directional trends that differ from the overall directional trends.

We illustrate the evolutionary pattern for resistance against VRC01 under the relaxed drift model in Figure 4.3. The branches of the maximum clade credibility tree are colored according to the inferred branch-specific mean drift rates, and the tree tips are labeled with subject identifications. We obtain a posterior mean estimate of 1.13 rate changes, with a posterior probability of 0.88 for exactly one rate change and probability greater than 0.99 for at least one rate change. Furthermore, the Bayes factor in favor of relaxed drift over constant drift is 359.04, providing very strong support for relaxed drift. As shown in Table 4.4, the displacement variance rate decreases as we move from no drift to a constant drift, and then to relaxed drift. Figure 4.3 shows a rate change occurring at the common ancestral node of samples from subjects P001 and P002. Apart from the two branches leading to tips P001 and P002 (which we refer to as "branch P001" and "branch P002," respectively), the branches have essentially identical mean drift rates of about 0.15 with 95% BCI (0.09,0.21). For the blue-colored branch P001, we have an estimated drift of 2.13 with 95% BCI (0.06, 4.89), and for the red-colored branch P002 the estimated drift is -1.18 with 95% BCI (-4.46, 0.22). Both estimated drift rates are drastically different from the parent branch rate, and their BCIs are also much wider. If either subject P001 or P002 is deleted from the data set, we infer a constant, significant drift rate of 0.15 over the entire evolutionary history. Notably, we do not infer any drift rate changes after deleting either P001 or P002.

Under the relaxed drift model, situations in which both child branches have drift rates that differ from the drift rate of the parent branch must be handled with care. In any given tree in the posterior sample, one of the two branches must inherit its drift rate from the parent branch. An examination of the posterior distribution of the rate change indicator at the parent node of tips P001 and P002 reveals that branch P001 inherits the parent branch drift rate with posterior probability of about 40% and branch P002 inherits the parent rate with the other 60% probability mass. Thus the posterior drift rate estimate for each of branches P001 and P002 averages over the cases where it inherits the parent drift rate and the cases where it differs from the parent rate. While the potentially large departures from the parent drift rate still come through in the "averaged" posterior estimates, it is of interest to find out how representative they are of the true drift rates. It is conceivable that the "inherited" portion of the posterior may bias the estimate, shifting the mean and widening the BCI. In the case of branch P002, for example, the posterior mass near the parent drift rate extends the BCI into the positive axis so that it includes zero. It is also conceivable that the "inherited" part of the posterior represents a part of the distribution that would show up even without the restrictions of the relaxed drift model.

To elucidate the true nature of the drift rate change that occurs at the parent of tips P001 and P002, we conduct a follow-up analysis. We introduce a new parameterization of our drift diffusion model that posits three unique drift rates: one each corresponding to branches P001 and P002, and another for all remaining branches in the phylogenetic tree. The results, presented in Table 4.5, are similar to the findings from the relaxed drift model. Notably, the 95% BCIs for the drift on branches P001 and P002 still contain the range of credible values for the parent drift rate. There are also some key differences between the two analyses. The distributions for the drift on branches P001 and P002 are bell-shaped rather than bimodal, and the 95% BCIs are wider than under the relaxed drift model. Unlike the

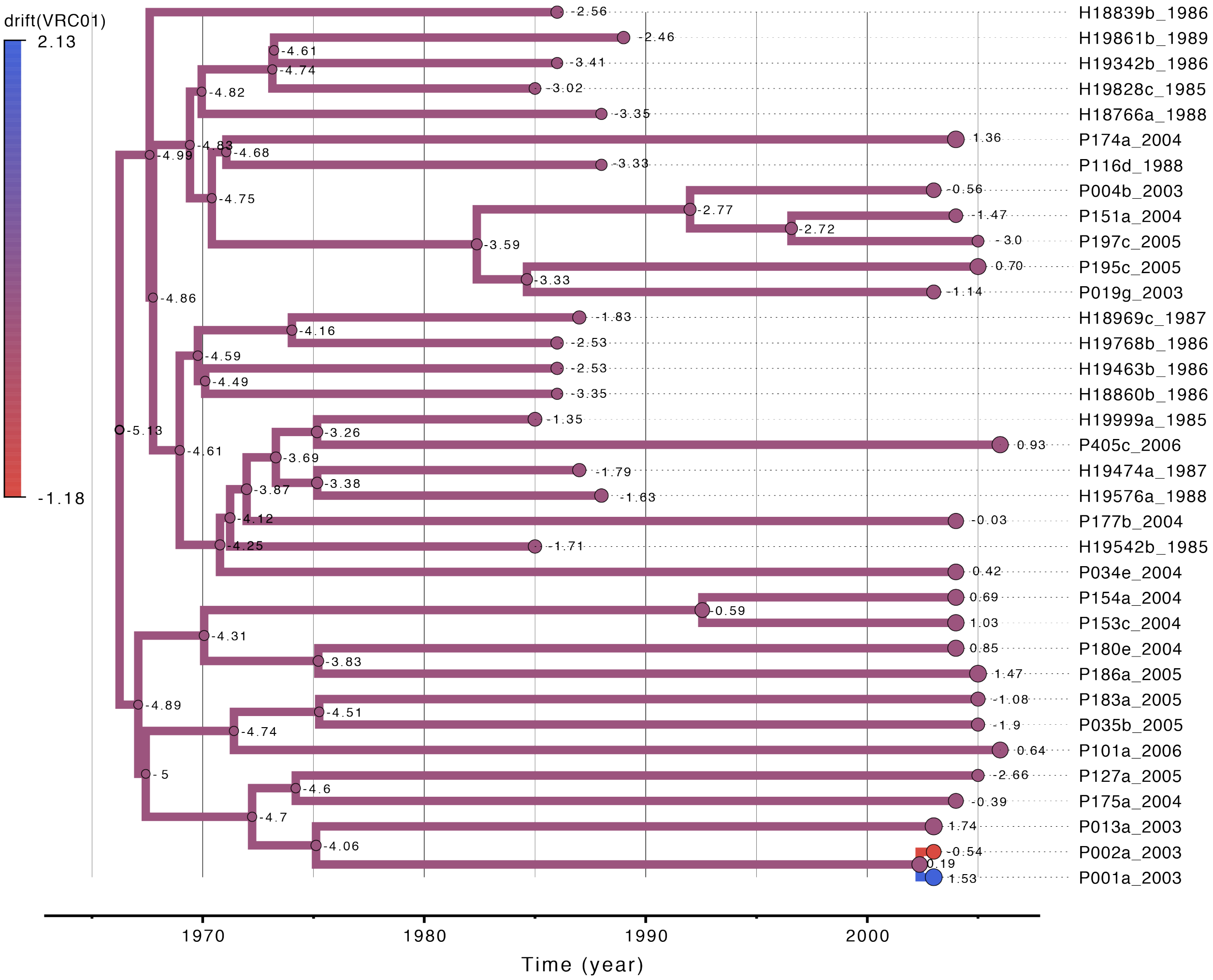

Figure 4.3: Maximum clade credibility tree depicting evolutionary pattern of HIV-1 resistance to neutralization by antibody VRC01. Subject identifiers corresponding to tree tips are listed to the right of the tree. The posterior mean drift is depicted using a color gradient along the branches. Positive drift rates correspond to a trend toward greater resistance to neutralization. The estimated posterior mean drift rate along the purple colored branches that make up most of the tree is 0.15. A drift rate change is inferred at the common ancestor of tips P001 and P002. The red colored branch leading to the tip P002 has an estimated posterior mean drift of -1.18, and the blue colored branch leading tip P001 has an estimated posterior mean drift of 2.13. Tree tips and internal nodes are annotated with observed and inferred $\log(IC_{50})$ values.

relaxed drift model estimate, the 95% BCI for the drift on branch P001 contains 0. However, it has a 0.95 posterior probability of being positive, so there is still some support that the drift on branch P001 is statistically significant.

An examination of the maximum clade credibility tree in Figure 4.3, annotated with observed $\log(\mathrm{IC}_{50})$ values and inferred ancestral trait realizations, clarifies why we infer a drift rate change. Consider node triples consisting of two nodes and their common parent node. The triple of tips P001, P002 and their common ancestor features a relatively large difference in trait values at the child nodes as well as relatively short branch lengths connecting nodes P001 and P002 to their parent. While there are other triples with child nodes possessing a comparable difference in $\log(\mathrm{IC}_{50})$ values, they have much longer branches leading from the parent node to the children. Similarly, while other triples feature relatively short branches connecting the parent to the children, the trait values at the child nodes do not differ as much. The unique combination of short branches coupled with a large difference between $\log(\mathrm{IC}_{50})$ values at the child nodes explains why the drift rate present on most of the tree may be incompatible with the triple of P001, P002 and their parent.

Although there is strong evidence of a drift rate change at the ancestral node of samples P001 and P002, the wide BCIs for the branch P001 and branch P002 drift rate estimates suggest that they are poorly informed by the data. The type of drift rate change that occurs at the ancestral node of tips P001 and P002 also remains unclear. Subjects P001 and P002 are a transmission couple and it appears that one person mounted a very different antibody response than the other to a highly similar virus. Further research may clarify the situation. Nevertheless, we infer a clear, significant drift towards increased resistance to neutralization by VRC01, and it is robust to deletion of either subject P001 or P002. At the population level, the phylogenetic structure of HIV is "starlike," featuring multiple co-circulating lineages, the dynamics of which generally reflects neutral epidemiological processes (Grenfell et al., 2004). It is therefore notable to find evidence of population level evolution towards increased resistance.

	Relaxed Drift			Fixed-Changes		
Branch	Mean	95% BCI	P(Drift $>$ 0)	Mean	95% BCI	P(Drift $>$ 0)
P001	2.13	(0.06, 4.89)	$>$ 0.99	2.27	(-0.36, 5.10)	0.95
P002	-1.18	(-4.46, 0.22)	0.60	-1.81	(-4.61, 1.03)	0.10
Other	0.15	(0.09, 0.21)	$>$ 0.99	0.14	(0.08, 0.20)	$>$ 0.99

Table 4.5: Rate changes in HIV-1 drift toward resistance to antibody VRC01. We report posterior means, 95% BCIs, and posterior probabilities that the drift > 0.

4.6 Discussion

Standard Brownian diffusion is a popular and, in many ways, natural starting point for modeling continuous trait evolution in a phylogenetic context. On the other hand, it is very restrictive and may not adequately describe the dynamics of the underlying evolutionary process. Development of non-Brownian models, such as mean-reverting Ornstein-Uhlenbeck processes, represents a promising avenue. However, substantial gains can also be made through building upon standard Brownian diffusion approaches. For example, the displacement along a branch is typically assumed to have variance equal to the product of the branch length and a diffusion variance rate matrix $\mathbf{P}^{-1}$, where $\mathbf{P}^{-1}$ does not vary along the phylogenetic tree. Lemey et al. (2010) demonstrate improvements by relaxing this homogeneity assumption via branch-specific diffusion rate scalars that yield a mixture of Brownian processes. Here, we show that progress towards a more realistic trait diffusion can be made by relaxing the assumption of a zero-mean displacement. Furthermore, the drift diffusion approach we consider is very general. Notably, the Ornstein-Uhlenbeck process is nested within the drift diffusion process defined by

$$\mathbf{Y}_i|\mathbf{Y}_{pa(i)} \sim N\left(\boldsymbol{\beta}_1(t_i)\mathbf{Y}_{pa(i)} + \boldsymbol{\beta}_2(t_i)\boldsymbol{\mu}_i, \boldsymbol{\Sigma}(t_i)\right). \tag{4.45}$$

Consider the special case where $\boldsymbol{\mu}_i = \boldsymbol{\mu}$ for every branch, and

$$\boldsymbol{\beta}_1(t_i) = e^{-\alpha t_i}, \quad \boldsymbol{\beta}_2(t_i) = 1 - e^{-\alpha t_i}, \quad \text{and} \quad \boldsymbol{\Sigma}(t_i) = \frac{\sigma^2}{2\alpha}\left[1 - e^{-2\alpha t_i}\right]. \tag{4.46}$$

This is equivalent to an Ornstein-Uhlenbeck process on a phylogenetic tree defined by the stochastic differential equation

$$d\mathbf{Y}_t = \alpha(\boldsymbol{\mu} - \mathbf{Y}_t)dt + \sigma d\mathbf{W}_t, \tag{4.47}$$

where $\mathbf{W}_t$ is a standard Brownian diffusion process. Here, $\boldsymbol{\mu}$ can be thought of as an optimal trait value, α represents the strength of selection towards $\boldsymbol{\mu}$, and σ^2 is the variance of the Brownian diffusion component. Such generality enables formal testing between a wide class of different Gaussian process models.

We introduce a flexible new Bayesian framework for phylogenetic trait evolution, modeling the evolutionary process as Brownian diffusion with a nontrivial drift. By allowing an estimable mean vector in the displacement distribution, we can account for and quantify a directional trend. However, imposing a constant drift rate can make for an unrealistic approximation of the underlying process. We overcome this limitation through the relaxed drift model. The relaxed drift model permits drift rate variation along a phylogenetic tree while maintaining model identifiability. Drift rates are generally passed on from parent branches to child branches, and variation is achieved by allowing at most one branch of any given pair of child branches to assume a different drift rate from their common parent branch.

The utility of incorporating drift into the diffusion is corroborated by our analyses of three viral examples. We apply our methodology to both geographic traits in a phylogeographic setting as well as phenotypic traits. Our phylogeographic analysis of the spread of HIV-1 in central Africa confirms the findings obtained by discrete phylogeographic inference (Faria et al., 2014). Drift diffusion models fit the data better than drift-neutral Brownian diffusion, and we uncover directional trends in the dispersal of the virus from its origin to sampling locations. We also see that drift rate variation characterizes real spatiotemporal diffusion processes. The absence of drift in the diffusion model can lead to conflation of the latent drift with other parameters, particularly the displacement variance. Our analysis of the spread of HIV-1 illustrates how inferred displacement variance rates can decrease with appropriate drift rate modeling, revealing a clearer, more detailed picture of dispersal dynamics.

While it is tempting to assume drift rate variation and seek out the additional insight it may provide, the data may not support multiple drift rates. This may be the case even when there is a significant constant drift, as we see in our analysis of the West Nile virus. Parameterizing the model to allow the maximal number of unique drift rates can result in numerous small rate changes that are a consequence of the modeling choice and are not necessarily driven by the data. Our relaxed drift framework overcomes this issue by inferring the locations and types of rate changes directly from the data as opposed to making a priori assumptions about the number of unique drift rates and their appropriate assignments. Bayesian stochastic search variable selection enables efficient exploration of all possible drift rate configurations.

Although our focus has been on drift, a major strength of our approach is its implementation in the larger Bayesian phylogenetic framework of BEAST. Through BEAST, we have access to a plethora of different models for molecular character substitution, demographic history and molecular clocks. Bayesian inference provides a natural framework for controlling for different sources of uncertainty in evolutionary models, including the phylogenetic tree and trait and sequence evolution parameters, and testing evolutionary hypotheses.

The gains from introducing drift into our real data examples are encouraging, and there is a need for continued development of more realistic trait evolutionary models. We anticipate that our drift diffusion approach will be useful in other scenarios not examined here, including antigenic drift in influenza. Antigenic drift is the process by which influenza viruses evolve to evade the immune system, and an understanding of its dynamics is essential to public health efforts. Bedford et al. (2014) have recently developed an integrated approach to mapping antigenic phenotypes that combines it with genetic information. It may be fruitful to model the diffusion of the antigenic phenotype in their framework with a relaxed drift.

While the relaxed drift model has proven to be flexible and useful, its identifiability restrictions may render it inappropriate for some evolutionary scenarios. For example, once a drift rate appears anywhere in the phylogenetic tree, the restrictions mandate that it must be passed on and "survive" until it reaches an external branch. Lack of support for the survival of a specific drift rate may not preclude its inclusion under relaxed drift. In the analysis of HIV-1 resistance to neutralization, for example, the parent branch of branches

P001 and P002 must pass on its drift rate to one of the two child branches in each tree in the posterior sample. Yet, the branch which is forced to inherit the parent drift rate alternates between P001 and P002 in the posterior sample, resulting in posterior drift distributions for both branches that differ from that of their parent drift. However, this unnatural mechanism of deflecting an unsupported drift rate may contribute to misleading drift rate estimates. It would be preferable to sidestep such problems by developing alternative models that accommodate drift rate variation while retaining identifiability.

Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement no. 278433-PREDEMICS and ERC Grant agreement no. 260864 and the National Institutes of Health (R01 AI107034, R01 HG006139, R01 LM011827, T32 AI007370 and 5T32AI007370-24) and the National Science Foundation (DMS 1264153).

4.7 Appendix: Identifiability

To ensure that our results are meaningful, it is important to understand the conditions under which our model is identifiable. For convenience, and without loss of generality, we assume here that traits are one-dimensional. Following the development in section 4.2.1, the observed traits $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_N)^t$ at the tips of the phylogeny τ are multivariate-normal distributed:

$$P\left(\mathbf{Y}|\mathbf{Y}_{2N-1}, \mathbf{P}, \mathbf{V}_\tau, \boldsymbol{\mu}_\tau\right) = \text{MVN}\left(\mathbf{Y}; \mathbf{Y}_{\text{root}} + \mathbf{T}\boldsymbol{\mu}_\tau, \mathbf{P}^{-1} \otimes \mathbf{V}_\tau\right). \tag{4.48}$$

Here, $\mathbf{Y}_{\text{root}}$ is an $N \times 1$ vector with the root trait $\mathbf{Y}_{2N-1}$ repeated N times. The vector $\boldsymbol{\mu}_\tau = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{2N-2})^t$ consists of branch-specific drift rates $\boldsymbol{\mu}_i$. The $N \times N$ variance matrix $\mathbf{V}_\tau$ is a deterministic function of τ and represents the contribution of the phylogenetic tree to the covariance structure. Its diagonal entries V_{ii} are equal to the distance in time between the tip $\mathcal{V}_i$ and the root node $\mathcal{V}_{2N-1}$, and off-diagonal entries V_{ij} correspond to the distance in time between the root node $\mathcal{V}_{2N-1}$ and the most recent common ancestor of tips $\mathcal{V}_i$ and $\mathcal{V}_j$. Finally, the $N \times (2N-2)$ matrix $\mathbf{T}$ is defined as follows: $T_{ij} = t_j$, the length of branch j, if branch j is part of the path from the external node i to the root, and $T_{ij} = 0$ otherwise. In other words, the ith row of $\mathbf{T}$ specifies the path of branches in τ connecting external node i to the root.

Let $\mathbf{V}(\boldsymbol{\mu}_\tau)$ denote the vector space of permissible values of $\boldsymbol{\mu}_\tau$ for our model. With respect to the drift, the model is identifiable if the equality

$$P\left(\mathbf{Y}|\mathbf{Y}_{2N-1}, \mathbf{P}, \mathbf{V}_\tau, \boldsymbol{\mu}_\tau\right) = P\left(\mathbf{Y}|\mathbf{Y}_{2N-1}, \mathbf{P}, \mathbf{V}_\tau, \boldsymbol{\mu}^*_\tau\right) \tag{4.49}$$

implies that $\boldsymbol{\mu}_\tau = \boldsymbol{\mu}^*_\tau$. Because the drift appears only in the mean of the distribution, we have an identifiability problem if the same mean $\mathrm{E}(\mathbf{Y})$ can be realized from different values of $\boldsymbol{\mu}_\tau$. In other words, if there exist $\boldsymbol{\mu}_\tau \neq \boldsymbol{\mu}^*_\tau$ such that $\mathbf{T}\boldsymbol{\mu}_\tau = \mathbf{T}\boldsymbol{\mu}^*_\tau$. This can happen if and only if the linear transformation $\mathbf{T}$ has a nontrivial kernel. We know that

$$\dim \mathbf{V}(\boldsymbol{\mu}_\tau) = \dim\ \ker(\mathbf{T}) + \dim\ \mathrm{range}(\mathbf{T}), \tag{4.50}$$

and we also know that for any phylogeny τ, $\mathbf{T}$ is of full rank because its rows are linearly independent. It follows that for the kernel of $\mathbf{T}$ to be trivial, we must have

$$\dim \mathbf{V}(\boldsymbol{\mu}_\tau) \leq N. \tag{4.51}$$

If we allow a unique drift rate on each branch of τ, we have $\mathbf{V}(\boldsymbol{\mu}_\tau) = \mathbb{R}^{\mathrm{2N-2}}$. Therefore we must take a different approach.

It is illuminating to look at identifiability from the perspective of linear equations. For a given $\boldsymbol{\mu}_\tau$, $\mathbf{T}$ maps $\boldsymbol{\mu}_\tau$ to an $N \times 1$ vector $\boldsymbol{\gamma}$:

$$\mathbf{T}\boldsymbol{\mu}_\tau = \boldsymbol{\gamma}. \tag{4.52}$$

Identifiability is then equivalent to the system (4.52) of N linear equations having a unique solution.

To achieve identifiability, we introduce the relaxed drift model. Starting with a drift rate on the unobserved branch leading to the root node and moving down the tree toward the external nodes, every time a branch splits into two branches, one of two things happens. Either both of the child branches inherit the drift rate from the parent branch, or exactly one of the child branches inherits the drift rate from the parent branch while the other gets a new drift rate. Both child branches taking on different drift rates than the parent branch is not permitted. To avoid confusion, we continue to denote the $2N-2$ branch-specific drift rates as $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{2N-2}$, with the understanding that they are not all unique. We let $\boldsymbol{\mu}_1^*, \ldots, \boldsymbol{\mu}_K^*$ denote the unique drift rates, where $K \leq N$.

Definition: We say that a row $\mathbf{T}_i = (T_{i1}, T_{i2}, \ldots, T_{i,2N-2})$ in $\mathbf{T}$ is $\boldsymbol{\mu}_k^* -$ *dominated* if its associated path from the root of τ to a tip ends with a branch with drift rate $\boldsymbol{\mu}_k^*$. We also refer to the sum $\mathbf{T}_i \boldsymbol{\mu}_\tau = \sum_{j=1}^{2N-2} T_{ij} \boldsymbol{\mu}_j$ and the path associated with $\mathbf{T}_i$ as $\boldsymbol{\mu}_k^* -$ *dominated.*

Note that each unique drift rate dominates at least one path.

Definition: An *initial branch* of the rate $\boldsymbol{\mu}_k^*$ is a branch whose parent branch has a different drift rate. The unobserved branch leading to the root node is also defined to be an initial branch.

Observe that every branch with drift $\boldsymbol{\mu}_k^*$ is an initial branch of $\boldsymbol{\mu}_k^*$ or a descendant of an initial branch of $\boldsymbol{\mu}_k^*$. A drift rate may have more than one initial branch. In order to quantify how deep into the tree τ a drift rate extends (starting from the tips and going toward the root), we make the following definition.

Definition: By a *descendant path* of a branch b, we mean a series of connected branches, starting with a child branch of b and ending with a branch leading to a tip. The *depth* of a branch b is equal to the maximal number of branches in descendant paths of b. The depth of a drift rate $\boldsymbol{\mu}_k^*$ is equal to the maximal depth of its initial branches.

For example, if $\boldsymbol{\mu}_k^*$ has one initial branch leading to a tip, then $\boldsymbol{\mu}_k^*$ has depth 0. If the number of unknowns K is less than the number of equations N in the system $\mathbf{T}\boldsymbol{\mu}_\tau = \boldsymbol{\gamma}$, a unique solution can be established by working with a reduced system. We form the reduced system by choosing K of the N rows in $\mathbf{T}$, say $\mathbf{T}_{i_1}, \ldots, \mathbf{T}_{i_K}$, such that each is dominated by a different drift rate. If a drift rate dominates more than one path, we choose a path containing a maximal depth initial branch of the rate for the reduced system.

Claim: The relaxed drift model is identifiable.

Proof: The reduced linear system

$$\sum_{j=1}^{2N-2} T_{i_1 j}\boldsymbol{\mu}_j = \boldsymbol{\gamma}_{i_1} \tag{4.53}$$

$$\sum_{j=1}^{2N-2} T_{i_2 j}\boldsymbol{\mu}_j = \boldsymbol{\gamma}_{i_2} \tag{4.54}$$

$$\cdots \tag{4.55}$$

$$\sum_{j=1}^{2N-2} T_{i_K j}\boldsymbol{\mu}_j = \boldsymbol{\gamma}_{i_K}, \tag{4.56}$$

consists of K equations and K variables. Therefore to show that the solution is unique, it suffices to show that the linear system is independent. To establish independence, it suffices to show that if

$$a_1 \sum_{j=1}^{2N-2} T_{i_1 j}\boldsymbol{\mu}_j + a_2 \sum_{j=1}^{2N-2} T_{i_2 j}\boldsymbol{\mu}_j \cdots + a_N \sum_{j=1}^{2N-2} T_{i_K j}\boldsymbol{\mu}_j = 0, \tag{4.57}$$

for some constants $a_1, \ldots, a_K$, then we must have

$$a_1 = a_2 = \cdots = a_K = 0. \tag{4.58}$$

Suppose (4.57) holds. The idea behind the proof is as follows: we consider all drift rates of depth 0, conclude that each sum in (4.57) dominated by a drift rate of depth 0 must have its corresponding coefficient $a_i = 0$, then consider all drift rates of depth 1, conclude that each sum in (4.57) dominated by a drift rate of depth 1 must have corresponding coefficient $a_i = 0$, and so on until we have gone through all possible depth values of drift rates in τ.

Suppose $\boldsymbol{\mu}_k^*$ has depth 0. Then $\boldsymbol{\mu}_k^*$ only appears in the single $\boldsymbol{\mu}_k^*-$dominated sum and cannot be canceled out by a linear combination of the other sums. This forces the coefficient a_i of the $\boldsymbol{\mu}_k^*-$dominated sum in (4.57) to be equal to zero. Having shown that any sum dominated by a drift rate of depth 0 must have a zero coefficient in (4.57), we can move on to the case of depth 1. Rather than handle the case of depth 1 separately, we present a general argument.

Suppose the coefficients of all sums in (4.57) that are dominated by drift rates of depth less than m have been shown to be zero. Consider drift rates of depth m. If $\boldsymbol{\mu}_k^*$ has depth

m, then it appears in the $\boldsymbol{\mu}_k^*-$dominated path, and it may appear in paths dominated by other rates. Suppose $\boldsymbol{\mu}_k^*$ appears in a path P_i dominated by a different rate, say $\boldsymbol{\mu}_i^*$. By construction of the reduced system, P_i contains an initial branch $\mathbf{b}_i$ of $\boldsymbol{\mu}_i^*$ of maximal depth. This means the depth of $\mathbf{b}_i$ is equal to the depth of $\boldsymbol{\mu}_i^*$. Because $\mathbf{b}_i$ is a descendant of a branch with rate $\boldsymbol{\mu}_k^*$, the depth of $\boldsymbol{\mu}_i^*$ must be less than the depth of $\boldsymbol{\mu}_k^*$. But sums dominated by rates of depth less than m have already been shown to have zero coefficients in (4.57). Thus the sum associated with P_i has coefficient zero. Because $\boldsymbol{\mu}_k^*$ appears in only one sum which is not already known to have a zero coefficient, the $\boldsymbol{\mu}_k^*-$dominated sum, it follows that in order for (4.57) to hold, the $\boldsymbol{\mu}_k^*-$dominated sum must also have a zero coefficient. Therefore sums dominated by drift rates of depth m must have zero coefficients in (4.57).

Invoking this argument until we have gone through all possible values of drift rate depth, it follows that $a_1 = a_2 = \cdots = a_K = 0$. ■

CHAPTER 5

Understanding Past Population Dynamics: Bayesian Coalescent-Based Modeling with Covariates

5.1 Introduction

The effective population size is an abstract parameter of fundamental importance in population genetics, evolutionary biology and infectious disease epidemiology. Wright (1931) introduces the concept of effective population size as the size of an idealized Fisher-Wright population that gains and loses genetic diversity at the same rate as the real population under study. The Fisher-Wright model is a classic forward-time model of reproduction that assumes random mating, no selection or migration, and non-overlapping generations. Coalescent theory (Kingman, 1982a,b) provides a probabilistic model for generating genealogies relating samples of individuals arising from a Fisher-Wright model of reproduction. Importantly, the coalescent elucidates the relationship between population genetic parameters and ancestry. In particular, the dynamics of the effective population size greatly inform the shapes of coalescent-generated genealogies. This opens the door for the inverse problem of coalescent-based inference of effective population size trajectories from gene genealogies.

While the coalescent was originally developed for constant-size populations, extensions that accommodate a variable population size (Slatkin and Hudson, 1991; Griffiths and Tavaré, 1994; Donnelly and Tavaré, 1995) provide a basis for estimation of the effective population size as a function of time (also called the demographic function). Early approaches assumed simple parametric forms for the demographic function, such as exponential or logistic growth, and provided maximum likelihood (Kuhner et al., 1998) or Bayesian (Drummond et al., 2002) frameworks for estimating the parameters that characterized the parametric forms. However, *a priori* parametric assumptions can be quite restrictive, and finding an appropriate parametric form for a given demographic history can be time consuming and computationally expensive. To remedy this, there has been considerable development of nonparametric methods to infer past population dynamics.

Nonparametric coalescent-based models typically approximate the effective population size as a piecewise constant or linear function. The methodology has evolved from fast but noisy models based on method of moments estimators (Pybus et al., 2000; Strimmer and Pybus, 2001), to a number of flexible Bayesian approaches, including multiple change-point models (Opgen-Rhein et al., 2005; Drummond et al., 2005; Heled and Drummond, 2008), and models that employ Gaussian process-based priors on the population trajectory (Minin et al., 2008; Gill et al., 2013; Palacios and Minin, 2013). Extending the basic methodological framework to incorporate a number of key features, including accounting for phylogenetic error (Drummond et al., 2005; Minin et al., 2008; Heled and Drummond, 2008; Gill et al., 2013), the ability to analyze heterochronous data (Pybus et al., 2000; Drummond et al., 2005; Minin et al., 2008; Heled and Drummond, 2008; Gill et al., 2013; Palacios and Minin, 2013), and simultaneous analysis of multilocus data (Heled and Drummond, 2008; Gill et al., 2013) has hastened progress.

In spite of all of these advances, there remains a need for further development of population dynamics inference methodology. One promising avenue is introduction of covariates into the inference framework. A central goal in demographic reconstruction is to gain insights into the association between past population dynamics and external factors (Ho and Shapiro, 2011). For example, Lorenzen et al. (2011) combine demographic reconstructions from ancient DNA with species distribution models and the human fossil record to elucidate how climate and humans impacted the population dynamics of woolly rhinoceros, woolly mammoth, wild horse, reindeer, bison and musk ox during the Late Quaternary period. Lorenzen et al. (2011) show that changes in megafauna abundance are idiosyncratic, with different species (and continental populations within species) responding differently to the effects of climate change, human encroachment and habitat redistribution. Lorenzen et al. (2011) identify climate change as the primary explanation behind the extinction of Eurasian musk ox and woolly rhinoceros, point to a combination of climatic and anthropogenic factors as the causes of wild horse and steppe bison decline, and observe that reindeer remain largely unaffected by any such factors. Similarly, Stiller et al. (2010) examine whether climatic changes were related to the extinction of the cave bear, and Finlay et al. (2007) consider the impact of domestication on the population expansion of bovine species. Comparison of external factors with past population dynamics is also a popular approach in epidemiological studies to explore hypotheses about the spread of viruses (Lemey et al., 2003; Faria et al., 2014).

In addition to the association between past population dynamics and potential driving factors, it is of fundamental interest is to assess the association between effective population size and census population size (Crandall et al., 1999; Liu and Mittler, 2008; Volz et al., 2009; Palstra and Fraser, 2012). For instance, Bazin et al. (2006) argue that in animals, diversity of mitochondrial DNA (mtDNA) is not reflective of population size, whereas allozyme diversity is. Atkinson et al. (2008) follow up by examining whether mtDNA diversity is a reliable predictor of human population size. The authors compare Bayesian Skyline (Drummond et al., 2005) effective population size reconstructions with historical estimates of census population sizes and find concordance between the two quantities in terms of relative regional population sizes.

Existing methods for population dynamics inference do not incorporate covariates directly into the model, and associations between the effective population size and potentially related factors are typically examined in post hoc fashions that ignore uncertainty in demographic reconstructions. We propose to fill this void by including external time series as covariates in a generalized linear model framework. We accomplish this task by building upon the the Bayesian nonparametric Skygrid model of Gill et al. (2013). The Skygrid is a particularly well-suited starting point among nonparametric coalescent-based models. In most other comparable models, the trajectory change-points must correspond to internal nodes of the genealogy, creating a hurdle for modeling associations with covariates that are measured at fixed times. The Skygrid bypasses such difficulties by allowing users to specify change-points, providing a more natural framework for our extension. Furthermore, the Skygrid's

Gaussian Markov random field (GMRF) smoothing prior is highly generalizable and affords a straightforward extension to include covariates.

We demonstrate the utility of incorporating covariates into demographic inference on four examples. First, we find striking similarities between the demographic and spatial expansion of raccoon rabies in North America. Second, we compare and contrast the epidemiological dynamics of dengue in Puerto Rico with patterns of viral diversity. Third, we examine the population history of the HIV-1 CRF02_AG clade in Cameroon and find that the effective population size is more reflective of HIV incidence than prevalence. Finally, we explore the relationship between musk ox population dynamics and climate change during the Late Quaternary period. Our extension to the Skygrid proves to be a useful framework for ascertaining the association between effective population size and external covariates while accounting for demographic uncertainty. Furthermore, we show that incorporating covariates into the demographic inference framework can improve estimates of effective population size trajectories, increasing precision and uncovering patterns in the population history that integrate the covariate data in addition to the sequence data.

5.2 Methods

5.2.1 Coalescent Theory

Coalescent theory forms the basis of our inference framework, and here we review the basic set-up. Consider a random sample of n individuals arising from a classic Fisher-Wright population model of constant size N_e. The coalescent (Kingman, 1982a,b) is a stochastic process that generates genealogies relating such a sample. The process begins at the sampling time of all n individuals, $t = 0$, and proceeds backward in time as t increases, successively merging lineages until all lineages have merged and we have reached the root of the genealogy, which corresponds to the most recent common ancestor (MRCA) of the sampled individuals. The merging of lineages is called a coalescent event and there are $n - 1$ coalescent events in all. Let t_k denote the time of the $(n-k)^{\text{th}}$ coalescent event for $k = 1, \ldots, n-1$ and

$t_n = 0$ denote the sampling time. Then for $k = 2, \ldots, n$, the waiting time $w_k = t_{k-1} - t_k$ is exponentially distributed with rate $\frac{k(k-1)}{2N_e}$.

Researchers have extended coalescent theory to model the effects of recombination (Hudson, 1983), population structure (Notohara, 1990), and selection (Krone and Neuhauser, 1997). We do not, however, incorporate any of these extensions here. The relevant extensions for our development generalize the coalescent to accommodate a variable population size (Griffiths and Tavaré, 1994) and heterochronous data (Rodrigo and Felsenstein, 1999). The latter occurs when one samples the n individuals at possibility different times.

Let $N_e(t)$ denote the effective population size as a function of time, where time increases into the past. Thus, $N_e(0)$ is the effective population size at the most recent sampling time, and $N_e(t')$ is the effective population size t' time units before the most recent sampling time. We also refer to $N_e(t)$ as the "demographic function" or "demographic model." Griffiths and Tavaré (1994) show that the waiting time w_k between coalescent events is given by the conditional density

$$P(w_k|t_k) = \frac{k(k-1)}{2N_e(w_k + t_k)} \exp\left[-\int_{t_k}^{w_k+t_k} \frac{k(k-1)}{2N_e(t)} dt\right]. \tag{5.1}$$

Taking the product of such densities yields the joint density of intercoalescent waiting times, and this fact can be exploited to obtain the probability of observing a particular genealogy given a demographic function.

5.2.2 Skygrid Demographic Model

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

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

$$P(\mathbf{g}|\boldsymbol{\theta}) = \prod_{i=1}^{m} P(g_i|\boldsymbol{\theta}). \tag{5.2}$$

To construct the likelihood of genealogy g_i, let t_{0_i} be the most recent sampling time of sequences contributing to genealogy i and t_{MRCA_i} be the time of the MRCA for locus i. Let x_{α_i} denote the minimal grid point greater than at least one sampling time in the genealogy, and x_{β_i} the greatest grid point less than at least one coalescent time. Let $u_{ik} = [x_{k-1}, x_k]$, $k = \alpha_i + 1, \ldots, \beta_i$, $u_{i\alpha_i} = [t_{0_i}, x_{\alpha_i}]$, and $u_{i(\beta_i+1)} = [x_{\beta_i}, t_{\text{MRCA}_i}]$. For each u_{ik} we let t_{kj}, $j = 1, \ldots, r_k$, denote the ordered times of the grid points and sampling and coalescent events in the interval. With each t_{kj} we associate an indicator ϕ_{kj} which takes a value of 1 in the case of a coalescent event and 0 otherwise. Finally, let v_{kj} denote the number of lineages present in the genealogy in the interval $[t_{kj}, t_{k(j+1)}]$. Following Griffiths and Tavaré (1994), the likelihood of observing an interval is

$$P(u_{ik}|\theta_k) \quad = \prod_{1 \le j < r_k : \phi_{kj} = 1} \frac{v_{kj}(v_{kj}-1)}{2\theta_k} \prod_{j=1}^{r_k - 1} \exp\left[-\frac{v_{kj}(v_{kj}-1)(t_{k(j+1)} - t_{kj})}{2\theta_k}\right], \tag{5.3}$$

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

The product of interval likelihoods (5.3) yields the likelihood of coalescent times given the sampling times with genealogy g_i. To obtain the likelihood of the genealogy, however, we must account for the specific lineages that merge and result in coalescent events. Let $P_*(u_{ik}|\theta_k)$ denote $P(u_{ik}|\theta_k)$ except with factors of the form $\frac{v_{kj}(v_{kj}-1)}{2\theta_k}$ replaced by $\frac{2(2-1)}{2\theta_k} = \frac{1}{\theta_k}$. Then

$$P(g_i|\boldsymbol{\theta}) = \prod_{k=\alpha_i}^{\beta_i+1} P_*(u_{ik}|\theta_k). \tag{5.4}$$

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

$$P(g_i|\boldsymbol{\theta}) = \prod_{k=\alpha_i}^{\beta_i+1} \left(\frac{1}{\theta_k}\right)^{c_{ik}} \exp\left[-\frac{SS_{ik}}{\theta_k}\right], \tag{5.5}$$

where the SS_{ik} are appropriate constants. Rewriting this expression in terms of $\gamma_k = \log(\theta_k)$, we arrive at

$$P(g_i|\boldsymbol{\gamma}) = \prod_{k=\alpha_i}^{\beta_i+1} e^{-\gamma_k c_{ik}} \exp[-SS_{ik}e^{-\gamma_k}] = \prod_{k=\alpha_i}^{\beta_i+1} \exp[-\gamma_k c_{ik} - SS_{ik}e^{-\gamma_k}]. \tag{5.6}$$

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

$$\begin{aligned} P(\mathbf{g}|\boldsymbol{\gamma}) &= \prod_{i=1}^{m} P(g_i|\boldsymbol{\gamma}) && (5.7)\\ &= \prod_{i=1}^{m} \prod_{k=\alpha_i}^{\beta_i+1} \exp[-\gamma_k c_{ik} - SS_{ik}e^{-\gamma_k}] && (5.8)\\ &= \exp\left[\sum_{k=1}^{M+1} \left[-\gamma_k c_k - SS_k e^{-\gamma_k}\right]\right] && (5.9) \end{aligned}$$

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

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

$$P(\boldsymbol{\gamma}|\tau) \propto \tau^{M/2} \exp\left[-\frac{\tau}{2}\sum_{i=1}^{M}(\gamma_{i+1} - \gamma_i)^2\right]. \tag{5.10}$$

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

$$P(\tau) \propto \tau^{a-1} e^{-b\tau}. \tag{5.11}$$

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

$$P(\boldsymbol{\gamma}, \tau | \mathbf{g}) \propto P(\mathbf{g}|\boldsymbol{\gamma}) P(\boldsymbol{\gamma}|\tau) P(\tau). \tag{5.12}$$

5.2.3 Incorporating Covariates

We can incorporate covariates into our inference framework by adopting a generalized linear model (GLM) approach. Let $Z_1, \ldots, Z_P$ be a set of P predictors. Each covariate Z_j is observed or measured at $M+1$ time points, $t_1, \ldots, t_M, t_{M+1}$. Here, $t_0 = 0$ is the most recent sequence sampling time, t_i denotes the units of time before t_0, and $t_0 < t_1 < \cdots < t_M < t_{M+1}$. Alternatively, the covariate may correspond to time intervals $[t_0, t_1], \ldots, [t_{M-1}, t_M], [t_M, t_{M+1}]$ rather than time points (for example, the yearly incidence or prevalence of viral infections). In any case, Z_{ij} denotes covariate Z_j at time point or interval i. Skygrid grid points are chosen to match up with measurement times (or measurement interval endpoints): $x_1 = t_1, \ldots, x_M = t_M$. Then $N_e(t) = \theta_k$ for $x_{k-1} \leq t \leq x_k$, $k = 1, \ldots, M$, and $N_e(t) = \theta_{M+1}$ for $t \geq x_M$. In our GLM framework, we model the effective population size on a given interval as a log-linear function of covariates

$$\gamma_k = \log \theta_k = \beta_1 Z_{k1} + \cdots + \beta_P Z_{kP} + w_k. \tag{5.13}$$

Here, we can impose temporal dependence by modeling $w = (w_1, \ldots, w_{M+1})$ as a zero-mean Gaussian process. Adopting this viewpoint, we propose the following GMRF smoothing prior on $\boldsymbol{\gamma}$:

$$P(\boldsymbol{\gamma}|\mathbf{Z}, \boldsymbol{\beta}, \tau) \propto \tau^{M/2} \exp\left[-\frac{\tau}{2}(\boldsymbol{\gamma} - \mathbf{Z}\boldsymbol{\beta})'\mathbf{Q}(\boldsymbol{\gamma} - \mathbf{Z}\boldsymbol{\beta})\right]. \tag{5.14}$$

In this prior, $\mathbf{Z}$ is an $(M+1)\times P$ matrix of covariates and $\boldsymbol{\beta}$ is a $P\times 1$ vector of coefficients representing the effect sizes for the predictors, quantifying their contribution to $\boldsymbol{\gamma}$. Precision $\mathbf{Q}$ is an $(M+1)\times(M+1)$ tri-diagonal matrix with off-diagonal elements equal to -1, $Q_{11} = Q_{M+1,M+1} = 1$, and $Q_{ii} = 2$ for $i = 2,\ldots,M$. Let $\boldsymbol{\gamma}_{-i}$ denote the vector obtained by excluding only the i^{th} component from vector $\boldsymbol{\gamma}$. Therefore, conditional on $\boldsymbol{\gamma}_{-i}$, γ_i depends only on its immediate neighbors. Let $\mathbf{Z}_i$ denote the i^{th} row of covariate matrix $\mathbf{Z}$. The individual components of γ have full conditionals

$$\begin{aligned}
\gamma_1|\boldsymbol{\gamma}_{-1} &\sim N\left(\mathbf{Z}_1'\boldsymbol{\beta} - \mathbf{Z}_2'\boldsymbol{\beta} + \gamma_2, \frac{1}{\tau}\right), && (5.15)\\
\gamma_i|\boldsymbol{\gamma}_{-i} &\sim N\left(\mathbf{Z}_i'\boldsymbol{\beta} + \frac{\gamma_{i-1}+\gamma_{i+1} - \mathbf{Z}_{i-1}'\boldsymbol{\beta} - \mathbf{Z}_{i+1}'\boldsymbol{\beta}}{2}, \frac{1}{2\tau}\right) && (5.16)\\
&\quad \text{for } i = 2,\ldots,M, \\
\gamma_{M+1}|\boldsymbol{\gamma}_{-(M+1)} &\sim N\left(\mathbf{Z}_{M+1}'\boldsymbol{\beta} - \mathbf{Z}_M'\boldsymbol{\beta} + \gamma_M, \frac{1}{\tau}\right). && (5.17)
\end{aligned}$$

As in the original Skygrid GMRF prior, the precision parameter τ governs the smoothness of the trajectory and is assigned a gamma prior

$$P(\tau) \propto \tau^{a-1}e^{-b\tau}. \tag{5.18}$$

To complete the model specification, we place a relatively uninformative multivariate normal prior $P(\boldsymbol{\beta})$ on the coefficients $\boldsymbol{\beta}$. This yields the posterior

$$P(\boldsymbol{\gamma},\boldsymbol{\beta},\tau|\mathbf{g},\mathbf{Z}) \propto P(\mathbf{g}|\boldsymbol{\gamma})P(\boldsymbol{\gamma}|\mathbf{Z},\boldsymbol{\beta},\tau)P(\boldsymbol{\beta})P(\tau). \tag{5.19}$$

5.2.4 Missing Covariate Data

It is important to have a mechanism for dealing with unobserved covariate values. This is particularly crucial because the population history timeline, which ranges from the most recent sampling time to the time of the MRCA, necessitates observations from a wide and *a priori* unknown time span. Let $\mathbf{Z}^{\text{obs}}$ denote the observed covariate values and $\mathbf{Z}^{\text{mis}}$ the missing covariate values, so that $\mathbf{Z} = (\mathbf{Z}^{\text{obs}}, \mathbf{Z}^{\text{mis}})$. The missing data can be treated as extra unknown parameters in a Bayesian model, and they can be estimated provided that there is a model that links them to the observed data and other model parameters. We have the factorization

$$P(\boldsymbol{\gamma}, \mathbf{Z}^{\text{mis}}|\mathbf{Z}^{\text{obs}}, \boldsymbol{\beta}, \tau) = P(\boldsymbol{\gamma}|\mathbf{Z}^{\text{obs}}, \mathbf{Z}^{\text{mis}}, \boldsymbol{\beta}, \tau)P(\mathbf{Z}^{\text{mis}}|\mathbf{Z}^{\text{obs}}, \boldsymbol{\beta}, \tau), \tag{5.20}$$

and the marginal density $P(\boldsymbol{\gamma}|\mathbf{Z}^{\text{obs}}, \boldsymbol{\beta}, \tau)$ can be recovered by integrating out the missing data. As a starting point, we assume a missing completely at random structure, meaning that the probability that a covariate value is missing is independent of observed trait values and other model parameters. For the priors on missing covariate values in (21), we can adopt uniform distributions over plausible ranges.

Alternatively, we can formulate a prior on the missing covariate data that makes use of the observed covariate values. Here, we focus on a common scenario where covariate j is observed at times $x_0, \ldots, x_K$ and unobserved at times $x_{K+1}, \ldots, x_M$. Thus, we can write $\mathbf{Z}_j^{\text{obs}} = (Z_{0j}, \ldots, Z_{Kj})'$ and $\mathbf{Z}_j^{\text{mis}} = (Z_{(K+1)j}, \ldots, Z_{Mj})'$. We model the joint distribution of the observed and missing covariate values as multivariate normal,

$$\begin{pmatrix} \mathbf{Z}_j^{\text{obs}} \\ \mathbf{Z}_j^{\text{mis}} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{21} & \mathbf{P}_{22} \end{pmatrix}^{-1} \right), \tag{5.21}$$

where

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{21} & \mathbf{P}_{22} \end{pmatrix} \tag{5.22}$$

is the precision matrix. To impose a correlation structure that enforces dependence between covariate values corresponding to adjacent times, we adopt a first-order random walk with full conditionals

$$Z_{0j}|Z_{-0j} \sim N\left(Z_{1j}, \frac{1}{\kappa}\right), \tag{5.23}$$

$$Z_{ij}|Z_{-ij} \sim N\left(\frac{Z_{(i-1)j} + Z_{(i+1)j}}{2}, \frac{1}{2\kappa}\right) \tag{5.24}$$

$$\text{for } i = 1, \ldots, M-1,$$

$$Z_{Mj}|Z_{-Mj} \sim N\left(Z_{(M-1)j}, \frac{1}{\kappa}\right). \tag{5.25}$$

Let $\mathbf{Z}^K$ denote a vector of dimension $M - K$ with every entry equal to Z_{Kj}. Then the distribution of missing covariate values conditional on observed covariate values is

$$P(\mathbf{Z}_j^{\text{mis}}|\mathbf{Z}_j^{\text{obs}}) \propto \kappa^{(M-K)/2} \exp\left(-\frac{\kappa}{2}(\mathbf{Z}_j^{\text{mis}} - \mathbf{Z}^K)'\mathbf{P}_{22}(\mathbf{Z}_j^{\text{mis}} - \mathbf{Z}^K)\right), \tag{5.26}$$

where

$$\mathbf{P}_{22} = \begin{pmatrix} -1 & 2 & -1 & & \\ & & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}. \tag{5.27}$$

5.2.5 Markov Chain Monte Carlo Sampling Scheme

We use MCMC sampling to approximate the posterior

$$P(\boldsymbol{\gamma}, \boldsymbol{\beta}, \tau | \mathbf{g}, \mathbf{Z}) \propto P(\mathbf{g}|\boldsymbol{\gamma})P(\boldsymbol{\gamma}|\mathbf{Z}, \boldsymbol{\beta}, \tau)P(\boldsymbol{\beta})P(\tau). \tag{5.28}$$

To sample $\boldsymbol{\gamma}$ and τ, we propose a fast-mixing, block-updating MCMC sampling scheme for GMRFs (Knorr-Held and Rue, 2002). Suppose we have current parameter values $(\boldsymbol{\gamma}^{(n)}, \tau^{(n)})$. First, consider the full conditional density

$$\begin{aligned} P(\boldsymbol{\gamma}|\mathbf{g}, \mathbf{Z}, \boldsymbol{\beta}, \tau) &\propto P(\mathbf{g}|\boldsymbol{\gamma})P(\boldsymbol{\gamma}|\mathbf{Z}, \boldsymbol{\beta}, \tau) \\ &\propto \exp\left[\sum_{k=1}^{M+1}(-\gamma_k c_k - SS_k e^{-\gamma_k})\right] \tau^{M/2} \exp\left[-\frac{\tau}{2}(\boldsymbol{\gamma} - \mathbf{Z}\boldsymbol{\beta})'\mathbf{Q}(\boldsymbol{\gamma} - \mathbf{Z}\boldsymbol{\beta})\right] \\ &= \tau^{M/2} \exp\left[-\frac{\tau}{2}(\boldsymbol{\gamma} - \mathbf{Z}\boldsymbol{\beta})'\mathbf{Q}(\boldsymbol{\gamma} - \mathbf{Z}\boldsymbol{\beta}) - \sum_{k=1}^{M+1}(\gamma_k c_k + SS_k e^{-\gamma_k})\right] \\ &= \tau^{M/2} \exp\left[-\frac{\tau}{2}\boldsymbol{\gamma}'\mathbf{Q}\boldsymbol{\gamma} + (\mathbf{Z}\boldsymbol{\beta})'\tau\mathbf{Q}\boldsymbol{\gamma} - \sum_{k=1}^{M+1}(\gamma_k c_k + SS_k e^{-\gamma_k})\right]. \end{aligned} \tag{5.29}$$

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

$$\begin{aligned} h_k(\gamma_k) &\approx h_k(\hat{\gamma_k}) + h_k'(\hat{\gamma_k})(\gamma_k - \hat{\gamma_k}) + \frac{1}{2}h_k''(\hat{\gamma_k})(\gamma_k - \hat{\gamma_k})^2 \\ &= SS_k e^{-\hat{\gamma_k}}\left(\frac{1}{2}\hat{\gamma_k}^2 + \hat{\gamma_k} + 1\right) \\ &\quad + \left[c_k - SS_k e^{-\hat{\gamma_k}} - SS_k e^{-\hat{\gamma_k}}\hat{\gamma_k}\right]\gamma_k \\ &\quad + \left[\frac{1}{2}SS_k e^{-\hat{\gamma_k}}\right]\gamma_k^2. \end{aligned} \tag{5.30}$$

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

$$\boldsymbol{\gamma}_{(n+1)} = \boldsymbol{\gamma}_{(n)} - [d^2 f(\boldsymbol{\gamma}_{(n)})]^{-1}(df(\boldsymbol{\gamma}_{(n)}))' \tag{5.31}$$

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

$$f(\boldsymbol{\gamma}) = -\frac{1}{2}\boldsymbol{\gamma}'\tau\mathbf{Q}\boldsymbol{\gamma} + (\mathbf{Z}\boldsymbol{\beta})'\tau\mathbf{Q}\boldsymbol{\gamma} - \sum_{k=1}^{M+1}(\gamma_k c_k + SS_k e^{-\gamma_k}) \tag{5.32}$$

with

$$df(\boldsymbol{\gamma}) = -\boldsymbol{\gamma}'\tau\mathbf{Q} + (\mathbf{Z}\boldsymbol{\beta})'\tau\mathbf{Q} - [c_1 - SS_1 e^{-\gamma_1}, ..., c_{M+1} - SS_{M+1}e^{-\gamma_{M+1}}] \tag{5.33}$$

and

$$d^2 f(\boldsymbol{\gamma}) = -\tau\mathbf{Q} - \text{diag}[SS_k e^{-\gamma_k}]. \tag{5.34}$$

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

$$\begin{aligned} P(\boldsymbol{\gamma}|\mathbf{g}, \mathbf{Z}, \boldsymbol{\beta}, \tau) \approx \tau^{M/2} \exp\Bigg[&-\frac{1}{2}\boldsymbol{\gamma}'[\tau\mathbf{Q} + \text{Diag}(SS_k e^{-\hat{\gamma_k}})]\boldsymbol{\gamma} + (\tau\mathbf{Q}\mathbf{Z}\boldsymbol{\beta})'\boldsymbol{\gamma} \\ &- \sum_{k=1}^{M+1}(c_k - SS_k e^{-\hat{\gamma_k}} - SS_k e^{-\hat{\gamma_k}}\hat{\gamma_k})\gamma_k\Bigg], \end{aligned} \tag{5.35}$$

where Diag$(\cdot)$ is a diagonal matrix.

Starting from current parameter values $(\boldsymbol{\gamma}^{(n)}, \tau^{(n)})$, we first generate a candidate value for the precision, $\tau^* = \tau^{(n)}f$, where f is drawn from a symmetric proposal distribution with density $P(f) \propto f + \frac{1}{f}$ defined on $[1/F, F]$. The tuning constant F controls the distance between the proposed and current values of the precision. Next, conditional on τ^*, we propose a new state $\boldsymbol{\gamma}^*$ using the Gaussian approximation (5.35) to the full conditional density $P(\gamma|\mathbf{g}, \mathbf{Z}, \boldsymbol{\beta}, \tau^*)$. In the final step, the candidate state $(\tau^*, \boldsymbol{\gamma}^*)$ is accepted or rejected according to the Metropolis-Hastings ratio (Metropolis et al., 1953; Hastings, 1970).

5.2.6 Genealogical Uncertainty

In our development thus far, we have assumed the genealogies $g_1, \ldots, g_m$ are known and fixed. However, in reality we observe sequence data rather than genealogies. It is possible to estimate genealogies beforehand from sequence data and then infer the effective population size from fixed genealogies. However, this ignores the uncertainty associated with phylogenetic reconstruction. Alternatively, we can jointly infer genealogies and population dynamics from sequence data by combining the estimation procedures into a single Bayesian framework.

We can think of the aligned sequence data $\mathbf{Y} = (Y_1, \ldots, Y_m)$ for the m loci as arising from continuous-time Markov chain (CTMC) models for molecular character substitution that act along the hidden genealogies. Each CTMC depends on a vector of mutational parameters Λ_i, that include, for example, an overall rate multiplier, relative exchange rates among characters and across-site variation specifications. We let $\boldsymbol{\Lambda} = (\Lambda_1, \ldots, \Lambda_m)$. We then jointly estimate the genealogies, mutational parameters, precision, and vector of effective population sizes through their posterior distribution

$$P(\mathbf{g}, \boldsymbol{\Lambda}, \tau, \boldsymbol{\gamma} | \mathbf{Y}) \propto \left[\prod_{i=1}^{m} P(Y_i | g_i, \Lambda_i) \right] P(\boldsymbol{\Lambda}) P(\mathbf{g} | \boldsymbol{\gamma}) P(\boldsymbol{\gamma} | \tau) P(\tau). \tag{5.36}$$

Here, the coalescent acts as a prior for the genealogies, and we assume that $\boldsymbol{\Lambda}$ and $\mathbf{g}$ are *a priori* independent of each other. Hierarchical models are however available to share information about $\boldsymbol{\Lambda}$ among loci without strictly enforcing that they follow the same evolutionary process (Suchard et al., 2003; Edo-Matas et al., 2011). We implement our models in the open-source software program BEAST (Drummond et al., 2012). The posterior distribution is approximated through MCMC methods. We combine our block-updating scheme for $\boldsymbol{\gamma}$ and τ with standard transition kernels available in BEAST to update the other parameters.

5.3 Results

5.3.1 Expansion in Epizootic Rabies Virus

Rabies is a zoonotic disease caused by the rabies virus, and is responsible for over 50,000 human deaths annually. In over 99% of human cases, the rabies virus is transmitted by dogs. However, there are a number of other important rabies reservoirs, such as bats and several terrestrial carnivore species, including raccoons (WHO, 2015b). Epizootic rabies among raccoons was first identified in the U.S. in Florida in the 1940s, and the affected area of the subsequent expansion was limited to the southeastern U.S. (Kappus et al., 1970). A second focus of rabies among raccoons emerged in West Virginia in the late 1970s due to the translocation of raccoons incubating rabies from the southeastern U.S. The virus spread rapidly along the mid-Atlantic coast and northeastern U.S. over the following decades, and is one of the largest documented outbreaks in the history of wildlife rabies (Childs et al., 2000).

Biek et al. (2007) examine the population dynamics of the rabies epizootic among raccoons in the northeastern U.S. starting in the late 1970s. In a spatiogenetic analysis, Biek et al. (2007) compare a coalescent-based Bayesian Skyline estimate (Drummond et al., 2005) of the demographic history to the spatial expansion of the epidemic. In a *post hoc* approach, the authors find very similar temporal dynamics between the effective population size and the 15-month moving average of the area (in square kilometers) of counties newly affected by the rabies outbreak each month. The effective population size exhibits stages of moderate and rapid growth, as well as plateau periods with little or no growth. Population expansion coincides with time periods during which the virus invades new area at a generally increasing rate. On the other hand, the effective population size shows little, if any, growth during periods when the virus invades new area at a declining rate. These trends can be seen in Figure 5.1, which depicts a Skygrid demographic reconstruction from sequence data along with the monthly area newly affected by the virus as a solid black line. Notably, Biek et al. (2007) demonstrate through their analysis that the largest contribution to the population expansion comes from the wave front, highlighting the degree to which the overall viral dynamics depend on processes at the wave front.

We build upon the analysis of Biek et al. (2007) by incorporating the spatiotemporal spread of rabies into the demographic inference model through the Skygrid. The data consist of 47 sequences with sampling dates between 1982 and 2004. As a covariate, we initially adopt the 15-month moving average of the log-transformed area of all counties newly affected by the raccoon rabies virus each month from 1977-1999 (Biek et al., 2007). We infer a posterior mean covariate effect size of 0.24 with a 95% Bayesian credibility interval (BCI) of (-0.77, 1.27), implying that there is not a significant association between the log effective population size and the covariate. This is not surprising, considering the patterns of growth and decline

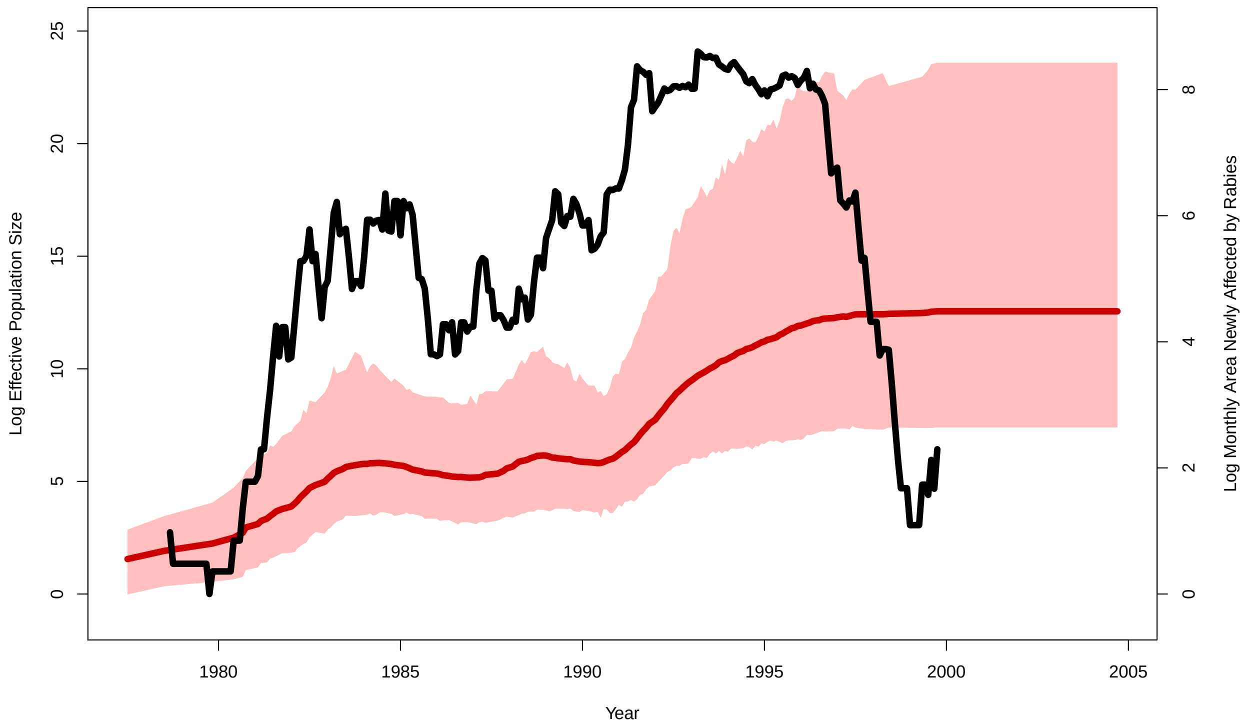

Figure 5.1: Demographic history of raccoon rabies epidemic in the northeastern United States. The black line represents the covariate, the 15-month moving average of the log–transformed area of all counties newly affected by the raccoon rabies virus each month. The solid red line is the posterior mean log effective population size trajectory estimated from sequence data. The 95% Bayesian credibility interval region for the trajectory is shaded in light red.

in the covariate compared with the essentially monotonic trend in the log effective population size (see Figure 5.1).

Graphically comparing the rate at which the virus invades new area with population dynamics clearly illustrates the relationship between the demographic and spatial expansion of the raccoon rabies outbreak. In modeling the association between the population dynamics and a covariate, however, we relate the covariate to the total effective population size (as opposed to the change in the effective population size). In this case, the cumulative affected area is a more suitable covariate than the newly affected area. We conduct an additional Skygrid analysis and use the log-transform of the cumulative area (in square kilometers) of counties affected by raccoon rabies at various time points between 1977 and 1999 as a covariate. 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 grid points to coincide with these change points.

The Skygrid analysis with the log cumulative affected area covariate yields a posterior mean estimate of 1.30 for the coefficient $\boldsymbol{\beta}$, with a 95% BCI of (0.18, 2.86), implying a significant, positive association between the effective population size of the raccoon rabies virus and the cumulative area affected by the outbreak. Skygrid demographic reconstructions of the epidemic are displayed in Figure 5.2. The red line is the posterior mean log effective population size trajectory inferred using the Skygrid without incorporation of the covariate. Its 95% BCI region is shaded in light red. The log effective population size trajectory from the Skygrid analysis with the log cumulative area covariate is represented by the blue line, and its 95% BCI region is shaded in light blue. The overlap between the two BCI regions is shaded purple. The log cumulative area affected by the rabies epidemic is depicted as a solid black line. Figure 5.2 shows great correspondence between the temporal dynamics of the demographic and spatial expansions. The period 1977-1984 is marked by a steady increase in effective population size and a rapid exponential increase in affected area. This is followed in the period 1984-1990 by a plateau in effective population size along with a much more modest rate of increase for the affected area. The affected area begins increasing at a greater rate around 1990 and then plateaus around 1996. Similary, 1990-1996 marks a stage

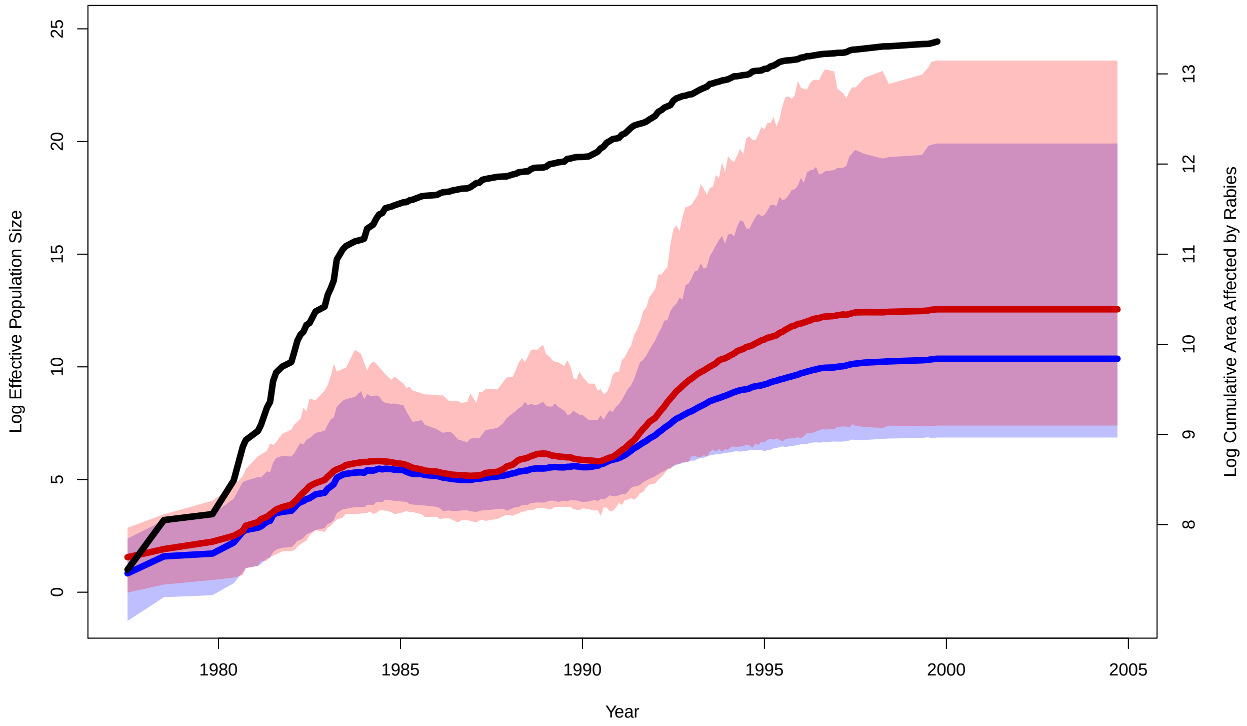

Figure 5.2: Demographic history of raccoon rabies epidemic in the northeastern United States. The black line represents the covariate, the log cumulative area of counties affected by raccoon rabies virus. The solid blue line is the posterior mean log effective population size trajectory from the Skygrid analysis with the covariate, and the solid red line is the posterior mean log effective population size trajectory from the Skygrid analysis without the covariate. The 95% Bayesian credibility interval regions for the trajectories are shaded in light blue for the analysis that includes the covariate and in light red for the analysis without the covariate. The overlap between the two Bayesian credibility interval regions is shaded in purple.

of demographic expansion, culminating in a period of stasis from 1996 on. The two effective population size curves are nearly identical from 1977-1990 and 1996-2004. From 1990-1996, both effective population size trajectories increase, but the rate of increase is more modest in the Skygrid estimate that incorporates the covariate data. Notably, the light blue BCI region inferred from the sequence and covariate data is narrower than and virtually entirely contained within the red BCI region inferred only from the sequence data. Thus including the covariate in this analysis not only yields an estimate consistent with what we infer from the sequence data alone, but also a more precise estimate.

5.3.2 Epidemic Dynamics in Dengue Evolution

Dengue is a mosquito-borne viral infection that causes a severe flu-like illness in which potentially lethal syndromes occasionally arise. Dengue is caused by the dengue virus, DENV, an RNA virus which comes in four antigenically distinct but closely related serotypes, DENV-1 through DENV-4. (WHO, 2015a). A recent estimate places the worldwide burden of dengue at 390 million infections per year (with 95% confidence interval 284-528 million), of which 96 million (67-136 million) manifest clinically (with any level of disease severity) (Bhatt et al., 2013). Dengue is found in tropical and sub-tropical climates throughout the world, mostly in urban and semi-urban areas (WHO, 2015a).

Dengue incidence records show patterns of periodicity with outbreaks every 3-5 years (Cummings et al., 2004; Adams et al., 2006; Bennett et al., 2010). Studies have shown that the epidemiological dynamics of dengue transmission in Puerto Rico are reflective of changes in the viral effective population size (Bennett et al., 2010; Carrington et al., 2005). Bennett et al. (2010) explore the dynamics of DENV-4 in Puerto Rico from 1981-1998. By *post hoc* comparing dengue isolate counts to effective population size estimates obtained using the Skyride model (Minin et al., 2008), Bennett et al. (2010) show that the pattern of cyclic epidemics is highly correlated with similar fluctuations in genetic diversity. We build upon their analysis by inferring the effective population size of DENV-4 in Puerto Rico with DENV-4 isolate counts as a covariate.

We analyze a data set of 75 DENV-4 sequences, compiled by Bennett et al. (2003) through sequencing randomly selected DENV-4 isolates from Puerto Rico from the U.S. Centers for Disease Control and Prevention (CDC) sample bank. The sampling dates include 1982 ($n = 14$), 1986/1987 ($n = 19$), 1992 ($n = 15$), 1994 ($n = 14$), and 1998 ($n = 13$). The covariate data consist of the number of DENV-4 isolates recorded over every 6-month period from 1981-1998. DENV-4 isolate counts are transformed via the map $x \mapsto \log(x + 1)$ (this specific logarithmic transformation is chosen to accommodate the transformation of isolate counts of zero).

Figure 5.3 presents the demographic and epidemiological patterns of DENV-4, with the top plot consisting of Skygrid effective population size estimates and the bottom showing a bar graph of transformed DENV-4 isolate counts. The blue and red curves in the top plot correspond to the posterior mean log effective population size trajectories from Skygrid analyses with and without the DENV-4 isolate count covariate, respectively. The 95% BCI regions for the trajectories are shaded in light blue for the analysis that includes the covariate and in light red for the analysis without the covariate. The overlap between the two BCI regions is shaded in purple. The demographic patterns are generally consistent with the isolate count fluctuations, and suggest a periodicity of 3-5 years. This concordance is supported by a positive, statistically significant estimate of the coefficient $\boldsymbol{\beta}$ relating the effective population size to isolate counts: a posterior mean of 0.90 with 95% BCI (0.36, 1.69).

While the two effective population size trajectories are similar, they do have some notable differences. The blue-colored trajectory inferred from both sequence and covariate data closely reflects the DENV-4 isolate count patterns, but the red-colored trajectory inferred entirely from sequence data diverges from the isolate count trends during certain periods. First, the red trajectory shows a dramatic increase in effective population size in 1981, consistent with a rise in DENV-4 isolates. However, the red trajectory decreases during 1982 while the DENV-4 isolate counts remain at a high level. Second, the period from late 1986 to late 1988 begins and ends with relative peaks in DENV-4 isolates, with a trough in between. By contrast, the red curve reaches a peak during the isolate trough and is on the

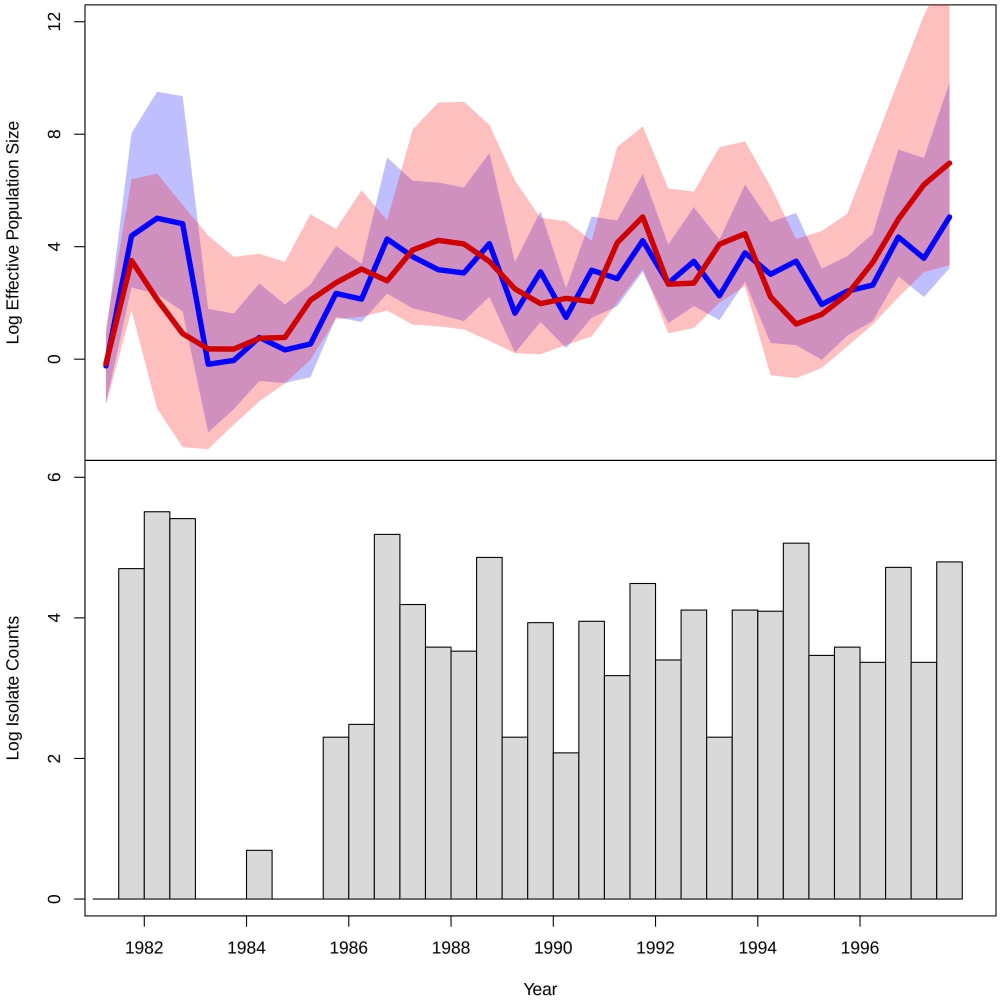

Figure 5.3: Population and epidemiological dynamics of DENV-4 virus in Puerto Rico. The top plot depicts Skygrid effective population size estimates. See Figure 5.2 for the legend explanation. The bars in the bottom plot represent DENV-4 isolate count covariate data.

decline during the late 1988 peak. Third, the red trajectory shows a trough in the effective population size during 1994 that occurs about a year before a similar trough in DENV-4 isolates. These discrepancies may be due to biased sampling in isolate counts and reflect limitations of epidemiological surveillance. Isolate counts are a rough measure of incidence, and their error rates are subject to accurate diagnostic rates by medical personnel, reporting rates, and the rate at which suspected cases are submitted for isolation (Bennett et al., 2010). On the other hand, the epidemiological trends are not necessarily incompatible with the effective population size trajectory estimated entirely from sequence data when the latter's uncertainty is taken into account. The blue-colored trajectory inferred from both sequence and isolate count data does not deviate from the isolate count data in the ways that the red trajectory does. However, the blue trajectory lies entirely inside the red 95% BCI region. Furthermore, apart from a 1.5 year period in 1981-82, the blue 95% BCI region is virtually entirely contained within, and is narrower than, the red 95% BCI region. Therefore, the Skygrid estimate that incorporates the DENV-4 isolate count covariate yields a demographic pattern that reflects epidemiological dynamics, and is more precise than, but not incompatible with, the effective population size estimate inferred only from sequence data.

5.3.3 Demographic History of the HIV-1 CRF02_AG Clade in Cameroon

Circulating recombinant forms (CRFs) are genomes that result from recombination of two or more different HIV-1 subtypes and that have been found in at least three epidemiologically unrelated individuals. Although CRF02_AG is globally responsible for only 7.7% of HIV infections (Hemelaar et al., 2011), it accounts for 60-70% of infections in Cameroon (Brennan et al., 2008; Powell et al., 2010).

We investigate the population history of the CRF02_AG clade in Cameroon by examining a multilocus alignment of 336 *gag*, *pol*, and *env* CRF02_AG gene sequences sampled between 1996 and 2004 from blood donors from Yaounde and Douala (Brennan et al., 2008). Faria et al. (2012) infer the effective population size from this data set with a parametric piecewise logistic growth-constant demographic model. Their results point to a period of exponential growth up until the mid 1990s, at which point the effective population size plateaus. Gill et al. (2013) follow up with a nonparameteric Skygrid analyis that reveals a monotonic growth in effective population size that peaks around 1997 and is then followed by a decline (rather than a plateau) that persists up until the most recent sampling time. We build upon these analyses by introducing two covariates: the yearly prevalence of HIV in Cameroon among adults ages 18-49, and the yearly HIV incidence rate in Cameroon among adults ages 18-49 (UNAIDS, 2015). UNAIDS prevalence and incidence estimates for Cameroon only go back to 1990, so we integrate out the missing covariate values as described in Section 5.2.4 by modeling the covariate values as a first-order random walk.

Figure 5.4 depicts the effective population size trajectory along with the HIV prevalence data. The prevalence increases up until 2000, stays constant for 4 years, and then declines slightly in 2004. The temporal pattern of the prevalence differs markedly from that of the demographic history, and this discordance is reflected in the GLM coefficient quantifying the prevalence effect size. The coefficient has a posterior mean of 0.85 with 95% BCI (-0.18, 2.03), indicating no significant association between the effective population size and prevalence.

The coefficient quantifying the effect size for the incidence rate covariate has a posterior mean of 9.20 with 95% BCI (1.43, 16.17), implying a significant association between the population history of the CRF02_AG clade and the HIV incidence rate among adults ages 18-49 in Cameroon. As shown in Figure 5.5, the effective population size and incidence rate display similar dynamics: both increase up until a peak around 1997, then decline. The posterior mean log effective population size and 95% BCI under the Skygrid model without covariates are virtually the same as the Skygrid estimates that incorporate the incidence data. This is in contrast to the previous examples we've seen, where inclusion of covariates affects effective population size estimates, and it may reflect the larger amount of sequence data relative to covariate data in this example. It is notable that in this example the effective population size is more reflective of incidence than prevalence. This is in accordance with expectations put forth by recent epidemiological modeling of infectious disease dynamics (Volz et al., 2009; Frost and Volz, 2010).

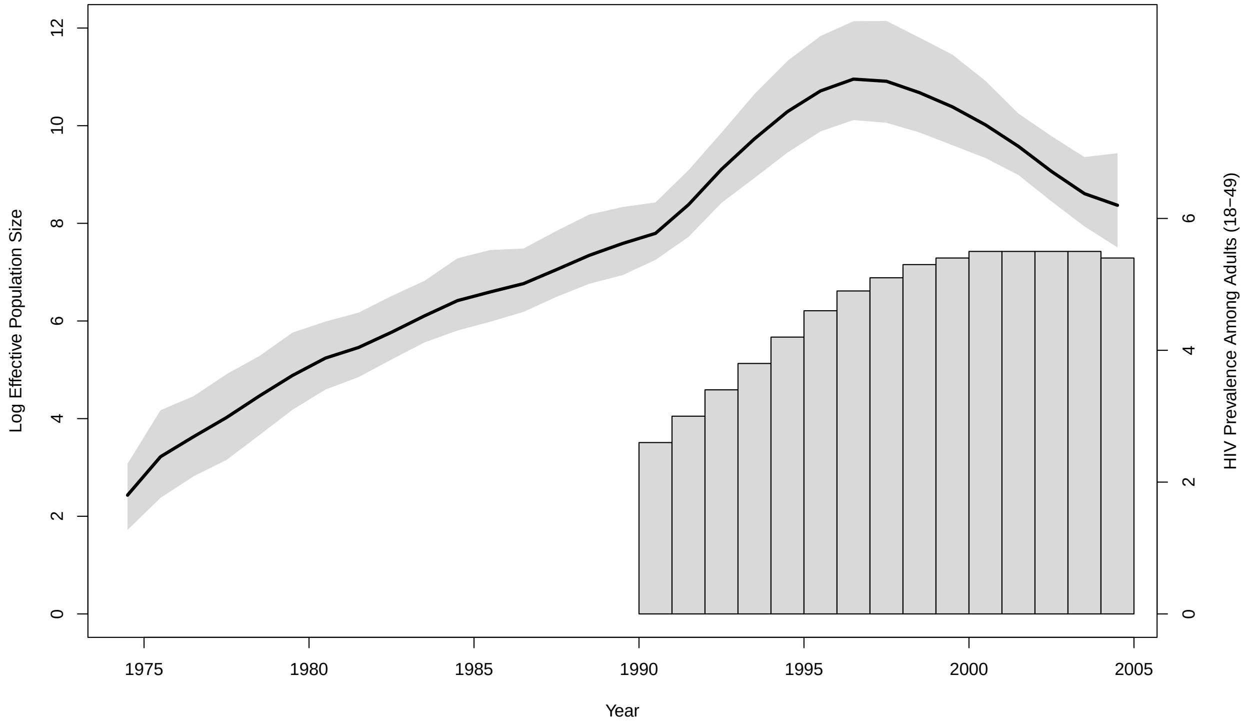

Figure 5.4: Demographic history of HIV-1 CRF02_AG clade in Cameroon. The solid black line is the posterior mean log effective population size trajectory, and its 95% Bayesian credibility interval region is shaded in gray. The bars represent HIV prevalence estimates for adults of ages 18-49 in Cameroon.

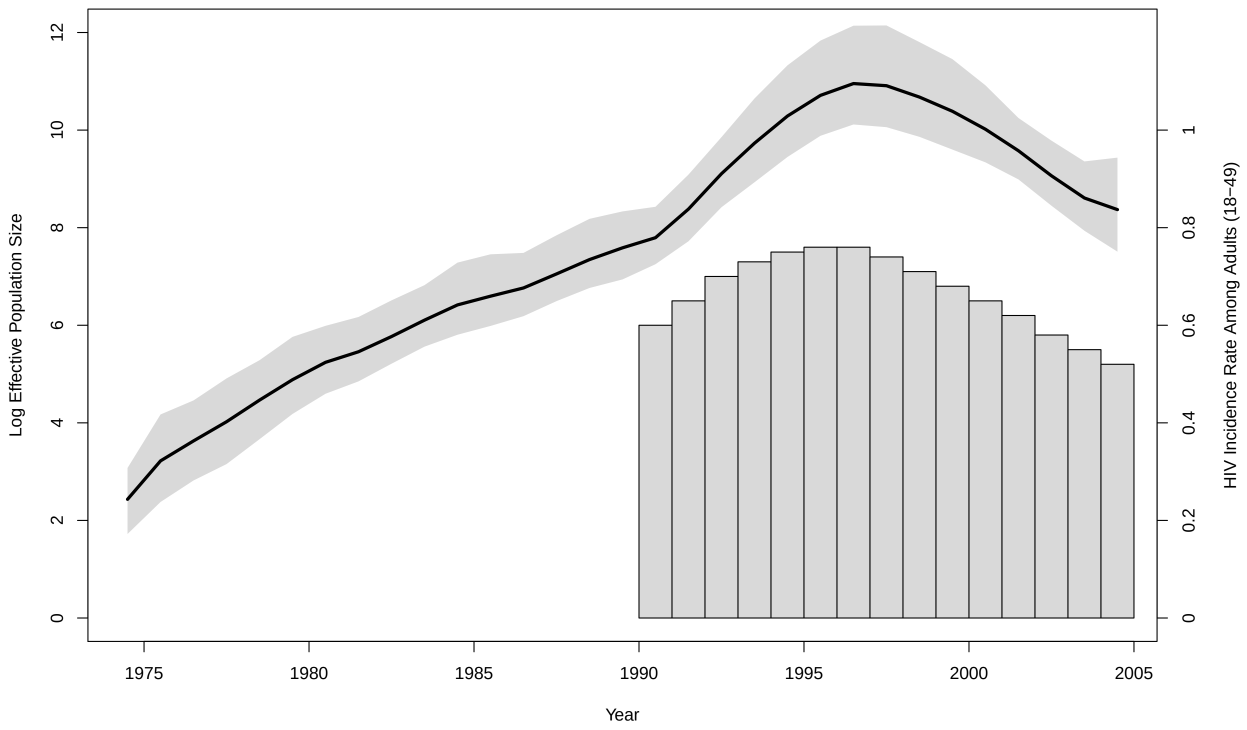

Figure 5.5: Demographic history of HIV-1 CRF02_AG clade in Cameroon. The solid black line is the posterior mean log effective population size trajectory, and its 95% Bayesian credibility interval region is shaded in gray. The bars represent HIV incidence rate estimates for adults of ages 18-49 in Cameroon.

5.3.4 Population Dynamics of Late Quaternary Musk Ox

Population decline and extinction of large-bodied mammals characterizes the Late Quaternary period (Barnosky et al., 2004; Lorenzen et al., 2011). The causes of these megafaunal extinctions remain poorly understood, and much of the debate revolves around the impact of climate change and humans (Stuart et al., 2004; Lorenzen et al., 2011). Demographic reconstructions from ancient DNA enable clarification of the roles of climatic and anthropogenic factors by providing a means to compare demographic patterns over geologically significant time scales with paleoclimatic and fossil records (Shapiro et al., 2004; Lorenzen et al., 2011).

Campos et al. (2010) employ the Skyride (Minin et al., 2008) and Bayesian Skyline (Drummond et al., 2005) models to reconstruct the population dynamics of musk ox dating back to the late Pleistocene era from ancient DNA sequences. The musk ox population was once widely distributed in the holarctic ecozone but is now confined to Greenland and the Arctic Archipelago, and Campos et al. (2010) explore potential causes of musk ox population decline. The authors find that the arrival of humans into relevant areas did not correspond to changes in musk ox effective population size. On the other hand, Campos et al. (2010) observe that time intervals during which musk ox populations increase generally correspond to periods of global climatic cooling, and musk ox populations decline during warmer and climatically unstable periods. Thus environmental change, as opposed to human presence, emerges as a more promising candidate as a driving force behind musk ox population dynamics.

We apply our extended Skygrid model to assess the relationship between the population history of musk ox and climate change. Oxygen isotope records serve as useful proxies for temperature in ancient climate studies. Here, we use ice core δ^{18}O data from the Greenland Ice Core Project (GRIP) (Johnsen et al., 1997; Dansgaard et al., 1993; GRIP Members, 1993; Grootes et al., 1993; Dansgaard et al., 1989). δ^{18}O is a measure of oxygen isotope composition. In the context of ice core data, lower δ^{18}O values correspond to colder polar temperatures. As a covariate, we adopt a mean δ^{18}O value, taking the average of δ^{18}O values corresponding to each 3,000-year interval. The sequence data consist of 682 bp of the mitochondrial control region, obtained from 149 radiocarbon dated specimens (Campos et al., 2010). The ages of the specimens range from the present to 56,900 radiocarbon (^{14}C) years before present (YBP). The sampling locations span the demographic range of ancient musk ox, with samples from the Taimyr Peninsula ($n = 54$), the Urals ($n = 26$), Northeast Siberia ($n = 12$), North America ($n = 14$) and Greenland ($n = 43$).

Figure 5.6 presents the posterior mean log effective population size trajectory (blue line) along with its 95% BCI region shaded in light blue. The δ^{18}O covariate values are represented by the red line. The plot shows a steady increase in effective population size up until about 60,000 YBP. During this period, the covariate pattern suggests a general trend of cooling, although there is considerable fluctuation in the mean δ^{18}O values. The effective population size plateaus from 60,000 to 55,000 YBP, then decreases from 55,000 to 40,000 YBP. During the aforementioned period of demographic decline, the covariate does not display any clear trends, fluctuating back and forth. The effective population size increases from about 40,000 to 25,000 YBP while the δ^{18}O covariate continues to fluctuate, although it undergoes a net decrease from the beginning to the end of the phase. Notably, the musk ox population reaches its peak diversity at around 25,000 YBP, coinciding with the Last Glacial Maximum. The period from 25,000 to 12,000 YBP is marked by a decrease in effective population size along with a postglacial warming. Finally, a demographic recovery and a very mild decline in the δ^{18}O covariate characterize the last 12,000 YBP. The patterns we note in Figure 5.6 are consistent with the observations of Campos et al. (2010). However, the GLM coefficient $\boldsymbol{\beta}$ has a posterior mean of -0.09 with a 95% BCI of (-0.50, 0.35), indicating that there is not a significant association between the log effective population size and the δ^{18}O covariate. This is not surprising upon further reflection. While the net changes in the covariate from the beginning to the end of the various monotonic phases of the population trajectory lend some support to the hypothesis of a negative relationship between the effective population size and δ^{18}O covariate, the pervasive covariate fluctuations render the relationship insignificant.

There are more than 5,000 δ^{18}O measurements in the GRIP data corresponding to different time points in the musk ox population history timeline. Our default approach is to specify Skygrid grid points so that the trajectory has as many piecewise constant segments

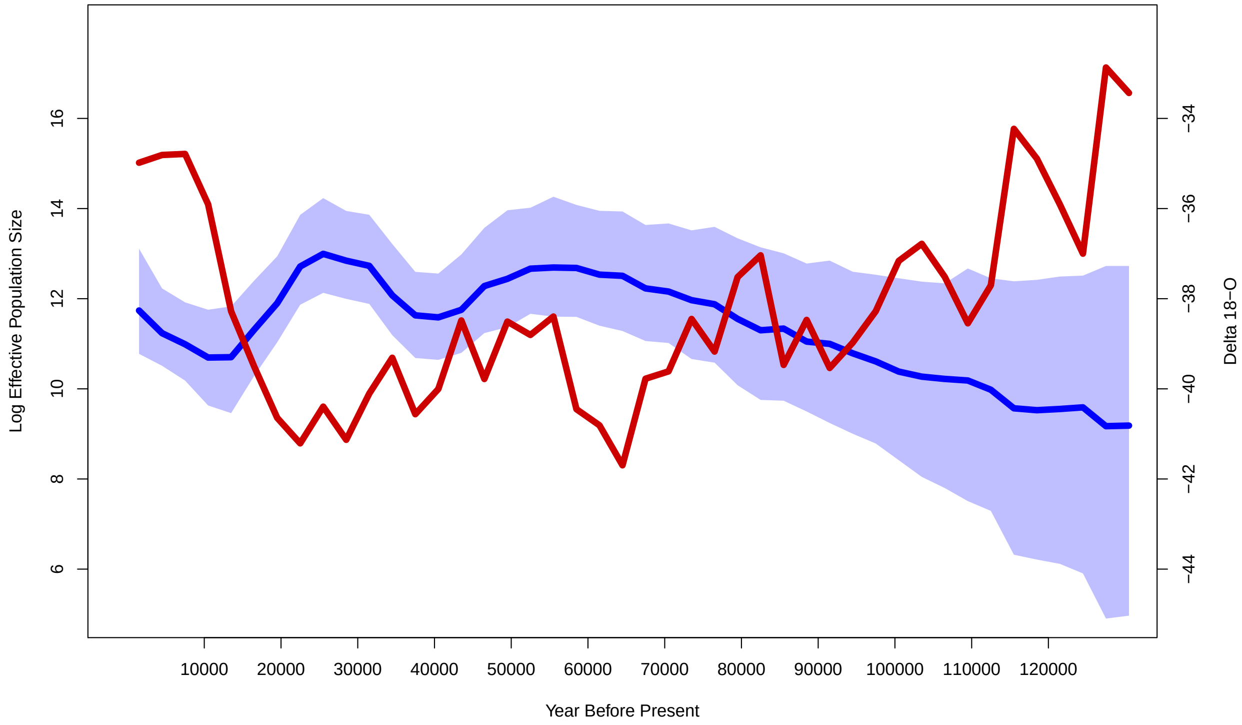

Figure 5.6: Demographic history of ancient musk ox. The axis is labelled according to years before present. The solid blue line is the posterior mean log effective population size trajectory, and its 95% Bayesian credibility interval region is shaded in light blue. The solid red line represents the δ^{18}O covariate. We do not infer a significant relationship between the effective population size and the covariate.

as there are covariate measurement times. To avoid having an inappropriately large number of change points, however, we've used the average of δ^{18}O values corresponding to each 3,000-year interval in the timeline as a covariate. Notably, adopting averages over intervals of lengths 1,000, 5,000, or 10,000 years as covariates yields the same basic outcome: the effect size of the covariate is not statistically significant.

While we do not infer a significant association between the log effective population size and δ^{18}O covariate values, this does not rule out climate change as a driving force behind musk ox population dynamics. The musk ox is known to be very sensitive to temperature and is not able to tolerate high summer temperatures (Tener, 1965). Using species distribution models, dated fossil remains and paleoclimatic data, Lorenzen et al. (2011) demonstrate a positive correlation between musk ox genetic diversity and its climate-driven range size over the last 50,000 years. The δ^{18}O data we use here do not account for geographic variability in temperature. Furthermore, we have not controlled for any potential confounders, such as range size or proportion of range overlap with humans. Nevertheless, our analysis serves as a precaution against oversimplification in the search for explanations of megafaunal population decline and extinctions. Incorporating additional covariate data into future studies may reveal a more complete, nuanced story of large mammal population dynamics during the Late Quaternary period.

5.4 Discussion

We present a novel coalescent-based Bayesian framework for estimation of effective population size dynamics from molecular sequence data and external covariates. We achieve this by extending the popular Skygrid model to incorporate covariates. In doing so, we retain the key elements of the Skygrid: a flexible, nonparametric demographic model, smoothing of the trajectory via a GMRF prior, and accommodation of sequence data from multiple genetic loci.

Effective population size is of fundamental interest in population genetics, infectious disease epidemiology, and conservation biology. It is crucial to identify explanatory factors, and to achieve a greater understanding of the association between the effective population size and such factors. In the context of viruses, it is important to assess the relationship between effective population size and epidemiological dynamics characterizing the number of infections and the spatiotemporal spread of an outbreak. Our extended Skygrid framework enables formal testing and characterization of such associations.

We showcase our methodology in four examples. Our analysis of the raccoon rabies epidemic in the northeastern United States uncovers striking similarities between the viral demographic expansion and the amount of area affected by the outbreak. We reconstruct a cyclic pattern for the effective population size of DENV-4 in Puerto Rico, coinciding with trends in viral isolate count data. Comparing the population history of the HIV-1 CRF02_AG clade in Cameroon with HIV incidence and prevalence data reveals a greater alignment with the HIV incidence rate than the prevalence rate. Finally, we consider the role of climate change in ancient musk ox population dynamics by using oxygen isotope data from the GRIP ice core as a proxy for temperature. We do not find a significant association, but our analysis demonstrates the need for a more thorough examination with additional covariates to follow up on previous investigations of the causes of ancient megafaunal population dynamics that consider a number of different factors.

Simultaneous inference of the effective population size and its association with covariates enables the uncertainty of the effective population size to be taken into account when assessing the association. Post hoc analyses comparing the mean effective population size trajectory with covariates (employing a standard linear regression approach, for example) are possible. However, such approaches may erroneously rule out significant associations by overemphasizing incompatibilities between the covariates and mean population trajectory. Furthermore, in the case of significant associations, regression coefficient estimates that disregard demographic uncertainty may have inflated precision.

Integrating covariates into the demographic inference framework not only enables testing and quantification of associations with the effective population size, it also provides additional information about past population dynamics. In two of our four examples, effective population size trajectories inferred from both sequence and covariate data differ markedly from trajectories inferred only from sequence data. In the rabies and dengue examples, the estimates based on sequence and covariate data are essentially consistent with with the estimates from the sequence data (in terms of the former having BCI regions almost entirely contained in the BCI regions of the latter), but more precise and more reflective of covariate trends.

Our extension of the Skygrid represents a first step toward a more complete understanding of past population dynamics, and the utility of the approach as demonstrated in the real data examples is promising. Our examples have only involved one or two covariates, but our implementation can support a large number of predictors. Furthermore, we plan to equip the Skygrid with efficient variable selection procedures to identify optimal subsets of predictors (George and McCulloch, 1993; Kuo and Mallick, 1998; Chipman et al., 2001). There is considerable potential for further development. For example, there is a prominent correspondence between spatial distribution and genetic diversity in the raccoon rabies example, and in previous studies of megafauna species (Lorenzen et al., 2011). We envision combining the Skygrid with phylogeographic inference models (Bloomquist et al., 2010) to simultaneously infer the wave-front of a population from sampling location data and use the wave-front as a predictor to model the effective population size. Attempts to infer associations between covariates and effective population size dynamics can be hampered by a scarcity of covariate data. Fortunately, there may exist measurements of the same covariates corresponding to different, but similar, genetic sequence data sets. We may, for example, have drug treatment data corresponding to several different HIV patients and wish to assess the relationship between the drug and intrahost HIV evolution. In such a setting, Bayesian hierarchical modeling could enable pooling of information from multiple data sets.

Acknowledgments

The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement no. 278433-PREDEMICS and ERC Grant agreement no. 260864 and the National Institutes of Health (R01 AI107034, R01 HG006139, R01 LM011827 and 5T32AI007370-24) and the National Science Foundation (DMS 1264153). R.B. was supported by NIH grant RO1 AI047498 and the RAPIDD programme of the Science and Technology Directorate of the Department of Homeland Security and NIH Fogarty International Centre.

CHAPTER 6

Future Work

6.1 Hierarchical Modeling of Past Population Dynamics

One of the Skygrid's major features is the ability to infer past population dynamics simultaneously from multilocus data. Combining data from several effectively unlinked genetic loci with the same demographic history can strikingly improve inference. We can also achieve gains by extending this principle of combining data to samples from different populations. Traditional modeling of demographic histories of multiple populations generally assumes that within-population demographic histories vary independently. However, there are numerous scenarios in which there are commonalities in the population dynamics of different populations. For example, within-host HIV population dynamics may be similar in different patients (Shriner et al., 2004; Lemey et al., 2006).

Data from multiple populations can be combined by enforcing strict equality of demographic histories of different populations, but such an assumption can be unrealistic and unnecessarily restrictive. We propose a hierarchical modeling approach to combine data and draw conclusions across multiple populations while still allowing different demographic histories for each population. Once again, the Skygrid framework is well-suited for our goal. The GRMF prior can be naturally extended to a hierarchical prior with an unknown estimable mean, and population-specific demographic histories can vary about this common mean. Formulating a hierarchical framework with a common "mean" trajectory is not straightforward when effective population size trajectories for different populations have different change-points. Fortunately, the Skygrid's user-specified change-point functionality enables exact alignment of trajectory change-points across populations.

6.1.1 Hierarchical Models for Multiply Observed Demographic Histories

Coalescent theory provides a striking connection between genealogical relationships and past population dynamics. Nevertheless, there remain significant challenges in inferring the effective population size as a function of time from a genealogy. A genealogy provides information about population dynamics through the times of its coalescent events. Therefore a dearth of coalescent events results in imprecise estimates of the demographic function. A genealogy relating n individuals has $n-1$ coalescent events, so the number of coalescent times in general can be increased by simply increasing the sample size n. However, even with large sample sizes, there are typically long stretches of time in the population history during which relatively few coalescent events occur. In a constant-size population, for example, coalescent events become increasingly rare as we move further back in time toward the MRCA. Increasing the sample size n will mitigate the problem to some extent, but most of the additional coalescent events occur during stretches of time where they are already plentiful. A more effective way to combat this problem is by incorporating data from additional effectively unlinked genetic loci that share the same demographic history. Samples corresponding to different loci are related by different genealogies, and increasing the number of loci provides additional coalescent events during times frames for which they are rare (Felsenstein, 2006; Heled and Drummond, 2008; Gill et al., 2013).

Inspired by the improvements from incorporating data from multiple loci into the inference framework, we seek to take advantage of the same philosophy in additional settings. There are a number of situations where samples are available from populations that exhibit similar, although not identical, demographic histories. Examples include HIV populations in different patients (Shriner et al., 2004; Lemey et al., 2006), and human influenza dynamics during different flu seasons (Minin et al., 2008). We pool information from different genealogies generated by similar demographic functions through hierarchical modeling. Hierarchical modeling provides a summary of the demographic history across populations while permitting the demographic history of each population to vary. This is accomplished by placing a hierarchical prior on the effective population size trajectory of each population. The hierarchical priors posit that the demographic functions of the different populations vary about a common unknown estimable mean demographic function.

Suppose we have samples from m different populations, and let $g_1, \ldots, g_m$ denote genealogies representing the ancestries of the m samples. We assume M grid points, $x_1, \ldots, x_M$, and we let $\boldsymbol{\gamma}_i = (\gamma_{i1}, \ldots, \gamma_{iM}, \gamma_{i(M+1)})$ denote the vector of log effective population sizes for population i. Importantly, we adopt the same grid points for each population. The likelihood $P(g_i|\boldsymbol{\gamma}_i)$ is constructed according to the Skygrid demographic model, as detailed in section 5.5.2. We can extend over $i = 1, \ldots, m$ populations simultaneously by assigning the $\boldsymbol{\gamma}_i$ the hierarchical prior

$$P(\boldsymbol{\gamma}_i|\boldsymbol{\gamma}, \tau) \propto \tau^{M/2} \exp\left[-\frac{\tau}{2}(\boldsymbol{\gamma}_i - \boldsymbol{\gamma})'\mathbf{Q}(\boldsymbol{\gamma}_i - \boldsymbol{\gamma})\right]. \tag{6.1}$$

Thus the effective population size across the different populations is captured in the mean $\boldsymbol{\gamma}$ of the prior. We assign $\boldsymbol{\gamma}$ a relatively uninformative multivariate-normal hyperprior.

6.1.2 Hierarchical Modeling for Multilocus Data

Our development of a hierarchical modeling approach thus far has been motivated by situations where we desire to combine information from genealogies that correspond to demographic histories that cannot be assumed to be identical. Population dynamics inference frameworks developed for multilocus data typically assume that the different loci share the same demographic history. This assumption is justified so long as the loci are unlinked. Hierarchical modeling may prove useful in the context of multilocus data by enabling detection of linkage if the inferred effective population size trajectories for different loci are divergent (Storz et al., 2002).

6.1.3 Hierarchical Modeling with Covariates

Attempts to infer associations between covariates and effective population size dynamics can be hampered by a scarcity of data. Fortunately, if there exist measurements of the same covariates $Z_1, \ldots, Z_P$ in different populations, hierarchical modeling may afford a solution.

Let $\mathbf{Z}^{(i)}$ denote the matrix of covariate measurements corresponding to population i. Also, let $\boldsymbol{\gamma}_i$ and $\boldsymbol{\beta}_i$ denote the vector of log effective population sizes and regression coefficients, respectively, for population i. Likelihoods $P(g_i|\boldsymbol{\gamma}_i)$ are constructed for each population as in 5.2.2, and the association between covariates and effective population size in each population is modeled as in 5.2.3 through priors:

$$P(\boldsymbol{\gamma}_i|\mathbf{Z}^{(i)},\boldsymbol{\beta}_i,\tau) \propto \tau^{M/2}\exp\left[-\frac{\tau}{2}(\boldsymbol{\gamma}_i-\mathbf{Z}^{(i)}\boldsymbol{\beta}_i)'\mathbf{Q}(\boldsymbol{\gamma}_i-\mathbf{Z}^{(i)}\boldsymbol{\beta}_i)\right]. \tag{6.2}$$

We assign each population-specific vector of regression coefficients $\boldsymbol{\beta}_i$ a hierarchical prior

$$\boldsymbol{\beta}_i|\sigma^2 \sim N(\boldsymbol{\beta},\sigma^2\mathbf{I}), \tag{6.3}$$

where $\mathbf{I}$ is the identity matrix. We place relatively uninformative hyperpriors on $\boldsymbol{\beta}$ and σ^2.

Bibliography

Adams B, Holmes E, Zhang C, Mammen M, Nimmannitya S, Kalayanarooj S, Boots M. 2006. Cross-protective immunity can account for the alternating epidemic pattern of dengue virus serotypes circulating in Bangkok. *Proceedings of the National Academy of Sciences*. 103:14234–14239.

Allicock O, Lemey P, Tatem A, Pybus O, Bennett S, Mueller B, Suchard M, Foster J, Rambaut A, Carrington C. 2012. Phylogeography and population dynamics of dengue viruses in the americas. *Molecular Biology and Evolution*. 29:1533–1543.

Aris-Brosou A, Yang Z. 2002. Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18s ribosomal RNA phylogeny. *Systematic Biology*. 51:703–714.

Atkinson Q, Gray R, Drummond A. 2008. mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. *Molecular Biology and Evolution*. 25:468–474.

Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV. 2012a. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol Biol Evol*. 29:2157–67.

Baele G, Li WLS, Drummond AJ, Suchard MA, Lemey P. 2012b. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Mol Biol Evol*. .

Barnosky A, Koch P, Feranec R, Wing S, Shabel A. 2004. Assessing the causes of Late Pleistocene extinctions on the continents. *Science*. 306:70–75.

Barouch D. 2008. Challenges in the development of an HIV-1 vaccine. *Nature*. 455:613–619.

Bartoszek K, Pienaar J, Mostad P, Andersson S, Hansen T. 2012. A phylogenetic comparative method for studying multivariate adaptation. *Journal of Theoretical Biology*. 314:204–215.

Bazin E, Glemin S, Galtier N. 2006. Population size does not influence mitochondrial genetic diversity in animals. *Science*. 312:570–572.

Bedford T, Suchard M, Lemey P, Dudas G, Gregory V, Hay A, McCauley J, Russell C, Smith D, Rambaut A. 2014. Integrating influenza antigenic dynamics with molecular evolution. *Elife*. 3:e01914.

Bennett S, Drummond A, Kapan D, Suchard M, Munoz-Jordan J, Pybus O, Holmes E, Gubler D. 2010. Epidemic dynamics revealed in dengue evolution. *Molecular Biology and Evolution*. 27:811–818.

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

Bhatt S, Gething P, Brady O, et al. (18 co-authors). 2013. The global distribution and burden of dengue. *Nature*. 496:504–507.

Biek R, Henderson J, Waller L, Rupprecht C, Real L. 2007. A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. *Proceedings of the National Academy of Sciences*. 104:7993–7998.

Bikandou B, Ndoundou-Nkodia M, Niama F, Ekwalanga M, Obengui O, Taty-Taty R, Parra H, Saragosti S. 2004. Genetic subtyping of gag and env regions of HIV type 1 isolates in the Republic of Congo. *AIDS Research and Human Retroviruses*. 20.

Bikandou B, Takehisa J, Mboudjeka I, et al. (18 co-authors). 2000. Genetic subtypes of HIV type 1 in Republic of Congo. *AIDS Research and Human Retroviruses*. 16:613–619.

Bloomquist EW, Lemey P, Suchard MA. 2010. Three roads diverged? routes to phylogeographic inference. *Trends in Ecology & Evolution*. 25:626–632.

Bouvin-Pley M, Morgand M, Moreau A, Jestin P, Simonnet C, Tran L, Goujard C, Meyer L, Barin F, Braiband M. 2013. Evidence for a continuous drift of the HIV-1 species towards higher resistance to neutralizing antibodies over the course of the epidemic. *PLoS Pathogens*. 9:e1003477.

Brennan C, Bodelle P, Coffey R, et al. (20 co-authors). 2008. The prevalence of diverse HIV-1 strains was stable in Cameroonian blood donors from 1996 to 2004. *Journal of Acquired Immune Deficiency Syndrome*. 49:432–439.

Brown R. 1828. A brief account of microscopial observations made in the months of June, July and August, 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. *Philos Mag.* 4:161–173.

Bunnik E, Euler Z, Welkers M, Boeser-Nunnink B, Grijsen M, Prins J, Schuitemaker H. 2010. Adaptation of HIV-1 envelope gp120 to humoral immunity at a population level. *Nature Medicine*. 16:995–997.

Burton D. 2002. Antibodies, viruses and vaccines. *Nature Reviews Immunology*. 2:706–713.

Burton D. 2004. HIV vaccine design and the neutralizing antibody problem. *Nature Immunology*. 5:233–236.

Butler M, King A. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *American Naturalist*. 164.

Campos P, Willerslev E, Sher A, et al. (20 co-authors). 2010. Ancient DNA analyses exclude humans as the driving force behind late plestocene musk ox (*Ovibos moschatus*) population dynamics. *Proceedings of the National Academy of Sciences*. 107:5675–5680.

Carrington C, Foster J, Pybus O, Bennett S, Holmes E. 2005. Invasion and maintenance of dengue virus type 2 and type 4 in the Americas. *Journal of Virology*. 79:14680–14687.

CDC. 2013. Centers for disease control and prevention, West Nile virus.

Childs J, Curns A, Dey M, Real L, Feinstein L, Bjornstad O. 2000. Predicting the local dynamics of epizootic rabies among raccoons in the United States. *Proceedings of the National Academy of Sciences*. 97:13666–13671.

Chipman H, George E, McCulloch R. 2001. The practical implementation of Bayesian model selection. *IMS Lecture Notes - Monograph Series*. 38:67–134.

Crandall K, Posada D, Vasco D. 1999. Effective population sizes: missing measures and missing concepts. *Animal Conservation*. 2:317–319.

Cummings D, Irizarry R, Huang N, Endy T, Nisalak A, Ungchusak K, Burke D. 2004. Travelling waves in the occurrence of dengue haemorrhagic fever in Thailand. *Nature*. 427:344–347.

Cybis G, Sinsheimer J, Bedford T, Mather A, Lemey P, Suchard M. 2015. Assessing phenotypic correlations through the multivariate phylogenetic latent liability model. *Annals of Applied Statistics*. 9:969–991.

Dansgaard W, Johnsen S, Clausen H, et al. (11 co-authors). 1993. Evidence for general instability of past climate from a 250 kyr ice-core record. *Nature*. 364:218–220.

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

Donnelly P, Tavaré S. 1995. Coalescents and genealogical structure under neutrality. *Annual Review of Genetics*. 29:401–421.

Drummond A, Ho S, Phillips M, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology*. 4:e88.

Drummond A, Nicholls G, Rodrigo A, Solomon W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*. 161:1307–1320.

Drummond A, Rambaut A, Shapiro B, Pybus O. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*. 22:1185–1192.

Drummond A, Suchard M. 2010. Bayesian random local clocks, or one rate to rule them all. *BMC Biology*. 8.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*. 29:1969–1973.

Edo-Matas D, Lemey P, Tom JA, Serna-Bolea C, van den Blink AE, van't Wout AB, Schuitemaker H, Suchard MA. 2011. Impact of CCR5delta32 host genetic background and disease progression on HIV-1 intrahost evolutionary processes: efficient hypothesis testing through hierarchical phylogenetic models. *Molecular Biology and Evolution*. 28:1605–16.

Euler Z, Bunnik E, Burger J, Boeser-Nunnink B, Grijsen M, Prins J, Schuitemaker H. 2011. Activity of broadly neutralizing antibodies, including PG9, PG16, and VRC01, against recently transmitted subtype B HIV-1 variants from early and late in the epidemic. *Journal of Virology*. 85:7236–7245.

Faria N, Rambaut A, Suchard M, et al. (14 co-authors). 2014. The early spread and epidemic ignition of HIV-1 in human populations. *Science*. 346:56–61.

Faria N, Suchard M, Abecasis A, Sousa J, Ndembi N, Bonfim I, Camacho R, Vandamme A, Lemey P. 2012. Phylodynamics of the HIV-1 CRF02_AG clade in Cameroon. *Infection, Genetics and Evolution*. 12:453–460.

Felsenstein J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics*. 25:471–492.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*. 13:93–104.

Felsenstein J. 1985a. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*. 39:783–791.

Felsenstein J. 1985b. Phylogenies and the comparative method. *The American Naturalist*. 125:1–15.

Felsenstein J. 2006. Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? *Mol Biol Evol.* 23:691–700.

Finlay E, Gaillard C, Vahidi S, Mirhoseini S, Jianlin H, Qi X, El-Barody M, Baird J, Healy B, Bradley D. 2007. Bayesian inference of population expansions in domestic bovines. *Biology Letters.* 3:449–452.

Freckleton R, Harvey P, Pagel M. 2002. Phylogenetic analysis and comparative data: A test and review of evidence. *The American Naturalist.* 160:712–726.

Frost S, Volz E. 2010. Viral phylodynamics and the search for an 'effective number of infections'. *Philosophical Transactions of the Royal Society B.* 365:1879–1890.

George E, McCulloch R. 1993. Variable selection via Gibbs sampling. *Journal of American Statistical Association.* 88:881–889.

Gill M, Lemey P, Faria N, Rambaut A, Shapiro B, Suchard M. 2013. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Molecular Biology and Evolution.* 30:713–724.

Grenfell B, Pybus O, Gog J, Wood J, Daly J, Mumford J, Holmes E. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science.* 303:327–332.

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

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, Jouzel J. 1993. Comparison of oxygen isotope records from the GISP2 and GRIP Greenland ice cores. *Nature.* 366:552–554.

Hansen T. 1997. Stabilizing selection and the comparative analysis of adaptation. *Evolution.* 52:1341–1351.

Harvey P, Pagel M. 1991. The comparative method in evolutionary biology. Oxford University Press.

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

Hasegawa M, Kishino H, Yano T. 1989. Estimation of branching dates among primates by molecular clocks of nuclear DNA which slowed down in Hominoidea. *Journal of Human Evolution*. 18:461–476.

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

Hein J, Schierup M, Wiuf C. 2005. Gene Genealogies, Variation and Evolution. USA: Oxford University Press.

Heled J, Drummond A. 2008. Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology*. 8:289.

Hemelaar J, Gouws E, Ghys PD, Osmanov S, WHO-UNAIDS Network for HIV Isolation and Characterisation. 2011. Global trends in molecular epidemiology of HIV-1 during 2000-2007. *AIDS*. 25:679–89.

Ho S, Shapiro B. 2011. Skyline-plot methods of estimating demographic history from nucleotide sequences. *Molecular Ecology Resources*. 11:423–434.

Hudson R. 1983. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*. 23:183–201.

Huelsenbeck J, Rannala B. 2003. Detecting correlation between characters in a comparative analysis with uncertain phylogeny. *Evolution*. 57:1237–1247.

Jeffreys H. 1935. Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society*. 31:203–222.

Jeffreys H. 1961. Theory of Probability. Oxford University Press.

Johnsen S, Clausen H, Dansgaard W, et al. (15 co-authors). 1997. The d18O record along the Greenland Ice Core Project deep ice core and the problem of possible Eemian climatic instability. *Journal of Geophysical Research*. 102:26397–26410.

Johnston M, Fauci A. 2007. An HIV vaccine: evolving concepts. *N. Engl. J. Med.* 356:2073–2080.

Jukes T, Cantor C. 1969. Mammalian Protein Metabolism, New York: Academic Press, chapter Evolution of protein molecules, pp. 21–123.

Kalish M, Robbins K, Pieniazek D, et al. (11 co-authors). 2004. Recombinant viruses and early global HIV-1 epidemic. *Emerging Infectious Diseases*. 10:1227–1234.

Kappus K, Bigler W, McLean R, Trevino H. 1970. The raccoon as an emerging rabies host. *Journal of Wildlife Diseases*. 6:507–509.

Kass R, Raftery A. 1995. Bayes factors. *Journal of the American Statistical Association*. 90:773–795.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*. 16:111–120.

Kingman J. 1982a. The coalescent. *Stochastic Processes and their Applications*. 13:235–248.

Kingman J. 1982b. On the genealogy of large populations. *Journal of Applied Probability*. 19:27–43.

Kita K, Ndembi N, Ekwalanga M, et al. (13 co-authors). 2004. Genetic diversity of HIV type 1 in Likasi, southeast of the Democratic Republic of Congo. *AIDS Research and Human Retroviruses*. 20:1352–1357.

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

Krone S, Neuhauser C. 1997. Ancestral processes with selection. *Theoretical Population Biology*. 51:210–237.

Kuhner M, Yamato J, Felsenstein J. 1998. Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*. 149:429–434.

Kuo L, Mallick B. 1998. Variable selection for regression models. *Sankhya B*. 60:65–81.

Lanciotti R, Roehrig J, Deubel V, et al. (24 co-authors). 1999. Origin of the West Nile virus responsible for an outbreak of encephalitis in the northeastern United States. *Science*. 286:2333–2337.

Lange K. 2002. Mathematical and Statistical Methods for Genetic Analysis. Statistics for Biology and Health. Springer, second edition.

Lartillot N, Poujol R. 2011. A phylogenetic model for investigating correlated evolution of subsitution rates and continuous phenotypic characters. *Molecular Biology and Evolution*. 28:729–744.

Lemey P, Pybus O, Wang B, Saksena N, Salemi M, Vandamme A. 2003. Tracing the origin and history of the HIV-2 epidemic. *Proceedings of the National Academy of Sciences*. 100:6588–6592.

Lemey P, Rambaut A, Drummond A, Suchard M. 2009a. Bayesian phylogeography finds its roots. *PLoS Computational Biology*. 5:e1000520.

Lemey P, Rambaut A, Pybus O. 2006. HIV evolutionary dynamics within and among hosts. *AIDS Reviews*. 8:125–140.

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

Lemey P, Suchard M, Rambaut A. 2009b. Reconstructing the initial global spread of a human influenza pandemic: a Bayesian spatial-temporal model for the global spread of H1N1. *PLoS Currents*. RRN1031.

Liu Y, Mittler J. 2008. Selection dramatically reduces effective population size in HIV-1 infection. *BMC Evolutionary Biology.* 8:133.

Lorenzen E, Nogues-Braco D, Orlando L, et al. (55 co-authors). 2011. Species-specific responses of Late Quaternary megafauna to climate and humans. *Nature.* 479:359–365.

Mascola J, Montefiori D. 2010. The role of antibodies in HIV vaccines. *Annual Review of Immunology.* 28:413–444.

Mello F, Araujo O, Lago B, Motta-Castro A, Moraes M, Gomes S, Bello G, Araujo N. 2013. Phylogeography and evolutionary history of hepatitis B virus genotype F in Brazil. *Virology Journal.* 10:236.

Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E. 1953. Equation of state calculation by fast computing machines. *Journal of Chemical Physics.* 21:1087–1092.

Minin V, Bloomquist E, Suchard M. 2008. Smooth skyride through a rough skyline: Bayesian coalescent based inference of population dynamics. *Molecular Biology and Evolution.* 25:1459–1471.

Niama F, Toure-Kane C, Vidal N, et al. (15 co-authors). 2006. HIV-1 subtypes and recombinants in the Republic of Congo. *Infection, Genetics and Evolution.* 6:337–343.

Notohara M. 1990. The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology.* 29:59–75.

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

Pagel M. 1999. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic Biology.* 48:612–622.

Palacios J, Minin V. 2013. Gaussian process-based Bayesian nonparametric inference of population size trajectories from gene genealogies. *Biometrics.* 69:8–18.

Palstra F, Fraser D. 2012. Effective/census population size ratio estimation: a compendium and appraisal. *Ecology and Evolution*. 2:2357–2365.

Powell R, Barengolts D, Mayr L, Nyambi P. 2010. The evolution of HIV-1 diversity in rural Cameroon and its implications in vaccine design and trials. *Viruses*. 2:639–654.

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

Pybus O, Suchard M, Lemey P, et al. (11 co-authors). 2012. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *PNAS*. 109:15066–71.

Rambaut A, Bromham L. 1998. Estimating divergence dates from molecular sequences. *Molecular Biology and Evolution*. 15:442–448.

Rambaut A, Pybus O, Nelson M, Viboud C, Taubenberger J, Holmes E. 2008. The genomic and epidemiological dynamics of human influenza A virus. *Nature*. 453:615–619.

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

Rodrigo A, Felsenstein J. 1999. Coalescent Approaches to HIV Population Genetics, Baltimore, MD: Johns Hopkins Universtiy Press, pp. 233–274.

Seetahal J, Velasco-Billa A, Allicock O, et al. (13 co-authors). 2013. Evolutionary history and phylogeography of rabies viruses associated with outbreaks in Trinidad. *PLoS Neglected Tropical Diseases*. 7:e2365.

Shapiro B, Drummond A, Rambaut A, et al. (27 co-authors). 2004. Rise and fall of the Beringian steppe bison. *Science*. 306:1561–1565.

Shriner D, Shankarappa R, Jansen M, Nickle D, Mittler J, Margolick J, Mullins J. 2004. Influence of random genetic drift on human immunodeficiency virus type 1 env evolution during chronic infection. *Genetics*. 166:1155–1164.

Sinsheimer J, Lake J, Little R. 1996. Bayesian hypothesis testing of four-taxon topologies using molecular sequence data. *Biometrics*. 52:193–210.

Slatkin M, Hudson R. 1991. Pairwise comparison of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*. 129:555–562.

Smith G, Vijaykrishna D, Bahl J, et al. (13 co-authors). 2009. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*. 459:1122–1125.

Stiller M, Baryshnikov G, Bocherens H, et al. (12 co-authors). 2010. Withering away-25,000 years of genetic decline preceded cave bear extinction. *Molecular Biology and Evolution*. 27:975–978.

Storz J, Beaumont M, Alberts S. 2002. Genetic evidence for long-term population decline in a savannah-dwelling primate: Inferences from a hierarchical bayesian model. *Molecular Biology and Evolution*. 19:1981–1990.

Strimmer K, Pybus O. 2001. Exploring the demographic history of DNA sequences using the generalized skyline plot. *Molecular Biology and Evolution*. 18:2298–2305.

Stuart A, Kosintsev P, Higham T, Lister A. 2004. Pleistocene to Holocene extinction dynamics in giant deer and wooly mammoth. *Nature*. 431:684–689.

Suchard M, Kitchen C, Sinsheimer J, Weiss R. 2003. Hierarchical phylogenetic models for analyzing multipartite sequence data. *Systematic Biology*. 52:649–664.

Suchard M, Weiss R, Sinsheimer J. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. *Molecular Biology and Evolution*. 18:1001–1013.

Suchard M, Weiss R, Sinsheimer J. 2005. Models for estimating Bayes factors with applications to phylogeny and tests of monophyly. *Biometrics*. 61:665–673.

Sullivan J, Swofford D. 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? *Systematic Biology*. 50:723–729.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*. 10:512–526.

Tavare S. 1986. Some probabilistic and statistical problems on the analysis of dna sequences. *Lectures on Mathematics in the Life Sciences*. 17:57–86.

Tener J. 1965. Muskoxen in Canada: a biological and taxonomic review. *Wildlife Service Monograph Series No. 2.* .

Thorne J, Kishino H, Painter I. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution*. 15:1647–1657.

UNAIDS. 2015. AIDSinfo. `http://aidsinfo.unaids.org/`.

UNAIDS/WHO. 2008. UNAIDS/WHO epidemiological fact sheets on HIV and AIDS, 2008 update. *UNAIDS/WHO.* .

Vidal N, Mulanga C, Bazepeo S, et al. (11 co-authors). 2005. Distribution of HIV-1 variants in the Democratic Republic of Congo suggests increase of subtype C in Kinshasa between 1997 and 2002. *Journal of Acquired Immune Deficiency Syndromes*. 40:456–462.

Vidal N, Peeters M, Mulanga-Kabeya C, Nzilambi N, Robertson D, Ilunga W, Sema H, Tshimanga K, Bongo B, Delaporte E. 2000. Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. *Journal of Virology*. 74:10498–10507.

Volz E, Pond SK, Ward M, Brown AL, Frost S. 2009. Phylodynamics of infectious disease epidemics. *Genetics*. 183:1421–1430.

Vrancken B, Lemey P, Rambaut A, Bedford T, Longdon B, Gunthard H, Suchard M. 2014. Simultaneously estimating evolutionary history and repeated traits phylogenetic signal: applications to viral and host phenotypic evolution. *Methods in Ecology and Evolution*. doi: 10.1111/2041-210X.12293.

Walker B, Burton D. 2008. Toward an AIDS vaccine. *Science*. 320:760–764.

Walker L, Phogat S, Chan-Hui P, et al. (22 co-authors). 2009. Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. *Science*. 326:285–289.

Weiss R, Clapham P, Cheingsong-Popov R, Dalgleish A, Carne C, Weller I, Tedder R. 1985. Neutralization of human T-lymphocyte virus type III by sera of AIDS and AIDS-risk patients. *Nature*. 316:69–72.

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

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

Wiener N. 1958. Nonlinear problems in random theory. Cambridge (MA): MIT Press.

Wright S. 1931. Evolution in Mendelian populations. *Genetics*. 16:97–159.

Yang C, Li M, Mokili J, et al. (11 co-authors). 2005. Genetic diversification and recombination of HIV type 1 group M in Kinshasa, Democratic Republic of Congo. *AIDS Research and Human Retroviruses*. 21:661–666.

Yang Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*. 10:13961401.

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

Yang Z. 1995. A space-time process model for the evolution of DNA sequences. *Genetics*. 139:993–1005.

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

Yang Z. 2006. Computational Molecular Evolution. Oxford Series in Ecology and Evolution. Oxford University Press.

Yoder A, Yang Z. 2000. Estimation of primate speciation dates using local molecular clocks. *Molecular Biology and Evolution*. 17:1081–1090.

Zhou T, Georgiev I, Wu X, et al. (18 co-authors). 2010. Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. *Science*. 329:811–817.