University of Cape Town

Identifying Predictors of Evolutionary Dispersion with Phylogeographic Generalised Linear Models

Author:
Tim Wolff-Piggott

Supervisors:
Dr Miguel Lacerda
University of Cape Town
Dr Ben Murrell
University of California, San Diego

University of Cape Town

A thesis submitted in partial fulfilment of the requirements
for the degree of Masters in Advanced Analytics and Decision Sciences

in the

Faculty of Science
Department of Statistical Sciences

March 2017

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

University of Cape Town

University of Cape Town

Abstract

Faculty of Science
Department of Statistical Sciences

Masters in Advanced Analytics and Decision Sciences

Identifying Predictors of Evolutionary Dispersion with Phylogeographic Generalised Linear Models

by Tim Wolff-Piggott

Discrete phylogeographic models enable the inference of the geographic history of biological organisms along phylogenetic trees. Frequently applied in the context of epidemiological modelling, phylogeographic generalised linear models were developed to allow for the evaluation of multiple predictors of spatial diffusion. The standard phylogeographic generalised linear model formulation, however, assumes that rates of spatial diffusion are a noiseless deterministic function of the set of covariates, admitting no other unobserved sources of variation. Under a variety of simulation scenarios, we demonstrate that the lack of a term modelling stochastic noise results in high false positive rates for predictors of spatial diffusion. We further show that the false positive rate can be controlled by including a random effect term, thus allowing unobserved sources of rate variation. Finally, we apply this random effects model to three recently published datasets and contrast the results of analysing these datasets with those obtained using the standard model. Our study demonstrates the prevalence of false positive results for predictors under the standard phylogeographic model in multiple simulation scenarios and, using empirical data from the literature, highlights the importance of a model accounting for random variation.

Acknowledgements

Profound thanks for the expertise and guidance of my supervisors Dr Miguel Lacerda and Dr Ben Murrell, and for the financial assistance of the National Research Foundation and the South African Statistical Association. To my friends Jason Fourie and Jonathan Rayner for their generous proofreading and invaluable input. To my family and partner for their constant love and support.

Contents

Chapter 1

Introduction

1.1 Problem Description

Large population sizes, short generation times, and the rapid accumulation of mutations (especially for RNA viruses), mean that the pace of viral evolution is swift (Dimmock, Easton, and Leppard, 2007). Moreover, viruses have the potential to spread over large geographic stretches, resulting in epidemics and pandemics, with serious consequences for public health (Lemey et al., 2014; Scotch et al., 2013). Scientists, however, have a suite of approaches and often rich datasets with which to decipher viral diffusion and mitigate its human cost. This is a complex and expanding area of research, in which the established body of knowledge is subject to ongoing, often novel tests.

Molecular data obtained via modern sequencing techniques can illuminate evolutionary histories through the reconstruction of phylogenetic trees. Modelling spatial diffusion also introduces the problem of inferring geographic states at ancestral nodes. Finally, researchers are challenged to incorporate sequence data for the empirical verification of spatial epidemiological hypotheses, following from the impetus to unify evolutionary and epidemiological dynamics in a modelling framework (Grenfell et al., 2004). Identifying predictors of evolutionary dispersion lies at the nexus of evolutionary, geographic and epidemiological inference, and poses a formidable challenge for the development of integrated frameworks.

1.2 Background to the Investigation

Traditionally, phylogeography is the study of the dynamics that inform the geographic distribution of genetic lines of descent (Avise, 2000). Viruses provide an apt setting for

phylogeographic study, as their rapid pace of evolution means that epidemic spread occurs on the same time-scale as the fixation of informative mutations (Holmes, 2004). This means that viral gene sequences can be informative of their spatial and temporal history. Accordingly, phylogeographic inference from viral sequences provides important insight into epidemiological processes (Lemey, Rambaut, Drummond, et al., 2009). Additionally, phylogeographic patterns in viral data can inform public health initiatives, for instance, vaccine design (Gaschen et al., 2002; Nickle et al., 2003).

The development of Bayesian phylogeography in particular has provided a flexible and integrated framework to test hypotheses about epidemiological dynamics and the spatial diffusion of viruses (Lemey, Rambaut, Drummond, et al., 2009). Jointly modelling molecular evolution and geographic diffusion, Bayesian phylogeography can naturally accommodate model-based approaches to assessing epidemiological predictors of diffusion patterns, expanding connectivities between public health and evolutionary biology (Talbi et al., 2010).

Phylogeographic generalised linear models were developed in a Bayesian setting to integrate spatial epidemiological inference and the reconstruction of viral evolutionary patterns (Lemey et al., 2012). These models provide a powerful approach to quantifying the contribution of epidemiological predictors to viral diffusion and have been applied in many subsequent studies, making inferences on viral evolution, diffusion and spatial epidemiology (Gräf et al., 2015; Lemey et al., 2014; Magee et al., 2015; Nelson et al., 2015).

1.3 Purpose of the Research

Despite recent advancements (Trovão et al., 2015), the established phylogeographic generalised linear model framework assumes that spatial diffusion rates are a deterministic function of the predictors, with no unobserved sources of variation. This minor dissertation (m.d.) assesses the implications of this assumption when predictors with no relation to diffusion rates are included in the model. In particular, the m.d. quantifies false positive rates when phylogeographic generalised linear models are used to test randomly generated predictors under several simulation scenarios.

Furthermore, an unobserved source of rate variation is introduced into the model in the form of a random effect term, and the m.d. evaluates the effectiveness of this approach in controlling the false positive rate. The implications of this investigation have particular relevance to the literature, as covariates with no relation to diffusion rates are often tested in the phylogeographic generalised linear model framework as potential predictors (Lemey

et al., 2012; Magee et al., 2015). With this in mind, the random effects model is applied to three recently published datasets from the literature. Predictor significance results from this analysis are compared with those obtained using the standard phylogeographic generalised linear model.

1.4 Layout of the Paper

The literature review provides an overview of the relevant literature on statistical phylogenetics. Particular attention is given to Bayesian phylogenetics. Discrete phylogeography and earlier approaches to incorporating epidemiological predictors into phylogeographic analyses are explored, leading to an in-depth discussion of phylogeographic generalised linear models. Chapter 3 details the methods employed in this investigation and touches on the significance of our research. Model specification and Markov chain Monte Carlo inference of evolutionary and phylogeographic parameters are described, with a detailed discussion of the simulation structure. Approaches to assessing convergence and the software utilised are also covered. Chapter 4 presents the empirical results of the simulations, primarily focusing on the false positive rates for randomly generated predictors observed across different phylogenetic trees and phylogeographic processes. Chapter 5 contrasts predictor significance using the standard phylogeographic generalised linear model with the model including random effects on three recently published datasets from the literature. Conclusions and recommendations for further research are outlined in Chapter 6.

Chapter 2

Literature review

2.1 Phylogenetics

Phylogenetics is the study of evolutionary relationships between organisms. Traces of evolutionary processes are present in the observable form and structure of organisms, but evolutionary relationships are often inferred using the fine-grained information contained in molecules that collectively make up the genome of a given organism. These molecules are predominantly comprised of deoxyribonucleic acid (DNA) and in some cases, ribonucleic acid (RNA) (Lemey, Salemi, and Vandamme, 2011). Nucleotides are the building blocks of DNA and RNA, and comprise one sugar, one phosphate group and one nitrogenous base. Figure 2.1 provides the abridged chemical structure of nucleotides.

Structurally, DNA consists of two sugar phosphate backbones linked by the pairing of complementary nitrogenous bases. DNA is formed through multiple nucleotides bonding in a double helix shape, with the complementary nitrogenous bases forming struts between pairs of helices. Figure 2.2 gives a schematic of the double helix shape and chemical structure of DNA. The bases cytosine (C), guanine (G), adenine (A) and thymine (T) constitute the informative component of DNA. Triplets of bases encode amino acids, and

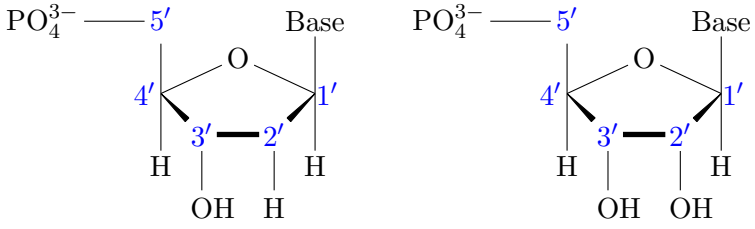


FIGURE 2.1: Nucleotides (comprising DNA and RNA respectively)

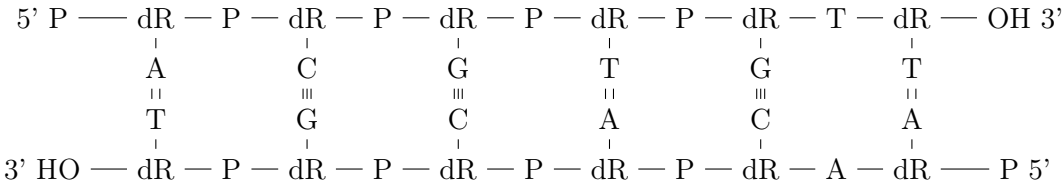


FIGURE 2.2: Double helix schematic and chemical structure of a DNA molecule.

for computational purposes DNA can be represented as linear sequences of characters on the state space of either bases or the amino acids encoded by the bases.

Over time, sequences undergo random changes, and exposure to evolutionary pressures determines whether those changes become fixed in the population. Even as sequences diverge over time, regions of functional or structural significance tend to be preserved (Xiong, 2006). Multiple sequence alignment arranges sequences so as to best reflect their common ancestry. After alignment, each site (character position or column in a multiple alignment) is assumed to have evolved from a common ancestor.

Although the globally optimal alignment can be found for any two sequences, in practice the computational burden of finding this optimal alignment is such that approaches are generally trade-offs between speed and accuracy (Needleman and Wunsch, 1970; Thompson, Higgins, and Gibson, 1994). The process of multiple alignment is closely connected to the reconstruction of the evolutionary histories of organisms or genes, which are represented by phylogenetic trees (see Figure 2.3). Phylogenetic trees are conventionally estimated from multiple alignments, though coestimation yields more accurate results in some cases (Notredame, Higgins, and Heringa, 2000; Redelings and Suchard, 2005).

Maximum parsimony was the preferred criterion for estimating phylogenetic trees for a number of years (Felsenstein, 2001). Under maximum parsimony, the phylogenetic tree

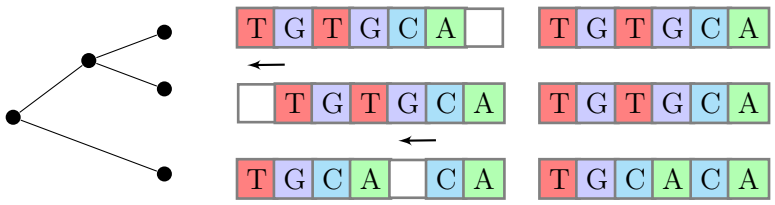


FIGURE 2.3: Multiple alignment and example inferred phylogeny.

that minimises the number of character changes across all sites is preferred. Maximum parsimony is susceptible to long branch attraction, where it fails to converge to the true phylogeny for certain tree structures. In the absence of branch length inequalities, however, maximum parsimony is computationally efficient and accurate (Felsenstein, 1978; Hillis, Huelsenbeck, and Swofford, 1994). Maximum parsimony methods are not based on an explicit model of evolution, and the emergence of pairwise distance methods and maximum likelihood methods enabled the formalisation of biological assumptions and the incorporation of uncertainty into phylogenetic tree estimation.

Building on the work of Kaplan and Langley (1979), Felsenstein (1981) modelled site-wise evolution down a phylogenetic tree by a reversible continuous time Markov chain (CTMC) on the state space of nucleotides $\{A, C, G, T\}$. Models of amino acid evolution have also been formulated, working with similar principles on the state space of amino acids (Whelan and Goldman, 2001). The model of molecular evolution determines the parametrisation of the generator matrix for the Markov process, which in turn informs the state transition probabilities. The generator matrix associated with the generalised time-reversible (GTR) model of Tavare (1986), for instance, is given by

$$\mathbb{Q} = \begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{bmatrix} * & \alpha\pi_C & \beta\pi_A & \gamma\pi_G \\ \alpha\pi_T & * & \delta\pi_A & \epsilon\pi_G \\ \beta\pi_T & \delta\pi_C & * & \eta\pi_G \\ \gamma\pi_T & \epsilon\pi_C & \eta\pi_A & * \end{bmatrix} \end{array}, \tag{2.1}$$

where $\boldsymbol{\pi} = (\pi_T, \pi_C, \pi_G, \pi_A)$ are the equilibrium base frequencies, and $\{\alpha, \beta, \gamma, \delta, \epsilon, \eta\}$ are the exchangeabilities, with a unique exchangeability for each pair of nucleotides.

Under time-reversibility, the likelihood of a multiple alignment can be computed without identifying the most recent common ancestor of the sampled sequences, and so we restrict consideration to the class of time-reversible models. Provided that substitution rates are held constant in time, the GTR model is the most general finite-state time-reversible model, and other models of nucleotide evolution are special cases obtained via applying parameter constraints to the GTR model. For instance, the HKY85 model due

to Hasegawa, Kishino, and Yano (1985) is obtained through imposing the restrictions $\beta = \gamma = \delta = \epsilon$ and $\alpha = \eta$ in the GTR model, distinguishing only between transitions and transversions as in Figure 2.4.

Felsenstein (1981) made an important inroad in the field of statistical phylogenetics when he introduced a computationally tractable maximum likelihood approach to estimate phylogenetic trees from DNA sequences. Transition probabilities, which are derived from the generator matrix for a given phylogenetic model, are essential to the likelihood formulation. In general, the transition probability matrix is obtained from the generator matrix through matrix exponentiation

$$P(t) = \exp(\mathbb{Q}t) = \sum_{n=0}^{\infty} \mathbb{Q}^n \frac{t^n}{n!}$$
$$\mathbb{Q}^n = \underbrace{\mathbb{Q} \times \mathbb{Q} \times \cdots \times \mathbb{Q}}_{n}.$$

In the phylogenetic context, t refers to branch lengths, generally measured in units of calender time or evolutionary time. $P_{ij}(t)$, the ijth entry in $P(t)$, represents the probability that starting in state i, after time t, the process is in state j.

Consider modelling nucleotide evolution over the phylogenetic tree in Figure 2.5. Let τ denote the topology, together with branch lengths $\boldsymbol{\nu}$ and a model of evolution parameterised by the vector $\boldsymbol{\theta} = (\boldsymbol{r}, \boldsymbol{\pi})$ of exchangeabilities and equilibrium frequencies. The likelihood of the parameters given a specific pattern X_i at site i (corresponding to the ith observed column in the multiple alignment) is then obtained from

$$L(\tau, \boldsymbol{\nu}, \boldsymbol{\theta} | X_i) = \sum_{S_0} \sum_{S_6} \sum_{S_7} \sum_{S_8} \pi_{S_0} P_{S_0 S_6}(v_6|\boldsymbol{\theta}) P_{S_6 S_1}(v_1|\boldsymbol{\theta}) P_{S_6 S_2}(v_2|\boldsymbol{\theta}) P_{S_0 S_8}(v_8|\boldsymbol{\theta})$$
$$P_{S_8 S_3}(v_3|\boldsymbol{\theta}) P_{S_8 S_7}(v_7|\boldsymbol{\theta}) P_{S_7 S_4}(v_4|\boldsymbol{\theta}) P_{S_7 S_5}(v_5|\boldsymbol{\theta}). \quad (2.2)$$

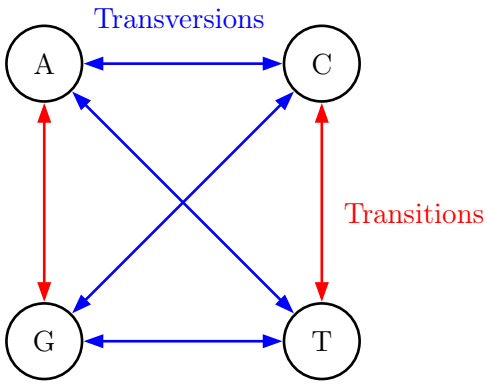


FIGURE 2.4: Transitions and transversions between nucleotides.

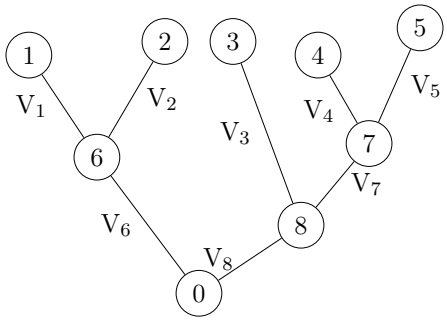


FIGURE 2.5: Phylogenetic tree in Felsenstein's (1981) derivation of the likelihood.

In practice the bases at the internal nodes 0, 6, 7 and 8 are unknown, and are represented symbolically by S_0, S_6, S_7 and S_8. π_{S_0} is the equilibrium frequency of S_0 and $P_{S_iS_j}(v_k|\boldsymbol{\theta})$ is the probability of transitioning from nucleotide S_i to nucleotide S_j over the branch length v_k, conditional on the parameters $\boldsymbol{\theta}$ for the model of evolution. The summation terms in the likelihood correspond to marginalisation over all the possible ancestral states.

Evaluation of this likelihood is computationally expensive. For a phylogenetic tree with n leaf nodes (and therefore $n-1$ interior nodes), the likelihood consists of 4^{n-1} terms, and therefore the number of terms is exponential in the number of nodes. Applying Felsenstein's tree pruning algorithm achieves a significant computational economy, as the likelihood becomes linear in n. According to this algorithm, the summation terms are moved rightwards, giving

$$L(\tau,\boldsymbol{\nu},\boldsymbol{\theta}|X_i) = \sum_{S_0}\pi_{S_0}\left[\sum_{S_6}P_{S_0S_6}(v_6|\boldsymbol{\theta})\,P_{S_6S_1}(v_1|\boldsymbol{\theta})\,P_{S_6S_2}(v_2|\boldsymbol{\theta})\right] \left[\sum_{S_8}P_{S_0S_8}(v_8|\boldsymbol{\theta})\,P_{S_8S_3}(v_3|\boldsymbol{\theta})\sum_{S_7}P_{S_8S_7}(v_7|\boldsymbol{\theta})\,P_{S_7S_4}(v_4|\boldsymbol{\theta})\,P_{S_7S_5}(v_5|\boldsymbol{\theta})\right]. \tag{2.3}$$

The expression is then evaluated in terms of conditional likelihoods.

Assuming independent evolution at different sites, the probability of the multiple alignment X given the model parameters can be expressed as the product across sites

$$L(\tau,\boldsymbol{\nu},\boldsymbol{\theta}|X) = p(X|\tau,\boldsymbol{\nu},\boldsymbol{\theta}) = \prod_i p(X_i|\tau,\boldsymbol{\nu},\boldsymbol{\theta}) = \prod_i L(\tau,\boldsymbol{\nu},\boldsymbol{\theta}|X_i). \tag{2.4}$$

The assumption of site-wise independence of evolution is not biologically realistic, but greatly facilitates computation of the likelihood. Hidden Markov models have been used

to incorporate the effects of neighbouring bases at each site on the pattern of substitution (Siepel and Haussler, 2004). Phylogenetic models on the amino acid state space tend to retain more accuracy than nucleotide models when there are dependencies between the three nucleotides within amino acid positions (Nasrallah, Mathews, and Huelsenbeck, 2010).

2.2 Bayesian Phylogenetics

The growth of statistical phylogenetics has been expedited by increasing computational power and the availability of accessible software packages for statistical analyses (Felsenstein, 2001). Bayesian phylogenetics is a burgeoning part of the field, and enables the incorporation of prior information, producing a posterior distribution that naturally expresses uncertainty in the topology, branch lengths and substitution model parameters. While uncertainty in phylogenetic inference using maximum likelihood methods can be assessed by bootstrapping, Bayesian analysis enjoys computational advantages (Larget and Simon, 1999). Bayesian phylogenetics is based on Bayes' formula

$$p\left(\tau, \boldsymbol{\nu}, \boldsymbol{\theta}|X\right) = \frac{p\left(X|\tau, \boldsymbol{\nu}, \boldsymbol{\theta}\right) p\left(\tau, \boldsymbol{\nu}, \boldsymbol{\theta}\right)}{p\left(X\right)} = \frac{L\left(\tau, \boldsymbol{\nu}, \boldsymbol{\theta}|X\right) p\left(\tau, \boldsymbol{\nu}, \boldsymbol{\theta}\right)}{p\left(X\right)}. \tag{2.5}$$

In the context of phylogenetics, X typically denotes a multiple sequence alignment, with the topology τ, branch lengths $\boldsymbol{\nu}$ and the parameter vector for the model of evolution $\boldsymbol{\theta}$ consisting of the exchangeabilities $\boldsymbol{r}$ and the equilibrium frequencies $\boldsymbol{\pi}$. Bayesian inference requires a prior distribution $p\left(\tau, \boldsymbol{\nu}, \boldsymbol{\theta}\right)$ for the parameters of interest, which can be specified as non-informative, or calibrated to reflect expert opinion and data from the literature. The likelihood $L\left(\tau, \boldsymbol{\nu}, \boldsymbol{\theta}|X\right)$ is specified as in Equation (2.4).

The marginal likelihood

$$p\left(X\right) = \sum_{\tau} \int_{\boldsymbol{\nu}} \int_{\boldsymbol{r}} \int_{\boldsymbol{\pi}} p\left(X|\tau, \boldsymbol{\nu}, \boldsymbol{r}, \boldsymbol{\pi}\right) p\left(\tau, \boldsymbol{\nu}, \boldsymbol{r}, \boldsymbol{\pi}\right) d\boldsymbol{\pi} d\boldsymbol{r} d\boldsymbol{\nu} \tag{2.6}$$

can be obtained in principle by marginalising over the parameter space. Since the exchangeabilities, equilibrium frequencies and branch lengths are continuous, evaluation of the marginal likelihood entails complicated multi-dimensional integration and summation.

Markov chain Monte Carlo (MCMC) approaches sample from the posterior distribution of the phylogenetic parameters without evaluation of the marginal likelihood, and were first proposed in the context of phylogenetics by Mau, Newton, and Larget (1999) and Li, Pearl, and Doss (2000). Bayesian phylogenetic inference based on MCMC methods

has become popular with software packages like MrBayes (Huelsenbeck and Ronquist, 2001) and BEAST (Drummond, Suchard, et al., 2012).

In Bayesian phylogenetics, MCMC is typically implemented via the Metropolis Hastings algorithm (Drummond and Rambaut, 2007). Consider a posterior distribution of interest $p(\theta|X)$. The Metropolis Hastings algorithm constructs a Markov chain with its equilibrium distribution equal to $p(\theta|X)$. The parameter θ is initialised at some value, and a proposal distribution q that will perturb the current parameter value is chosen, typically favouring nearby values. Given the last parameter value θ_{t-1}, a new value is proposed by drawing from the distribution $q(\theta_t|\theta_{t-1})$, and this state is accepted with probability

$$\min\left(1, \frac{p(\theta_t|X)}{p(\theta_{t-1}|X)} \times \frac{q(\theta_{t-1}|\theta_t)}{q(\theta_t|\theta_{t-1})}\right) = \min\left(1, \frac{L(\theta_t|X)\,p(\theta_t)}{L(\theta_{t-1}|X)\,p(\theta_{t-1})} \times \frac{q(\theta_{t-1}|\theta_t)}{q(\theta_t|\theta_{t-1})}\right). \quad (2.7)$$

Note that the potentially intractable marginal likelihood $p(X)$ falls away, and if a symmetric proposal distribution is chosen, the acceptance probability simplifies to

$$\min\left(1, \frac{p(\theta_t|X)}{p(\theta_{t-1}|X)}\right) = \min\left(1, \frac{L(\theta_t|X)\,p(\theta_t)}{L(\theta_{t-1}|X)\,p(\theta_{t-1})}\right). \quad (2.8)$$

Intuitively, the Metropolis Hastings algorithm samples more frequently from regions of high posterior density. For instance, consider a Metropolis Hastings sampler on the simple posterior distribution in Figure 2.6, and suppose that a symmetric proposal distribution has proposed the indicated changes to states θ_A and θ_B. According to Equation (2.8), the proposed state change from θ_A is always accepted, while the proposed state change from θ_B is accepted with probability 2/3.

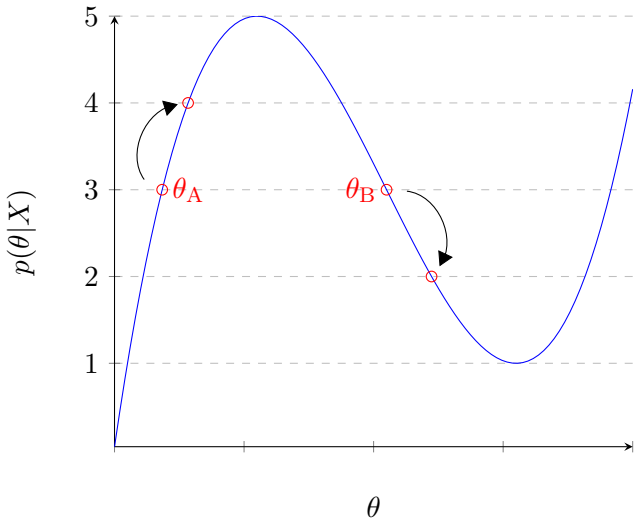


FIGURE 2.6: Metropolis Hastings algorithm schematic.

In the phylogenetic context, the distribution of interest is the multi-dimensional joint posterior distribution. A Metropolis Hastings sampler on this distribution works on the same principles, but components of the proposal distribution must be specified for each individual parameter (these are referred to as operators). Successive states in the joint parameter space may sometimes differ in only a few parameters, as changes may only be proposed and accepted for a subset of parameters at each state (Drummond and Bouckaert, 2015).

The use of a burn-in period, where the first b states in the chain are discarded, removes dependence on the initial values specified for the chain. Although successive steps in the Metropolis Hastings algorithm are highly correlated, under thinning, where only states at intervals of k iterations are accepted, accepted states constitute approximately independent samples from the posterior distribution. Functions of the sampled values are then used to estimate and construct credibility intervals for the parameters of interest.

2.2.1 Bayesian Stochastic Search Variable Selection (BSSVS)

Bayesian stochastic search variable selection (BSSVS) is a widely-applied approach to variable selection (Lee et al., 2003; O'Hara and Sillanpää, 2009). The initial formulation of George and McCulloch (1993) embedded a multiple regression model in a hierarchical Bayes normal mixture model, with subset choices governed by latent variables. MCMC methods were then used to sample from the posterior distribution on the set of possible predictor selections, which naturally lends the method towards applications in the MCMC context.

Drawing on the work of George and McCulloch, Kuo and Mallick (1998) introduced a simpler implementation of BSSVS through embedding indicator variables in the regression equation

$$\begin{aligned} y_i &= \delta_1\beta_1 x_{i1} + \cdots + \delta_p\beta_p x_{ip} + \epsilon_i \\ \epsilon_i &\overset{iid}{\sim} N\left(0, \sigma^2\right), \end{aligned} \tag{2.9}$$

where y_i is the ith observation on some response variable of interest, x_{ij} is the ith observation on the jth predictor of the response, β_j is the regression coefficient for the jth predictor and δ_j is an indicator variable that determines if the jth predictor is kept in the model, for $i \in \{1, ..., n\}$ and $j \in \{1, ..., p\}$. Prior distributions are specified on the regression coefficients, indicator variables and the standard deviation of the error σ.

The joint space $\left(\boldsymbol{\delta}, \boldsymbol{\beta}, \sigma^2\right)$ is explored using MCMC. If an accepted state has $\delta_j = 0$, then the jth predictor is omitted from the model for that state. Different predictors could

either be retained or discarded at each iteration. Variable selection is made on the basis of posterior support for indicator values equal to one.

2.2.2 Bayes Factors

Bayes factors quantify evidence in support of competing hypotheses within the framework of Bayesian hypothesis testing. The Bayes factor for a test of two hypotheses H_0 and H_1 follows from Bayes' theorem, which defines the posterior probabilities for the two hypotheses according to

$$p\left(\mathrm{H}_k|\mathbf{D}\right) = \frac{p\left(\mathbf{D}|\mathrm{H}_k\right)p\left(\mathrm{H}_k\right)}{p\left(\mathbf{D}|\mathrm{H}_1\right)p\left(\mathrm{H}_1\right) + p\left(\mathbf{D}|\mathrm{H}_2\right)p\left(\mathrm{H}_2\right)}, \tag{2.10}$$

where $p\left(\mathbf{D}|\mathrm{H}_k\right)$ is the likelihood of the data under hypothesis H_k for $k = 1, 2$. Given prior probabilities $p\left(\mathrm{H}_1\right)$ and $p\left(\mathrm{H}_0\right) = 1 - p\left(\mathrm{H}_1\right)$, the Bayes factor is defined simply as the multiplicative factor through which the posterior odds of H_1 are obtained from the prior odds

$$\frac{p\left(\mathrm{H}_1|\mathbf{D}\right)}{p\left(\mathrm{H}_0|\mathbf{D}\right)} = \underbrace{\frac{p\left(\mathbf{D}|\mathrm{H}_1\right)}{p\left(\mathbf{D}|\mathrm{H}_0\right)}}_{\text{Bayes factor } \mathrm{B}_{10}} \times \frac{p\left(\mathrm{H}_1\right)}{p\left(\mathrm{H}_0\right)}. \tag{2.11}$$

Clearly from Equation (2.11), the Bayes factor B_{10} also corresponds to the posterior odds of the hypothesis H_1 divided by its prior odds

$$\begin{aligned} \mathrm{B}_{10} &= \frac{p\left(\mathbf{D}|\mathrm{H}_1\right)}{p\left(\mathbf{D}|\mathrm{H}_0\right)} \Big/ \frac{p\left(\mathrm{H}_1\right)}{p\left(\mathrm{H}_0\right)} \\ &= \frac{\tilde{p}}{1-\tilde{p}} \Big/ \frac{\tilde{q}}{1-\tilde{q}}, \end{aligned} \tag{2.12}$$

where $\tilde{p}$ is the posterior probability of H_1 and $\tilde{q}$ is the prior probability of H_1.

Kass and Raftery (1995) established widely-cited guidelines to the significance of the evidence presented by different values of the Bayes factors, reproduced here in Table 2.1.

TABLE 2.1: Kass and Raftery (1995) guidelines for classification of Bayes factors.

$2\log_e\left(\mathrm{B}_{10}\right)$	B_{10}	Evidence against H_0
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
> 10	> 150	Very strong

2.3 The Coalescent

The coalescent is a stochastic process that models the distribution of ancestral histories of organisms (Kingman, 1982). These ancestral histories are assumed to arise from population genetics models such as the Wright-Fisher process (Fisher, 1930; Wright, 1931) and can be represented in the form of phylogenetic trees. The expected frequencies with which different phylogenetic trees arise under the Wright-Fisher process are obtained from the coalescent (Drummond, Nicholls, et al., 2002). In the context of Bayesian phylogenetic inference, the coalescent can therefore define the prior distribution for the unknown phylogenetic topology.

The simplest form of the Wright-Fisher model is based on a constant population size N, non-overlapping generations, and random reproduction among individuals (Drummond and Bouckaert, 2015). For any two members of the current generation, the probability of a shared common ancestor in the previous generation is $1/N$, and it follows that the probability of a common ancestor t generations in the past is

$$p\left(t\right) = \frac{1}{N}\left(1 - \frac{1}{N}\right)^{t-1}. \tag{2.13}$$

The number of generations in the past it takes for two members of the current generation to find a common ancestor therefore follows the geometric distribution with parameter $1/N$.

A more general case considers coalescence among a subset of k lineages within the population of size N. The probability that k different members of a given generation have k different ancestors in the previous generation is

$$\begin{aligned} &\frac{N-1}{N} \times \frac{N-2}{N} \times \cdots \times \frac{N-k+1}{N} \\ &= 1 - \binom{k}{2}\frac{1}{N} + O\left(1/N^2\right). \end{aligned} \tag{2.14}$$

This is precisely the probability that no coalescent event occurs in the previous generation. Each member must choose a parent from the previous generation: the first can choose without restriction from the population of N; the second chooses a different parent, with $N-1$ options, and so on. When N is large relative to k, the $O\left(1/N^2\right)$ term is negligible, and the probability of a coalescent event in a given generation is therefore simply $\binom{k}{2}\frac{1}{N}$.

Similarly to Equation (2.13), the number of generations in the past it takes for two of the k lineages to coalesce is approximately geometrically distributed, with a probability of coalescence in any generation of $\binom{k}{2}\frac{1}{N}$. The geometric distribution can be approximated

by the exponential distribution through the use of a scaling factor M, which allows for representation of the coalescent in terms of continuous waiting times to coalescence (Hein, Schierup, and Wiuf, 2004). The scaling factor converts coalescent time (in generations j) to continuous time $t = j/M$ and the resulting exponential process has rate parameter $a = (\binom{k}{2}\frac{1}{N}) \times M$.

When trees with dated leaf nodes are modelled, and the coalescent is used to define a prior distribution on the topology, the scaling factor is chosen as $M = 1/\rho$, where ρ is the number of calender units per generation. The object of inference in this context is in fact $\theta = N\rho$, where the factor θ converts from coalescent time to the calender time implied by the dating of the tree (Drummond, Nicholls, et al., 2002). The resulting exponential distribution on the waiting time to first coalescence among k individuals has rate parameter $\binom{k}{2}\frac{1}{\theta}$.

Assuming temporally spaced sequences, the probability density of a given phylogenetic tree under the coalescent process is then calculated according to

$$\begin{aligned} f\left(g|\theta\right) &= \frac{1}{\theta^{n-1}} \prod_{i=2}^{2n-1} \exp\left(-\binom{k_i}{2}\frac{1}{\theta} \times (t_i - t_{i-1})\right) \\ &= \frac{1}{\theta^{n-1}} \prod_{i=2}^{2n-1} \exp\left(-k_i(k_i-1)/2\theta \times (t_i - t_{i-1})\right), \end{aligned} \tag{2.15}$$

where g is a rooted binary tree with n leaf nodes and $n-1$ ancestral nodes (each of which corresponds to a coalescence of two ancestral lineages). In the time interval $t_i - t_{i-1}$ preceding the time of a leaf node i, there is no coalescence with probability $\exp\left(-k_i(k_i-1)/2\theta \times (t_i - t_{i-1})\right)$, where k_i is the number of lineages present in the time interval $t_i - t_{i-1}$. Similarly, the length of time preceding an ancestral node i is exponentially distributed with rate parameter $\binom{k_i}{2}\frac{1}{\theta}$. The formulation in Equation (2.15) enables an efficient prior specification for phylogenetic trees assumed to arise under a coalescent process, and the subsequent inference of the coalescent population size parameter.

2.4 The Molecular Clock

Zuckerkandl and Pauling (1962) first observed an approximately linear relationship between the genetic distance and estimated time since divergence from a common ancestor for haemoglobin proteins in gorillas and humans. Further research led to the formalisation of a molecular clock hypothesis, which postulated a constant basic rate of evolution for any given protein across different phylogenetic lineages (Zuckerkandl and Pauling, 1965). Empirical observation in the intervening decades has presented substantial evidence for

rate variation over time, challenging the strict molecular clock hypothesis (Ayala, 1997; Britten, 1986). This led to the development of Bayesian relaxed clock models, which allow molecular rate variation among lineages (Drummond, Ho, et al., 2006).

The strict molecular clock provides a firm foundation for general simulations, with straightforward theoretical implications, and exploration of relaxed clock methods is beyond the scope of this study. When the phylogenetic topology under consideration is a time-tree (that is, a tree with branch lengths in units of calender time), the molecular clock rate μ specifies the linear relationship between the time-scale and the branch-length scale, measured in expected number of substitutions per site per unit time (Rambaut, 2000). Differences in tip dates therefore provide a vital source of information to infer the clock rate μ (Lemey, Salemi, and Vandamme, 2011).

2.5 Phylogeography

Historically, population geneticists have investigated and modelled the relationship between population structure and genetic processes. Early models of population structure incorporated geographic distance into the modelling of genetic processes. For instance, in the stepping-stone model (Kimura and Weiss, 1964), gene exchange is only possible between adjacent populations. More recently, coalescent theory has integrated mathematical population genetics models with tree-based phylogeographic approaches (Hein, Schierup, and Wiuf, 2004).

Phylogeography centres on the study of the processes underlying the geographic distribution of genetically related organisms (Avise, 2000). The question that informed seminal research in the field was how the formation of distinct species was affected by the geographic distribution of genealogical lineages (Lemey, 2010). Highly informative DNA sequence data were collected from populations in the 1970s and population geneticists realised that phylogenetic trees constructed with DNA sequence data could inform the study of population history. Phylogenetic trees describing DNA ancestry were inferred, leading to the growth of phylogeography as a tree-based approach (Hey and Machado, 2003).

Early research in phylogeography inferred phylogenetic trees and interpreted the branching structure in light of the geographic distribution of the leaf nodes (Cann, Stoneking, and Wilson, 1987; Vigilant et al., 1991). However, it did not formally infer locations at all ancestral nodes, nor did it model the historical processes that produced the observed geographic distributions. The absence of a formal statistical framework also complicated tests of hypotheses. Nested clade analysis (NCA) emerged as a major methodology used

to test hypotheses about geographic structure in gene populations (Templeton, Routman, and Phillips, 1995). Independent studies have however identified high false positive rates and raised concerns about the ability of NCA to distinguish signal from stochastic noise in the context of phylogeography (Knowles and Maddison, 2002; Petit, 2008).

Maximum parsimony methods provide a framework within which states at ancestral nodes of phylogenetic trees can be inferred, and have been applied in the context of phylogeography to reconstruct ancestral locations (Sullivan, Markert, and Kilpatrick, 1997). Parsimony, however, neglects to consider uncertainty in mapping states to ancestral nodes, as well as in the phylogenetic tree itself. Although phylogenetic uncertainty in parsimony methods may be addressed through bootstrapping, uncertainty in the inferred states at internal nodes cannot be quantified in this manner (Ronquist, 2004).

Pagel, Meade, and Barker (2004) introduced a formal Bayesian framework within which ancestral states of phylogenetic trees could be estimated. This framework accounts for uncertainty in the phylogeny as well as in the ancestral states, while also providing for a Bayesian approach to test comparative hypotheses. Within this approach, discrete spatial change is modelled as a CTMC down the phylogenetic tree.

Taking the phylogenetic tree as given, the likelihood for the phylogeographic process is constructed similarly to Equation (2.2). Instead of conditioning on a column in a multiple alignment X, the likelihood for a phylogeographic CTMC relies on the observed locations at the leaf nodes Y, which are drawn from a finite set $\{L_1, \ldots, L_k\}$. The phylogeographic CTMC is parametrised by the vector $\mathbf{\Lambda}$ of exchangeabilities and equilibrium frequencies. As in Equation (2.3), the likelihood is evaluated using Felsenstein's tree pruning algorithm.

The phylogeographic CTMC formulation differs from the phylogenetic CTMC in that only one column of data – the observed locations at the leaf nodes – informs the likelihood, instead of the numerous columns that constitute a multiple alignment. The single column of observed locations may provide inadequate information to accurately infer all of the parameters $\mathbf{\Lambda}$ that populate the generator matrix of the phylogeographic process (Lemey, 2010).

Aside from obtaining more data, the concern of inadequate information can be addressed through specifying informative prior distributions on the parameters $\mathbf{\Lambda}$. Lemey, Rambaut, Drummond, et al. (2009) also apply BSSVS to the parameters $\mathbf{\Lambda}$, identifying the transition rates with the highest posterior support and obtaining a parsimonious parametrisation of the generator matrix.

In practice, the phylogenetic tree – specified by τ and $\boldsymbol{\nu}$ – is not known, and the rate parameters governing the phylogeographic CTMC $\mathbf{\Lambda}$ need to be inferred together with the

rate parameters governing the phylogenetic CTMC $\boldsymbol{\theta}$. Lemey, Rambaut, Drummond, et al. (2009) model the phylogenetic and phylogeographic CTMCs as independent, conditional on the unobserved phylogeny $\{\tau, \boldsymbol{\nu}\}$, and express the posterior distribution for the whole model as

$$\begin{aligned} p\left(\tau, \boldsymbol{\nu}, \boldsymbol{\theta}, \boldsymbol{\Lambda} | X, Y\right) &\propto p\left(X, Y | \tau, \boldsymbol{\nu}, \boldsymbol{\theta}, \boldsymbol{\Lambda}\right) p\left(\tau, \boldsymbol{\nu}, \boldsymbol{\theta}, \boldsymbol{\Lambda}\right) \\ &= p\left(X | \tau, \boldsymbol{\nu}, \boldsymbol{\theta}\right) p\left(Y | \tau, \boldsymbol{\nu}, \boldsymbol{\Lambda}\right) p\left(\tau, \boldsymbol{\nu}\right) p\left(\boldsymbol{\theta}\right) p\left(\boldsymbol{\Lambda}\right), \end{aligned} \tag{2.16}$$

where X is the multiple alignment, and Y the observed locations as above. MCMC is then used to sample from the joint posterior distribution.

2.6 Phylogeographic Generalised Linear Models

Pandemics and epidemics are phenomena of serious concern, and the genomes of some rapidly spreading and evolving pathogens have been extensively sequenced, providing an expansive body of data for study (Lemey et al., 2014; Nelson et al., 2015). In addition to reconstructing the phylogenies and inferring the spatial histories of viruses, the puzzle of the dynamics underlying their spread and data on possible predictors of diffusion posed a vital challenge in computational biology that naturally generalised to other phylogeographic contexts.

Modelling the spread of the dog rabies virus in North Africa, Talbi et al. (2010) approached the problem of identifying drivers of spatial diffusion by fixing the rates in the generator matrix for the phylogeographic process according to covariate values. Model performance was then evaluated in terms of marginal log-likelihoods, relative to a baseline model assuming equal rates in the phylogeographic CTMC (Lemey, 2010). For instance, equal rates could be contrasted with rates inversely proportional to pairwise distance for the phylogeographic CTMC.

The approach by which rates in the generator matrix are fixed can be applied to evaluate predictors that are expressible as pairwise quantities for each pair of geographic locations. However, under this approach, the model for each predictor (as characterised by the rate parametrisation of the phylogeographic CTMC) needs to be fitted independently, precluding more complicated scenarios that incorporate differing contributions of multiple predictors.

The problem of testing hypotheses that consist of varying contributions from multiple predictors of spatial diffusion, ideally simultaneously, found an elegant solution in the application of a generalised linear models (GLM) approach. First proposed by Lemey et al. (2012) to identify predictors of the migration history of human influenza H3N2,

transition rates for the phylogeographic CTMC are modelled as log-linear functions of the logarithms of potential predictors of spatial diffusion

$$\log \Lambda_{ij} = \beta_1 \delta_1 \log (x_{ij1}) + \beta_2 \delta_2 \log (x_{ij2}) + \cdots + \beta_p \delta_p \log (x_{ijp}) . \tag{2.17}$$

β_k quantifies the effect size of the logarithm of the kth predictor, and δ_k is an indicator variable that determines whether the kth predictor is included in the model, for $k \in \{1, ..., p\}$. The log transformation applied to the predictor values x_{ijk} removes range constraints on non-negative predictors like distance and population size. Λ_{ij} is the transition rate between the ith and jth location – the ijth element in the generator matrix for the phylogeographic process

$$\mathbf{\Lambda}_{K \times K} = \begin{bmatrix} * & \Lambda_{12} & \dots & \Lambda_{1K} \\ \Lambda_{21} & * & \dots & \Lambda_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \Lambda_{K1} & \Lambda_{K2} & \dots & * \end{bmatrix} . \tag{2.18}$$

As components in a Bayesian phylogenetic model, prior distributions are specified on the parameters governing the phylogeographic process $\{\boldsymbol{\beta}, \boldsymbol{\delta}\}$, as well as the other phylogenetic parameters, which are then sampled via MCMC. The indicator variables enable BSSVS, and predictor significance is determined by Bayes factors for indicator values equal to one.

Trovão et al. (2015) introduced a novel phylogeographic GLM parametrisation, including a random effect for each rate in the phylogeographic process

$$\log \Lambda_{ij} = \beta_1 \delta_1 x_{ij1} + \beta_2 \delta_2 x_{ij2} + \cdots + \beta_p \delta_p x_{ijp} + \epsilon_{ij}, \tag{2.19}$$

where ϵ_{ij} is the random effect for the transition rate Λ_{ij}. Trovão et al. (2015) specified a hierarchical prior on the random effects, with $\epsilon_{ij} \sim N\left(0, \sigma^2\right)$ and the precision of the random effects $1/\sigma^2$ gamma-distributed with $\alpha = 0.001$ and $\beta = 0.001$ *a priori*.

Trovão et al. (2015) introduced their novel GLM formulation for the purpose of modelling the epidemic expansion of Influenza A H5N1. The authors implemented random effects to identify exceptions to diffusion modelled using a particular predictor, which they argue may need an additional effect to be adequately explained. In their paper, diffusion rates were modelled as a function of geographic distance.

In order to determine exceptions to distance-based diffusion, Trovão et al. (2015) focused on consistently large random effects, which were identified through a statistic that

summarises the probability that a given random effect is largest among all random effects

$$p\left(|\epsilon_{ij}| = \max_{0 \leq i,j \leq K} |\epsilon_{ij}|\right),$$

where $i \neq j$ are geographic states. The formulation incorporating random effects is the only alteration of the phylogeographic GLM in Equation (2.17), and to date, has not been applied subsequently in the literature.

Chapter 3

Methods

This chapter details the methods employed in this m.d., and proceeds as follows: firstly, the significance of our research, with reference to the literature, is described. Next, the general structure of the simulations, designed to assess false positive rates for the standard and random effects phylogeographic GLMs, is illustrated. All simulations were performed within a Bayesian phylogeographic framework using BEAST (Drummond, Suchard, et al., 2012), with parameters inferred via MCMC. The specific simulation settings with reference to choice of phylogeographic processes, phylogenetic tree structures and number of covariates are detailed, followed briefly by the model assumptions made when inferring parameters in the simulations. MCMC calibration is discussed, covering prior and proposal distributions for the model parameters, as well as convergence diagnostics. Lastly, the software used in the analysis is described.

3.1 Significance of this research

The standard Bayesian phylogeographic GLM introduced by Lemey et al. (2012) has been applied successfully by multiple authors to identify epidemiological predictors, and is the dominant formulation in the literature. Lemey et al. (2014) applied phylogeographic GLMs using covariates based on human transportation data to predict the global diffusion of human influenza H3N2. The same phylogeographic GLM formulation was also applied by Magee et al. (2015) to identify drivers of the spread of avian influenza H5N1 in Egypt, as well as by Nelson et al. (2015) to identify predictors of the global spread of Influenza A viruses in swine.

Modelling the epidemic expansion of Influenza A H5N1, Trovão et al. (2015) introduced random effects into the phylogeographic GLM to identify transition rates in the phylogeographic CTMC for which distance was not an adequate predictor. The authors argued

that the posterior estimated effects for transition rates that deviate from distance-based diffusion patterns should be consistently high relative to other random effects. If distance adequately modelled the diffusion process, Trovão et al. (2015) argued that all posterior random effects estimates should be close to zero.

We argue that even in the presence of significant predictors of diffusion, some random variation is expected in the transition rates between pairs of location states after conditioning on the predictors. That is, transition rates are not deterministic functions of the predictors. If uninformative predictors are used to model diffusion rates in the GLM without random effects, these covariates may be identified as important predictors. Each covariate has some probability of explaining part of the variation in the transition rates by chance. Such covariates would be spuriously identified as significant in a model that does not incorporate stochastic noise in the transition rates.

The importance of random effects has been overlooked in the literature. For instance, Gräf et al. (2015) modelled the contribution of epidemiological predictors to the spread of HIV-1 subtype C in Brazil using the GLM without random effects shortly after the publication of Trovão et al. (2015).

This research quantifies the degree to which the current phylogeographic GLM formulation is susceptible to false positive results through simulations under different choices of phylogenetic tree structures, phylogeographic processes, and different numbers of covariates. While varying these model choices, support for spurious predictors of phylogeographic transition rates is assessed. We also contrast these results with the support for spurious predictors under the phylogeographic GLM with random effects, as introduced by Trovão et al. (2015).

3.2 Simulation structure

The simulations were structured to illustrate conditions under which the GLM formulation as in Lemey et al. (2012) – without random effects – would be susceptible to false positives; i.e. variables that are not associated with the transition rates being spuriously identified as important predictors. Phylogenetic tree structures were chosen firstly to emphasise conditions under which false positives could arise, and secondly to roughly emulate real-world data.

Figure 3.1 gives a schematic of the simulation structure. For a given set of simulations, a topology and set of branch lengths were generated and fixed independently of the sequence data. Given the independence of molecular evolution and geographic diffusion

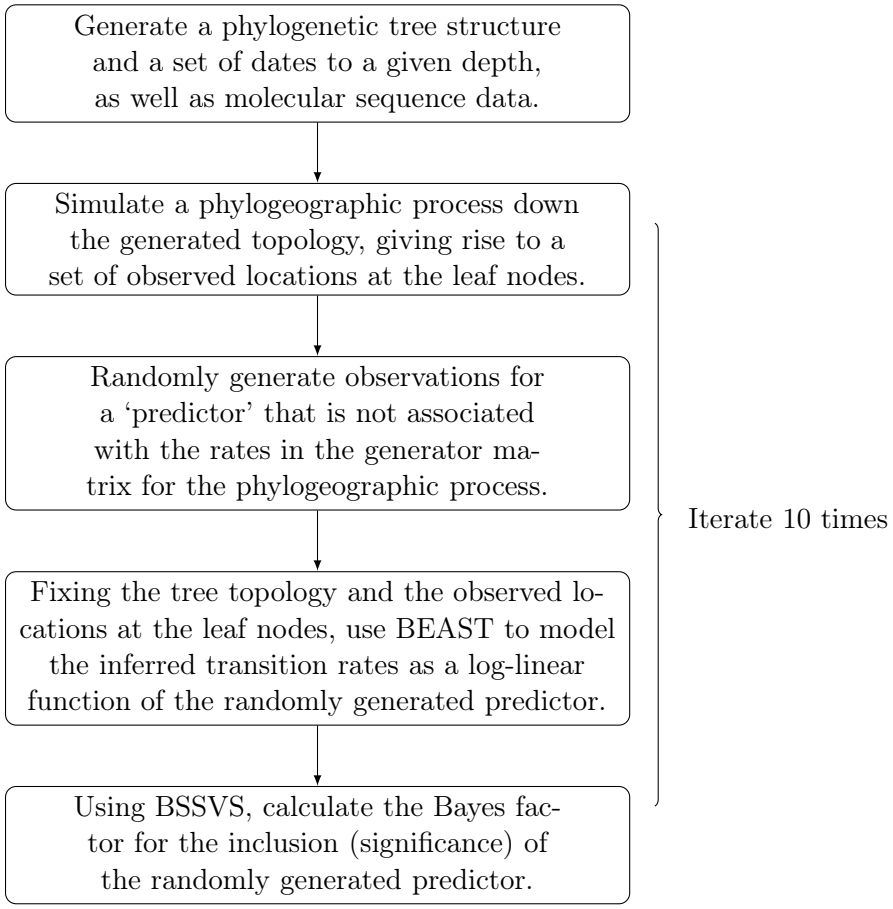


FIGURE 3.1: Schematic for simulation method.

(conditional on the phylogenetic tree, which is fixed in our simulations), molecular evolution is of limited interest and does not influence our phylogeographic model. With this consideration, sequence data was simulated according to the HKY85 model of Hasegawa, Kishino, and Yano (1985), a widely-applied model of nucleotide evolution and sufficiently general for our purposes. The HKY85 model is governed by the generator matrix

$$\mathbb{Q} = \begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{bmatrix} * & \kappa\pi_C & \pi_A & \pi_G \\ \kappa\pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & \kappa\pi_G \\ \pi_T & \pi_C & \kappa\pi_A & * \end{bmatrix} \end{array}, \tag{3.1}$$

which allows for unequal base frequencies and uses the κ parameter to distinguish between transitions and transversions. In the simulations, the base frequencies were fixed according to $\pi_T = 0.4$, $\pi_C = 0.3$, $\pi_A = 0.2$, $\pi_G = 0.1$, with $\kappa = 2$. Each simulated molecular sequence had a length of 1600 nucleotide positions, informed by the average

sequence length in the datasets of Gräf et al. (2015), Magee et al. (2015), and Trovão et al. (2015), all of which are recently published applications of phylogeographic generalised linear models.

A K-state phylogeographic process determined by a generator matrix $\mathbf{\Lambda}$ (the form of which will be provided in the next section) was then simulated over the tree, producing a set of observed locations at the leaf nodes. A set of $K\,(K-1)$ observations (corresponding to the off-diagonal transition rates in $\mathbf{\Lambda}$) on either a single spurious predictor or on multiple spurious predictors of the transition rates was randomly generated.

Keeping the topology and branch lengths fixed, the simulated geographic states were used to recover the generator matrix $\mathbf{\Lambda}$, which was modelled without random effects according to

$$\log \Lambda_{ij} = \beta_1 \delta_1 x_{ij1} + \beta_2 \delta_2 x_{ij2} + \cdots + \beta_p \delta_p x_{ijp}, \tag{3.2}$$

where x_{ijk} is the observation on the kth randomly generated predictor for Λ_{ij}. The parameters $\{\boldsymbol{\delta}, \boldsymbol{\beta}\}$, and hence $\mathbf{\Lambda}$, were then inferred via MCMC in BEAST (Drummond, Suchard, et al., 2012). In the random effects model, the generator matrix $\mathbf{\Lambda}$ was modelled according to

$$\log \Lambda_{ij} = \beta_1 \delta_1 x_{ij1} + \beta_2 \delta_2 x_{ij2} + \cdots + \beta_p \delta_p x_{ijp} + \epsilon_{ij}. \tag{3.3}$$

The parameters $\{\boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\epsilon}\}$, and hence $\mathbf{\Lambda}$, were again inferred via MCMC in BEAST. The generator matrix $\mathbf{\Lambda}$ was recovered and predictor support quantified via Bayes factors using both the standard model and the random effects model. This process was repeated for each simulation scenario with 10 different sets of randomly generated observations on the spurious predictors, and 10 different sets of observed location states arising under the phylogeographic process governed by $\mathbf{\Lambda}$. Finally, the results for the two different models were summarised.

3.3 Informative phylogenetic tree

Figure 3.2 displays the informative tree structure used in the first round of simulations. The phylogenetic tree is a time-tree. The node labels refer to the assigned tip dates, where '2000.1', for instance, denotes January 2000. Each internal node is paired with a branch of length 0 that produces a leaf node. This informative phylogenetic tree structure was chosen such that the observed states at certain pre-defined leaf nodes were highly informative of the states at the internal nodes.

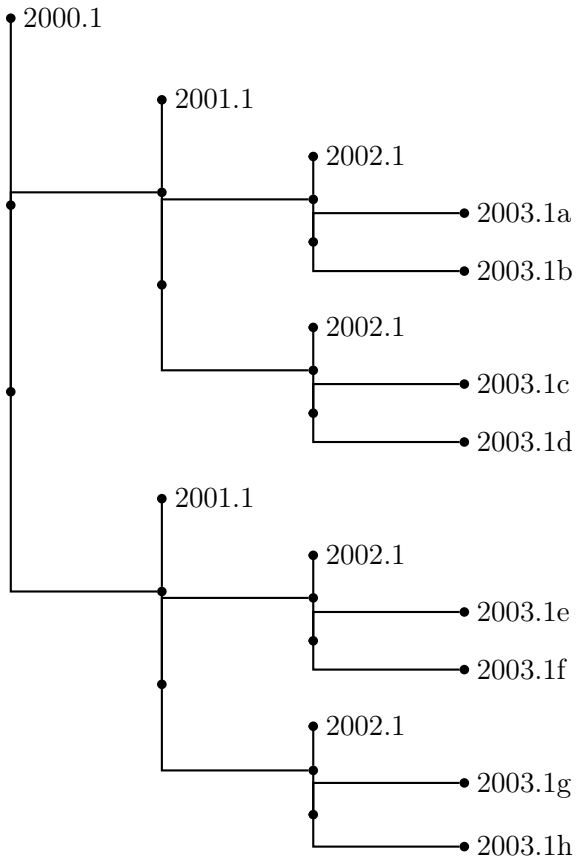


FIGURE 3.2: Informative phylogenetic tree structure (to three generations).

Figure 3.3 provides a more fine-grained view of a hypothetical two-state phylogeographic process on the state space $\{A, B\}$, over the informative tree structure. The phylogeographic process gives rise to the pictured observed states at the leaf nodes, and the unobserved states S_1 and S_2 at the internal nodes. The observed state A at the leaf node arising from the length 0 branch imposes $S_1 = S_2 = A$ for the unobserved states at the internal nodes, as the length 0 branch precludes the possibility of a state change.

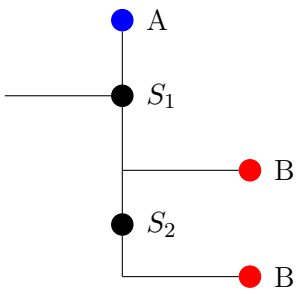


FIGURE 3.3: Fine-grained view of informative phylogenetic tree.

The pictured subset provides evidence to infer a non-zero transition rate Λ_{AB}. If the same pattern – where state A is observed at the blue leaf nodes attached to length 0 branches, and state B is observed at the red leaf nodes resulting from non-zero branches – were to be repeated over the entire tree, the observed states would give strong evidence for a non-zero transition rate Λ_{AB} and a zero transition rate Λ_{BA}.

One of the distinguishing features of BEAST – in which inference of the phylogeographic transition rates $\mathbf{\Lambda}$ was performed via MCMC – is its focus on rooted trees incorporating a time scale (Drummond and Rambaut, 2007). The branch lengths for a given topology are determined by the dates assigned to each leaf node. Each phylogenetic tree structure considered in the simulations was assigned an earliest date of January 2000, and branched at yearly intervals.

3.3.1 Two geographic states

A two-state phylogeographic process was simulated over the informative phylogenetic tree structure with a depth of 7 years. The phylogeographic CTMC was governed by the generator matrix

$$\mathbf{\Lambda} = \begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{pmatrix} -0.2 & 0.2 \\ 0 & 0 \end{pmatrix} \end{array}, \tag{3.4}$$

which was specified to produce strong signal for Λ_{AB} at the leaf nodes in the final generation; that is, the generator matrix should produce mostly observed states A at the leaf nodes of length 0 branches, and mostly observed states B at the leaf nodes in the final generation. This effectively sets most internal nodes to state A, while most terminal nodes would be in state B.

The informative tree structure simulated to a depth of 7 years had 255 leaf nodes, so the simulated phylogeographic process produced 255 observed states. For each simulation, two independent $N(0, 1)$ random variables were generated as the observed values of the spurious predictor of the transition rates $\{\Lambda_{AB}, \Lambda_{BA}\}$.

3.3.2 Five geographic states

A five-state phylogeographic process was simulated over the informative phylogenetic tree structure with a depth of 8 years. The depth of the tree was increased from 7 years as the five-state process requires the inference of 20 transition rates as compared to only 2 transition rates for the two-state process. A greater tree depth results in a greater

number of leaf nodes and observed locations with which to infer the transition rates. The phylogeographic CTMC was governed by the generator matrix[1]

$$\mathbf{\Lambda} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} -0.3 & 0.0 & 0.0 & 0.3 & 0.0 \\ 0.0 & -1.0 & 0.0 & 0.8 & 0.2 \\ 0.0 & 0.0 & -0.9 & 0.9 & 0.0 \\ 0.1 & 1.9 & 0.5 & -2.6 & 0.0 \\ 0.0 & 2.1 & 3.3 & 0.0 & -5.4 \end{pmatrix} \end{array}. \tag{3.5}$$

The entries of the infinitesimal rate matrix in Equation (3.5) were sampled such that all states are mutually accessible; i.e. such that any state i has a non-zero probability of transitioning into any other state j within some time period for $i, j \in \{A, B, C, D, E\}$. The transition rates were randomly sampled from an exponential distribution with rate parameter equal to 1. Each entry had a 50% chance of being set equal to zero, and different random seeds were used until the resulting generator matrix was irreducible.

The informative tree structure to a depth of 8 generations had 511 leaf nodes, so the simulated phylogeographic process produced 511 observed states. For each simulation, the value of the spurious covariate for each of the twenty transition rates in the generator matrix was sampled randomly from the standard normal distribution.

3.4 Influenza-like phylogenetic tree

The influenza-like phylogenetic tree emulates the ladder-like tree structure typical of the influenza virus. Figure 3.4 illustrates the real-world structure of a specific region of the influenza A (H3N2) HA gene, displaying a strong backbone lineage, where branches off this backbone persist for only a short space of time before dying out (Volz, Koelle, and Bedford, 2013).

The ladder-like tree structure of the influenza virus is characteristic of strong selection, where the population is repeatedly replaced by selective sweeps, and Figure 3.5 displays the idealised form of a phylogenetic tree under such selection. For the idealised influenza-like tree structure, all branches off the backbone produce no further branching, and end in leaf nodes.

For phylogeographic simulations on the influenza-like phylogenetic tree, far greater tree depth (as compared to the informative tree structure, or a more balanced branching

[1] Row D in the generator matrix does not sum to 0, as the generator matrix entries have been rounded to a single decimal place for display purposes.

structure in general) was required to produce enough state observations to infer the phylogeographic process with reasonable accuracy.

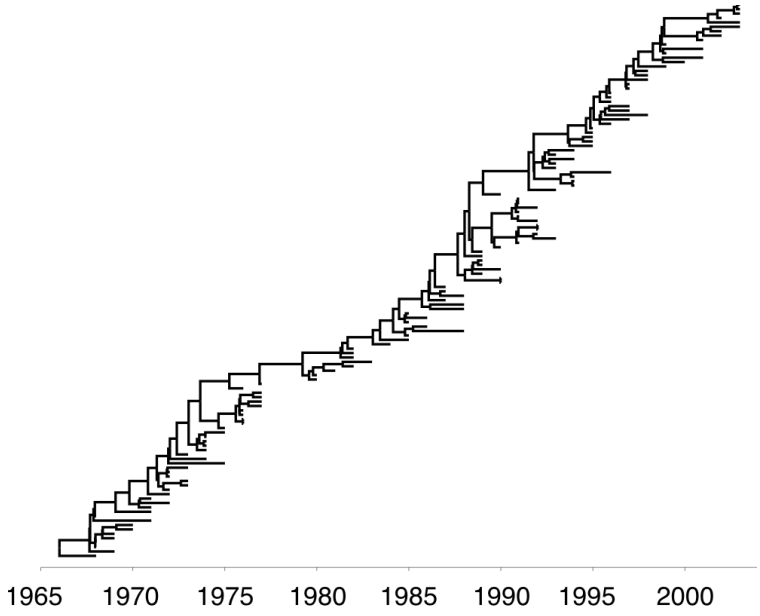


FIGURE 3.4: Phylogenetic tree for the HA1 region of the HA gene of influenza A (H3N2) from viruses sampled between 1968 and 2002 (Volz, Koelle, and Bedford, 2013).

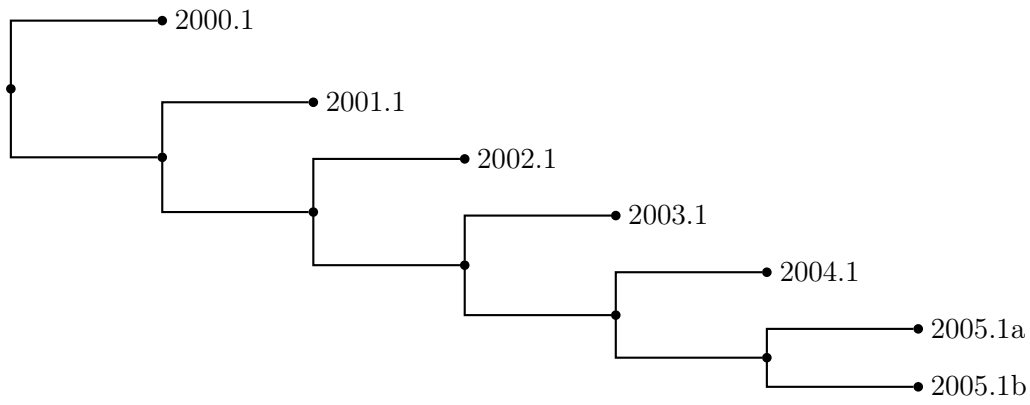


FIGURE 3.5: Influenza-like phylogenetic tree structure (to six generations).

3.4.1 Two geographic states

A two-state phylogeographic process was simulated over the influenza-like phylogenetic tree structure to a depth of 900 years. The phylogeographic CTMC was governed by the generator matrix

$$\mathbf{\Lambda} = \begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{pmatrix} -1.2 & 1.2 \\ 0.3 & -0.3 \end{pmatrix} \end{array}. \tag{3.6}$$

This generator matrix was parametrised to favour transitions from A to B relative to transitions from B to A. The transition rate Λ_{BA} is non-zero to create more state diversity, otherwise a state transition $A \to B$ along the backbone of the tree at any stage would result in only state B being observed at every subsequent leaf node.

The simulated phylogeographic process produced 901 observed states, corresponding to the number of leaf nodes for the influenza-like tree structure to a depth of 900 years. The two observations on the spurious predictor of the transition rates $\{\Lambda_{AB}, \Lambda_{BA}\}$ were generated as independent $N(0, 1)$ random variables.

3.4.2 Five geographic states

A five-state phylogeographic process was simulated over the influenza-like phylogenetic tree structure to a depth of 900 years. The 900 year depth for the two-state process was a conservative choice, and the same tree specification provided adequate evidence to infer the rates for a five-state process. The phylogeographic CTMC was governed by the irreducible generator matrix given by Equation (3.5).

The simulated phylogeographic process resulted in 901 observed states. For each simulation, the twenty transition rates in $\mathbf{\Lambda}$ were modelled as log-linear functions of twenty independently generated observations of a $N(0, 1)$ random variable.

3.5 Multiple covariates

The preceding simulations on the informative and influenza-like phylogenetic tree structures have all considered a single spurious predictor of the transition rates in the generator matrix $\mathbf{\Lambda}$ for the phylogeographic CTMC. Many applications in the literature, however, evaluate multiple predictors of the transition rates (Lemey et al., 2014; Magee et al.,

2015). Simulations were therefore performed to ascertain the effect of multiple spurious predictors of the transition rates for a phylogeographic process.

The multiple covariate simulations were performed using the influenza-like tree structure to a depth of 900 generations. A phylogeographic process governed by the generator matrix in Equation (3.5) (such that all states are mutually accessible) was simulated over the phylogenetic tree.

A phylogeographic state was observed at each of the 901 leaf nodes. For each simulation, twenty observations on each of five $N(0,1)$ spurious predictors of the off-diagonal transition rates in $\mathbf{\Lambda}$ were independently generated.

3.6 Model specification for parameter inference

The phylogenetic model for the simulations was specified in terms of nucleotides, and the model of substitution assumed was the HKY85 model of Hasegawa, Kishino, and Yano (1985). An asymmetric phylogeographic process was assumed to have generated the observed location states. For both the CTMC modelling nucleotide substitution and the phylogeographic CTMC a strict molecular clock was assumed. Specific phylogenetic tree structures were fixed for the model, and neither the tree topology nor branch lengths were inferred.

3.7 MCMC calibration

3.7.1 Prior distributions

Each effect size was modelled *a priori* with $\beta \sim N(0,4)$. The indicator variables were distributed as $\delta \overset{iid}{\sim} \text{Bernoulli}(q)$, which implied that for p predictors $\sum_{k=1}^{p} \delta_k \sim \text{Binomial}(p,q)$. The prior distribution on the indicator variables $\boldsymbol{\delta}$ was specified such that there was a prior mass of 0.5 on no predictors being included in the model, and so

$$q = 1 - 0.5^{1/p}. \tag{3.7}$$

The prior distributions on $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ were identical for the GLMs both with and without random effects. The model with random effects, however, required an additional prior parametrisation for the random effects $\boldsymbol{\epsilon}$. Following the hierarchical formulation of Trovão et al. (2015), each random effect was modelled as $\epsilon \sim N\left(0, \sigma^2\right)$. In turn, the precision of the normal distribution $1/\sigma^2$ was distributed *a priori* according to a gamma

distribution with shape and rate parameters equal to 0.001. This is a vague distribution frequently specified on variance parameters, and is characterised by an approximately uniform density over most of its range, with a peak close to zero (Lambert et al., 2005). Under such a vague prior distribution, the posterior distribution is quickly updated to reflect the information contained in the empirical observations.

A log-normal prior was specified on the transition/transversion rate ratio, such that $\log \kappa \sim N\left(0, 1.25^2\right)$. The equilibrium state frequencies for the phylogenetic and phylogeographic CTMCs were distributed *a priori* as Dirichlet $(1, 1, 1, 1)$. The phylogenetic tree was assumed to arise from a coalescent process *a priori*, and the population size parameter N introduced in Equation (2.13) was estimated from the data. A one-on-X prior was specified on N – one-on-X priors are improper distributions that allocate a density of $1/X$ to the value X, for $0 \leq X \leq \infty$. However, when modelling the coalescent constant population size parameter, a one-on-X prior corresponds to a Jeffrey's prior and leads to a proper posterior distribution (Wu and Drummond, 2011).

The phylogenetic clock rate was distributed *a priori* as Gamma $(\alpha = 0.001, \beta = 0.001)$. An approximation of a conditional reference prior – which is vague, and is derived in Ferreira and Suchard (2008) – was specified on the phylogeographic clock rate. Although both the phylogenetic and phylogeographic clock rates could be derived *a priori* from the exchangeabilities and equilibrium frequencies of the known processes that generated the data, vague priors were used in order to emulate analyses in more general scenarios, such as those characterised by a lack of prior information on these rates. This choice was made as convergence was achieved using vague distributions, and there is possible value in the increased generalisability of this approach for future investigations.

3.7.2 Proposal distributions and operators

In BEAST, operators are functions that implement the proposal distributions for the Metropolis Hastings MCMC algorithm, used to infer phylogenetic and phylogeographic parameters. Operators propose new parameter values and typically change only a small subset of the model parameters for a given state. Some operators are derived from probability distributions, and sample from the conditional distribution of a subset of parameters, given the values of the other parameters. These operators constitute efficient Gibbs samplers, but are only applicable when the conditional distribution of the subset of parameters has a recognisable form (Geman and Geman, 1984). Some of the model parameters for the simulations adhere to this condition, thus motivating the use of these operators in some cases, to be specified shortly.

The operators used to propose parameter values in the simulations are listed below, and their application to this investigation will be described subsequently. In this m.d., operators are generally described according to the convention of Drummond and Bouckaert (2015).

- A scale operator is an operator that picks a random number s and multiplies the values of its parameter by s. A specific scale factor s can also be passed to the operator as an argument. It is typically applied to parameters such as clock rates.

- A delta exchange operator is typically applied to parameter values that are constrained to a fixed sum– such as equilibrium state frequencies for a CTMC. Given a multi-dimensional parameter, the operator chooses two of its values, picks a random number δ that will not violate the parameter constraints and adds δ to the first selected value, while subtracting δ from the second selected value. A specific δ can also be passed to the operator as an argument.

- A bit flip operator chooses a random component of a boolean-valued parameter (such as the indicator variables $\boldsymbol{\delta}$) and changes its value from true to false or from false to true.

- A random walk operator chooses a random value of a real-valued parameter and changes the value by a random amount. In the phylogeographic context, random walk operators are applied to effect sizes $\boldsymbol{\beta}$, among other parameters. A window size W needs to be passed as an argument to the operator; the new proposed value is drawn from within a radius of $2 \times W$ from the current parameter value.

- A multivariate normal operator is applied in the context of a GLM. Given a set of effect sizes and the corresponding design matrix X, and assuming $\boldsymbol{\beta}$ gives the current values for the coefficients, the new values $\boldsymbol{\beta}^*$ are proposed according to the distribution:

$$\boldsymbol{\beta}^* \sim MVN\left(\boldsymbol{\beta}, \alpha\left(X^\top X\right)^{-1}\right),$$

  where α is a variance scalar that the operator tunes itself. This proposal is motivated by potentially high correlations between predictors, resulting in changes to single coefficients at each time step producing high autocorrelation times and impeding the efficiency of the Markov chain. The Multivariate Normal operator was formulated by Lemey et al. (2012) for use with phylogeographic GLMs.

- A normal gamma precision Gibbs (NGPG) operator implements a Gibbs sampler for a normal-gamma model. Given a hierarchical model where $\epsilon|\sigma^2 \sim N\left(0, \sigma^2\right)$

and $1/\sigma^2 \sim \text{Gamma}\,(\alpha, \beta)$, with $K(K-1)$ observed random errors, the operator proposes values of $1/\sigma^2$ from its conditional distribution

$$1/\sigma^2|\alpha, \beta, \boldsymbol{\epsilon} \sim \text{Gamma}\left(\alpha + K(K-1)/2, \beta + 0.5 \times \sum_{i=1}^{K(K-1)} \epsilon_i\right).$$

In general, operators do not propose changes to every subset of parameters at each state. In addition to its unique tuning arguments, each operator is passed a weight, which stipulates how frequently it is applied relative to the other operators. Parameters that take longer to converge are sampled more frequently.

The following elaborates on the application of the aforementioned operators to this investigation. A delta exchange operator was allocated to the equilibrium state frequencies for the phylogenetic CTMC, and similarly a delta exchange operator proposed changes to the equilibrium state frequencies for the phylogeographic CTMC. Scale operators were specified on the clock rates for the phylogenetic and phylogeographic CTMCs, as well as on the transition/transversion ratio κ and the constant coalescent population size parameter.

A bit flip operator was specified on the indicator variables $\boldsymbol{\delta}$, both for the models with and without random effects. A random walk operator was allocated to the effect sizes $\boldsymbol{\beta}$. A multivariate normal operator proposed additional changes to the effect sizes $\boldsymbol{\beta}$, such that both a random walk operator and a multivariate normal operator were used to propose changes to the effect sizes for all simulations. Similarly, a random walk operator as well as a multivariate normal operator were used to propose changes in the estimated random effects $\boldsymbol{\epsilon}$ for each simulation. Finally, an NGPG operator was specified on the precision $1/\sigma^2$ of the normal distribution for the random effects.

Table 3.1 presents all operators used, along with their tuning settings and relative weights. The tuning parameters were dynamically adjusted in BEAST as the MCMC ran for optimal efficiency.

TABLE 3.1: Operators on model parameters

		Operator arguments	
Parameter	Operator type	Tuning parameter	Weight
Phylogenetic equilibrium frequencies	Delta exchange	$\delta = 0.01$	1
Phylogeographic equilibrium frequencies	Delta exchange	$\delta = 0.75$	1
Phylogenetic clock rate	Scale	$s = 0.75$	30
Phylogeographic clock rate	Scale	$s = 0.75$	30
Transition/transversion ratio	Scale	$s = 0.75$	1
Coalescent population size	Scale	$s = 0.75$	30
Indicator variables	Bit flip	–	3
Effect sizes	Random walk	$W = 0.5$	1
Effect sizes	Multivariate normal	$\alpha = 1$	5
Random effects	Random walk	$W = 0.5$	4
Random effects	Multivariate normal	$\alpha = 1$	14
Random effects precision	NGPG	–	12

3.7.3 Convergence

This section elaborates on the diagnostics used to monitor the convergence of the Markov chains used to infer the model parameters.

Each chain had a minimum length of 1 million states. Different chain lengths were trialled and convergence was monitored to inform this choice. Thinning was applied through logging states every 1000 steps, or, for longer chains, whichever interval would produce 1000 samples. A burn-in period of 10% of the retained states was specified for each chain. Trace plots of the parameters were examined and reflected a random scatter about an approximately constant mean and variance.

Figure 3.6 gives an example of a trace plot for a (δ, β) pair in one of the simulations, with the burn-in period indicated in red. All sampled parameters were additionally confirmed to have effective sample sizes in excess of 200, and were therefore equivalent to a minimum of 200 independent samples from the posterior distribution.

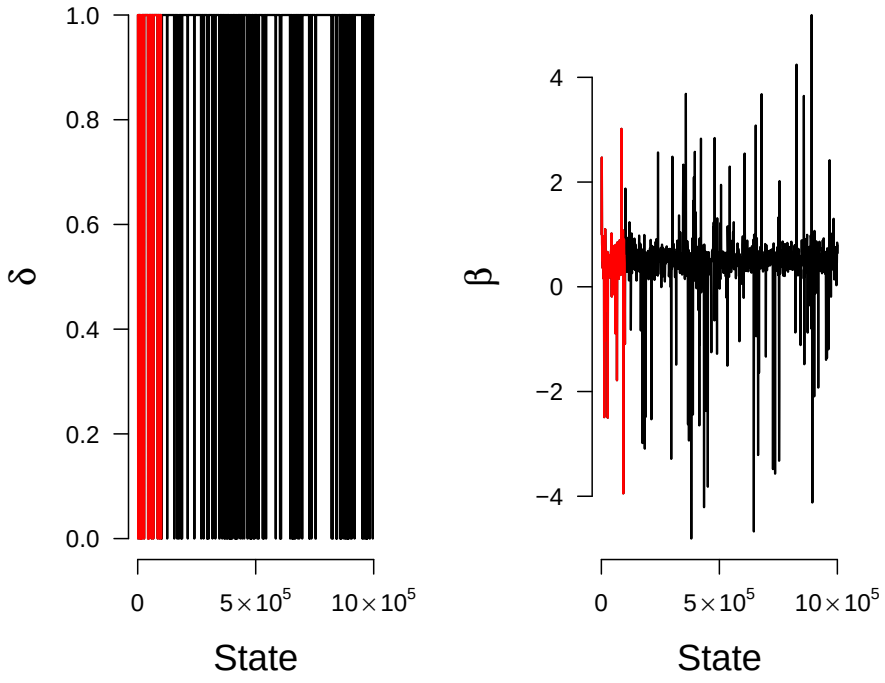


FIGURE 3.6: Trace plot for posterior (δ, β) pair.

3.8 Software

All MCMC simulations were run in BEAST (Drummond, Suchard, et al., 2012), which is currently the only standardised software package that has implemented phylogeographic generalised linear models. Computations in BEAST were performed using facilities provided by the University of Cape Town's ICTS High Performance Computing team[2]. Molecular data was simulated in R using PhyloSim (Sipos et al., 2011). The informative tree structures were generated in Python, and the phylogeographic processes were simulated in R using the APE: Analyses of Phylogenetics and Evolution library (Paradis, Claude, and Strimmer, 2004).

[2]The ICTS High Performance Computing team has requested to be referenced via their URL http://hpc.uct.ac.za.

Chapter 4

Simulations

This chapter presents the results of the simulations that were carried out. For every simulation, the Bayes factor support for the inclusion of spurious predictors in the models with and without random effects is presented visually. Support for the spurious predictor is quantified on the vertical axis, while the horizontal axis distinguishes between support under the model without random effects and support under the model including random effects. Each simulation corresponds to its own unique set of observations on the spurious predictor (or predictors, in the case of the multiple covariate simulations), and its own unique set of observed location states. A simulation is represented by a pair of points on the plot, connected by a straight line.

The Bayes factor support is obtained according to Equation (2.12), where the hypothesis H_1 corresponds to $\delta = 1$, so the Bayes factor is simply the posterior odds that $\delta = 1$, divided by the prior odds that $\delta = 1$. Results are plotted and tabulated on a $2\log_e(B_{10})$ scale instead of a raw Bayes factors B_{10} scale. The transformation narrows the large range of the plot in the presence of strong predictors, for ease of visual interpretation. The guidelines for positive and strong evidence that $\delta = 1$ (see Table 2.1) are marked on the plots in orange and red respectively.

The chapter proceeds as follows: firstly, predictor significance results for the two-state and five-state simulations over the informative tree structure are presented. Next, results for the two-state and five-state simulations over the influenza-like phylogenetic tree structure are discussed. Finally, the results for the simulations using multiple covariates are presented, and the chapter is concluded with a summary of all of the simulations.

4.1 Informative phylogenetic tree

Figure 4.1 plots the results for the two-state simulations and the five-state simulations over the informative phylogenetic tree, respectively.

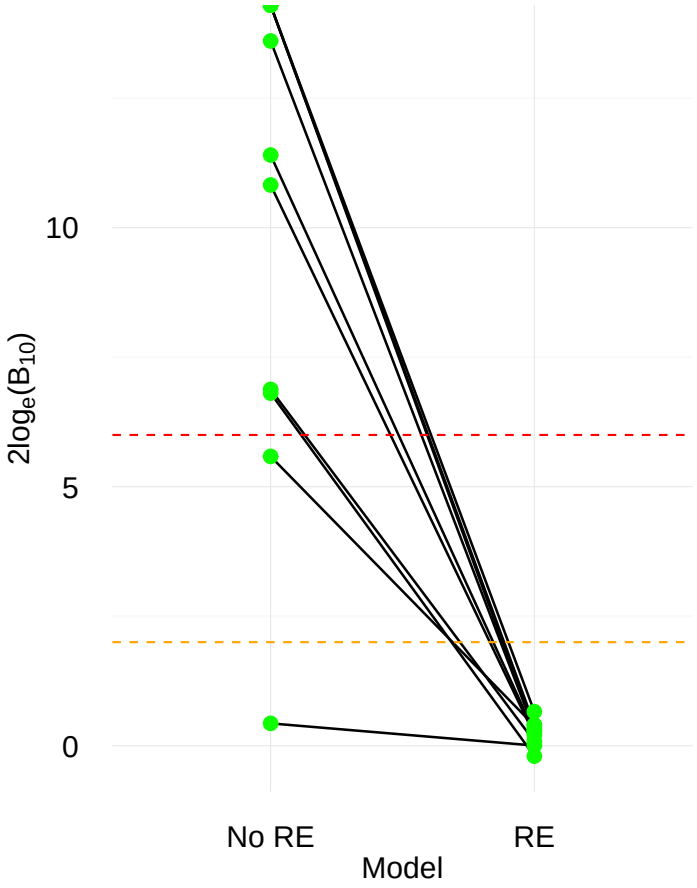
In the two-state simulations, the model without random effects provided very strong evidence that $\delta = 1$ in six of the simulations, and strong evidence that $\delta = 1$ in an additional two simulations (according to the guidelines in Table 2.1). These results are spurious, as the predictors that registered as significant were randomly generated, independently of the phylogeographic transition rates. Only one of the simulations without random effects correctly provided no evidence that $\delta = 1$, while every simulation for the model with random effects provided no evidence for the inclusion of the spurious predictor.

The five-state simulations without random effects identified very strong evidence for the inclusion of the spurious predictor in three of the simulations, and positive evidence for the inclusion of the spurious predictor in one of the simulations. The model with random effects identified no evidence for the inclusion of the spurious predictor across all simulations. Table 4.1 gives the transformed Bayes factors for both the two-state and five-state simulations.

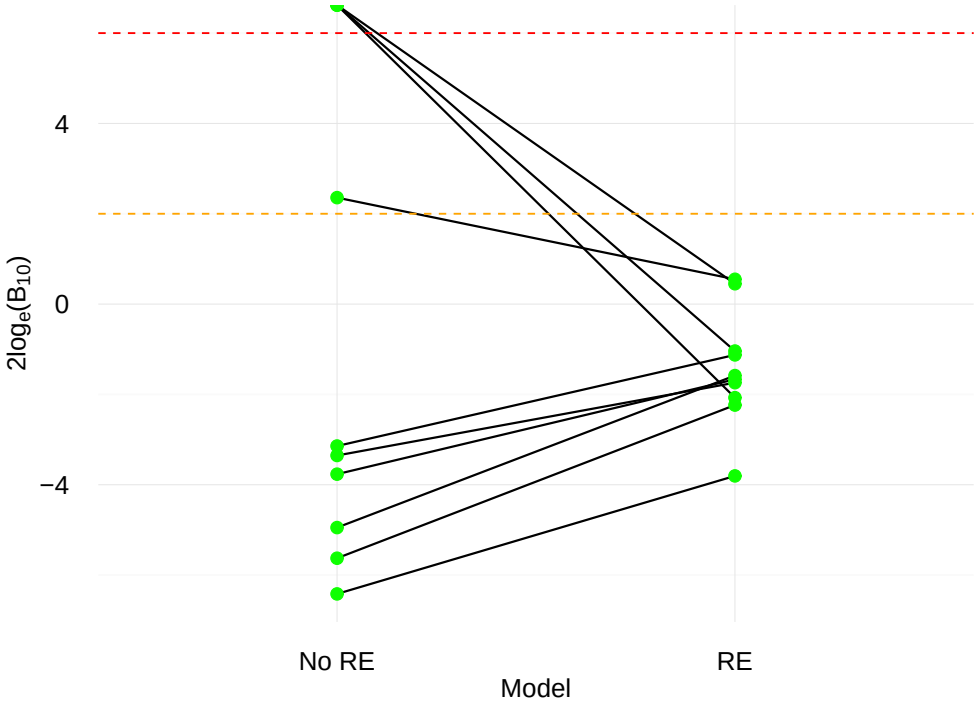
The difference between the false positive rates for the model without random effects in the two-state and five-state simulations can be described using the following reasoning. In the two-state simulations, the standard phylogeographic GLM selects an effect size β_1 and indicator variable value δ_1 that, together with two observations on a randomly generated covariate, replicate the generator matrix according to

$$\begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{pmatrix} -1.2 & 1.2 \\ 0.3 & -0.3 \end{pmatrix} \end{array} = \begin{pmatrix} -\exp(\beta_1\delta_1 x_{12}) & \exp(\beta_1\delta_1 x_{12}) \\ \exp(\beta_1\delta_1 x_{21}) & -\exp(\beta_1\delta_1 x_{21}) \end{pmatrix}, \tag{4.1}$$

where x_{12} and x_{21} are the observations of the random predictor corresponding to rate Λ_{AB} and rate Λ_{BA} respectively. Since the simulations are based on distinct rates Λ_{AB} and Λ_{BA}, the phylogeographic GLM simply has to choose δ_1 and β_1 to approximate the pattern seen in the rates, which is achievable in the vast majority of cases through selecting a reasonable effect size and $\delta_1 = 1$. As the number of entries in the rate matrix increases (for example, in the five-state simulations), the pattern in the rates becomes more complex and increasingly difficult to approximate given a set of observations on the random covariate. Thus, with greater probability, no effect size β_1 that is able to reproduce the pattern in the rates can be found, and the model correctly infers $\delta_1 = 0$.



(a) Two-state phylogeographic process



(b) Five-state phylogeographic process

FIGURE 4.1: $2\log_e(\mathrm{B}_{10})$ values for two-state and five-state phylogeographic processes over the informative phylogenetic tree structure, contrasting predictor significance under the standard and random effects models. The orange line denotes the lower bound for a positive predictor and the red line denotes the lower bound for a strong predictor (Kass and Raftery, 1995).

TABLE 4.1: Bayes factor support for spurious predictors of phylogeographic transition rates (informative tree).

	Two-state process		Five-state process	
Simulation	No random effects	Random effects	No random effects	Random effects
1	10.83	0.21	∞	-1.04
2	13.60	0.30	∞	0.45
3	∞	0.66	-6.42	-3.80
4	5.59	0.39	-4.95	-1.58
5	∞	0.32	-3.14	-1.13
6	6.81	-0.20	∞	-2.07
7	6.88	0.08	-5.63	-2.24
8	∞	0.41	-3.77	-1.67
9	11.40	0.26	-3.35	-1.74
10	0.43	0.00	2.36	0.55

$2\log_e(B_{10})$ values between 0 and 2 provide minimal evidence for the covariate, values between 2 and 6 provide positive evidence, values between 6 and 10 provide strong evidence, and values greater than 10 provide very strong evidence (Kass and Raftery, 1995). ∞ denotes an infinite Bayes factor, which arises when very strong predictor support is provided, and the predictor is included in the model at every MCMC state after burn-in.

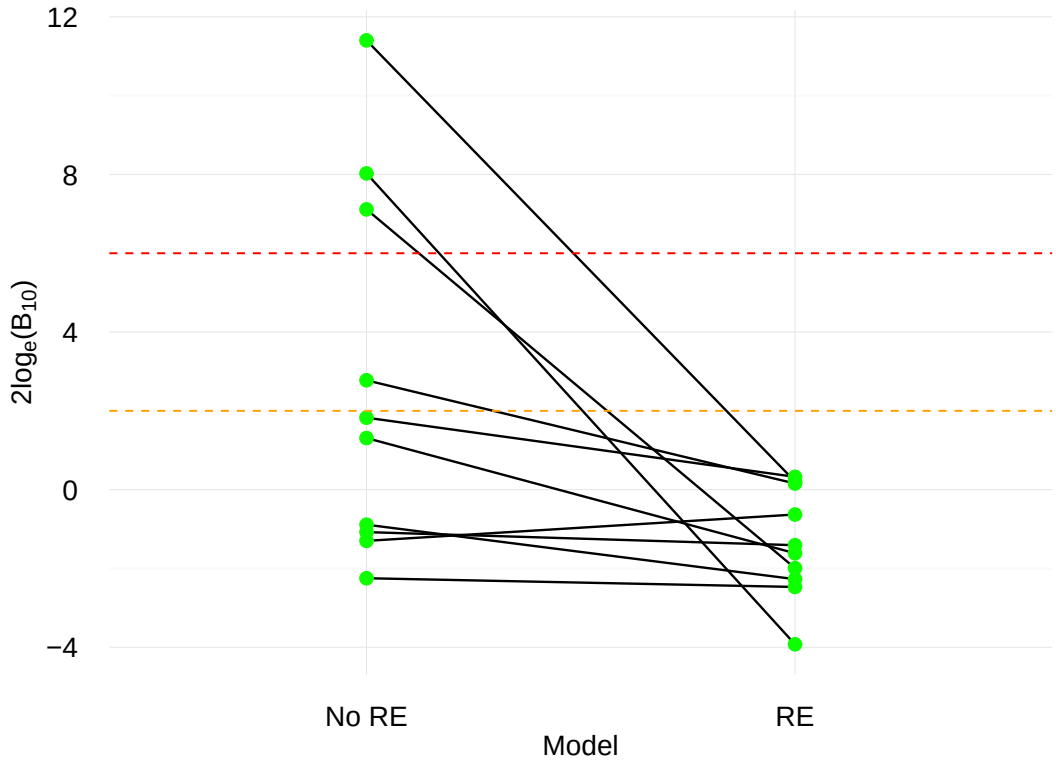
Altogether, substantial evidence in support of spurious predictors was provided by the standard model in both the two-state and five-state simulations, with particularly strong evidence provided in the two-state simulations over the informative phylogenetic tree structure. Additionally, for the standard model, the spread of Bayes factors in the five-state simulations was larger than that for the two-state simulations. The random effects model did not produce any false positives in either the two-state or five-state simulations.

4.2 Influenza-like phylogenetic tree

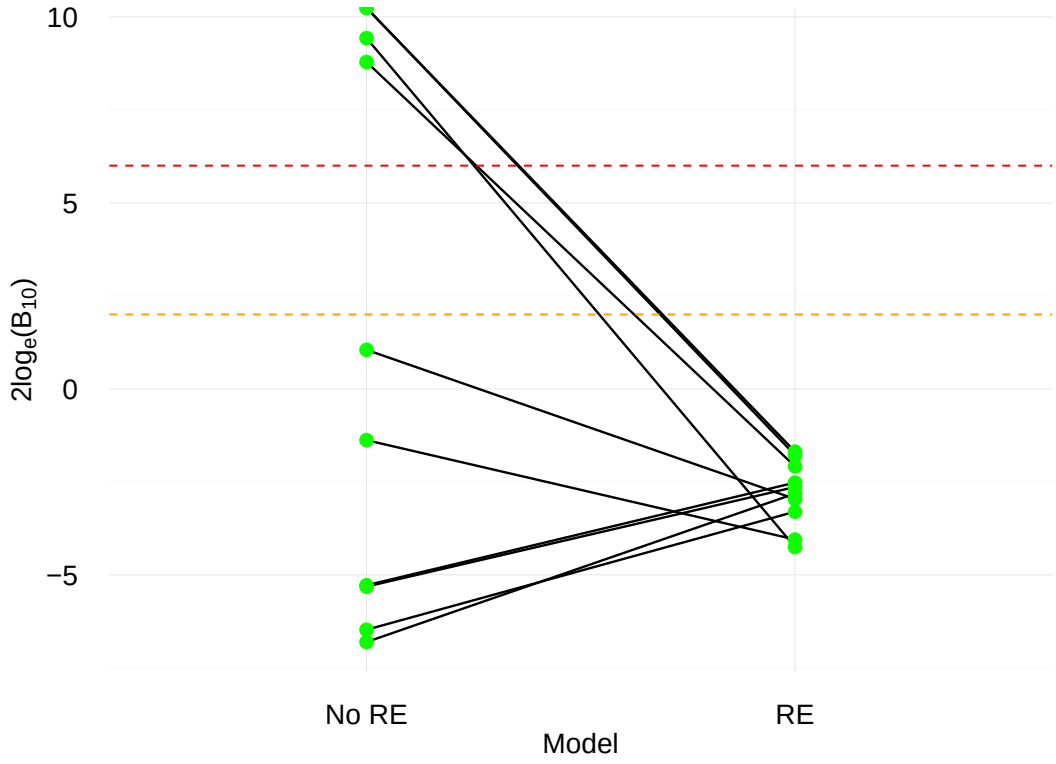
The results for the two-state and five-state simulations over the influenza-like phylogenetic tree are plotted in Figure 4.2. Table 4.2 gives the transformed Bayes factors for all simulations, across both the two-state and five-state phylogeographic processes.

The two-state process consistently identified strong and positive evidence for the inclusion of the spurious predictor, though the support was weaker than that in the informative tree simulations. The five-state simulations also produced a substantial false positive rate, and a greater spread in Bayes factors for the model without random effects, similar to the pattern seen in the informative tree simulations.

As outlined for the informative phylogenetic tree, in general, false positive rates would be expected to decrease with the number of location states. However, given the limited number of simulations, random variation in the predictor values and observed location



(a) Two-state phylogeographic process



(b) Five-state phylogeographic process

FIGURE 4.2: $2\log_e(B_{10})$ values for two-state and five-state phylogeographic processes over the influenza-like phylogenetic tree structure, contrasting predictor significance under the standard and random effects models. The orange line denotes the lower bound for a positive predictor and the red line denotes the lower bound for a strong predictor (Kass and Raftery, 1995).

states mean that this is not a deterministic rule. In fact, for the influenza-like phylogenetic tree, the five-state simulations provided slightly more evidence for the inclusion of the spurious predictor than the two-state simulations. The five-state process provided strong or very strong evidence for the spurious covariate in four of the simulations, while the two-state process provided strong or very strong evidence for the spurious predictor in three of the simulations, and positive or near-positive evidence in two of the simulations.

When the observations of the spurious predictor are mostly orthogonal to the transition rates, the model without random effects correctly rejects the predictor. However, as soon as the covariate correlates slightly more with the transition rates, the spurious predictor is incorrectly identified as highly predictive. The inclusion of a term modelling stochastic noise entirely removes this phenomenon.

TABLE 4.2: Bayes factor support for spurious predictors of phylogeographic transition rates (influenza-like tree).

	Two-state process		Five-state process	
Simulation	No random effects	Random effects	No random effects	Random effects
1	8.03	-3.93	∞	-1.78
2	-0.89	-2.27	-5.28	-2.52
3	-2.25	-2.47	-5.32	-2.64
4	1.31	-1.61	8.79	-2.08
5	11.40	0.21	∞	-1.69
6	-1.08	-1.41	1.05	-2.96
7	7.11	-1.99	-6.81	-2.82
8	2.78	0.16	9.43	-4.25
9	1.83	0.32	-1.38	-4.05
10	-1.30	-0.63	-6.48	-3.30

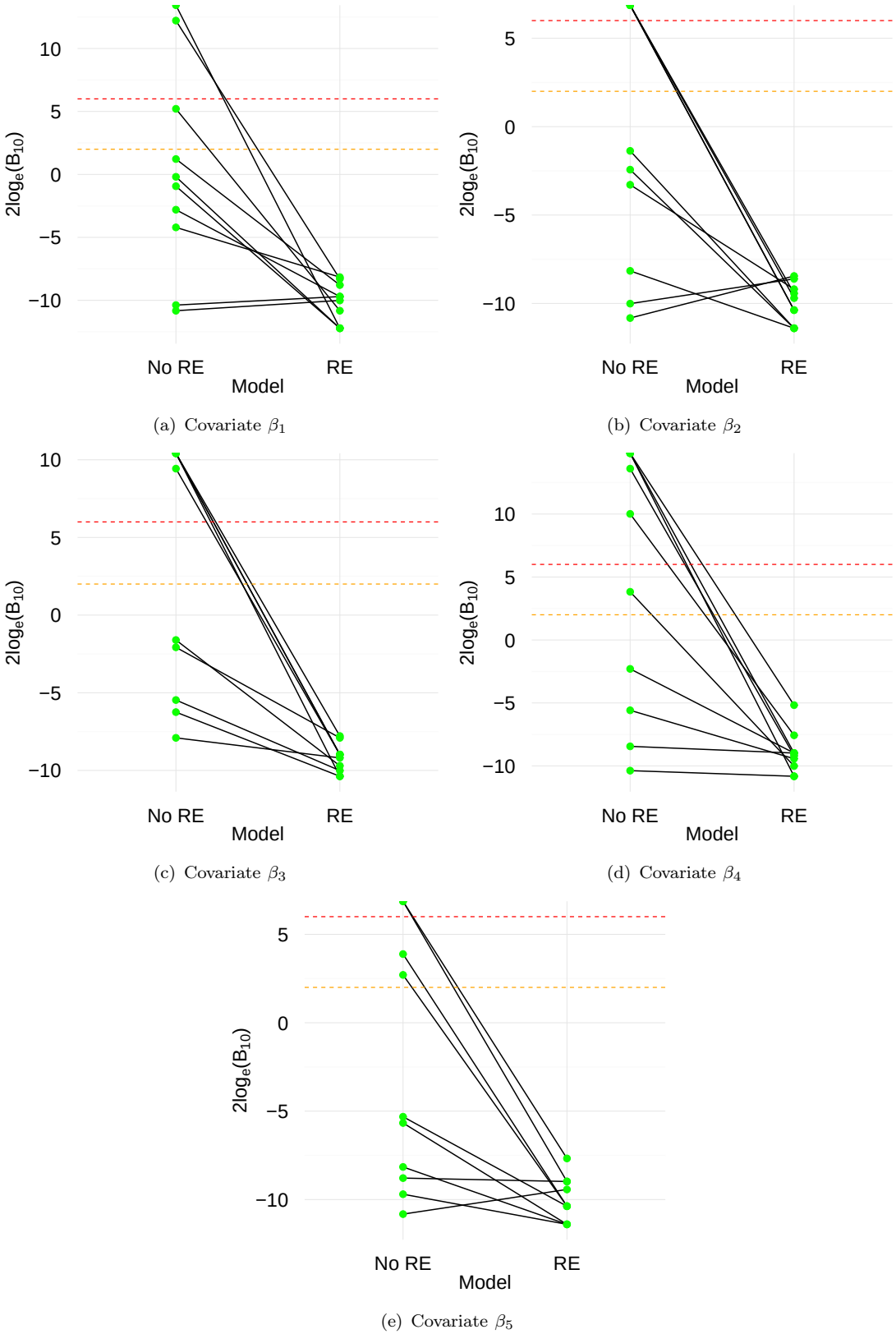
$2\log_e(B_{10})$ values between 0 and 2 provide minimal evidence for the covariate, values between 2 and 6 provide positive evidence, values between 6 and 10 provide strong evidence, and values greater than 10 provide very strong evidence (Kass and Raftery, 1995). ∞ denotes an infinite Bayes factor, which arises when very strong predictor support is provided, and the predictor is included in the model at every MCMC state after burn-in.

4.3 Multiple covariates

Figure 4.3 presents the results for the five-state simulations over the influenza-like phylogenetic tree, using five randomly generated covariates. The plots in the figure quantify the Bayes factor support for each of the five spurious predictors separately, according to both the models with and without random effects, across all simulations. The transformed Bayes factors for the simulations according to each covariate are tabulated in Table A.1 and Table A.2.

Using the standard model, a median of four simulations provided strong evidence for the inclusion of each of the five covariates. Without random effects, the simulations in general identified either evidence against the inclusion of a given covariate, or very strong evidence for its inclusion, resulting in the widest spread of Bayes factors across all simulation structures.

This behaviour is very similar to that seen in the five-state simulations over the influenza-like tree, and can be explained using similar reasoning. With the inclusion of five spurious covariates, for each simulation there is a greater probability that a covariate will be identified as highly predictive. This observation corresponds to the empirical results, which exhibited a 90% false positive rate: nine of the ten simulations without random effects identified at least one covariate as highly predictive, and half of the simulations without random effects identified multiple covariates as highly predictive. In comparison, the simulations using the model with random effects had a 0% false positive rate: no simulation using the random effects model provided evidence for the inclusion of a spurious predictor.



(a) Covariate β_1

(b) Covariate β_2

(c) Covariate β_3

(d) Covariate β_4

(e) Covariate β_5

FIGURE 4.3: $2\log_e(\mathrm{B}_{10})$ values for five-state phylogeographic processes over the influenza-like phylogenetic tree structure, contrasting predictor significance for sets of five randomly generated covariates under the standard and random effects models. The orange line denotes the lower bound for a positive predictor and the red line denotes the lower bound for a strong predictor (Kass and Raftery, 1995).

4.4 Summary of simulation results

Multiple simulations were performed on two different phylogenetic tree structures: informative and influenza-like. Both two-state and five-state phylogeographic processes were simulated, with the crucial aim of quantifying support for spurious predictors of phylogeographic transition rates according to the GLM formulations with and without random effects. Predictor support was measured using Bayes factors, the evidence provided by which was classified according to the guidelines of Kass and Raftery (1995) in Table 2.1.

The model with random effects did not identify evidence for the inclusion of spurious predictors in any of the simulation cases. In fact, in the simulations with multiple covariates, the model with random effects identified substantial evidence against the inclusion of the spurious predictors in many simulations.

The model without random effects was susceptible to identifying false positives across all simulation structures. Particularly strong and consistent evidence for the inclusion of the spurious predictor was provided in the two-state case using the informative phylogenetic tree structure. The five-state cases identified strong evidence for the inclusion of the spurious covariate, with less frequency than the two-state case for the informative phylogenetic tree, but with slightly increased frequency relative to the two-state case when applied to the influenza-like phylogenetic tree. Each of the five-state models also exhibited a greater spread of Bayes factors than the two-state models. In general, the false positive rate is expected to decrease with the number of location states, due to the decreased probability of a randomly generated predictor correlating by chance with twenty transition rates compared to two transition rates.

In the multiple covariate simulations, the model with random effects had a 0% false positive rate, while the model without random effects had a 90% false positive rate. That is, no covariate was identified as highly predictive across any of the simulations with random effects, while at least one spurious covariate registered as significant in nine of the ten simulations without random effects. This demonstrates that in the model without random effects, false positives are more likely with multiple spurious covariates, while the model with random effects is robust against false positives, even in the presence of multiple spurious covariates.

Altogether, the simulations demonstrate that false positives are a serious concern in the standard phylogeographic GLM formulation without random effects, exacerbated in the presence of multiple spurious covariates. Conversely, the model with random effects proved robust to false positives.

Chapter 5

Applications

In this chapter, three datasets from the literature that make use of phylogeographic generalised linear models are analysed, and the support for various epidemiological predictors using the random effects model and the standard model is contrasted. Gräf et al. (2015) and Magee et al. (2015) modelled the diffusion of HIV-1 Subtype C in Brazil and Influenza A Subtype H5N1 in Egypt, respectively, as functions of multiple epidemiological covariates under the standard phylogeographic generalised linear model. We apply the random effects model to these datasets, contrasting the results to those obtained using the standard model. Trovão et al. (2015) introduced the random effects model applied in this m.d., so it is of interest to contrast the behaviour of their dataset under the standard model. These datasets are recent contributions to the literature, having all been published in 2015.

5.1 HIV-1 Subtype C in Brazil

Gräf et al. (2015) modelled the phylogeographic history of the HIV-1 Subtype C epidemic in Brazil. The authors' methodology is summarised below.

Sequence data were sampled from 22 Brazilian locations, and these constituted the state space over which the phylogeographic process was modelled. The authors analysed several sub-sampled datasets, making results more robust against sample biases. 1552 *pol* sequences and 621 *env* sequences were originally collected across all locations. The data was then down-sampled to approximately 500 sequences for each gene, and again sub-sampled according to two schemes: Rand10, where a maximum of 10 sequences were sampled per location, and Rand20 where a maximum of 20 sequences were sampled per location. For each of these schemes the authors generated 3 random sub-samplings, creating the datasets {Rand10A, Rand10B, Rand10C} and {Rand20A, Rand20B, Rand20C}.

For each dataset, the authors evaluated a set of epidemiological predictors according to the phylogeographic GLM without random effects given in Equation (3.2). In their analysis, they tested geographic distance, sample size, population size, HIV prevalence, HIV-1C prevalence, and HIV-1C population size, each stratified into origin and destination, across all phylogeographic transition rates. Predictor significance was evaluated using Bayes factors, with a 50% prior mass on no epidemiological predictors being included in the model.

In our analyses we re-ran the phylogeographic models for the Brazilian HIV-1C *pol* sequences on the complete dataset and all six sub-sampled datasets, replicating the Bayes factor results for predictors evaluated via the GLM without random effects, then re-parametrising the phylogeographic process according to the GLM with random effects in Equation (3.3) and assessing predictor support. Table 5.1 and Table 5.2 display the original Bayes factors published in Gräf et al. (2015), as well as the replicated Bayes factors and the Bayes factors for the predictors when random effects were included, for the complete and Rand10A datasets. The results for the remaining datasets are given in Tables A.3 to A.7.

Across all datasets, the replicated Bayes factors that we simulated closely approximated the original Bayes factors. The inclusion of random effects did not cause any predictors to fall out of significance, or bring any new predictors into significance in any of the datasets. For the Rand10A, Rand10C and Rand20C datasets, the inclusion of random effects substantially decreased the Bayes factor for the predictor destination sample size, but it remained a very strong predictor in the model.

TABLE 5.1: Bayes factor support for an explanatory role in the Brazilian HIV-1C *pol* sequence diffusion process: complete dataset.

Predictor	Gräf et al. (2015)	Replicated	Random effects
Geographic distance	0.1	0.1	0.1
Origin sample size	∞	∞	∞
Destination sample size	∞	∞	∞
Origin HIV population size	0.6	0.6	0.6
Destination HIV population size	0.0	0.0	0.0
Origin HIV prevalence	0.2	0.3	0.2
Destination HIV prevalence	0.0	0.0	0.1
Origin HIV-1C population size	0.3	0.3	0.3
Destination HIV-1C population size	0.0	0.0	0.0
Origin HIV-1C prevalence	0.3	0.2	0.2
Destination HIV-1C prevalence	0.0	0.0	0.0

B_{10} values between 1 and 3 provide minimal evidence for the covariate, values between 3 and 20 provide positive evidence, values between 20 and 150 provide strong evidence, and values greater than 150 provide very strong evidence (Kass and Raftery, 1995). ∞ denotes an infinite Bayes factor, which arises when very strong predictor support is provided, and the predictor is included in the model at every MCMC state after burn-in.

TABLE 5.2: Bayes factor support for an explanatory role in the Brazilian HIV-1C *pol* sequence diffusion process: Rand10A dataset.

Predictor	Gräf et al. (2015)	Replicated	Random effects
Geographic distance	0.0	0.0	0.0
Origin sample size	0.5	0.6	0.5
Destination sample size	6583.3	4603.2	605.9
Origin HIV population size	1.4	1.2	1.2
Destination HIV population size	0.0	0.0	0.0
Origin HIV prevalence	6.1	6.2	7.0
Destination HIV prevalence	0.0	0.0	0.0
Origin HIV-1C population size	14.4	13.0	14.9
Destination HIV-1C population size	0.0	0.1	0.1
Origin HIV-1C prevalence	7.2	7.8	5.7
Destination HIV-1C prevalence	0.0	0.0	0.0

5.2 Influenza A Subtype H5N1 in Egypt

Magee et al. (2015) studied the spread of the H5N1 influenza A virus in Egypt, using phylogeographic GLMs to model viral diffusion and assess a number of epidemiological predictors of this diffusion. The authors used the dataset of Scotch et al. (2013), comprising 226 haemagglutinin sequences of the H5N1 influenza virus variant subclade 2.2.1.1. The spatial data was divided into 20 discrete location states, which correspond to governorates in Egypt.

H5N1 is transmitted from avian species to humans and is highly pathogenic, posing a substantial threat to both human and animal health. Numerous factors across a variety of categories (such as animal populations and climatic measurements) could influence the spread of the H5N1 influenza A virus, and Magee et al. (2015) used the phylogeographic GLM framework to evaluate multiple drivers of viral diffusion simultaneously. The authors incorporated 22 log-transformed predictors into a phylogeographic GLM with the form given in Equation (3.2). The covariates were all pairwise measures, and each was stratified into either a measure at the geographic origin or destination for each transition rate.

We reproduced the original analysis under the standard model, then analysed the dataset using the random effects model defined by Equation (3.3). These results are presented according to governorate of origin and destination in Table 5.3 and Table 5.4 respectively. In our analysis we did not have access to the same set of empirical trees used in the original paper, and inferred a set of empirical trees separately from the data, resulting

in substantial deviations of the replicated results from the Bayes factors in the original study.

TABLE 5.3: Bayes factor support for an exploratory role in H5N1 diffusion in Egypt (governorate of origin).

Predictor	Magee et al. (2015)	Replicated	Random effects
Distance	0.46	0.16	0.97
Latitude	9.51	2.37	1.55
Longitude	20.35	3.80	3.38
Human density	15.08	5.02	10.91
Avian counts	80.28	984.51	38.22
Human counts	12.69	1.78	5.62
Avian density	22.87	5.82	13.70
Chicken density	15.63	3.61	8.34
Turkey density	5.50	0.80	2.35
Duck density	13.20	2.09	65.99
Goose density	20.24	5.56	14.74
Pigeon density	21.45	5.04	12.77
No motif density	16.78	4.53	6.10
Elevation	14.99	11.93	4.05
Precipitation	13.64	4.85	4.41
Temperature	7.13	0.75	1.84
Humidity	9.21	0.81	7.79

B_{10} values between 1 and 3 provide minimal evidence for the covariate, values between 3 and 20 provide positive evidence, values between 20 and 150 provide strong evidence, and values greater than 150 provide very strong evidence (Kass and Raftery, 1995).

As observed in the original analysis, predictors from the governorate of destination were not identified as significant, with the exception of avian counts. The authors identified avian counts for the governorate of origin and the governorate of destination as being strong and very strong predictors of spatial diffusion respectively, and argue that this was likely as a result of sampling differentiation between locations. Interestingly, in the analysis under the random effects model, destination avian counts was identified as a weakly positive predictor instead of a very strong predictor. Another noteworthy contrast between the standard and random effects models was presented by origin duck density, which Magee et al. (2015) found to be a positive predictor of spatial diffusion. In comparison, the model incorporating random effects identified origin duck density as a strong predictor.

The large number of correlated predictors in this dataset made the analysis challenging. After run times in excess of a week, a small subset of model parameters still had effective sample sizes substantially below 100. The indicator variables and effect sizes associated with duck density at governorate of origin were among the parameters that failed to converge. Figure 5.1 displays the trace plot for the effect size of duck density

at governorate of origin, which clearly has not converged to an approximate random scatter about a constant mean. This indicates non-convergence of the Markov chain, or a possible bimodal posterior distribution.

TABLE 5.4: Bayes factor support for an exploratory role in H5N1 diffusion in Egypt (governorate of destination).

Predictor	Magee et al. (2015)	Replicated	Random effects
Distance	0.46	0.16	0.97
Latitude	0.13	0.09	1.07
Longitude	0.08	0.15	0.57
Human density	0.37	0.13	1.83
Avian counts	28058.39	42352.51	5.90
Human counts	0.16	0.13	0.98
Avian density	0.62	0.39	0.84
Chicken density	0.51	0.19	1.12
Turkey density	0.11	0.10	0.92
Duck density	0.46	0.14	4.45
Goose density	0.73	0.65	0.98
Pigeon density	0.59	0.89	1.20
No motif density	0.67	0.19	3.11
Elevation	0.29	0.15	0.52
Precipitation	0.08	0.10	0.87
Temperature	0.13	0.07	0.64
Humidity	0.13	0.08	0.67

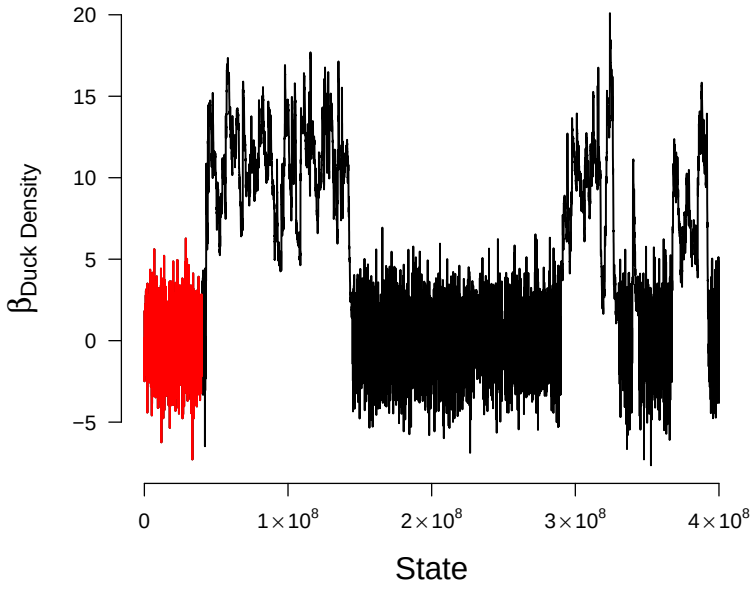


FIGURE 5.1: Trace plot for duck density (governorate of origin).

5.3 Influenza A Subtype H5N1 in Asia and Russia

The final dataset that was analysed was that of Trovão et al. (2015), in which the random effects phylogeographic GLM parametrisation was introduced. The authors studied the epidemic expansion of influenza A H5N1 across regions in Russia and Asia, using a dataset constructed from 806 HA and NA sequences sampled from different avian hosts. For the modelling with phylogeographic GLMs, the spatial data consisted of 19 discrete location states.

In our analysis, diffusion rates were modelled as a function of geographic distance, as well as host density at the origin and destination locations. This differed from the approach of Trovão et al. (2015), who only used geographic distance as a predictor. Before inclusion in the model, all predictors were log-transformed and standardised, as in Trovão et al. (2015). Our analysis consisted of re-evaluating the phylogeographic GLM with random effects published in Trovão et al. (2015), using all of the aforementioned predictors. We then re-parametrised the GLM according to the standard formulation without random effects in Equation (3.2), and contrasted predictor support under the two models.

Table 5.5 gives the Bayes factors associated with each of the three predictors of the spread of influenza A H5N1. As in Trovão et al. (2015), geographic distance was an extremely well-supported predictor, and was always included in both the models with and without random effects. Support for origin and destination host densities presented a more interesting contrast. In the random effects model, both covariates had Bayes factors of approximately zero, suggesting no evidence for their inclusion. However, in the model without random effects, although support for origin host density was negligible, destination host density was associated with a Bayes factor of 5.95, which constituted positive support under the Kass and Raftery (1995) guidelines in Table 2.1.

TABLE 5.5: Bayes factor support for an exploratory role in Influenza A H5N1 diffusion in Russia and Asia.

Predictor	Random effects	No random effects
Distance	∞	∞
Origin host density	0.25	0.18
Destination host density	0.15	5.95

B_{10} values between 1 and 3 provide minimal evidence for the covariate, values between 3 and 20 provide positive evidence, values between 20 and 150 provide strong evidence, and values greater than 150 provide very strong evidence (Kass and Raftery, 1995). ∞ denotes an infinite Bayes factor, which arises when very strong predictor support is provided, and the predictor is included in the model at every MCMC state after burn-in.

Although the contrast in support for destination host density under the two phylogeographic GLM formulations is not as striking as that for covariates in the simulations, it does present a case in a real-world dataset where the two models disagree over predictor support. The model with random effects identified no support for the origin and destination host density covariates, and ultimately the analysis in Trovão et al. (2015) included only geographic distance as a predictor of transition rates. However, modelling under the standard GLM formulation without random effects (depending on author interpretation), could suggest the inclusion of destination host density as a predictor of geographic transition rates.

5.4 Summary of applications

The standard and random effects phylogeographic GLM formulations were applied to contrast predictor support in three recently published datasets from the phylogeographic GLM literature.

For the dataset of Gräf et al. (2015), a set of 11 predictors of the transition rates between 22 location states was assessed. In this dataset, predictor significance under the two models generally corresponded closely, with substantial decreases in support for certain highly-significant predictors when random effects were included. Notably, none of these decreases altered the classification of the support according to the Kass and Raftery (1995) guidelines in Table 2.1.

The Magee et al. (2015) dataset modelled transition rates between 20 location states as a function of 33 covariates. Although inference of a different set of empirical trees resulted in the reproduced Bayes factors (without random effects) deviating from those in the original paper, the inclusion of random effects resulted in significant changes in support for some predictors. The most notable change in predictor support was for destination avian counts, which was classified as a very strong predictor without random effects, but only as a positive predictor under the random effects model. It should be noted, however, that some parameters in this analysis failed to converge, likely due to high correlations between covariates and the large size of the dataset.

Finally, the Trovão et al. (2015) dataset was used to test 3 predictors of the transition rates between 19 location states. Notably, destination host density was identified as a positive predictor under the model without random effects, but not in the model including random effects.

As an aside, the current specification for the random effects model requires the precision of the random effects to be estimated from the data, and adds $K(K-1)$ random effects

to the model for a K-state phylogeographic model. On the real world datasets, in order to obtain adequate effective sample sizes for the additional model parameters, run-times of between 4 and 8 times longer than those of the model without random effects were required. For the Magee et al. (2015) dataset, this translated into run-times spanning multiple days, although it should be noted that 33 covariates and 380 transition rates constitutes a large dataset in this application, and so this may be deemed acceptable.

Chapter 6

Conclusions and Recommendations

In Chapter 4, our simulation results across a range of phylogenetic trees and phylogeographic processes established that the standard phylogeographic GLM is susceptible to false positive results on covariates not associated with the phylogeographic transition rates. The false positive rate decreased with the number of discrete location states for the influenza-like phylogenetic tree, but was still substantial for a five-state phylogeographic process (spurious covariates were identified as strong or very strong predictors in between 30% and 40% of simulations). The false positive rate was positively associated with the number of spurious covariates in the model, and for a five-state phylogeographic process with five spurious covariates, at least one spurious covariate was identified as highly predictive in 90% of the simulations. In every simulation setting, the phylogeographic GLM including random effects was entirely robust to spurious predictors, with a false positive rate of 0%.

In Chapter 5, predictor significance according to the standard and random effects models was contrasted for the recently published datasets of Gräf et al. (2015), Magee et al. (2015) and Trovão et al. (2015). For the Gräf et al. (2015) data, the models including and excluding random effects generally corresponded in their classification of predictor significance across multiple sub-sampled datasets. However, both the Magee et al. (2015) and Trovão et al. (2015) datasets presented cases where the model without random effects identified substantial evidence for the inclusion of certain predictors, while the evidence for their inclusion according to the model with random effects was negligible. This demonstrates potential susceptibility of the standard phylogeographic GLM to false positive results for covariates in real world datasets.

Outside of controlled simulation scenarios, cases where the model without random effects suggests a significant predictor but the model with random effects suggests negligible support could either be interpreted as corresponding to insignificant predictors, or as the

random effects model lacking the power to identify truly significant covariates. In the applications to published datasets, predictor significance detections tended to be upheld for the random effects model. This suggests that there is not necessarily a substantial decrease in power as compared to the standard phylogeographic GLM. A detailed power simulation contrasting the two models would clarify this question, and stands as a future avenue for research.

Our simulation and application results provide evidence that spurious covariates pose a substantial challenge to the standard phylogeographic GLM formulation. The model including random effects proposed by Trovão et al. (2015) effectively accounts for random variation in the model and addresses the concern of spurious predictors. Although the models correspond closely in predictor support in many cases, without random effects or a similar approach to account for stochasticity in the transition rates, significant predictors determined using the standard phylogeographic GLM may in fact correspond to spurious covariates.

Appendices

Appendix A. Supplementary Material

TABLE A.1: Bayes factor support for predictors $\beta_1, \ldots, \beta_4$ in multiple covariate simulations.

	β_1		β_2	
Simulation	No random effects	Random effects	No random effects	Random effects
1	1.22	-8.79	-3.29	-9.19
2	-4.21	-8.16	-10.83	-8.45
3	-10.83	-10.01	-10.01	-8.61
4	-2.80	-9.70	∞	-10.38
5	-10.38	-9.70	-8.16	-11.40
6	∞	-12.22	∞	-10.38
7	5.21	-10.83	-2.43	-11.40
8	-0.94	-12.22	∞	-9.43
9	12.22	-8.30	∞	-9.70
10	-0.19	-12.22	-1.37	-11.40
	β_3		β_4	
Simulation	No random effects	Random effects	No random effects	Random effects
1	-7.90	-9.19	∞	-8.98
2	-5.47	-10.01	∞	-10.83
3	∞	-7.79	-5.59	-9.43
4	∞	-10.38	3.82	-10.01
5	-6.25	-10.38	-8.45	-8.98
6	-2.07	-7.90	-2.30	-8.98
7	∞	-8.98	10.01	-7.57
8	-1.60	-9.70	-10.38	-10.83
9	9.43	-8.98	∞	-5.18
10	∞	-8.98	13.60	-9.19

$2\log_e(\mathrm{B}_{10})$ values between 0 and 2 provide minimal evidence for the covariate, values between 2 and 6 provide positive evidence, values between 6 and 10 provide strong evidence, and values greater than 10 provide very strong evidence (Kass and Raftery, 1995). ∞ denotes an infinite Bayes factor, which arises when very strong predictor support is provided, and the predictor is included in the model at every MCMC state after burn-in.

TABLE A.2: Bayes factor support for predictor β_5 in multiple covariate simulations.

Simulation	No random effects	Random effects
1	2.71	-10.38
2	-8.16	-11.40
3	-8.79	-8.98
4	3.88	-10.38
5	-10.83	-9.43
6	∞	-8.98
7	∞	-7.68
8	-9.70	-11.40
9	-5.32	-10.38
10	-5.67	-11.40

TABLE A.3: Bayes factor support for an explanatory role in the Brazilian HIV-1C *pol* sequence diffusion process: Rand10B dataset.

Predictor	Gräf et al. (2015)	Replicated	Random effects
Geographic distance	0.0	0.0	0.0
Origin sample size	0.4	0.4	0.4
Destination sample size	5758.4	2458.8	3062.3
Origin HIV population size	1.1	1.1	1.1
Destination HIV population size	0.0	0.0	0.0
Origin HIV prevalence	10.8	10.3	10.2
Destination HIV prevalence	0.0	0.0	0.0
Origin HIV-1C population size	22.4	22.9	23.0
Destination HIV-1C population size	0.1	0.1	0.1
Origin HIV-1C prevalence	1.3	1.5	1.4
Destination HIV-1C prevalence	0.0	0.0	0.0

B_{10} values between 1 and 3 provide minimal evidence for the covariate, values between 3 and 20 provide positive evidence, values between 20 and 150 provide strong evidence, and values greater than 150 provide very strong evidence (Kass and Raftery, 1995). ∞ denotes an infinite Bayes factor, which arises when very strong predictor support is provided, and the predictor is included in the model at every MCMC state after burn-in.

TABLE A.4: Bayes factor support for an explanatory role in the Brazilian HIV-1C *pol* sequence diffusion process: Rand10C dataset.

Predictor	Gräf et al. (2015)	Replicated	Random effects
Geographic distance	0.1	0.1	0.1
Origin sample size	0.4	0.4	0.4
Destination sample size	1674.5	2598.9	677.4
Origin HIV population size	1.6	1.7	1.8
Destination HIV population size	0.0	0.0	0.0
Origin HIV prevalence	16.1	13.9	15.0
Destination HIV prevalence	0.0	0.0	0.0
Origin HIV-1C population size	9.6	10.6	10.1
Destination HIV-1C population size	0.1	0.1	0.1
Origin HIV-1C prevalence	4.1	4.4	4.1
Destination HIV-1C prevalence	0.0	0.0	0.0

TABLE A.5: Bayes factor support for an explanatory role in the Brazilian HIV-1C *pol* sequence diffusion process: Rand20A dataset.

Predictor	Gräf et al. (2015)	Replicated	Random effects
Geographic distance	0.0	0.0	0.0
Origin sample size	0.4	0.3	0.4
Destination sample size	∞	∞	∞
Origin HIV population size	1.2	1.2	1.1
Destination HIV population size	0.0	0.0	0.0
Origin HIV prevalence	17.3	15.5	16.6
Destination HIV prevalence	0.0	0.0	0.0
Origin HIV-1C population size	18.1	19.0	18.4
Destination HIV-1C population size	0.0	0.0	0.0
Origin HIV-1C prevalence	1.1	1.2	1.1
Destination HIV-1C prevalence	0.0	0.0	0.0

TABLE A.6: Bayes factor support for an explanatory role in the Brazilian HIV-1C *pol* sequence diffusion process: Rand20B dataset.

Predictor	Gräf et al. (2015)	Replicated	Random effects
Geographic distance	0.0	0.0	0.0
Origin sample size	0.4	0.4	0.4
Destination sample size	∞	∞	∞
Origin HIV population size	1.1	1.1	1.3
Destination HIV population size	0.0	0.0	0.0
Origin HIV prevalence	13.1	12.7	12.6
Destination HIV prevalence	0.0	0.0	0.0
Origin HIV-1C population size	23.4	22.7	19.3
Destination HIV-1C population size	0.0	0.0	0.0
Origin HIV-1C prevalence	1.4	1.7	2.3
Destination HIV-1C prevalence	0.0	0.0	0.0

TABLE A.7: Bayes factor support for an explanatory role in the Brazilian HIV-1C *pol* sequence diffusion process: Rand20C dataset.

Predictor	Gräf et al. (2015)	Replicated	Random effects
Geographic distance	0.0	0.0	0.0
Origin sample size	0.4	0.4	0.4
Destination sample size	∞	∞	34623.7
Origin HIV population size	1.2	1.1	1.1
Destination HIV population size	0.0	0.0	0.0
Origin HIV prevalence	7.9	7.7	7.0
Destination HIV prevalence	0.0	0.0	0.0
Origin HIV-1C population size	38.8	41.6	45.3
Destination HIV-1C population size	0.0	0.0	0.1
Origin HIV-1C prevalence	1.0	0.9	1.0
Destination HIV-1C prevalence	0.0	0.0	0.0

Appendix B. List of Figures

Appendix C. List of Tables

Bibliography

[1] John C. Avise. *Phylogeography – The History and Formation of Species*. Harvard University Press, 2000.

[2] Francisco J. Ayala. "Vagaries of the molecular clock". In: *Proceedings of the National Academy of Sciences* 94.15 (July 1997), pp. 7776–7783. ISSN: 0027-8424, 1091-6490.

[3] Roy J. Britten. "Rates of DNA sequence evolution differ between taxonomic groups". In: *Science (New York, N.Y.)* 231.4744 (Mar. 1986), pp. 1393–1398. ISSN: 0036-8075.

[4] Rebecca L. Cann, Mark Stoneking, and Allan C. Wilson. "Mitochondrial DNA and human evolution". In: *Nature* 325.6099 (Jan. 1987), pp. 31–36. ISSN: 0028-0836. DOI: 10.1038/325031a0.

[5] Nigel Dimmock, Andrew Easton, and Keith Leppard. *Introduction to Modern Virology*. 6th. Blackwell Publishing, 2007.

[6] Alexei J. Drummond and Remco R. Bouckaert. *Bayesian Evolutionary Analysis with BEAST*. Cambridge University Press, June 2015. ISBN: 978-1-107-01965-2.

[7] Alexei J. Drummond, Simon Y. W. Ho, et al. "Relaxed Phylogenetics and Dating with Confidence". In: *PLOS Biology* 4.5 (Mar. 2006), e88. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0040088.

[8] Alexei J. Drummond, Geoff K. Nicholls, et al. "Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data". In: *Genetics* 161.3 (July 2002), pp. 1307–1320. ISSN: 0016-6731, 1943-2631.

[9] Alexei J. Drummond and Andrew Rambaut. "BEAST: Bayesian evolutionary analysis by sampling trees". In: *BMC Evolutionary Biology* 7 (2007), p. 214. ISSN: 1471-2148. DOI: 10.1186/1471-2148-7-214.

[10] Alexei J. Drummond, Marc A. Suchard, et al. "Bayesian phylogenetics with BEAUti and the BEAST 1.7". In: *Molecular Biology and Evolution* 29.8 (Aug. 2012), pp. 1969–1973. ISSN: 1537-1719. DOI: 10.1093/molbev/mss075.

[11] Joseph Felsenstein. "Cases in which Parsimony or Compatibility Methods Will be Positively Misleading". In: *Systematic Zoology* 27.4 (1978), pp. 401–410. ISSN: 0039-7989. DOI: 10.2307/2412923.

[12] Joseph Felsenstein. "Evolutionary trees from DNA sequences: A maximum likelihood approach". In: *Journal of Molecular Evolution* 17.6 (1981), pp. 368–376. ISSN: 0022-2844, 1432-1432. DOI: 10.1007/BF01734359.

[13] Joseph Felsenstein. "The Troubled Growth of Statistical Phylogenetics". In: *Systematic Biology* 50.4 (2001), pp. 465–467. ISSN: 1063-5157.

[14] Marco A. R. Ferreira and Marc A. Suchard. "Bayesian analysis of elapsed times in continuous-time Markov chains". In: *Canadian Journal of Statistics* 36.3 (Sept. 2008), pp. 355–368. ISSN: 1708-945X. DOI: 10.1002/cjs.5550360302.

[15] Ronald A. Fisher. *The Genetical Theory of Natural Selection: A Complete Variorum Edition*. Google-Books-ID: sT4lIDk5no4C. OUP Oxford, 1930. ISBN: 978-0-19-850440-5.

[16] Brian Gaschen et al. "Diversity Considerations in HIV-1 Vaccine Selection". In: *Science* 296.5577 (June 2002), pp. 2354–2360. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1070441.

[17] Stuart Geman and Donald Geman. "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images". In: *IEEE transactions on pattern analysis and machine intelligence* 6.6 (June 1984), pp. 721–741. ISSN: 0162-8828.

[18] Edward I. George and Robert E. McCulloch. "Variable Selection via Gibbs Sampling". In: *Journal of the American Statistical Association* 88.423 (Sept. 1993), pp. 881–889. ISSN: 0162-1459. DOI: 10.1080/01621459.1993.10476353.

[19] Tiago Gräf et al. "Contribution of Epidemiological Predictors in Unraveling the Phylogeographic History of HIV-1 Subtype C in Brazil". In: *Journal of Virology* 89.24 (Dec. 2015), pp. 12341–12348. ISSN: 0022-538X, 1098-5514. DOI: 10.1128/JVI.01681-15.

[20] Bryan T. Grenfell et al. "Unifying the Epidemiological and Evolutionary Dynamics of Pathogens". In: *Science* 303.5656 (Jan. 2004), pp. 327–332. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1090727.

[21] Masami Hasegawa, Hirohisa Kishino, and Taka-aki Yano. "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA". In: *Journal of Molecular Evolution* 22.2 (1985), pp. 160–174. ISSN: 0022-2844.

[22] Jotun Hein, Mikkel Schierup, and Carsten Wiuf. *Gene Genealogies, Variation and Evolution: A primer in coalescent theory*. Google-Books-ID: QBC_SFOamksC. Oxford University Press, USA, Dec. 2004. ISBN: 978-0-19-154615-0.

[23] Jody Hey and Carlos A. Machado. "The study of structured populations — new hope for a difficult and divided science". In: *Nature Reviews Genetics* 4.7 (July 2003), pp. 535–543. ISSN: 1471-0056. DOI: 10.1038/nrg1112.

[24] David M. Hillis, John P. Huelsenbeck, and David L. Swofford. "Hobgoblin of phylogenetics?" In: *Nature* 369.6479 (June 1994), pp. 363–364. ISSN: 0028-0836. DOI: 10.1038/369363a0.

[25] Edward C. Holmes. "The phylogeography of human viruses". In: *Molecular Ecology* 13.4 (Apr. 2004), pp. 745–756. ISSN: 0962-1083.

[26] John P. Huelsenbeck and Fredrik Ronquist. "MRBAYES: Bayesian inference of phylogenetic trees". In: *Bioinformatics* 17.8 (Aug. 2001), pp. 754–755. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/17.8.754.

[27] Norman Kaplan and Charles H. Langley. "A new estimate of sequence divergence of mitochondrial DNA using restriction endonuclease mappings". In: *Journal of Molecular Evolution* 13.4 (Dec. 1979), pp. 295–304. ISSN: 0022-2844, 1432-1432. DOI: 10.1007/BF01731370.

[28] Robert E. Kass and Adrian E. Raftery. "Bayes Factors". In: *Journal of the American Statistical Association* 90.430 (June 1995), pp. 773–795. ISSN: 0162-1459. DOI: 10.1080/01621459.1995.10476572.

[29] Motoo Kimura and George H. Weiss. "The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance". In: *Genetics* 49.4 (Apr. 1964), pp. 561–576. ISSN: 0016-6731.

[30] John F. C. Kingman. "The coalescent". In: *Stochastic Processes and their Applications* 13.3 (Sept. 1982), pp. 235–248. ISSN: 0304-4149. DOI: 10.1016/0304-4149(82)90011-4.

[31] L. Lacey Knowles and Wayne P. Maddison. "Statistical phylogeography". In: *Molecular Ecology* 11.12 (Dec. 2002), pp. 2623–2635. ISSN: 0962-1083.

[32] Lynn Kuo and Bani Mallick. "Variable Selection for Regression Models". In: *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)* 60.1 (1998), pp. 65–81. ISSN: 0581-5738.

[33] Paul C. Lambert et al. "How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS". In: *Statistics in Medicine* 24.15 (Aug. 2005), pp. 2401–2428. ISSN: 1097-0258. DOI: 10.1002/sim.2112.

[34] Bret Larget and Donald L. Simon. "Markov Chasin Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees". In: *Molecular Biology and Evolution* 16.6 (June 1999), p. 750. ISSN: 0737-4038, 1537-1719.

[35] Kyeong Eun Lee et al. "Gene selection: a Bayesian variable selection approach". In: *Bioinformatics* 19.1 (Jan. 2003), pp. 90–97. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/19.1.90.

[36] Philippe Lemey. *Phylogenetic Diffusion Models and their Applications in viral epidemiology*. University of Leuven, 2010.

[37] Philippe Lemey, Andrew Rambaut, Alexei J. Drummond, et al. "Bayesian Phylogeography Finds Its Roots". In: *PLoS Comput Biol* 5.9 (Sept. 2009), e1000520. DOI: 10.1371/journal.pcbi.1000520.

[38] Philippe Lemey, Marco Salemi, and Anne-Mieke Vandamme. *The Phylogenetic Handbook*. 2nd ed. Cambridge University Press, 2011.

[39] Philippe Lemey et al. "The seasonal flight of influenza: a unified framework for spatiotemporal hypothesis testing". In: *arXiv:1210.5877 [q-bio]* (Oct. 2012).

[40] Philippe Lemey et al. "Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2". In: *PLOS Pathog* 10.2 (Feb. 2014), e1003932. ISSN: 1553-7374. DOI: 10.1371/journal.ppat.1003932.

[41] Shuying Li, Dennis K. Pearl, and Hani Doss. "Phylogenetic Tree Construction Using Markov Chain Monte Carlo". In: *Journal of the American Statistical Association* 95.450 (June 2000), pp. 493–508. ISSN: 0162-1459. DOI: 10.1080/01621459.2000.10474227.

[42] Daniel Magee et al. "Combining Phylogeography and Spatial Epidemiology to Uncover Predictors of H5N1 Diffusion". In: *Archives of virology* 160.1 (Jan. 2015), pp. 215–224. ISSN: 0304-8608. DOI: 10.1007/s00705-014-2262-5.

[43] Bob Mau, Michael A. Newton, and Bret Larget. "Bayesian phylogenetic inference via Markov chain Monte Carlo methods". In: *Biometrics* 55.1 (Mar. 1999), pp. 1–12. ISSN: 0006-341X.

[44] Chris A. Nasrallah, David H. Mathews, and John P. Huelsenbeck. "Quantifying the Impact of Dependent Evolution among Sites in Phylogenetic Inference". In: *Systematic Biology* (Nov. 2010), syq074. ISSN: 1063-5157, 1076-836X. DOI: 10.1093/sysbio/syq074.

[45] Saul B. Needleman and Christian D. Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". In: *Journal of Molecular Biology* 48.3 (Mar. 1970), pp. 443–453. ISSN: 0022-2836. DOI: 10.1016/0022-2836(70)90057-4.

[46] Martha I. Nelson et al. "Global migration of influenza A viruses in swine". In: *Nature Communications* 6 (Mar. 2015), p. 6696. ISSN: 2041-1723. DOI: 10.1038/ncomms7696.

[47] David C. Nickle et al. "Consensus and Ancestral State HIV Vaccines". In: *Science* 299.5612 (Mar. 2003), pp. 1515–1518. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.299.5612.1515c.

[48] Cédric Notredame, Desmond G Higgins, and Jaap Heringa. "T-coffee: a novel method for fast and accurate multiple sequence alignment1". In: *Journal of Molecular Biology* 302.1 (Sept. 2000), pp. 205–217. ISSN: 0022-2836. DOI: 10.1006/jmbi.2000.4042.

[49] Robert B. O'Hara and Mikko J. Sillanpää. "A review of Bayesian variable selection methods: what, how and which". In: *Bayesian Analysis* 4.1 (Mar. 2009), pp. 85–117. ISSN: 1936-0975, 1931-6690. DOI: 10.1214/09-BA403.

[50] Mark Pagel, Andrew Meade, and Daniel Barker. "Bayesian estimation of ancestral character states on phylogenies". In: *Systematic Biology* 53.5 (Oct. 2004), pp. 673–684. ISSN: 1063-5157. DOI: 10.1080/10635150490522232.

[51] Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. "APE: Analyses of Phylogenetics and Evolution in R language". In: *Bioinformatics (Oxford, England)* 20.2 (Jan. 2004), pp. 289–290. ISSN: 1367-4803.

[52] Rémy J. Petit. "The coup de grâce for the nested clade phylogeographic analysis?" In: *Molecular Ecology* 17.2 (Jan. 2008), pp. 516–518. ISSN: 0962-1083. DOI: 10.1111/j.1365-294X.2007.03589.x.

[53] Andrew Rambaut. "Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies". In: *Bioinformatics* 16.4 (Apr. 2000), pp. 395–399. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/16.4.395.

[54] Benjamin D. Redelings and Marc A. Suchard. "Joint Bayesian estimation of alignment and phylogeny". In: *Systematic Biology* 54.3 (June 2005), pp. 401–418. ISSN: 1063-5157. DOI: 10.1080/10635150590947041.

[55] Fredrik Ronquist. "Bayesian inference of character evolution". In: *Trends in Ecology & Evolution* 19.9 (Sept. 2004), pp. 475–481. ISSN: 0169-5347. DOI: 10.1016/j.tree.2004.07.002.

[56] Matthew Scotch et al. "Phylogeography of influenza A H5N1 clade 2.2.1.1 in Egypt". In: *BMC Genomics* 14 (2013), p. 871. ISSN: 1471-2164. DOI: 10.1186/1471-2164-14-871.

[57] Adam Siepel and David Haussler. "Combining phylogenetic and hidden Markov models in biosequence analysis". In: *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 11.2-3 (2004), pp. 413–428. ISSN: 1066-5277. DOI: 10.1089/1066527041410472.

[58] Botond Sipos et al. "PhyloSim - Monte Carlo simulation of sequence evolution in the R statistical computing environment". In: *BMC Bioinformatics* 12 (2011), p. 104. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-104.

[59] Jack Sullivan, Jeffrey A. Markert, and C. William Kilpatrick. "Phylogeography and molecular systematics of the Peromyscus aztecus species group (Rodentia: Muridae) inferred using parsimony and likelihood". In: *Systematic Biology* 46.3 (Sept. 1997), pp. 426–440. ISSN: 1063-5157.

[60] Chiraz Talbi et al. "Phylodynamics and Human-Mediated Dispersal of a Zoonotic Virus". In: *PLOS Pathogens* 6.10 (Oct. 2010), e1001166. ISSN: 1553-7374. DOI: 10.1371/journal.ppat.1001166.

[61] Simon Tavare. "Some probabilistic and statistical problems in the analysis of DNA sequences". In: *Some mathematical questions in biology / DNA sequence analysis edited by Robert M. Miura* (1986).

[62] Alan R. Templeton, Eric Routman, and Christopher A. Phillips. "Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, Ambystoma tigrinum." In: *Genetics* 140.2 (June 1995), pp. 767–782. ISSN: 0016-6731, 1943-2631.

[63] Julie D. Thompson, Desmond G. Higgins, and Toby J. Gibson. "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". In: *Nucleic Acids Research* 22.22 (Nov. 1994), pp. 4673–4680. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/22.22.4673.

[64] Nídia Sequeira Trovão et al. "Bayesian Inference Reveals Host-Specific Contributions to the Epidemic Expansion of Influenza A H5N1". In: *Molecular Biology and Evolution* 32.12 (Dec. 2015), pp. 3264–3275. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msv185.

[65] Linda Vigilant et al. "African populations and the evolution of human mitochondrial DNA". In: *Science (New York, N.Y.)* 253.5027 (Sept. 1991), pp. 1503–1507. ISSN: 0036-8075.

[66] Erik M. Volz, Katia Koelle, and Trevor Bedford. "Viral Phylodynamics". In: *PLOS Computational Biology* 9.3 (Mar. 2013), e1002947. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1002947.

[67] Simon Whelan and Nick Goldman. "A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach". In: *Molecular Biology and Evolution* 18.5 (May 2001), pp. 691–699. ISSN: 0737-4038, 1537-1719.

[68] Sewall Wright. "Evolution in Mendelian Populations". In: *Genetics* 16.2 (Mar. 1931), pp. 97–159. ISSN: 0016-6731.

[69] Chieh-Hsi Wu and Alexei J. Drummond. "Joint inference of microsatellite mutation models, population history and genealogies using transdimensional Markov Chain Monte Carlo". In: *Genetics* 188.1 (May 2011), pp. 151–164. ISSN: 1943-2631. DOI: 10.1534/genetics.110.125260.

[70] Jin Xiong. *Essential Bioinformatics*. 1st ed. Cambridge University Press, 2006.

[71] Emile Zuckerkandl and Linus Pauling. "Evolutionary Divergence and Convergence in Proteins". In: *Evolving Genes and Proteins*. Academic Press, 1965, pp. 97–166. ISBN: 978-1-4832-2734-4.

[72] Emile Zuckerkandl and Linus Pauling. "Molecular disease, evolution, and genetic heterogeneity." In: *Horizons in Biochemistry* (Jan. 1962), pp. 189–225.