RH: Understanding Past Population Dynamics

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

Mandev S. Gill[1], Philippe Lemey[2], Shannon N. Bennett[3], Roman Biek[4] and Marc A. Suchard[5,6,7]

[1]*Department of Statistics, Columbia University, New York, United States*

[2]*Department of Microbiology and Immunology, Rega Institute, KU Leuven, Leuven, Belgium*

[3]*Department of Microbiology, California Academy of Sciences, San Francisco, United States*

[4]*Boyd Orr Centre for Population and Ecosystem Health, Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom*

[5]*Department of Biomathematics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, United States*

[6]*Department of Human Genetics, David Geffen School of Medicine at UCLA, Universtiy of California, Los Angeles, United States*

[7]*Department of Biostatistics, Jonathan and Karin Fielding School of Public Health, University of California, Los Angeles, United States*

Corresponding author: Marc A. Suchard, Departments of Biostatistics, Biomathematics, and Human Genetics, University of California, Los Angeles, 695 Charles E. Young Dr., South, Los Angeles, CA 90095-7088, USA; E-mail: `msuchard@ucla.edu`

Abstract.— Effective population size characterizes the genetic variability in a population and is a parameter of paramount importance in population genetics and evolutionary biology. Kingman's coalescent process enables inference of past population dynamics directly from molecular sequence data, and researchers have developed a number of flexible coalescent-based models for Bayesian nonparametric estimation of the effective population size as a function of time. Major goals of demographic reconstruction include identifying driving factors of effective population size, and understanding the association between the effective population size and such factors. Building upon Bayesian nonparametric coalescent-based approaches, we introduce a flexible framework that incorporates time-varying covariates that exploit Gaussian Markov random fields to achieve temporal smoothing of effective population size trajectories. To approximate the posterior distribution, we adapt efficient Markov chain Monte Carlo algorithms designed for highly structured Gaussian models. 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. 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.
(Keywords: Coalescent, Effective Population Size, Phylogenetics, Phylodynamics, Population Genetics, Gaussian Markov Random Fields)

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 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 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.

Methods

We begin with an overview of coalescent theory and follow with a detailed development of the Skygrid inference framework before presenting its extension that incorporates external covariate data. Readers interested in previewing our approach to include covariates may skip to the section *Incorporating Covariates*. However, we encourage readers who are unfamiliar with the Skygrid to proceed in order.

Coalescent Theory

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 the n individuals are sampled at two or more 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

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

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.

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:

$$P(\mathbf{g}|\boldsymbol{\theta}) = \prod_{i=1}^{m} P(g_i|\boldsymbol{\theta}). \tag{2}$$

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 $u_{i(\beta_i+1)} = [x_{\beta_i}, t_{\mathrm{MRCA}_i}]$. For each u_{ik} we let t_{kj}, $j = 1, \ldots, r_k$, denote the ordered times of the grid points and sampling and coalescent events in the interval. With each t_{kj} we associate an indicator 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

$$P(u_{ik}|\theta_k) = \prod_{1 \le j < r_k : \phi_{kj}=1} \frac{v_{kj}(v_{kj}-1)}{2\theta_k} \prod_{j=1}^{r_k-1} \exp\left[-\frac{v_{kj}(v_{kj}-1)(t_{k(j+1)} - t_{kj})}{2\theta_k}\right], \tag{3}$$

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

The product of interval likelihoods (3) yields the likelihood of coalescent times given the sampling times associated with genealogy g_i. However, identical coalescent times can arise from distinct genealogies. Immediately prior to a coalescent time $t_{k(j+1)}$, there are v_{kj} distinct lineages and therefore $\frac{v_{kj}(v_{kj}-1)}{2}$ different pairs of lineages that can merge and result in a coalescent event at time $t_{k(j+1)}$. The different possible mergings correspond to different genealogies. To obtain the likelihood of a particular genealogy we must account for the fact that a specific pair of lineages must merge at each coalescent time. 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

$$P(g_i|\boldsymbol{\theta}) = \prod_{k=\alpha_i}^{\beta_i+1} P_*(u_{ik}|\theta_k). \tag{4}$$

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

$$P(g_i|\boldsymbol{\theta}) = \prod_{k=\alpha_i}^{\beta_i+1} \left(\frac{1}{\theta_k}\right)^{c_{ik}} \exp\left[-\frac{SS_{ik}}{\theta_k}\right], \tag{5}$$

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

$$P(g_i|\boldsymbol{\gamma}) = \prod_{k=\alpha_i}^{\beta_i+1} e^{-\gamma_k c_{ik}} \exp[-SS_{ik}e^{-\gamma_k}] = \prod_{k=\alpha_i}^{\beta_i+1} \exp[-\gamma_k c_{ik} - SS_{ik}e^{-\gamma_k}]. \tag{6}$$

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

$$\begin{aligned} P(\mathbf{g}|\boldsymbol{\gamma}) &= \prod_{i=1}^{m} P(g_i|\boldsymbol{\gamma}) && (7)\\ &= \prod_{i=1}^{m}\prod_{k=\alpha_i}^{\beta_i+1} \exp[-\gamma_k c_{ik} - SS_{ik}e^{-\gamma_k}] && (8)\\ &= \exp\left[\sum_{k=1}^{M+1} \left[-\gamma_k c_k - SS_k e^{-\gamma_k}\right]\right] && (9) \end{aligned}$$

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 $\boldsymbol{\gamma}$:

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

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

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

In the 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

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

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, $s_1, \ldots, s_M, s_{M+1}$. Here, s_i denotes the units of time before the most recent sequence sampling time $s_0 = 0$, and $s_0 < s_1 < \cdots < s_M < s_{M+1}$. Alternatively, the covariate may correspond to time intervals $[s_0, s_1], \ldots, [s_{M-1}, s_M], [s_M, s_{M+1}]$ 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): $x_1 = s_1, \ldots, x_M = s_M$. 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

$$\gamma_k = \log \theta_k = \beta_1 Z_{k1} + \cdots + \beta_P Z_{kP} + w_k. \tag{13}$$

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 $\boldsymbol{\gamma}$:

$$P(\boldsymbol{\gamma}|\mathbf{Z}, \boldsymbol{\beta}, \tau) \propto \tau^{M/2} \exp\left[-\frac{\tau}{2}(\boldsymbol{\gamma} - \mathbf{Z}\boldsymbol{\beta})'\mathbf{Q}(\boldsymbol{\gamma} - \mathbf{Z}\boldsymbol{\beta})\right]. \tag{14}$$

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

$$\gamma_1|\boldsymbol{\gamma}_{-1} \sim N\left(\mathbf{Z}_1'\boldsymbol{\beta} - \mathbf{Z}_2'\boldsymbol{\beta} + \gamma_2, \frac{1}{\tau}\right), \tag{15}$$

$$\gamma_i|\boldsymbol{\gamma}_{-i} \sim N\left(\mathbf{Z}_i'\boldsymbol{\beta} + \frac{\gamma_{i-1} + \gamma_{i+1} - \mathbf{Z}_{i-1}'\boldsymbol{\beta} - \mathbf{Z}_{i+1}'\boldsymbol{\beta}}{2}, \frac{1}{2\tau}\right) \tag{16}$$

$$\text{for } i = 2, \ldots, M,$$

$$\gamma_{M+1}|\boldsymbol{\gamma}_{-(M+1)} \sim N\left(\mathbf{Z}_{M+1}'\boldsymbol{\beta} - \mathbf{Z}_M'\boldsymbol{\beta} + \gamma_M, \frac{1}{\tau}\right). \tag{17}$$

As in the original Skygrid GMRF prior, the precision parameter τ governs the smoothness

of the trajectory and is assigned a gamma prior

$$P(\tau) \propto \tau^{a-1} e^{-b\tau}. \tag{18}$$

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

$$P(\boldsymbol{\gamma}, \boldsymbol{\beta}, \tau | \mathbf{g}, \mathbf{Z}) \propto P(\mathbf{g}|\boldsymbol{\gamma}) P(\boldsymbol{\gamma}|\mathbf{Z}, \boldsymbol{\beta}, \tau) P(\boldsymbol{\beta}) P(\tau). \tag{19}$$

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

$$P(\boldsymbol{\gamma}, \mathbf{Z}^{\text{mis}} | \mathbf{Z}^{\text{obs}}, \boldsymbol{\beta}, \tau) = P(\boldsymbol{\gamma}|\mathbf{Z}^{\text{obs}}, \mathbf{Z}^{\text{mis}}, \boldsymbol{\beta}, \tau) P(\mathbf{Z}^{\text{mis}}|\mathbf{Z}^{\text{obs}}, \boldsymbol{\beta}, \tau), \tag{20}$$

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 covariate values and other model parameters. For the priors on missing covariate values in (20), 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 $x_1, \ldots, x_K$ and unobserved at times $x_{K+1}, \ldots, x_{M+1}$. Thus, we can write $\mathbf{Z}_j^{\text{obs}} = (Z_{1j}, \ldots, Z_{Kj})'$ and $\mathbf{Z}_j^{\text{mis}} = (Z_{(K+1)j}, \ldots, Z_{(M+1)j})'$. We model the joint distribution of the observed and missing covariate values as multivariate normal,

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

where

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

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

$$Z_{1j}|Z_{-1j} \sim N\left(Z_{2j}, \frac{1}{\kappa}\right), \tag{23}$$

$$Z_{ij}|Z_{-ij} \sim N\left(\frac{Z_{(i-1)j} + Z_{(i+1)j}}{2}, \frac{1}{2\kappa}\right) \tag{24}$$

$$\text{for } i = 2, \ldots, M,$$

$$Z_{(M+1)j}|Z_{-(M+1)j} \sim N\left(Z_{Mj}, \frac{1}{\kappa}\right). \tag{25}$$

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

$$P(\mathbf{Z}_j^{\text{mis}}|\mathbf{Z}_j^{\text{obs}}) \propto \kappa^{(M-K)/2} \exp\left(-\frac{\kappa}{2}(\mathbf{Z}_j^{\text{mis}} - \mathbf{Z}^K)'\mathbf{P}_{22}(\mathbf{Z}_j^{\text{mis}} - \mathbf{Z}^K)\right), \tag{26}$$

where

$$\mathbf{P}_{22} = \begin{pmatrix} -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}. \tag{27}$$

This technique of positing a random walk covariate distribution and recovering appropriate conditional distributions can also be employed for other missing data patterns.

Markov Chain Monte Carlo Sampling Scheme

We use MCMC sampling to approximate the posterior

$$P(\boldsymbol{\gamma}, \boldsymbol{\beta}, \tau | \mathbf{g}, \mathbf{Z}) \propto P(\mathbf{g}|\boldsymbol{\gamma}) P(\boldsymbol{\gamma}|\mathbf{Z}, \boldsymbol{\beta}, \tau) P(\boldsymbol{\beta}) P(\tau). \tag{28}$$

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

$$\begin{aligned} P(\boldsymbol{\gamma}|\mathbf{g}, \mathbf{Z}, \boldsymbol{\beta}, \tau) &\propto P(\mathbf{g}|\boldsymbol{\gamma}) P(\boldsymbol{\gamma}|\mathbf{Z}, \boldsymbol{\beta}, \tau) \\ &\propto \exp\left[\sum_{k=1}^{M+1} (-\gamma_k c_k - S S_k e^{-\gamma_k})\right] \tau^{M/2} \exp\left[-\frac{\tau}{2}(\boldsymbol{\gamma} - \mathbf{Z}\boldsymbol{\beta})' \mathbf{Q} (\boldsymbol{\gamma} - \mathbf{Z}\boldsymbol{\beta})\right] \\ &= \tau^{M/2} \exp\left[-\frac{\tau}{2}(\boldsymbol{\gamma} - \mathbf{Z}\boldsymbol{\beta})' \mathbf{Q} (\boldsymbol{\gamma} - \mathbf{Z}\boldsymbol{\beta}) - \sum_{k=1}^{M+1} (\gamma_k c_k + S S_k e^{-\gamma_k})\right] \\ &= \tau^{M/2} \exp\left[-\frac{\tau}{2}\boldsymbol{\gamma}' \mathbf{Q} \boldsymbol{\gamma} + (\mathbf{Z}\boldsymbol{\beta})' \tau \mathbf{Q} \boldsymbol{\gamma} - \sum_{k=1}^{M+1} (\gamma_k c_k + S S_k e^{-\gamma_k})\right]. \end{aligned} \tag{29}$$

Let $h_k(\gamma_k) = (\gamma_k c_k + S S_k e^{-\gamma_k})$. We can approximate each term $h_k(\gamma_k)$ by a second-order

Taylor expansion about, say, $\hat{\gamma}_k$:

$$\begin{aligned} h_k(\gamma_k) &\approx h_k(\hat{\gamma}_k) + h_k'(\hat{\gamma}_k)(\gamma_k - \hat{\gamma}_k) + \frac{1}{2}h_k''(\hat{\gamma}_k)(\gamma_k - \hat{\gamma}_k)^2 \\ &= SS_k e^{-\hat{\gamma}_k}\left(\frac{1}{2}\hat{\gamma}_k{}^2 + \hat{\gamma}_k + 1\right) \\ &\quad + \left[c_k - SS_k e^{-\hat{\gamma}_k} - SS_k e^{-\hat{\gamma}_k}\hat{\gamma}_k\right]\gamma_k \\ &\quad + \left[\frac{1}{2}SS_k e^{-\hat{\gamma}_k}\right]\gamma_k^2. \end{aligned} \tag{30}$$

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:

$$\boldsymbol{\gamma}_{(n+1)} = \boldsymbol{\gamma}_{(n)} - [d^2 f(\boldsymbol{\gamma}_{(n)})]^{-1}(df(\boldsymbol{\gamma}_{(n)}))' \tag{31}$$

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

$$f(\boldsymbol{\gamma}) = -\frac{1}{2}\boldsymbol{\gamma}'\tau\mathbf{Q}\boldsymbol{\gamma} + (\mathbf{Z}\boldsymbol{\beta})'\tau\mathbf{Q}\boldsymbol{\gamma} - \sum_{k=1}^{M+1}(\gamma_k c_k + SS_k e^{-\gamma_k}) \tag{32}$$

with

$$df(\boldsymbol{\gamma}) = -\boldsymbol{\gamma}'\tau\mathbf{Q} + (\mathbf{Z}\boldsymbol{\beta})'\tau\mathbf{Q} - [c_1 - SS_1 e^{-\gamma_1}, ..., c_{M+1} - SS_{M+1}e^{-\gamma_{M+1}}] \tag{33}$$

and

$$d^2 f(\boldsymbol{\gamma}) = -\tau\mathbf{Q} - \mathrm{diag}[SS_k e^{-\gamma_k}]. \tag{34}$$

Replacing the terms $h_k(\gamma_k)$ with their Taylor expansions yields the following second-order

Gaussian approximation to the full conditional density $P(\boldsymbol{\gamma}|\mathbf{g}, \mathbf{Z}, \boldsymbol{\beta}, \tau)$:

$$P(\boldsymbol{\gamma}|\mathbf{g}, \mathbf{Z}, \boldsymbol{\beta}, \tau) \approx \tau^{M/2} \exp\left[-\frac{1}{2}\boldsymbol{\gamma}'[\tau\mathbf{Q} + \text{Diag}(SS_k e^{-\hat{\gamma}_k})]\boldsymbol{\gamma} + (\tau\mathbf{Q}\mathbf{Z}\boldsymbol{\beta})'\boldsymbol{\gamma} - \sum_{k=1}^{M+1}(c_k - SS_k e^{-\hat{\gamma}_k} - SS_k e^{-\hat{\gamma}_k}\hat{\gamma}_k)\gamma_k\right], \quad (35)$$

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 (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).

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 $\mathbf{\Lambda} = (\Lambda_1, \ldots, \Lambda_m)$. We then jointly estimate the genealogies, mutational parameters, covariate effect size coefficients, precision, and vector of effective population sizes through their posterior distribution

$$P(\mathbf{g}, \mathbf{\Lambda}, \boldsymbol{\beta}, \tau, \boldsymbol{\gamma} | \mathbf{Y}, \mathbf{Z}) \propto \left[\prod_{i=1}^{m} P(Y_i | g_i, \Lambda_i) \right] P(\mathbf{\Lambda}) P(\mathbf{g} | \boldsymbol{\gamma}) P(\boldsymbol{\gamma} | \mathbf{Z}, \boldsymbol{\beta}, \tau) P(\boldsymbol{\beta}) P(\tau). \quad (36)$$

Here, the coalescent acts as a prior for the genealogies, and we assume that $\mathbf{\Lambda}$ and $\mathbf{g}$ are *a priori* independent of each other. Hierarchical models are however available to share information about $\mathbf{\Lambda}$ 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 v1 (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. The extended Skygrid model will be included in the next official release of BEAST v1. In the meantime, it can be accessed by users through the BEAST v1 development branch source code, which is available at `https://github.com/beast-dev/beast-mcmc/`. Example BEAST XML input files are available as part of the Supplementary Material online (`http://datadryad.org/resource/doi:10.5061/dryad.mj0hn`).

Empirical Examples

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 areas 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 areas at a declining rate. 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 observe the same trends in a Skygrid demographic reconstruction based on the Biek et al.

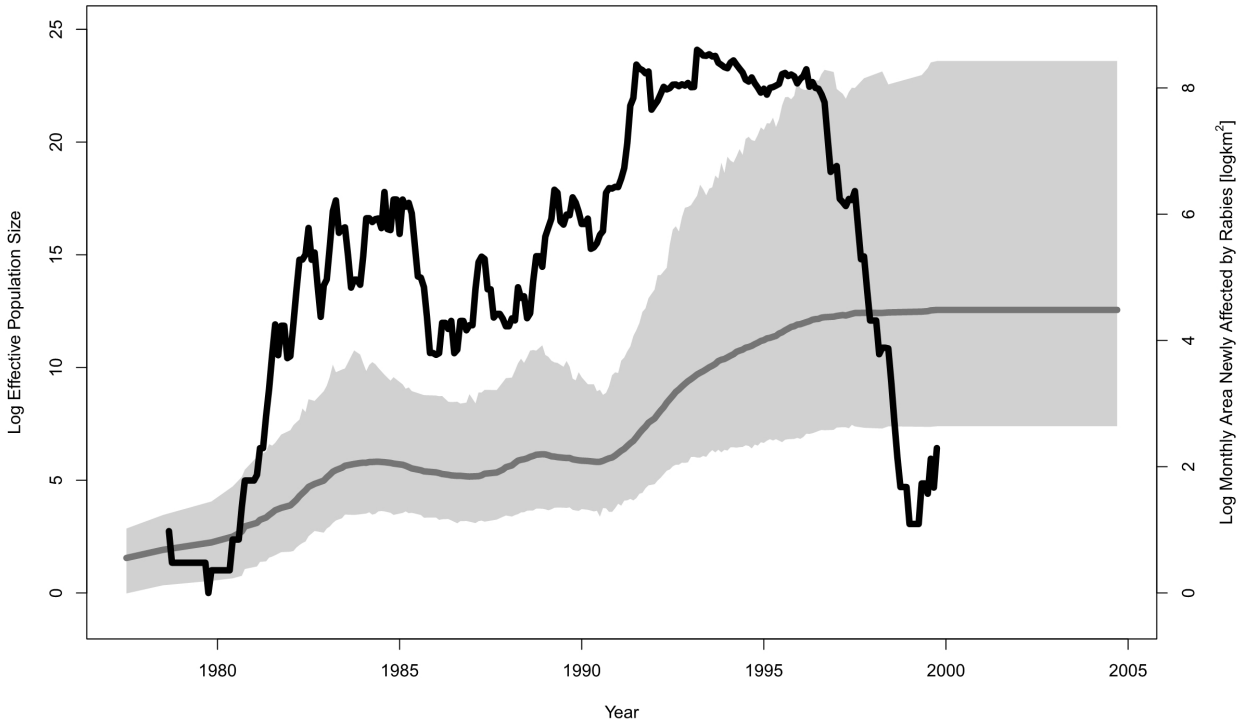

Figure 1: Skygrid demographic reconstruction of raccoon rabies epidemic in the northeastern United States. The gray line is the posterior mean log effective population size trajectory estimated only from sequence data without incorporating covariate data. The shaded gray region is the 95% Bayesian credibility interval region for the log effective population size. 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.

(2007) sequence data (Figure 1).

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 sequence data consist of 47 sequences sampled from rabid raccoons between 1982 and 2004. They encompass the complete rabies nucleoprotein (N) genes as well as large portions of the glycoprotein (G) genes. 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 in the covariate compared with the essentially monotonic trend in the log effective population size (see Figure 1).

Graphically comparing the rate at which the virus invades new areas 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 (Figure 2). Periods of demographic expansion are marked by relatively rapid rates of increase in the affected area, while plateaus in the effective population size coincide with more modest rates of increase in the affected area. The effective population size trajectory estimated from both sequence and covariate data displays nearly identical patterns to the trajectory estimated only from sequence data, except from 1990-1996, when its rate of increase is more modest.

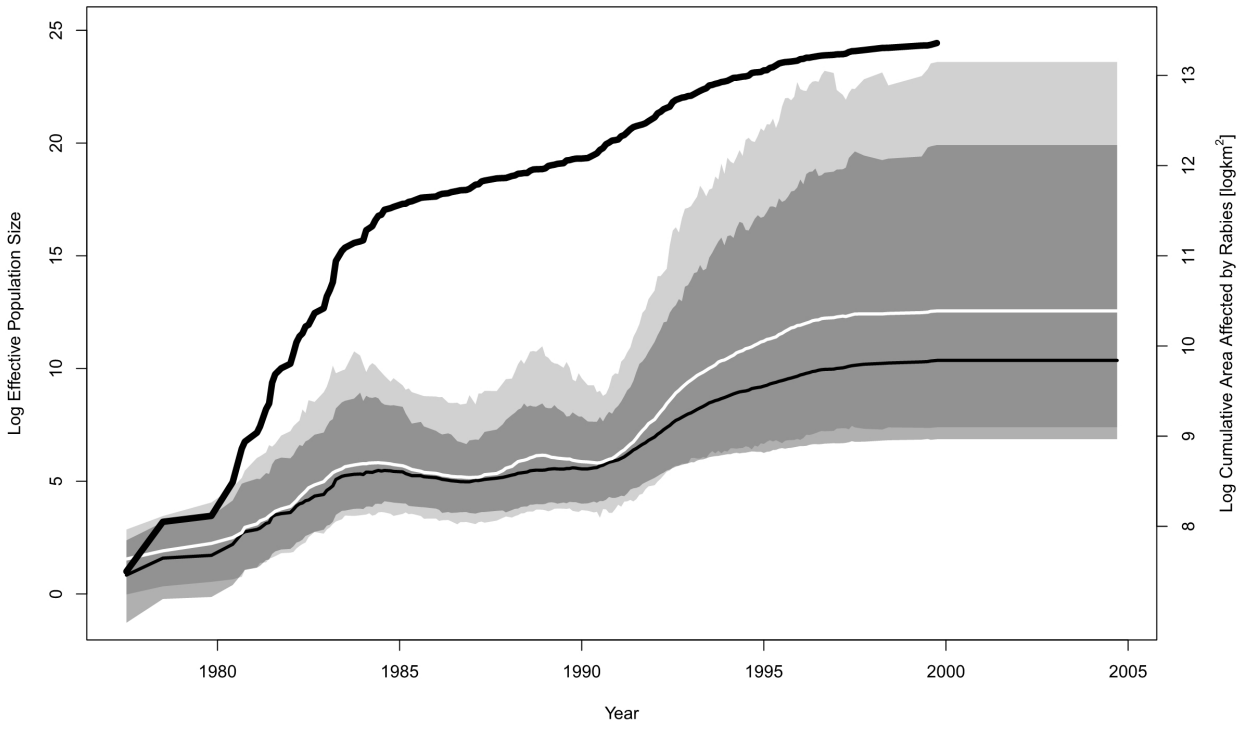

Figure 2: Demographic history of raccoon rabies epidemic in the northeastern United States. The black line that extends outside the shaded regions represents the covariate, the log cumulative area of counties affected by raccoon rabies virus. The black line contained within the shaded regions is the posterior mean log effective population size trajectory from the Skygrid analysis with the covariate, and the surrounding shaded dark gray region is its 95% Bayesian credibility interval (BCI) region. The white line is the posterior mean log effective population size trajectory from the Skygrid analysis without the covariate, and the surrounding shaded light gray region is its 95% BCI region. The two BCI regions overlap considerably, and the dark gray BCI region is almost entirely contained within the light gray BCI region.

Notably, the dark gray BCI region inferred from the sequence and covariate data is narrower than and virtually entirely contained within the light gray 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.

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 often 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.Each sequence contains gene regions amounting to 40% of the viral genome, including all structural genes (capsid: C; membrane: M; and envelope: E), a subset of nonstructural genes (NSI, $NS2A$, and $NS4B$), and the noncoding $3'$ NTR region. 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).

The patterns in the Skygrid demographic reconstructions are generally consistent with the isolate count fluctuations, and suggest a periodicity of 3-5 years (Figure 3). 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).

The effective population size trajectory inferred from both sequence and covariate data is similar to the trajectory estimated only from sequence data, but there are some notable differences. The black-colored estimate that incorporates covariate data closely reflects the DENV-4 isolate count patterns, but the white-colored trajectory inferred entirely from sequence data diverges from the isolate count trends during certain periods. First, the white trajectory shows a dramatic increase in effective population size in 1981, consistent with a rise in DENV-4 isolates. However, the white 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 white curve reaches a peak during the isolate trough and is on

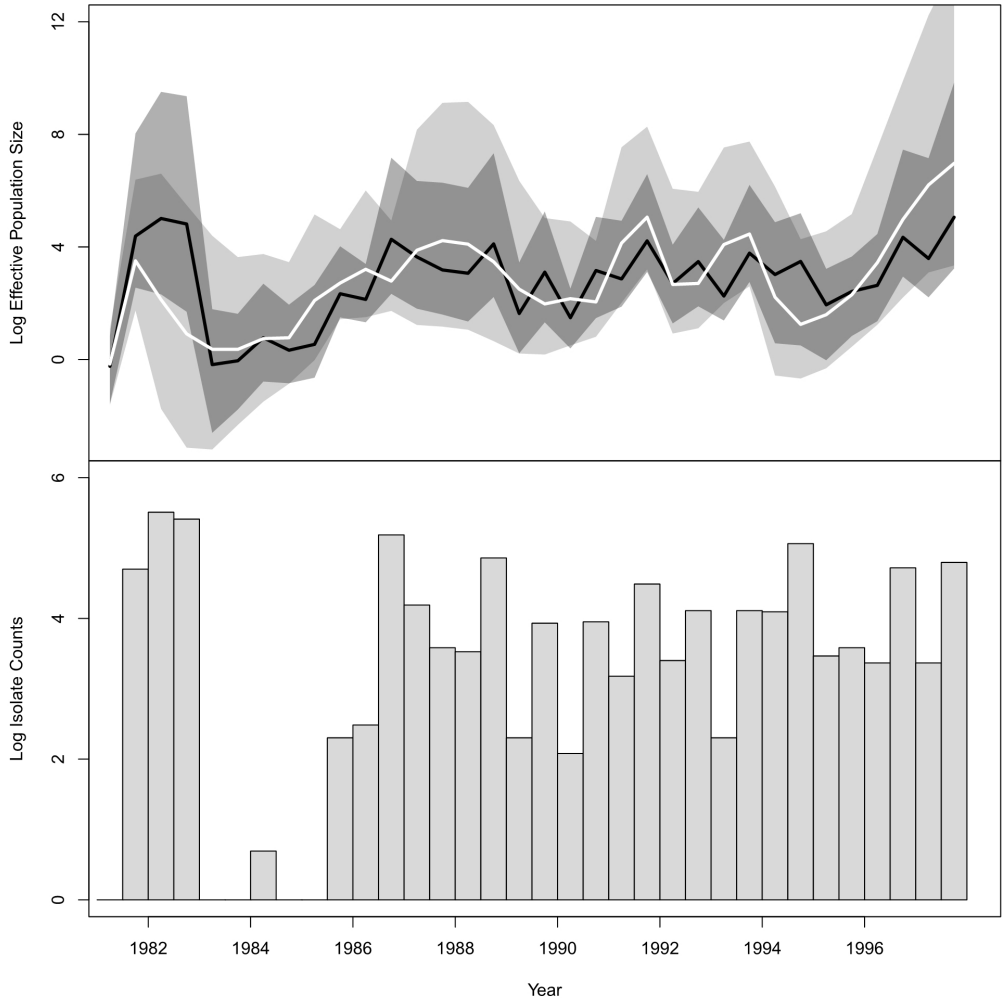

Figure 3: Population and epidemiological dynamics of DENV-4 virus in Puerto Rico. The top plot depicts Skygrid effective population size estimates. The black line is the posterior mean log effective population size trajectory from the Skygrid analysis with the covariate, and the surrounding shaded dark gray region is its 95% Bayesian credibility interval (BCI) region. The white line is the posterior mean log effective population size trajectory from the Skygrid analysis without the covariate, and the surrounding shaded light gray region is its 95% BCI region. The two BCI regions overlap considerably, and the dark gray BCI region is almost entirely contained within the light gray BCI region. The bars in the bottom plot represent DENV-4 isolate count covariate data.

the decline during the late 1988 peak. Third, the white 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 black-colored trajectory inferred from both sequence and isolate count data does not deviate from the isolate count data in the ways that the white trajectory does. However, the black trajectory lies entirely inside the light gray 95% BCI region. Furthermore, apart from a 1.5 year period in 1981-82, the dark gray 95% BCI region is virtually entirely contained within, and is narrower than, the light gray 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.

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 (26) by modeling the covariate values as a first-order random walk.

The HIV prevalence increases up until 2000, stays constant for 4 years, and then declines slightly in 2004. This differs markedly from the effective population size temporal pattern (Figure 4), 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. The effective population size and incidence rate display similar dynamics: both increase up until a peak around 1997, then decline (Figure 5). 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

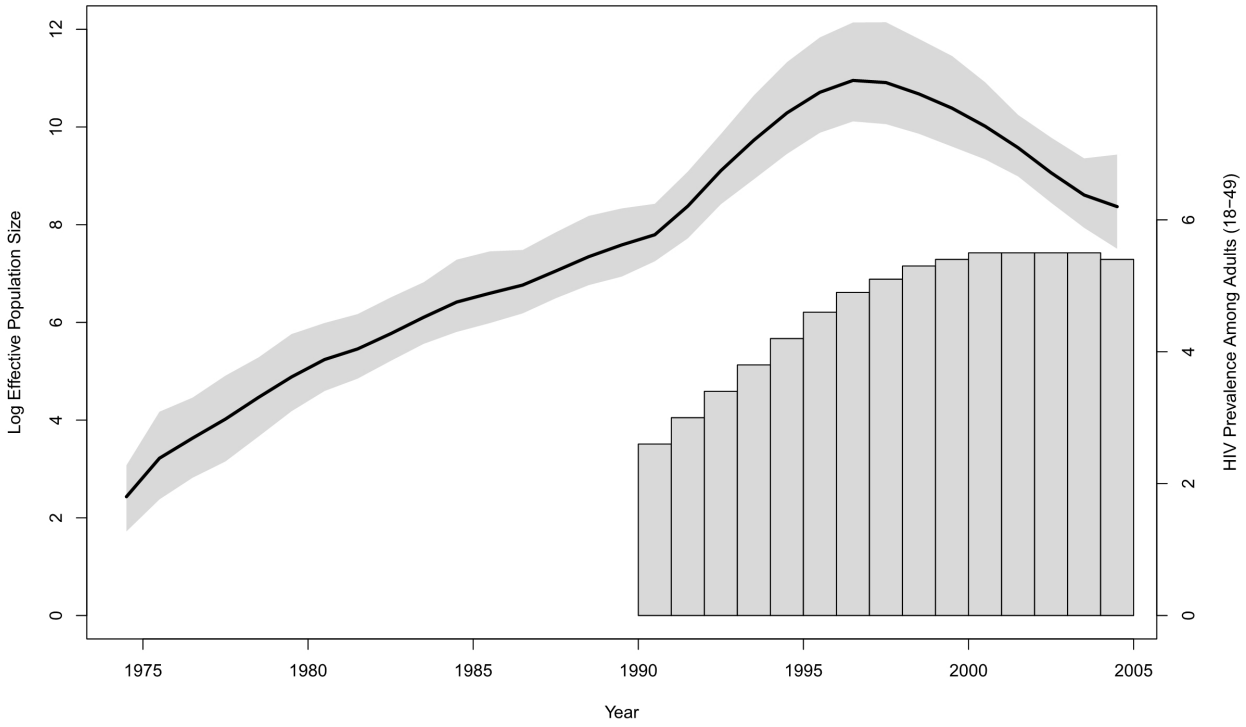

Figure 4: Demographic history of HIV-1 CRF02_AG clade in Cameroon. The black line is the posterior mean log effective population size trajectory, and its 95% Bayesian credibility interval region is shaded in gray. The bars represent HIV prevalence estimates for adults of ages 18-49 in Cameroon.

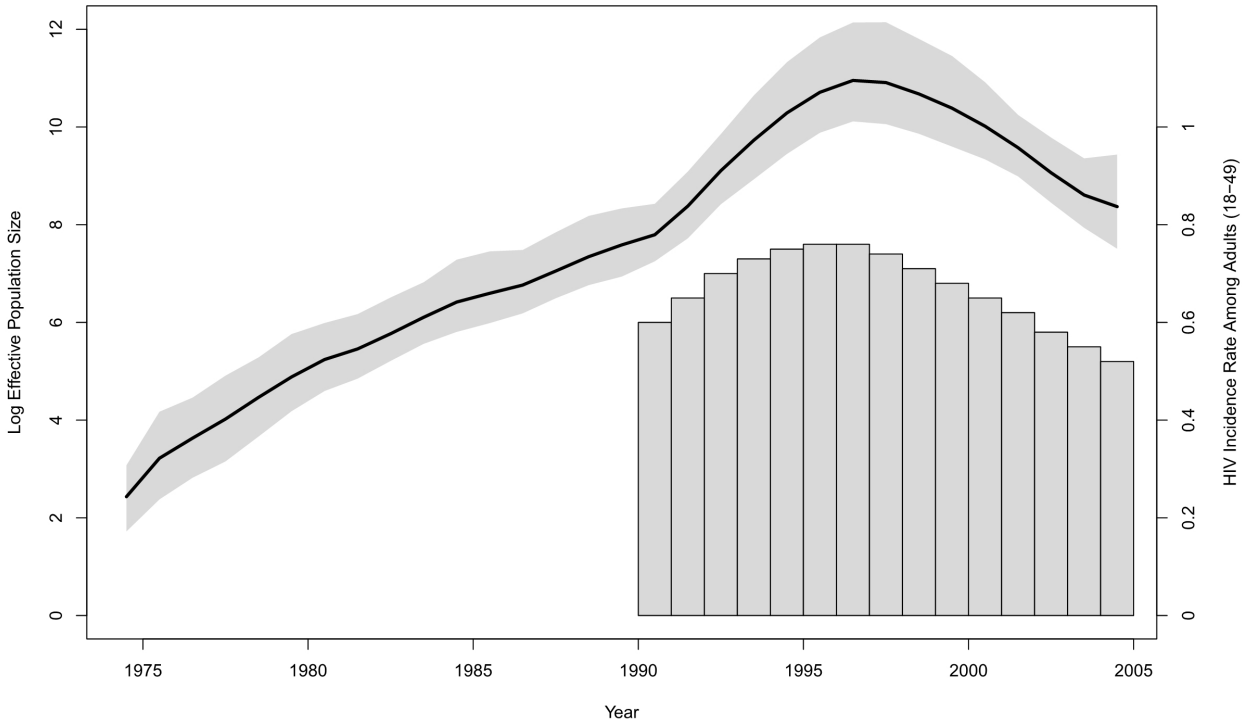

Figure 5: Demographic history of HIV-1 CRF02_AG clade in Cameroon. The black line is the posterior mean log effective population size trajectory, and its 95% Bayesian credibility interval region is shaded in gray. The bars represent HIV incidence rate estimates for adults of ages 18-49 in Cameroon.

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).

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 ($n = 43$).

During each time period that coincides with a monotonically increasing effective population size, the δ^{18}O covariate undergoes a net decrease (Figure 6), which suggests a general trend of cooling. On the other hand, periods of monotonic demographic decline coincide with either a covariate increase (indicative of a warming climate) or covariate fluctuations without any clear trends (suggesting climatic instability). These patterns are consistent with the observations of Campos et al. (2010). However, the covariate effect size 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. The net change in the covariate from the beginning to the end of each monotonic phase of the population trajectory lends some support to the hypothesis of a negative relationship between the effective population size

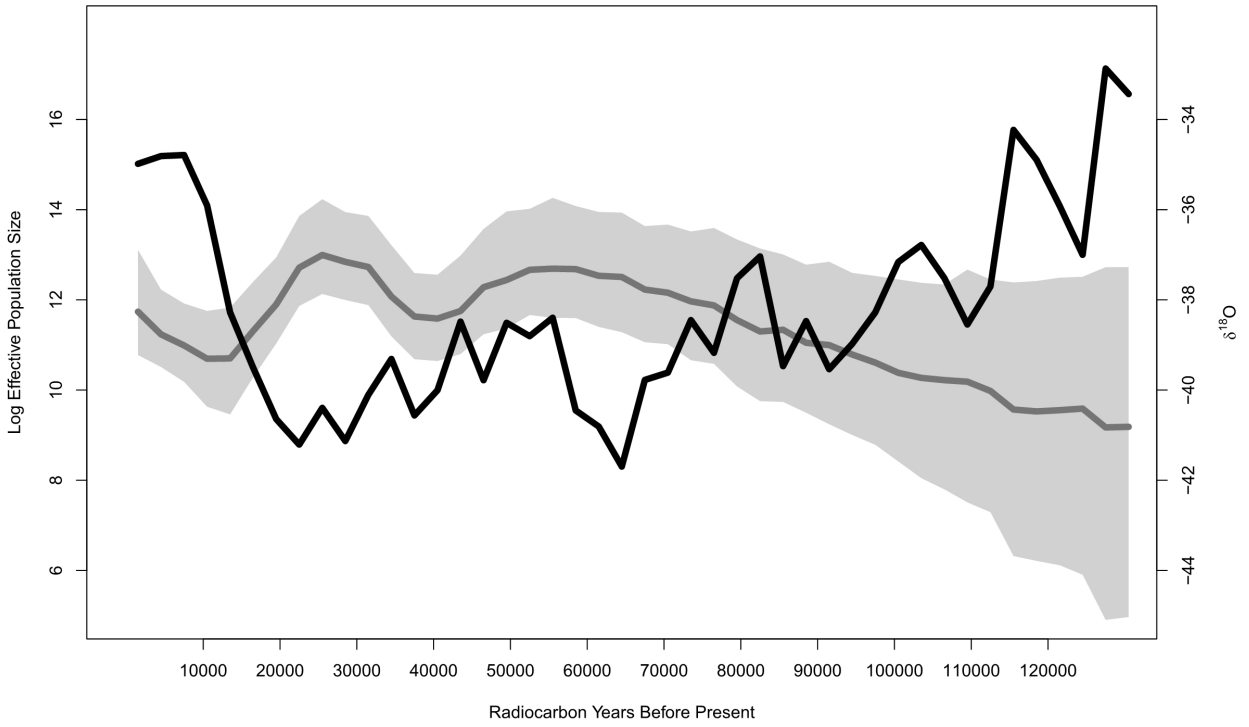

Figure 6: Demographic history of ancient musk ox. The axis is labelled according to radiocarbon years before present. The gray line is the posterior mean log effective population size trajectory, and its 95% Bayesian credibility interval region is shaded in light gray. The black line represents the δ^{18}O covariate. We do not infer a significant relationship between the effective population size and the covariate.

and the δ^{18}O covariate. However, there are numerous fluctuations in the covariate value during most of the aforementioned phases that 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 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 population structure, range size or proportion of range overlap with humans. If significant population structure exists, then appropriate geographic coverage of the sampling will also be important. 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. Finally, the sequence data in our analysis consist entirely of mitochondrial DNA. Including data from additional genetic loci may enhance our understanding of musk ox demographic history and provide some clarification.

Performance and Mixing

To confirm sufficient mixing within MCMC chains in our empirical examples, we monitor effective sample size (ESS) estimates of model parameters and adopt chain lengths that yield ESS estimates greater than 200 for the effective population size, precision, and covariate effect size parameters. We summarize performance in terms of ESS per minute (Table 1). Furthermore, we demonstrate the improvement in mixing by reporting the fold-increase in ESS per minute that the block-updating MCMC algorithm affords over

more basic Metropolis-Hastings transition kernels. The block-updating scheme exploits the structure of the GMRF smoothing prior. Under the more basic approach, we consider a random walk transition kernel for effective population size parameters that proposes new values by adding a random value within a specified window size to the current parameter value. For the precision, we generate candidate values by multiplying the current parameter value by a random scaling factor drawn from a specified window size. The block-updating algorithm consistently outperforms the random walk and rescaling transition kernels. Notably, the MCMC chain generated under the more basic transition kernels fails to generate sufficient ESS after 100 million iterations in the case of the rabies example. All analyses were conducted on a 2.7 GHz Intel Core i5 processor with 8 GB of RAM.

Table 1: Mixing of model parameters in terms of effective sample size (ESS) estimates per minute and fold-improvement in mixing due to a block-updating MCMC algorithm. For effective population size parameters, we report min-max range of ESS per minute. Fold-improvement due to block-updating is relative to more basic transition kernels.

	ESS per min.			Fold-improvement	
Example	Eff. pop. size	Precision	Effect size	Eff. pop. size	Precision
Rabies	12.6 - 53.0	35.7	33.1	165.6 - 252.0 ×	649.7 ×
Dengue	2.2 - 36.2	16.7	2.0	3.5 - 4.2 ×	22.9 ×
HIV	0.3 - 4.4	4.3	1.1	1.2 - 2.4 ×	4.2 ×
Musk ox	5.1 - 66.7	19.1	13.0	1.6 - 3.3 ×	5.3 ×

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 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.

It is possible that, in the presence of a statistically significant association between a covariate and the effective population size, the demographic trajectory estimated from sequence and covariate data will exhibit patterns inconsistent with the estimate based strictly on sequence data during a portion of the evolutionary history. This prospect raises concerns that a strong association between a covariate and the effective population size during one time period could cause the demographic history to be poorly estimated during another time period. However, such a scenario will correspond to one of two situations. First, the inconsistency between the two demographic reconstructions occurs for a relatively brief period of time. Second, the inconsistency occurs during a period for which the sequence data provide relatively little information about the population dynamics. Importantly, adding covariates to the model will not distort an originally precise demographic estimate. In our analysis of HIV population dynamics in Cameroon, for example, there is a strong association between the prevalence covariate and demographic history up until the late 1990s that nevertheless does not yield a significant effect size. The

sequence data are highly informative about the population dynamics during the early 2000s and do not allow for a significant effect size, which would result in a demographic estimate that diverges from the sequence data-based estimate during this period. In general, we recommend performing a sensitivity analysis by estimating the effective population size both with and without covariates and taking note of the duration and nature of inconsistencies between the two estimates. Also, Bayes factors (Jeffreys 1935, 1961) can be employed to formally compare the fit of different Skygrid models to observed data $(\mathbf{Y}, \mathbf{Z})$. A Bayes factor quantifies the evidence in favor of model M_1 over model M_0 by taking the ratio of marginal likelihoods:

$$BF_{10} = \frac{P(\mathbf{Y}, \mathbf{Z}|M_1)}{P(\mathbf{Y}, \mathbf{Z}|M_0)} = \frac{P(M_1|\mathbf{Y}, \mathbf{Z})}{P(M_0|\mathbf{Y}, \mathbf{Z})} \Big/ \frac{P(M_1)}{P(M_0)} \,. \tag{37}$$

The more general Skygrid model that incorporates covariates includes the more basic Skygrid model as the special case where the effect size $\boldsymbol{\beta} = 0$, affording straightforward computation of Bayes factors.

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 relevant measures of a population's

geographic distribution from sampling location data and use them as predictors to model the effective population size. Such approaches would need to rely on appropriate sampling not only through time, but also through geographic space. 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. Finally, it may be fruitful to develop inference frameworks similar to the Skygrid that are based on generalized coalescent models that incorporate population structure (Notohara 1990), recombination (Hudson 1983), and selection (Krone and Neuhauser 1997) to account for different reproductive phenomena and model their associations with external covariates.

Supplementary Material

BEAST XML input files for the data analysis are available from the Dryad digital repository: `http://datadryad.org/resource/doi:10.5061/dryad.mj0hn`.

Funding

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

748 programme of the Science and Technology Directorate of the Department of Homeland
749 Security and NIH Fogarty International Centre.

750 Acknowledgments

751 We would like to thank the editors, Frank Anderson and Laura Kubatko, as well as
752 David Rasmussen and an anonymous reviewer for constructive comments that helped
753 improve the manuscript.

754 *

755 References

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

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

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

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

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

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

Bhatt, S., P. Gething, O. Brady, J. Messina, A. Farlow, C. Moyes, J. Drake, J. Brownstein, A. Hoen, O. Sankoh, M. Myers, D. George, T. Jaenisch, G. Wint, C. Simmons, T. Scott, J. Farrar, and S. Hay. 2013. The global distribution and burden of dengue. Nature 496:504–507.

Biek, R., J. Henderson, L. Waller, C. Rupprecht, and L. Real. 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.

Bloomquist, E. W., P. Lemey, and M. A. Suchard. 2010. Three roads diverged? routes to phylogeographic inference. Trends in Ecology & Evolution 25:626–632.

Brennan, C., P. Bodelle, R. Coffey, S. Devare, A. Golden, J. Hackett Jr., B. Harris, V. Holzmayer, K. Luk, G. Schochetman, P. Swanson, J. Yamaguchi, A. Vallari, N. Ndembi, C. Ngansop, F. Makamche, D. Mbanya, L. Gurtler, L. Zekeng, and L. Kaptue. 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.

Campos, P., E. Willerslev, A. Sher, L. Orlando, E. Axelsson, A. Tikhonov, K. Aaris-Sorenson, A. Greenwood, R. Kahlke, P. Kosintsev, T. Krakhmalnaya, T. Kuznetsova, P. Lemey, R. MacPhee, C. Norris, K. Shepherd, M. Suchard, G. Zazula, B. Shapiro, and M. Gilbert. 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.

794 Carrington, C., J. Foster, O. Pybus, S. Bennett, and E. Holmes. 2005. Invasion and
795 maintenance of dengue virus type 2 and type 4 in the Americas. Journal of Virology
796 79:14680–14687.

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

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

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

804 Cummings, D., R. Irizarry, N. Huang, T. Endy, A. Nisalak, K. Ungchusak, and D. Burke.
805 2004. Travelling waves in the occurrence of dengue haemorrhagic fever in Thailand.
806 Nature 427:344–347.

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

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

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

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

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

820 Drummond, A. J., M. A. Suchard, D. Xie, and A. Rambaut. 2012. Bayesian phylogenetics
821 with beauti and the beast 1.7. Molecular biology and evolution 29:1969–1973.

822 Edo-Matas, D., P. Lemey, J. A. Tom, C. Serna-Bolea, A. E. van den Blink, A. B. van't
823 Wout, H. Schuitemaker, and M. A. Suchard. 2011. Impact of CCR5delta32 host genetic
824 background and disease progression on HIV-1 intrahost evolutionary processes: efficient
825 hypothesis testing through hierarchical phylogenetic models. Molecular Biology and
826 Evolution 28:1605–16.

827 Faria, N., A. Rambaut, M. Suchard, G. Baele, T. Bedford, M. Ward, A. Tatem, J. Sousa,
828 N. Arinaminpathy, J. Pepin, D. Posada, M. Peeters, O. Pybus, and P. Lemey. 2014. The
829 early spread and epidemic ignition of HIV-1 in human populations. Science 346:56–61.

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

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

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

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

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

Griffiths, R. and S. Tavaré. 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., M. Stuiver, J. White, S. Johnsen, and J. Jouzel. 1993. Comparison of oxygen isotope records from the GISP2 and GRIP Greenland ice cores. Nature 366:552–554.

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

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

Hemelaar, J., E. Gouws, P. D. Ghys, S. Osmanov, and 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. and B. Shapiro. 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.

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., H. Clausen, W. Dansgaard, N. Gundestrup, C. Hammer, U. Andersen, K. Andersen, C. Hvidberg, D. Dahl-Jensen, J. Steffensen, H. Shoji, A. Sveinbjornsdottir, J. White, J. Jouzel, and D. Fisher. 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.

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

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.

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

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

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

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

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

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

Lorenzen, E., D. Nogues-Braco, L. Orlando, J. Weinstock, J. Binladen, K. Marske, A. Ugan, M. Borregaard, M. Gilbert, R. Nielsen, S. Ho, T. Goebel, K. Graf, D. Byers, J. Stenderup, M. Rasmussen, P. Campos, J. Leonard, K. Koepfli, D. Froese, G. Zazula, T. Stafford, K. Aaris-Sorensen, P. Batra, A. Haywood, J. Singarayer, P. Valdes, G. Boeskorov, J. Burns, S. Davydov, J. Haile, D. Jenkins, P. Kosintsev, T. Kuznetsova, X. Lai, L. Martin, H. McDonald, D. Mol, M. Meldgaard, K. Munch, E. Stephan, M. Sablin, R. Sommer, T. Sipko, E. Scott, M. Suchard, A. Tikhonov, R. Willerslev, R. Wayne, A. Cooper, M. Hofreiter, A. Sher, B. Shapiro, C. Rahbek, and E. Willerslev. 2011. Species-specific responses of Late Quaternary megafauna to climate and humans. Nature 479:359–365.

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

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

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

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

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

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

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

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

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

Shapiro, B., A. Drummond, A. Rambaut, M. Wilson, P. Matheus, A. Sher, O. Pybus, M. Gilbert, I. Barnes, J. Binladen, E. Willerslev, A. Hansen, G. Baryshnikov, J. Burns, S. Davydov, J. Driver, D. Froese, C. Harington, G. Keddie, P. Kosintsev, M. Kunz, L. Martin, R. Stephenson, J. Storer, R. Tedford, S. Zimov, and A. Cooper. 2004. Rise and fall of the Beringian steppe bison. Science 306:1561–1565.

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

Stiller, M., G. Baryshnikov, H. Bocherens, A. Grandal-d'Anglade, B. Hilpert, S. Munzel, R. Pinhasi, G. Rabeder, W. Rosendahl, E. Trinkaus, M. Hofreiter, and M. Knapp. 2010. Withering away-25,000 years of genetic decline preceded cave bear extinction. Molecular Biology and Evolution 27:975–978.

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

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

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

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

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

Volz, E., S. K. Pond, M. Ward, A. L. Brown, and S. Frost. 2009. Phylodynamics of infectious disease epidemics. Genetics 183:1421–1430.

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/.

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