Title
Bayesian Phylogenetic Inference for Viral Dispersal Process

Author
Gao, Jiansi

Publication Date
2022

Peer reviewed|Thesis/dissertation

Bayesian Phylogenetic Inference for Viral Dispersal Process

By

JIANSI GAO

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Population Biology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Brian R. Moore, Chair

Bruce Rannala

John P. Huelsenbeck

Samuel Díaz-Muñoz

Committee in Charge

2022

CONTENTS

Abstract iii
Acknowledgments vi

1 The Impact of Prior Misspecification on Bayesian Phylodynamic Inference of Biogeographic History **1**
Introduction 1
Theoretical Concerns and Proposed Solutions 3
Empirical Consequences 9
Discussion 16
Supplementary Material 19

2 PrioriTree: an Interactive Web Utility for Specifying Priors and Assessing Their Impacts in BEAST Biogeographic Analysis **90**
Introduction 90
Features 91
Availability and Implementation 96
Supplementary Material: `PrioriTree` Manual 97

3 New Phylogenetic Models Incorporating Interval-Specific Dispersal Dynamics Improve Inference of Disease Spread **151**
Introduction 152
Extending Phylodynamic Models 153
Simulation Study 161
Empirical Application 163
Discussion 168
Supplementary Material 173

ABSTRACT

Bayesian Phylogenetic Inference for Viral Dispersal Process

Phylogenies have been increasingly used in studying the spatial and temporal dynamics of infectious disease outbreaks; this *phylodynamic* approach encompasses a suite of methods for inferring various aspects of pathogen biology, including: (1) patterns of variation in demography through time; (2) the history of geographic spread either over continuous space or among a set of discrete-geographic areas, and; (3) the interaction between demography and geographic history.

This dissertation focuses on the discrete-geographic phylodynamic methods, which have been used extensively to understand the spatial and temporal spread of infectious disease outbreaks, and have played a central role for inferring key aspects of the COVID-19 pandemic, such as the geographic location and time of origin of the disease, the rates and geographic routes by which it spread, and the efficacy of various mitigation measures to limit its geographic expansion. These phylodynamic methods adopt an explicitly probabilistic approach that model the process of pathogen dispersal among a set of discrete-geographic areas (*e.g.*, cities, states, countries) over the branches of the pathogen phylogeny. The observations include the times and locations of pathogen sampling, and the genomic sequences of the sampled pathogens. These data are used to estimate the parameters of discrete-geographic phylodynamic models, which include a dated phylogeny of the pathogen samples, the average dispersal rate among all areas, and the relative dispersal rates (the dispersal rate between each pair of areas). Inference under these models is performed within a Bayesian statistical framework.

Although these phylodynamic models provide a powerful tool for understanding pathogen spread, they contain many parameters that must be inferred from minimal information (*i.e.*, the single geographic area in which each pathogen occurs). As a result, inferences under these models are inherently sensitive to our prior assumptions about the model parameters. In Chapter 1, I (and co-authors) demonstrate that the priors on the average dispersal rate and the number of dispersal routes, implemented as defaults in `BEAST` (and assumed in the vast majority of empirical studies) make strong and biologically unrealistic assumptions about the underlying

dispersal process. I present empirical evidence demonstrating that these priors are strongly disfavored by real data, and that these priors strongly (and adversely) distort central conclusions of epidemiological studies, including the importance of dispersal routes for the spread of pathogens and the ancestral area in which a given epidemic originated. I conclude this chapter by offering strategies and introducing an interactive web utility, `PrioriTree`, to help researchers avoid these issues.

Chapter 2 presents `PrioriTree` in detail. This utility is designed to help researchers follow the strategies I explored and recommended in Chapter 1 more easily. Specifically, it provides a suite of functions to allow users to interactively set up `BEAST` discrete-geographic phylodynamic analyses with visualized priors, and specify `BEAST` analyses and and summarize the results for assessing prior sensitivity and model fit. Apart from generating `BEAST` analysis scripts and figures summarizing the analyses, `PrioriTree` also dynamically generates a description of the associated methods to facilitate transparent and explicit communications in empirical biogeographic studies regarding what exact priors are used, how they are chosen, and how their impacts are assessed, eventually enhancing the reproducibility of biogeographic studies.

Virtually all discrete-geographic phylodynamic studies are based on models that assume that pathogen dispersal dynamics—including the average and relative rates of pathogen dispersal—remain constant over time. However, the dispersal dynamics of emerging pathogens (*e.g.*, SARS-CoV-2) may have been impacted by the initiation (or alteration or cessation) of intervention measures. Moreover, pathogen dispersal processes may inevitably vary over time due to temporal variation of human travel dynamics even without the impact of intervention measures.

In Chapter 3, I (and co-authors) (1) extend discrete-geographic phylodynamic models to allow both the average and relative dispersal rates to vary independently across pre-specified time intervals; (2) enable stochastic mapping under these interval-specific models to infer the number and timing of pathogen dispersal events between areas, and; (3) develop posterior-predictive statistics to assess the absolute fit of discrete-geographic phylodynamic models to empirical datasets. I first validate the new methods using simulations, and then apply them to a SARS-CoV-2 dataset from the early phase of the COVID-19 pandemic. These analyses reveal that: (1) under simulation, failure to accommodate interval-specific variation in the study data will severely bias parameter estimates; (2) in practice, the interval-specific models can signifi-

cantly improve the relative and absolute fit to empirical data; and (3) the increased realism of these interval-specific models provides qualitatively different inferences regarding key aspects of the COVID-19 pandemic—revealing significant temporal variation in global viral dispersal rates, viral dispersal routes, and the number of viral dispersal events between areas—and alters interpretations regarding the efficacy of intervention measures to mitigate the spread of SARS-CoV-2.

Together, this dissertation serves as a careful and thorough exploration of various aspects of the phylodynamic methods for inferring pathogen dispersal process, and represents an advance in the conceptual and statistical framework of Bayesian phylogenetic inference.

ACKNOWLEDGMENTS

First and foremost, I am deeply indebted to my major advisor Brian R. Moore, a tremendous mentor and colleague. At some crucial moments in our lives, a single person can play a decisive role in our career path decision; Brian is that person for me. It was nine years ago when I first met Brian, the summer after my junior year in college. I traveled from the other side of the world for a summer internship under Brian's supervision. To my greatest surprise, Brian took almost an entire month that summer teaching me about phylogenetics, which I had virtually no exposure to previously. After that month-long personal phylogenetic instruction, I found the subject that I would like to develop my research career upon; I knew I would like to come back Davis for graduate school, and I did. Looking back, I always feel extremely fortunate for this decision; throughout graduate school, Brian has always been tremendously patient and supportive to me. Without Brian's (both academic and non-academic) guidance and advice, not only the work presented in this dissertation would not be possible, but I would also not be the person or researcher I am today.

I am also immensely grateful to my co-advisor Bruce Rannala. I have been very fortunate to have him as a next-door office neighbor; despite not being my official advisor for the early years of my graduate school, his door has always been open to me for a conversation. I greatly appreciate his mentorship and sponsorship during my PhD, without which this dissertation would have been impossible.

I am also especially thankful to Mike May, an unofficial advisor, great colleague, and close friend. I was immensely fortunate to have him as a senior graduate student in the Moore lab for the early years of my PhD, and as a postdoc in the Rannala lab in recent years. I first met Mike the same day I first met Brian nine years ago. Mike, then a second-year PhD student, was already very knowledgeable about phylogenetics, statistics, and evolutionary biology in general. We have had countless discussions about statistical phylogenetics while on our way grabbing a quick lunch on campus, during coffee breaks, or after my regular attention-catching sentence: "Can I ask a question?''. These discussions have contributed tremendously on every single phylogenetic study I have done, and my understanding of phylogenetics in general.

I am extremely lucky to have them—Mike, Bruce, and Brian—as collaborators and co-authors on the chapters presented in this dissertation.

My qualifying exam and dissertation committees—Peter Wainwright, Phil Ward, Sam Díaz-Muñoz, and John P. Huelsenbeck—have been incredibly supportive and inspiring. I am greatly thankful for their patience and guidance that helped me remain focus and directed in completing this dissertation, and for taking their time reading this (rather long) dissertation and providing many helpful comments that have greatly improved the dissertation.

Apart from my official advisors and committee, I have also received advice, encouragement, and support from many unofficial mentors in the broader phylogenetic community, whom I was lucky enough to know and interact with during workshops, conferences, or other academic occasions. Among them, I especially thank Sebastian Höhna for welcoming me into the RevBayes developer community, and for his invaluable suggestions on many of my projects (including the ones presented in the following chapters). I also wish to thank Ziheng Yang: even I have only had a limited number of chance talking with him in person about my projects, I have learned extensively about statistical phylogenetics from his books throughout years.

I am more than grateful to the mentors whom I was fortunate enough to interact with as a teaching assistant. I have taught many times during my graduate school, but the first few times teaching was really challenging to me. I wish to thank Joel Ledford and Geoff Benn for their patience and guidance in training me to become a qualified BIS2C TA (and hopefully a fine one in some of the later times I taught the course). I was also extremely lucky to work with Michael Turelli, one of the greatest scientists I have ever met, as a TA for EVE103; I have learned many cool stories about himself and many other great evolutionary biologists from that experience. In addition, I really appreciate the experience of organizing and teaching the Bodega Phylogenetic Workshop, where I had the chance to interact with many students with really cool phylogenetic research projects and to discuss my research projects with some of the instructors.

I am grateful to many other members of the Moore and Rannala labs as well, including but not limited to Andrew Magee, Nikolai Vetr, Edie Espejo, Sneha Chakraborty, and Anna Nagel. I wish to express my gratitude to Andrew especially: we joined the Moore lab and started to learn phylogenetics and develop phylogenetic research projects at about the same time (me as a first-year graduate student while Andrew as a senior undergraduate student); I have benefited greatly from countless conversations—including many on the projects presented here—with him on phylogenetics over the years.

I am also indebted to the broader PBGG&CPB community for embracing me into the community in the first place, and for all their support, encouragement, and commiseration. My cohort have been extremely kind to me; I really love to be around with these fine people. The numerous cohort activities—game nights, retreats, hiking trips, Thanksgiving dinners, etc.—we had together are truly unforgettable (and I can not be thankful enough to them for possibly saving my life once by pulling me back from being too close to an elephant seal on the beach). I also owe a debt of gratitude to the EVE office staff (especially Sherri Mann), who have helped me countless times navigate logistical and programmatic hurdles of graduate school.

Last but not least, I would like to thank my parents for raising me to value education and fostering my curiosity. Their patience, encouragement, and support throughout years are indispensable and invaluable. This dissertation is dedicated to them.

My graduate school has been a rather long but extremely rewarding and heartwarming journey, and there is no way that I can express all my gratitude—many people I have inevitably forgotten to include in this list—but I will never forget how lucky I have been and will always be grateful for all the helps and kindness I have received in this journey. Thank you all!

Chapter 1

The Impact of Prior Misspecification on Bayesian Phylodynamic Inference of Biogeographic History

Abstract.—Epidemiology has been transformed by the advent of Bayesian phylodynamic models that allow researchers to infer the biogeographic history of pathogens over a set of discrete geographic areas (Lemey et al. 2009; Edwards et al. 2011). These biogeographic models provide powerful tools for understanding the spatial dynamics of epidemics, but contain many parameters that are inferred from minimal information (*i.e.*, the single geographic area in which each pathogen occurs). Consequently, inferences under these biogeographic models may be sensitive to our prior assumptions about the model parameters. Here, we demonstrate that the priors assumed in empirical phylodynamic studies make strong and biologically unrealistic assumptions about the underlying biogeographic process. We provide empirical evidence that these unrealistic priors strongly (and adversely) impact commonly reported aspects of epidemiological studies, including: (1) the relative rates of dispersal between areas; (2) the importance of dispersal routes for the spread of pathogens; (3) the number of dispersal events between areas, and; (4) the ancestral area in which a given epidemic originated. We offer strategies to avoid these problems, and develop tools to help researchers specify more biologically reasonable prior models that will realize the full potential of phylodynamic methods to understand pathogen biology and, ultimately, inform surveillance and monitoring policies to mitigate the impacts of disease.

INTRODUCTION

Phylogenies are increasingly used to study epidemiological dynamics; this *phylodynamic* approach is used to infer various aspects of pathogen biology, including patterns of variation

in demography through time and biogeographic history. The approach developed by Lemey *et al.* (Lemey et al. 2009; Edwards et al. 2011)—implemented in the `BEAST` software package Drummond et al. (2012)—is now the standard approach used to infer key aspects of the biogeographic history of pathogen epidemics, including: (1) the area in which an epidemic first originated; (2) the dispersal routes by which the pathogen spread among geographic areas, and; (3) the number of dispersal events between areas.

Under this approach, biogeographic history involves changes among a set of discrete areas (*e.g.*, cities, states, countries) over the branches of the pathogen phylogeny. Biogeographic history is modeled as a probabilistic process with parameters that specify the average rate of pathogen dispersal, and the relative rate of pathogen dispersal between each pair of geographic areas. Inference under these biogeographic models is performed within a Bayesian statistical framework. Bayesian inference requires that we specify a *prior probability* distribution for each parameter of the biogeographic model (reflecting our beliefs about the corresponding parameter values *before* evaluating the data at hand); the priors are updated by the information in our data (the observed geographic area from which each pathogen was sampled) to provide a *posterior probability* distribution for each of the model parameters (reflecting our beliefs about the parameter values *after* evaluating our data).

These biogeographic models contain many parameters that must be inferred from minimal information; the data are limited to a single observation (*i.e.*, the area in which each pathogen occurs). Accordingly, biogeographic inference under this approach may be sensitive to the assumed priors. Here, we demonstrate that the priors on the average dispersal rate and the number of dispersal routes implemented as defaults in `BEAST` (and used in most empirical studies; Fig. 1.1) make strong and biologically unrealistic assumptions about the underlying biogeographic process. We present empirical evidence demonstrating that these priors are strongly disfavored by real data, and that these priors strongly (and adversely) distort central conclusions of epidemiological studies. Finally, we offer strategies—and provide tools—to help researchers specify more biologically reasonable priors that will enhance the potential of phylodynamic methods to understand pathogen biology.

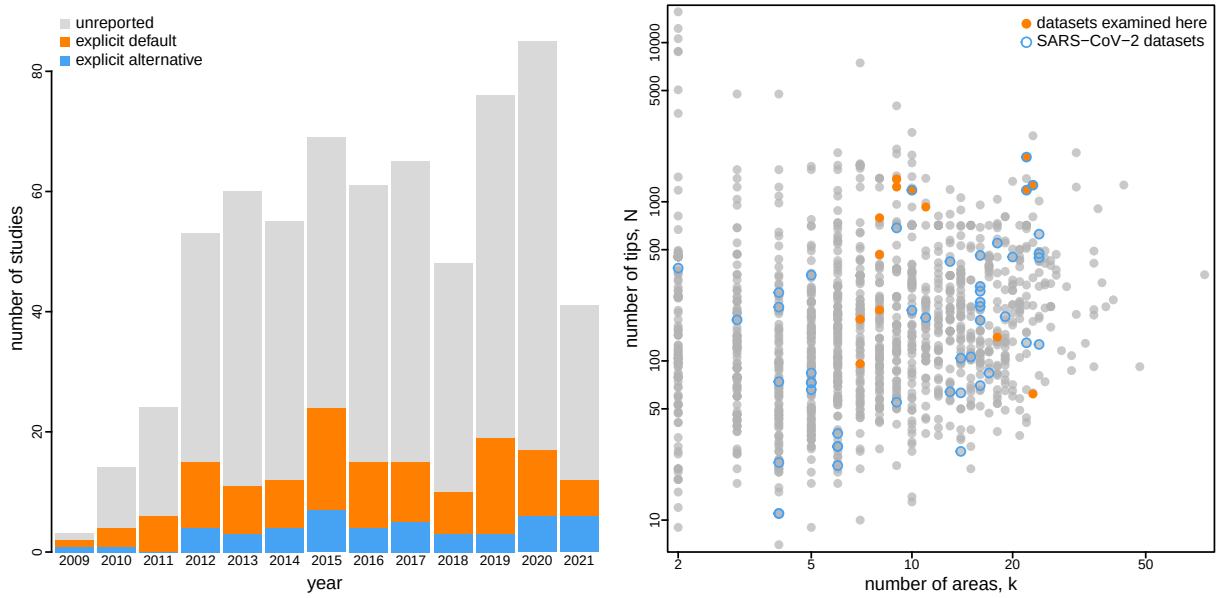

Figure 1.1: Empirical phylodynamic studies of pathogen biogeographic history. The bar plot at left depicts the choice of priors on the average dispersal rate and/or number of dispersal routes in the 651 published empirical studies (obtained from Google Scholar on June 30, 2021) that have inferred biogeographic history using the approach of (Lemey et al. 2009). The vast majority of these studies explicitly (orange) or implicitly (gray) specified default priors on these parameters; only 7.2% of published studies explicitly used non-default priors on the average dispersal rate and/or number of dispersal routes (blue). The right panel characterizes the size of published empirical datasets in terms of the number of discrete geographic areas (x-axis) and the number of tips (y-axis). Orange dots indicate the empirical datasets that we have examined in our study, and blue circled dots indicate datasets from SARS-CoV-2 studies.

Theoretical Concerns and Proposed Solutions

We assume that the tips of the study phylogeny occur in one of k discrete geographic areas. For clarity, we assume that the study phylogeny with divergence times is known. (In practice, the biogeographic history and the study phylogeny are usually inferred simultaneously; see Supplemental Material.) We first briefly describe the model proposed by Lemey *et al.* (Lemey et al. 2009) to infer biogeographic history; we then discuss theoretical concerns related to the priors on the parameters of that model; finally, we suggest alternative priors to address the concerns.

The Model

The model describes the evolution of the biogeographic history over the tree, Ψ, as a continuous-time Markov chain (CTMC). For a biogeographic history with k discrete areas, this stochastic process is fully specified by a $k \times k$ instantaneous-rate matrix, Q, where an element of the matrix, q_{ij}, is the instantaneous rate of change between state i and state j (*i.e.*, the instantaneous rate of dispersal from area i to area j). In principle, we may wish to treat each element

of this matrix as a free parameter to be estimated from the data. In practice, k is typically large, such that the biogeographic model includes many parameters, while the data are limited to a single observation (the geographic location of each tip), which raises concerns about our ability to estimate each parameter in the matrix. This concern motivated Lemey *et al.* (Lemey et al. 2009) to develop a Bayesian approach to simplify the biogeographic model. This is accomplished by specifying each element, q_{ij}, of the instantaneous-rate matrix, Q, as:

$$q_{ij} = r_{ij}\delta_{ij},$$

where r_{ij} is the relative rate of dispersal between areas i and j, and δ_{ij} is an indicator variable that takes one of two states (0 or 1). When $\delta_{ij} = 1$, the instantaneous dispersal rate for the corresponding element, q_{ij}, is simply $q_{ij} = r_{ij}$. Conversely, when $\delta_{ij} = 0$, the instantaneous dispersal rate for the corresponding element, q_{ij}, is zero, effectively removing that parameter from the biogeographic model. For a given Q matrix there is a vector of δ_{ij} and a vector of r_{ij}. Each unique vector of δ_{ij}—*i.e.*, $\boldsymbol{\delta}$, a string of zeros and ones for each of the possible pairwise dispersal routes between the k geographic areas—corresponds to a unique biogeographic model (Fig. 1.2). By convention, we rescale the Q matrix such that the expected number of dispersal events in one time unit is equal to the parameter μ (Yang 2014).

The original method (Lemey et al. 2009) assumes that instantaneous-rate matrix, Q, is symmetric, where $q_{ij} = q_{ji}$ (*i.e.*, $r_{ij} = r_{ji}$ and $\delta_{ij} = \delta_{ji}$). Accordingly, this model assumes that the instantaneous rate of dispersal from area i to area j is equal to the dispersal rate from area j to area i. For a dataset with k areas, the symmetric model has $\binom{k}{2}$ dispersal-route indicators and up to $\binom{k}{2}$ relative-rate parameters. A subsequent extension (Edwards et al. 2011) allows the Q matrix to be asymmetric, *i.e.*, q_{ij} and q_{ji} are not constrained to be equal. Accordingly, this model allows the rate of dispersal from area i to area j to be different from the rate of dispersal from area j to area i. For a dataset with k areas, the asymmetric model has $k \times (k-1)$ dispersal-route indicators and up to $k \times (k-1)$ relative-rate parameters.

We estimate the parameters of these biogeographic models in a Bayesian framework. Following Bayes' theorem, the joint posterior probability distribution of the model parameters is

(Bayes 1763):

$$\overbrace{P(\boldsymbol{r}, \delta, \mu \mid G, \Psi)}^{\text{posterior distribution}} = \frac{\overbrace{P(G \mid \boldsymbol{r}, \delta, \mu, \Psi)}^{\text{likelihood}} \overbrace{P(\boldsymbol{r})P(\delta)P(\mu)}^{\text{prior distribution}}}{\underbrace{P(G \mid \Psi)}_{\text{marginal likelihood}}},$$

where $\boldsymbol{r}$ is a vector that contains all of the relative-rate parameters, δ is a vector that contains all of the dispersal-route indicators, μ is the average rate of dispersal, Ψ is the phylogeny, and G is the observed geographic data. The likelihood function is equal to the probability of the observed geographic data, G, given the biogeographic model, Q, and the phylogeny, Ψ. The joint prior probability distribution reflects our beliefs about the model parameters before evaluating the geographic data at hand; the prior is updated by the information in the geographic data via the likelihood function to produce the joint posterior distribution, which reflects our beliefs about the model parameters after observing the geographic data. When there is little information in the data to update the assumed priors, our posterior estimates are apt to be quite sensitive to the chosen priors; this phenomenon is referred to as *prior sensitivity*.

The denominator of Bayes theorem is called the marginal likelihood, and is the likelihood function averaged over all possible values of the parameters in proportion to their prior probability (*i.e.*, it is the average probability of the data under the model). We approximate the joint posterior probability distribution using Markov chain Monte Carlo, which samples parameter values with a frequency proportional to their posterior probabilities, including in this case the dispersal routes defined by δ. Below, we detail our theoretical concerns regarding the priors on the number of dispersal routes, δ, and average dispersal rate, μ.

Prior on the Number of Dispersal Routes

Recall that each vector, δ, specifies a unique configuration of dispersal routes, which corresponds to a unique biogeographic model. The total number of dispersal routes for a given biogeographic model is denoted Δ. For a given value of Δ, there may be multiple distinct biogeographic models (*e.g.*, for $\Delta = 2$, there are three distinct symmetric models; Fig. 1.2). Lemey *et al.* (Lemey et al. 2009) impose a prior on biogeographic models by: (1) placing a prior on the total number of dispersal routes, Δ, and; (2) assuming that all biogeographic models with a given value of Δ are equiprobable. For example, the three distinct biogeographic models with $\Delta = 2$ depicted in Fig. 1.2 are assumed to have equal prior probability. Together, these assump-

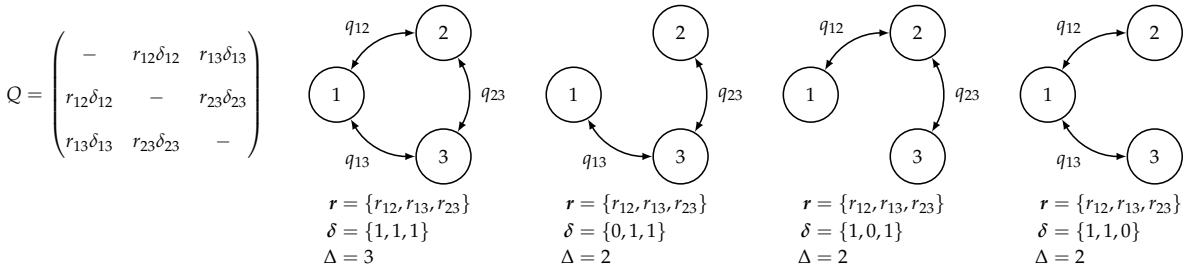

Figure 1.2: Biogeographic models for k = 3 geographic areas. The approach of Lemey *et al.* (Lemey et al. 2009) models the evolution of biogeographic range using a continuous-time Markov chain (CTMC). The CTMC is completely described by the instantaneous-rate matrix, Q, where each element q_{ij} specifies the instantaneous rate of dispersal between areas i and j. Each element, q_{ij}, is a function of the relative-rate parameter, r_{ij}, and a dispersal-route indicator, δ_{ij} (left panel). The dispersal-route indicator, δ_{ij}, is 1 when the corresponding dispersal route exists, and 0 when it does not exist. Alternative biogeographic models are specified by different configurations of dispersal routes. In the first model, all possible dispersal routes exist; the remaining models have two viable dispersal routes, corresponding to different vectors of dispersal-route indicators, δ (right panels). The total number of dispersal routes for a given biogeographic model is Δ. Note that there may be multiple distinct biogeographic models with an equal number of dispersal routes, Δ (*e.g.*, the three distinct models depicted here for which $\Delta = 2$). The models depicted are all symmetric, *i.e.*, they assume that the rate of dispersal from area i to area j is equal to the rate of dispersal from area j to area i.

tions induce a prior probability that a given dispersal route between areas i and j exists, *i.e.*, that $\delta_{ij} = 1$.

For the symmetric model, Lemey *et al.* (Lemey et al. 2009) specify an *offset* Poisson prior on the total number of dispersal routes, Δ. That is, the prior on Δ assigns zero probability

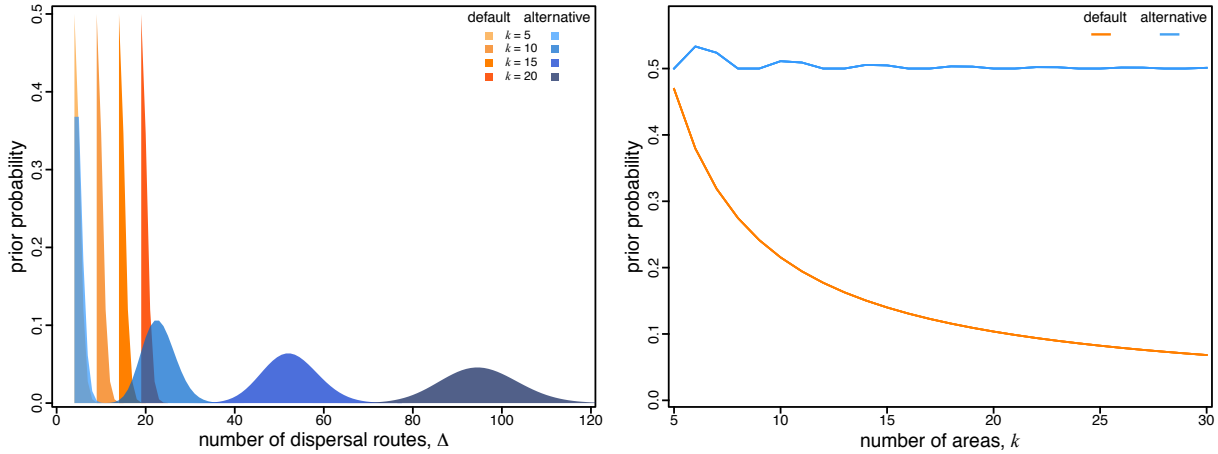

Figure 1.3: Prior on dispersal routes under the symmetric biogeographic model. The left panel illustrates the default (orange) and alternative (blue) prior distributions on the total number of dispersal routes, Δ, as a function of the number of areas, k. The default-prior distributions are highly focused on the minimal number of dispersal routes, $k - 1$, whereas the alternative-prior distributions are centered on an intermediate number of dispersal routes (*i.e.*, the expected number of dispersal routes is about half the maximum number for a given value of k). The right panel illustrates the prior probability under the default (orange) and alternative (blue) prior models that a given dispersal route exists (*i.e.*, that $\delta_{ij} = 1$) as a function of the total number of geographic areas, k. Under the default-prior model, the probability that a given dispersal route exists drops rapidly for datasets with a moderately large (and common; *c.f.*, Fig. 1.1) number of geographic areas; by contrast, under the alternative-prior model, this probability remains relatively constant for all values of k.

to all biogeographic models with fewer than $k-1$ dispersal routes; this reflects the constraint that a dataset with k geographic areas cannot be realized under a CTMC with fewer than $k-1$ non-zero q_{ij} values (*i.e.*, dispersal routes).[1] The prior on the number of dispersal routes *greater than* $k-1$ is described by a Poisson prior with rate parameter, λ. Lemey *et al.* (Lemey et al. 2009) express an explicit prior preference for biogeographic models with the minimal number of dispersal routes. Specifically, by default, $\lambda = \ln(2)$, which places 50% of the prior probability on biogeographic models with the minimum number of dispersal routes, $\Delta = (k-1)$ (Fig. 1.3, left panel). For the asymmetric model (Edwards et al. 2011), the number of dispersal routes is assumed to be drawn from a Poisson prior with rate λ. In this case, λ is specified such that the expected number of dispersal routes is $k-1$ (Figure S.1.1; Note that this prior does not enforce a minimum number of dispersal routes).

Recall that the number of dispersal-route indicators grows rapidly as a function of the number of areas, k; however, the prior expected number of dispersal routes grows linearly as a function of k. Consequently, the prior probability that any given dispersal route exists rapidly decreases as k increases (Fig. 1.3, right panel). For large (and common; *c.f.*, Fig. 1.1) values of k, the default prior on Δ results in an extremely informative prior on models with the minimum number of dispersal routes.

In the experiments below, we specify alternative and more diffuse priors on Δ, where the expected number of dispersal routes is about half the maximum number; this results in a relatively flat prior probability that any given dispersal route exists for all values of k (Fig. 1.3). Specifically, for the symmetric model, we specify an offset (*i.e.*, by $k-1$) Poisson prior on Δ with λ specified so that the expected number of dispersal routes is about half of the maximum number, $\binom{k}{2}$, for a dataset with k areas. For the asymmetric model, we specify a Poisson prior distribution on Δ with $\lambda = \binom{k}{2}$, which represents a prior belief that half of all possible dispersal routes are included in the biogeographic model.

Prior on the Average Dispersal Rate

Recall that the rate matrix, Q, is rescaled so that the average rate of dispersal between all areas is μ. For a tree of length T (*i.e.*, the sum of the durations of all branches in the tree), the expected number of dispersal events is $\mu \times T$. Therefore, the prior on μ represents our prior belief about

[1] We note that the real constraint on the symmetric model is that it must be *irreducible*, *i.e.*, it must be possible to reach each area from every other area either directly or indirectly. A model with fewer than $k-1$ dispersal routes cannot be irreducible; however, a model with at least $k-1$ dispersal routes is not guaranteed to be irreducible.

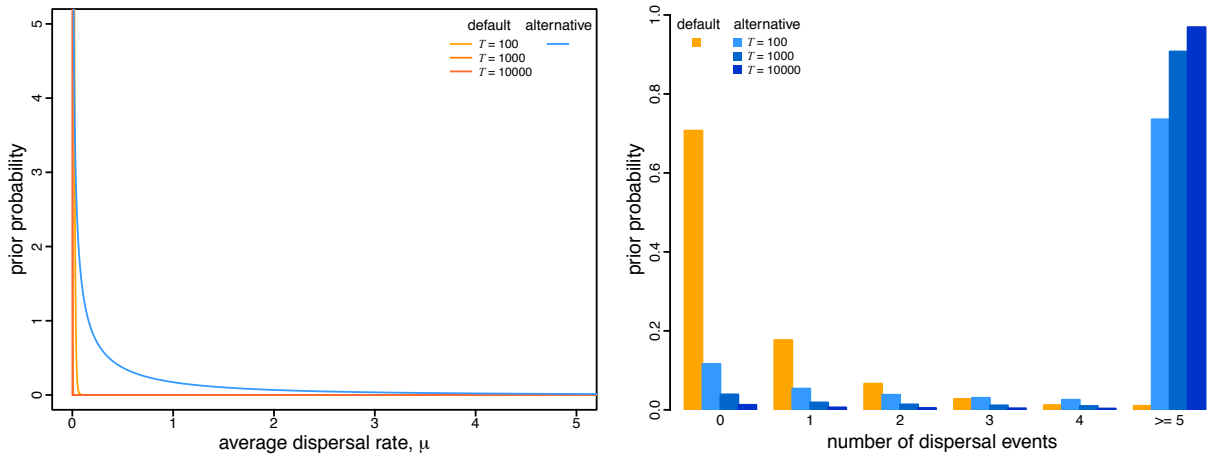

Figure 1.4: Prior on the average dispersal rate and the implied prior on the number of dispersal events. The left panel illustrates the default (orange) and alternative (blue) prior distributions on the average dispersal rate as a function of the duration of the biogeographic history, T. The default-prior distributions are highly focused on extremely low average dispersal rates, whereas the alternative-prior distribution is more permissive of higher rates. The right panel illustrates the implied prior distribution on the total number of dispersal events under the default (orange) and alternative (blue) prior models. Under the default-prior model, the expected number of dispersal events is 0.5, independent of the duration of the biogeographic history, T, whereas under the alternative-prior model, the expected number of dispersal events sensibly increases with the duration of the biogeographic history.

the number of dispersal events over the tree. By default, μ is assigned a gamma prior with shape parameter $\alpha = 0.5$ and rate parameter $\beta = T$. (Note that the gamma prior on the average dispersal rate is referred to as the CTMC-rate reference prior in the `BEAUTi` program used to generate input files for `BEAST` analyses.) The gamma distribution has a mean of α/β; therefore this prior expresses the belief that the average rate of dispersal is $0.5/T$ (Fig. 1.4, left panel).

Because the expected number of dispersal events is $\mu \times T$, the prior expected number of dispersal events under this prior is 0.5, independent of the duration of the entire biogeographic history (*i.e.*, the tree length, T), or the number of areas, k, in which the pathogen occurs. Similarly, the prior distribution on the number of dispersal events is independent of T and k: the 95% prior interval is $[0, 3]$ dispersal events, which implies that we would be very surprised if a biogeographic history of any duration with any number of areas involved more than three dispersal events (see Fig. 1.4, right panel). Logically, however, a biogeographic history that includes k areas minimally requires $k - 1$ dispersal events. Therefore, this prior becomes increasingly unreasonable as k grows to large (and common; *c.f.*, Fig. 1.1) values.

In the experiments below, we specify a more diffuse prior on the dispersal rate, μ. Specifically, we specify a more permissive exponential prior on μ; this prior has a rate parameter θ, and a mean of $1/\theta$. To address concerns about the potential impact of assuming a fixed value of θ on posterior estimates, we treat the mean of the exponential prior, $1/\theta$, as a random variable

to be estimated from the data. Specifically, we specify a gamma prior on $1/\theta$; this gamma hyperprior has shape parameter $\alpha = 0.5$ and rate parameter $\beta = 0.5$ (enforcing the shape and rate parameters to be equal ensures that the resulting prior on μ is proper). The resulting prior—known as the *K*-distribution (Jakeman and Pusey 1978)—is more diffuse than the default prior on μ (Fig. 1.4, right panel), as is the resulting prior distribution on the number of dispersal events (Fig. 1.4, left panel). Importantly, this alternative prior distribution on the number of dispersal events scales with the duration of the entire biogeographic history, T.

Empirical Consequences

In this section, we explore the empirical consequences arising from our theoretical concerns with the informative default priors on the number of dispersal routes and the average dispersal rate. We collected eleven datasets from published empirical studies, and reanalyzed each under a suite of biogeographic models, including all combinations of: (1) symmetric and asymmetric Q matrices; (2) default and alternative priors on the number of dispersal routes; and (3) default and alternative priors on the average dispersal rate. For each dataset, we estimated the marginal likelihood for each of the eight candidate models, and used Bayes factors to understand the impacts of both priors on the fit of the model to these datasets. For each dataset, we also estimated the joint posterior distribution for all eight models to explore the impact of the default and alternative priors on estimates of commonly reported biogeographic inferences. We describe the details of these analyses in the Supplemental Material.

The Impact of Prior Choice on Model Fit.

Our concern regarding the default priors is that they represent strongly informative and biologically unrealistic beliefs about the process that generated the data. Accordingly, the fit of these prior models to empirical data should be relatively low compared to more biologically reasonable prior models. Following Lemey *et al.* (Lemey et al. 2009), we assessed the relative fit of these competing prior models to the data by Bayes factor, which is computed using the estimated marginal likelihood (*i.e.*, the probability of the data under a given prior model). Because Bayes factors represent the relative fit of competing models to the data, they are often used as a way of selecting among alternative models.

Bayes-factor comparisons of all eight candidate models indicate that the default prior on the number of dispersal routes *and* the default prior on the average dispersal rate are both

biologically implausible (Table S.1.2). In all cases, the alternative prior on the average dispersal rate strongly outperformed the default prior; for all but one case (for which the Bayes factor was equivocal), the alternative prior on the number of dispersal routes outperformed the default prior.

Bayes factors assess the *relative* fit of competing prior models to the datasets; we also assessed the *absolute* fit of the prior models to the data using posterior-predictive simulation (Gelman et al. 1996; Bollback 2002). This approach is based on the following premise: if a given model provides an adequate description of the process that gave rise to our observed data, then we should be able to use that model to simulate new datasets that resemble our observed data. Results of the posterior-predictive simulations corroborate our findings based on Bayes-factor comparisons: in all cases, the alternative-prior models provide an adequate fit to the empirical datasets, whereas the default-prior models are generally inadequate (Figures S.1.2–S.1.3; Tables S.1.3–S.1.4).

Table 1.1: The relative fit of the default- and alternative-prior models. We inferred marginal likelihoods for each dataset under two models: one with both default priors, the other with both alternative priors. For each combination of priors, we assumed the preferred biogeographic model (*i.e.*, with a symmetric or asymmetric rate matrix). Marginal-likelihood estimates for the default- and alternative-prior models are listed in the first two columns ($\pm$ SD among four replicates); $2\ln \mathrm{BF}$ between the two models are listed in the third column. The default-prior models are decisively rejected for all datasets (*i.e.*, $2\ln \mathrm{BF} \gg 10$; Kass and Raftery 1995).

Da	Default	Alternative	$2\ln \mathrm{BF}$
1	-190.48 ± 0.10	-147.32 ± 0.11	86.33
2	-144.08 ± 0.05	-128.80 ± 0.09	30.56
3	-214.89 ± 0.12	-173.90 ± 0.23	81.98
4	-106.37 ± 0.09	-91.47 ± 0.12	29.79
5	-1176.37 ± 0.24	-1037.60 ± 0.26	277.54
6	-1310.88 ± 0.43	-1164.63 ± 0.24	292.50
7	-837.68 ± 0.27	-726.79 ± 0.17	221.78
8	-2875.30 ± 1.42	-2275.52 ± 0.39	1199.55
9	-2334.51 ± 0.94	-1872.97 ± 0.32	923.08
10	-309.90 ± 0.59	-258.14 ± 0.29	103.52
11	-2525.85 ± 1.11	-2160.57 ± 0.65	730.56
12	-1989.18 ± 2.22	-1721.17 ± 1.03	536.01
13	-1747.10 ± 1.75	-1531.50 ± 1.40	431.21
14	-1368.66 ± 2.70	-1221.43 ± 0.82	294.45

*Dataset sources: 1 (Dash et al. 2015); 2–4 (Wilfert et al. 2016); 5–7 (Faria et al. 2014); 8–9 (Bedford et al. 2015); 10 (Yao et al. 2015); 11 (Gao et al. 2022); 12 (Alpert et al. 2021), and; 13–14 (Candido et al. 2020).

Given that both default priors—on the number of dispersal routes and the average dispersal rate—negatively impact the relative and absolute fit of biogeographic models to our empirical datasets, we focus on two candidate prior models: one model with both default priors, and one model with both alternative priors. For both the default- and alternative-prior models, we identified the preferred biogeographic model (*i.e.*, symmetric or asymmetric). In every case, the default-prior models were decisively disfavored relative to the alternative-prior models (Table 1.1). Below, we explore the impact of these prior models on commonly reported biogeographic inferences.

The Impact of Prior Choice on Pairwise Dispersal Rates.

We first explored estimates of the model parameters that comprise the Q matrix—*i.e.*, $\boldsymbol{r}$, δ, and μ—under the default-prior model to those estimated under the alternative-prior model. Although these parameters are seldom (if ever) reported in empirical studies, they are the actual basis of commonly reported aspects of biogeographic history, *i.e.*, commonly reported inferences are a function of these Q-matrix parameters. The left two panels of Fig. 1.5 compare posterior-mean estimates of Q under the default- and alternative-prior models for the deformed-wing virus dataset (Wilfert et al. 2016); the choice of prior model strongly impacts estimates of the dispersal rates between many areas. We also summarized the impact of the default- and alternative-prior models across all datasets (Fig. 1.5, right panel), demonstrating

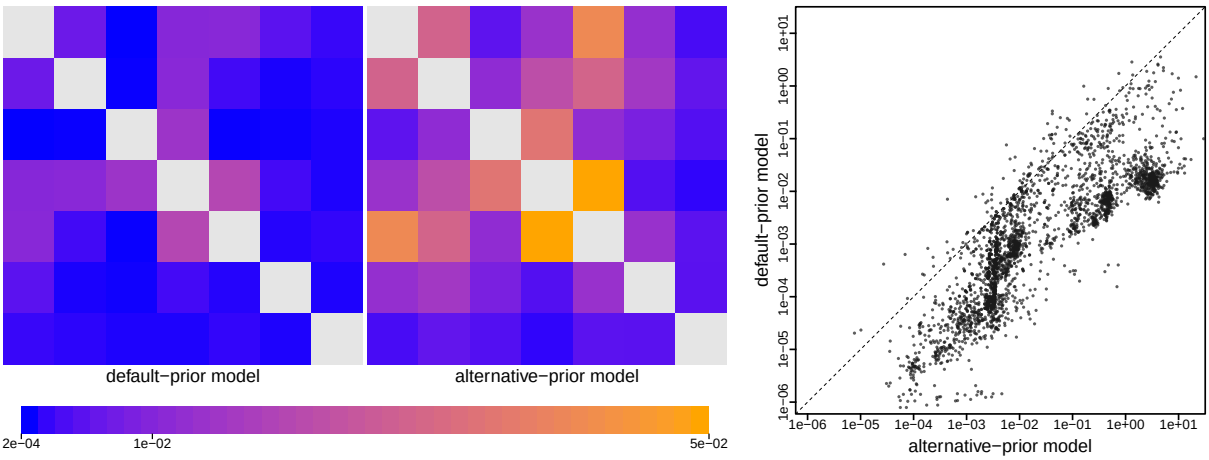

Figure 1.5: The impact of prior choice on estimates of pairwise dispersal rates. Heatmaps summarize posterior-mean estimates of the instantaneous rate of dispersal between each pair of geographic areas, q_{ij}, estimated for the deformed-wing virus dataset (Wilfert et al. 2016) under the default (left) and alternative (middle) prior models. At right, we summarize dispersal-rate estimates for each pair of areas across all eleven empirical datasets. Note that dispersal-rate estimates under the default-prior model are consistently lower than those estimated under the alternative-prior model.

the pervasive impact of the default priors on estimates of the Q-matrix parameters. Perhaps unsurprisingly—given that the default priors imply fewer dispersal routes and a lower number of dispersal events—posterior-mean estimates under the default-prior models are systematically much lower than those inferred under the alternative-prior models.

The Impact of Prior Choice on the Inferred Support for Dispersal Routes.

Empirical studies often focus on the evidential support for dispersal routes between each pair of geographic areas; these inferences are intended to identify the set of dispersal routes that were important in the geographic spread of the pathogen. This involves computing Bayes factors for each of the dispersal-rate parameters in the biogeographic model. Above, we computed Bayes factors for models as the difference in their log marginal likelihoods; an alternative (but equivalent) formulation is to compute the ratio of the posterior and prior odds for two alternative models. For each dispersal-rate parameter in the Q matrix, we compute the Bayes factor as:

$$\mathrm{BF}_{ij} = \frac{P(\delta_{ij} = 1 \mid G)}{P(\delta_{ij} = 0 \mid G)} \div \frac{P(\delta_{ij} = 1)}{P(\delta_{ij} = 0)},$$

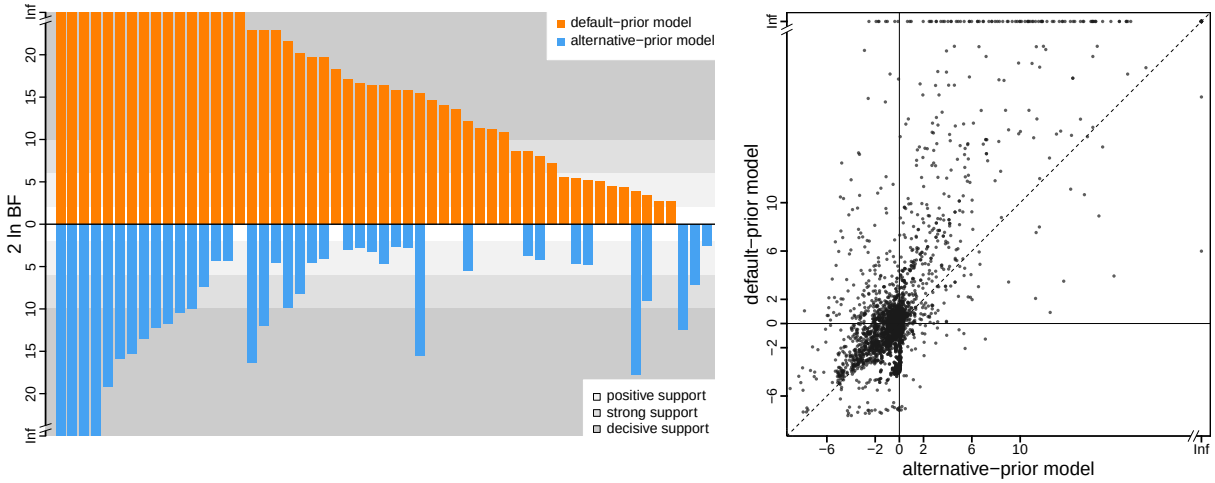

Figure 1.6: The impact of prior choice on the inferred support for dispersal routes. The left panel compares the evidential support for dispersal routes under the default (orange) and alternative (blue) prior models for the H3N2 influenza virus dataset (Bedford et al. 2015). Each bar indicates the $2\ln\mathrm{BF}$ (Bayes factor) for the corresponding dispersal route between two geographic areas; only "significant" dispersal routes (*i.e.*, $2\ln\mathrm{BF} > 2$) are figured. Some dispersal routes identified as significant under the default-prior model have no support under the alternative-prior model, and *vice versa*. Additionally, the rank order of dispersal routes according to their Bayes-factor support differs between the default- and alternative-prior models. The right panel plots the $2\ln\mathrm{BF}$ for each dispersal route under the default (y-axis) alternative (x-axis) prior models across all empirical datasets. Note that, under the alternative-prior model, many dispersal routes have equivocal Bayes-factor support (*i.e.*, $-2 \leq 2\ln\mathrm{BF} \leq 2$); conversely, Bayes factors under the default-prior model tend to be larger than those under the alternative-prior model (dots above the diagonal indicate greater support under the default-prior model compared to the alternative-prior model).

where $P(\delta_{ij} = 1)$ is the prior probability that the dispersal route exists, and $P(\delta_{ij} = 1 \mid G)$ is the posterior probability that the dispersal route exists, which is computed as the fraction of MCMC samples for which $\delta_{ij} = 1$. This formulation of the Bayes factor captures the degree to which our beliefs in the existence of a dispersal route changed after observing the geographic data. Because the default-prior model focuses on biogeographic models with a small number of dispersal routes, the prior probability that each dispersal route exists is correspondingly small. As a result, we expect the default-prior model to increase the apparent Bayes-factor support for individual dispersal routes.

Our analyses of the H3N2 influenza virus dataset (Bedford et al. 2015) illustrate the impact of the default- and alternative-prior models on the inferred support for dispersal routes (Fig. 1.6, left panel). Specifically, Bayes factors inferred under the default-prior model are much higher than those inferred under the alternative-prior model; *e.g.*, of the 38 dispersal routes that are decisively supported under the default-prior model (*i.e.* where $2 \ln \mathrm{BF} \geq 10$), only 15 of those routes are decisively supported under the alternative-prior model. Additionally, the rank order of these decisively supported dispersal routes differs markedly under the two prior models. The impact of prior choice on the estimated support for individual dispersal routes is pervasive across all of the empirical datasets (Fig. 1.6, right panel). The scale of the Bayes factors inferred under the default-prior model is much higher than that under the alternative-prior model, which is consistent with the fact that the prior ratio is smaller for a given dispersal route under the default-prior model.

The Impact of Prior Choice on the Inferred Biogeographic History.

Empirical studies frequently report summaries that are based on the conditional probability distribution of biogeographic histories over the tree. The distribution of histories depends on—*i.e.*, is *conditioned* on—the instantaneous-rate matrix, Q, the biogeographic data, G, and the phylogeny, Ψ. Conceptually, for a given tree and rate matrix, we imagine simulating a geographic history over the tree from the root to its tips, where the rate matrix specifies the waiting times between dispersal events. We can construct the conditional distribution of biogeographic histories by simulating a large number of individual histories, and retaining only those histories that realize the observed geographic areas at the tips, G. This conditional distribution contains all of the information required to compute two commonly reported summaries: the ancestral areas at internal nodes of the tree, and the number of dispersal events between ge-

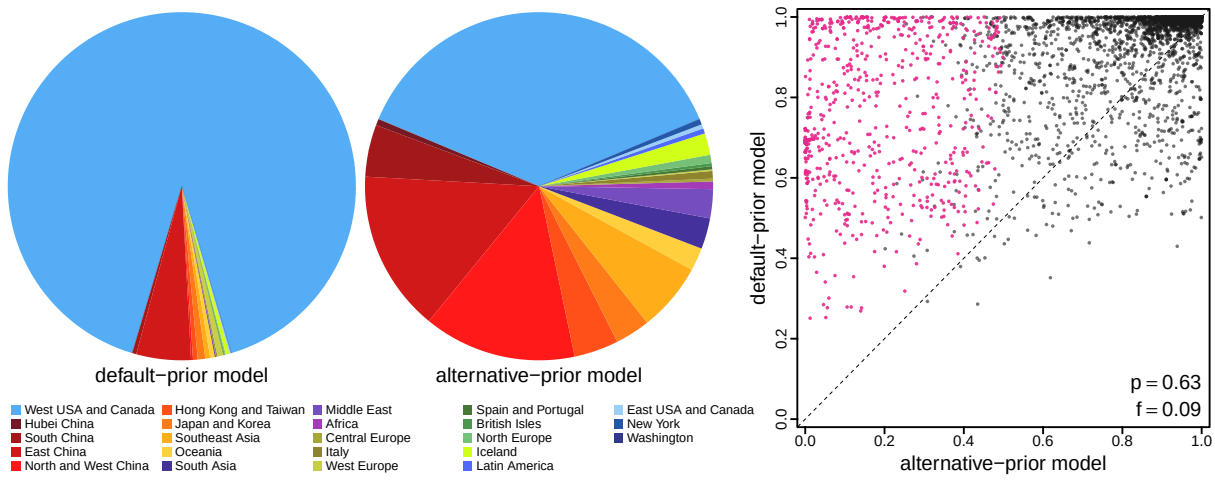

Figure 1.7: The impact of prior choice on ancestral-area estimates. The left panel compares the posterior probabilities for the geographic source of the Washington state outbreak clade of SARS-CoV-2 inferred under the default- and alternative-prior models for the SARS-CoV-2 Global dataset (Gao et al. 2022). Under the default-prior model, the virus is inferred to be introduced from Western North America to the Washington state with very high probability (90.8%); under the alternative-prior model, it is almost equally probable that SARS-CoV-2 was introduced to Washington from either Western North America (37.4%) or China (subarea combined, 38.7%). The right panel plots the posterior probability of the most probable ancestral area under the default-prior model for each node across all datasets (y-axis) against the corresponding posterior probability of that area under the alternative-prior model (x-axis). The summary statistic p denotes the fraction of internal nodes that are shared under the default- and alternative-prior models; f is the fraction of shared nodes where the MAP ancestral area differs under the default- and alternative-prior models. Note that the posterior probability of the MAP ancestral area inferred under the default-prior model is generally higher than that under the alternative-prior model.

ographic areas. Because these summaries depend on the rate matrix, which in turn is sensitive to the choice of prior (Fig. 1.5), we expect the choice of prior model to have a corresponding influence on these summaries. We detail the impacts of default- and alternative-prior models on each of these commonly reported summaries below.

Many studies aim to infer the probability that a pathogen occurred in each of the k geographic areas at internal nodes of the tree (including the root), *e.g.*, to infer the point of origin of an epidemic. The probability that a given node was in a particular area is simply the fraction of conditional histories where the node was in that area. Our reanalysis of the SARS-CoV-2 Global dataset (Gao et al. 2022) reveals that the choice of prior model may exert a strong impact on estimates of ancestral areas. For example, the most probable ancestral area to the "Washington state outbreak clade"—the earliest documented community COVID-19 outbreak in the United States (Bedford et al. 2020; Worobey et al. 2020)—is Western North America (posterior probability 90.8%) under the default-prior model, while Western North America (posterior probability 37.4%) and China (posterior probability 38.7% combining subareas) are effectively equally probable under the alternative-prior model (Fig. 1.7, left panel). This impact is prevalent across the 14 datasets. The choice of prior not only impacted the inferred probability of the

most probable area at an internal node, but in some fraction of cases (≈9%) also changed the identity of the most probable ancestral area (Fig. 1.7, right panel).

Empirical phylodynamic studies often infer the number of dispersal events between each pair of areas, *e.g.*, to understand whether a given area is a major source of viral outbreaks. A given geographic history includes the number of dispersal events between each pair of areas; therefore, we can compute the average number of dispersal events between each pair of areas as the posterior-mean number of events over the conditional distribution of histories. The choice of prior model exerts a strong influence on estimates of the number of dispersal events. For example, our analyses of the SARS-CoV-2 Brazil dataset (Candido et al. 2020) inferred São Paulo to be the only major source of SARS-CoV-2 dispersal within Brazil under the default-prior model, as 84.6% of the domestic dispersal events originated from it (compared to the second largest source, Rio de Janeiro, with only 3.7%). Conversely, six areas are inferred to

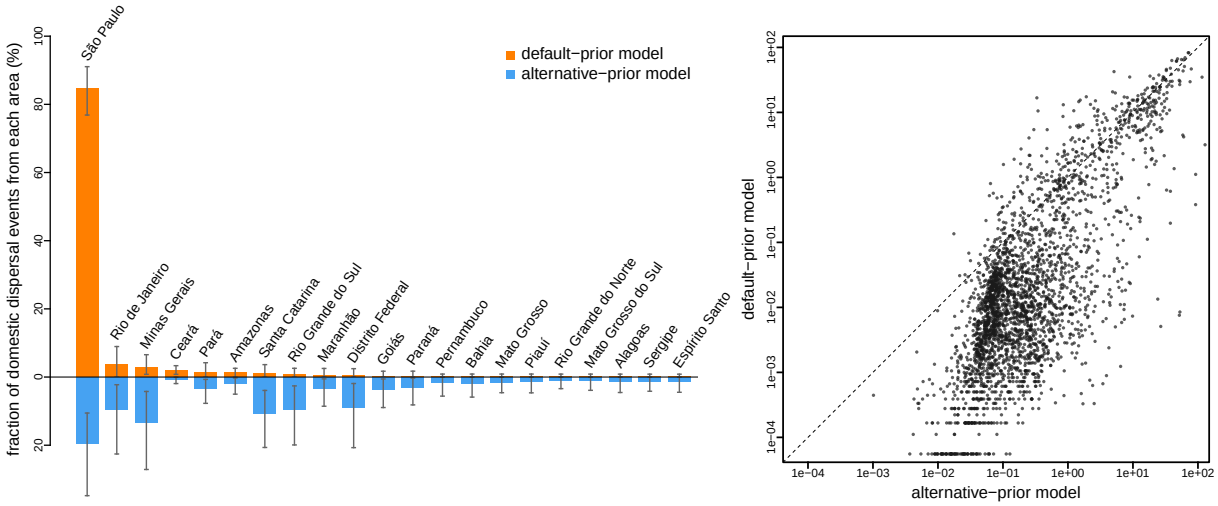

Figure 1.8: The impact of prior choice on the inferred number of dispersal events between areas. The left panel compares the number of dispersal events inferred under the default (orange) and alternative (blue) prior models for the SARS-CoV-2 Brazil dataset (Candido et al. 2020). Each bar indicates the estimated fraction of domestic dispersal events originated from each area in Brazil (mean [bar height] and 95% credible interval [whiskers]). Under the default-prior model, São Paulo is inferred to be the only major source of SARS-CoV-2 dispersal within Brazil as 84.6% of the domestic dispersal events originated from it (compared to the second largest source, Rio de Janeiro, with only 3.7%). Conversely, under the alternative-prior model, we inferred that only 19.4% of the domestic dispersal events originated from São Paulo, while five other areas each occupies a fraction that is $\gg$5%, including two areas in Southeast Brazil (Minas Gerais [13.2%] and Rio de Janeiro [9.4%]), two areas in South Brazil (Santa Catarina [10.7%] and Rio Grande do Sul [9.5%]), and one area in Central-West Brazil (Distrito Federal [9.0%]). Note that the rank order of dispersal routes according to their inferred fraction of dispersal events differs between the default- and alternative-prior models. The right panel plots the number of dispersal events across each dispersal route inferred under the default (y-axis) and alternative (x-axis) prior models across all empirical datasets. On average, the inferred number of dispersal events under the alternative-prior model is larger than that inferred under the default-prior model. As for the SARS-CoV-2 Brazil dataset, the number of dispersal events over the dispersal routes that are inferred with very small number of dispersal events are inferred to be generally much larger under the alternative-prior model.

have $\gg$5% of the domestic dispersal events originated from each of them, among which São Paulo is the source of only 19.4% of all the domestic dispersal events (Fig. 1.8, left panel). The impact of prior choice on the inferred number of dispersal events was pervasive across all of our empirical datasets. As might be expected from the default prior on the number of events, we infer a larger number of dispersal events under the alternative-prior model (Fig. 1.8, right panel).

DISCUSSION

The development of Bayesian biogeographic models has the potential to transform our ability to study pathogen biology. The complexity of these biogeographic models is both an asset and a liability. It is an asset because these complex models provide the ability to describe complex biogeographic processes. It is a liability because the biogeographic inference under these models inherently involves minimal information (*i.e.*, the single geographic area in which each pathogen occurs), which renders the posterior estimates sensitive to the choice of priors. Moreover, the complexity of these biogeographic models obscures the biological interpretation of their parameters, which makes it difficult to formulate biologically sensible prior beliefs about these parameter values. We suspect this underlies the fact that the vast majority of empirical phylodynamic biogeographic studies ($\approx$93%) have assumed default priors.

In the present study, we have demonstrated that the default priors on the average dispersal rate and the number of dispersal routes implemented in `BEAST` imply biologically unrealistic assumptions about the biogeographic process (Figs. 1.3 and 1.4). We have presented empirical evidence demonstrating that these default priors are in fact biologically unrealistic, *i.e.*, they are strongly disfavored by all of the empirical datasets that we evaluated (Tables 1.1 and S.1.2). We have also demonstrated the consequences of these strongly misinformative priors; their use qualitatively changes our understanding of many key aspects of pathogen biogeographic history, including inferences of relative dispersal rate between areas, the number of significant dispersal routes, the ancestral geographic areas, and the number of dispersal events between areas (Figures 1.5–1.8).

The results of our study highlight the need to adopt—and offer insights on—best practices for empirical phylodynamic studies. For all of the empirical datasets in our study (which are typical examples of empirical datasets, see Fig. 1.1), the choice of prior had a strong impact on

commonly reported biogeographic inferences. For any given empirical dataset, however, the impact of prior choice remains an open question. To be clear, the alternative priors explored in our study should not be treated as a panacea; rather, empirical biogeographic analyses should consider the biological meaning of alternative prior models, and assess their impact on a case-by-case basis. This first requires establishing a clear connection between the parameters of the biogeographic model and the implied process of biogeographic evolution. To this end, we have attempted to clarify the biological interpretation of the parameters (and their corresponding priors) of these biogeographic models. The next step is to evaluate the empirical impact of alternative prior models. For example, empirical biogeographic studies could adopt a robust Bayesian approach, *i.e.*, to assess the sensitivity of biogeographic inferences to alternative prior choices. To estimate the posterior under a set of k candidate prior models requires only a modest k-fold increase in computation.

The results of our study also highlight priorities for developers of statistical phylodynamic methods. A short-term priority is to provide tools for empirical users to visualize the biological implications of various prior models, to develop prior models that are motivated by biological processes, and to specify such prior models in inferences with popular phylogenetic programs (*e.g.*, `BEAST`). We have developed such a tool, `PrioriTree` (https://github.com/jsigao/prioritree), as an early attempt; `PrioriTree` allows users to explore the biological consequences under various biogeographic models and customizable prior distributions, meanwhile generating readily runnable `BEAST` XML script on the fly. This tool also provides an accessible way for users to configure further `BEAST` inferences—*e.g.*, marginal likelihood estimation—to evaluate the empirical impact of alternative prior models, following the procedure adopted in this study. We describe the details of `PrioriTree` in Chapter 2.

We are optimistic that rigorous empirical application of current phylodynamic models, and focused efforts to develop novel phylodynamic approaches, will help advance our understanding of pathogen biology and minimize the impact of infectious disease.

Data and Code Availability

The sequence, sampling time and geography data used in this study, as well as the phylogenies we marginalized over or conditioned on in the biogeographic inference, are maintained in the GitHub repository (https://github.com/jsigao/prior_misspecification_phylodynamic_

biogeography) and archived in the Dryad repository (https://datadryad.org/stash/share/7Rd5kdTh7V66w9XefTuSeoui0LLv6LAwcY_5buMwUZU). Our repositories also contain BEAST XML scripts used to perform the phylodynamic analyses and R scripts used to post process the analyses and perform posterior-predictive simulation.

SUPPLEMENTARY MATERIAL

Supplemental Figures and Tables for the Main Text

Priors on the Number of Dispersal Routes: Asymmetric Model

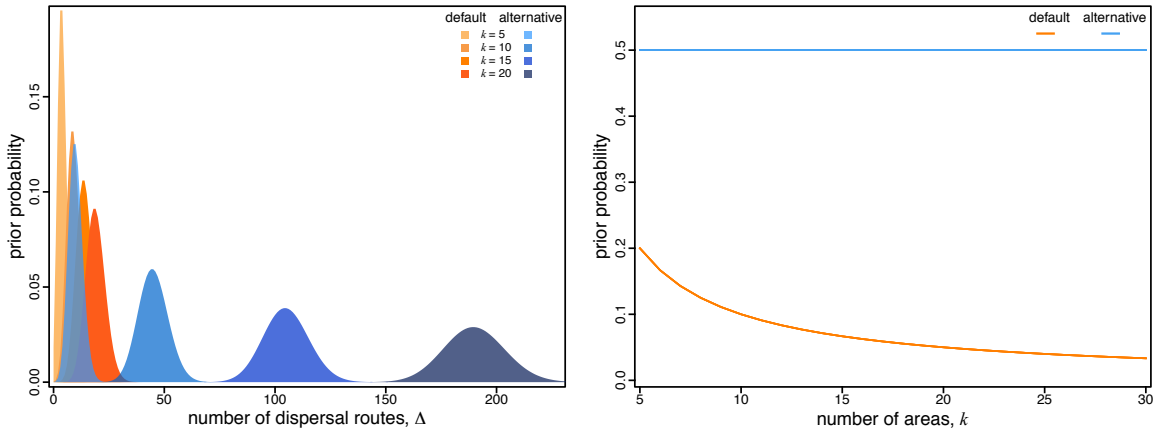

Figure S.1.1: Prior probability on dispersal routes under the asymmetric geographic model. The left panel illustrates the default (orange) and alternative (blue) prior distributions on the total number of dispersal routes, Δ, as a function of the number of areas, k. The default-prior distributions are highly focused on the minimal number of dispersal routes, $(k-1)$, whereas the alternative-prior distributions are centered on an intermediate number of dispersal routes (*i.e.*, the expected number of dispersal routes is half the maximum number). The right panel illustrates the prior probability under the default (orange) and alternative (blue) prior models that a given dispersal route exists (*i.e.*, $\delta_{ij} = 1$) as a function of the total number of areas, k. Under the default-prior model, the probability that a given dispersal route exists drops rapidly for moderately large (and common) values of k, whereas under the alternative-prior model, this probability remains constant for all values of k.

Table S.1.1: Default and alternative prior specifications.

Parameter	Model	Default	Alternative
Number of dispersal routes, Δ	Symmetric	$[\Delta - (k-1)] \sim \text{Pois}(\ln 2)$	$[\Delta - (k-1)] \sim \text{Pois}(\lceil \frac{k(k-5)}{4} + 1 \rceil)$
	Asymmetric	$\Delta \sim \text{Pois}(k-1)$	$\Delta \sim \text{Pois}(\frac{k(k-1)}{2})$
Average dispersal rate, μ	—	$\mu \sim \Gamma(0.5, T)$ (*i.e.*, CTMC-rate reference)	$\mu \sim \text{Exp}(1/\lambda)$ $\lambda \sim \Gamma(0.5, 0.5)$

Assessing the Fit of Default- and Alternative-Prior Models to Empirical Datasets

Using Bayes Factors to Assess Relative Model Fit

Table S.1.2: Assessing the relative fit of all eight candidate biogeographic models to the 14 empirical datasets. For each dataset we computed the marginal likelihood for the eight candiate models corresponding to all possible combinations of: (1) default and alternative priors on the average rate of dispersal, μ; (2) default and alternative priors on the number of dispersal routes, Δ, and; (3) symmetric and asymmetric biogeographic models. For each dataset, we computed twice the log Bayes factors ($2 \ln \mathrm{BF}$) between each model and the best model (*i.e.*, the model with the highest marginal likelihood; blue cells). For each dataset, N is the number of sequences (*i.e.*, tips), and k is the number of biogeographic areas.

			$P(\mu)$ default				$P(\mu)$ alternative			
			$P(\Delta)$ default		$P(\Delta)$ alternative		$P(\Delta)$ default		$P(\Delta)$ alternative	
Dataset*	N	k	Q_s	Q_a	Q_s	Q_a	Q_s	Q_a	Q_s	Q_a
1	62	23	−86.33	−91.04	−47.80	−47.03	−10.54	−22.53	0.00	−1.78
2	209	8	−32.52	−30.56	−25.91	−23.27	−9.12	−4.55	−1.88	0.00
3	183	7	−84.72	−96.15	−81.45	−89.27	−18.42	0.00	−18.50	−2.74
4	96	7	−29.79	−32.50	−25.20	−26.02	−2.72	−4.85	0.00	−1.08
5	792	8	−277.54	−284.58	−275.52	−279.72	−36.33	−8.96	−26.09	0.00
6	927	10	−343.58	−292.50	−340.53	−291.05	−46.51	−8.46	−42.63	0.00
7	466	8	−264.37	−221.78	−262.28	−217.02	−6.47	−11.93	−3.29	0.00
8	1391	9	−1223.03	−1199.55	−1125.60	−1101.86	−70.98	−55.88	−9.77	0.00
9	1240	9	−957.42	−923.08	−870.67	−840.93	−114.60	−54.48	−51.26	0.00
10	141	18	−122.46	−103.52	−82.88	−78.19	−22.35	−17.00	−0.87	0.00
11	1271	23	−799.28	−730.56	−685.88	−687.87	−191.91	−8.26	−79.71	0.00
12	1908	22	−659.32	−536.01	−515.86	−493.10	−252.06	−22.00	−142.19	0.00
13	1182	10	−486.13	−431.21	−456.26	−410.96	−126.13	−33.07	−80.22	0.00
14	1182	22	−308.19	−294.45	−267.17	−264.91	−137.99	−8.57	−45.28	0.00

* Dataset sources: 1) Dengue virus from Dash et al. (2015); 2–4) Deformed wing virus from Wilfert et al. (2016); 5–7) HIV from Faria et al. (2014); 8–9) Seasonal Influenza viruses from Bedford et al. (2015); 10) Rabies virus from Yao et al. (2015); 11) SARS-CoV-2 (Global) from Gao et al. (2022); 12) SARS-CoV-2 (B.1.1.7 USA) from Alpert et al. (2021), and; 13–14) SARS-CoV-2 (Brazil) from Candido et al. (2020).

Table S.1.3: Assessing the adequacy of all eight candidate biogeographic models for the 14 empirical datasets using the parsimony statistic. Each row summarizes model adequacy for a given dataset (numbered as described in Table S.1.2). For each dataset, N is the number of sequences (*i.e.*, tips), and k is the number of biogeographic areas. Each column lists the posterior-predictive p-values for a given prior model (notation follows Table S.1.2) based on the parsimony summary statistic. Red p-values indicate that the model is inadequate at the 95% level.

			$P(\mu)$ default				$P(\mu)$ alternative			
			$P(\Delta)$ default		$P(\Delta)$ alternative		$P(\Delta)$ default		$P(\Delta)$ alternative	
Dataset	N	k	Q_s	Q_a	Q_s	Q_a	Q_s	Q_a	Q_s	Q_a
1	62	23	**0.00**	**0.00**	**0.00**	**0.00**	0.93	0.33	0.74	0.64
2	209	8	**0.01**	**0.01**	**0.01**	**0.01**	0.51	0.43	0.47	0.43
3	183	7	**0.00**	**0.00**	**0.00**	**0.00**	0.71	0.10	0.64	0.11
4	96	7	**0.01**	**0.01**	**0.01**	**0.00**	0.46	0.27	0.43	0.25
5	792	8	**0.00**	**0.00**	**0.00**	**0.00**	0.08	0.51	0.09	0.58
6	927	10	**0.00**	**0.00**	**0.00**	**0.00**	0.21	0.36	0.23	0.39
7	466	8	**0.00**	**0.00**	**0.00**	**0.00**	0.07	0.08	0.08	0.08
8	1391	9	**0.00**	**0.00**	**0.00**	**0.00**	0.68	0.15	0.47	0.11
9	1240	9	**0.00**	**0.00**	**0.00**	**0.00**	0.75	0.34	0.57	0.28
10	141	18	**0.00**	**0.00**	**0.00**	**0.00**	0.62	0.13	0.51	0.32
11	1271	23	**0.00**	**0.00**	**0.00**	**0.00**	0.50	0.45	0.14	0.43
12	1908	22	**0.00**	**0.01**	**0.00**	**0.00**	0.78	0.57	0.85	0.88
13	1182	10	**0.00**	**0.00**	**0.00**	**0.00**	0.42	0.52	0.31	0.58
14	1182	22	**0.00**	**0.00**	**0.00**	**0.00**	0.94	0.85	0.89	0.95

Table S.1.4: Assessing the adequacy of all eight candidate biogeographic models for the 14 empirical datasets using the tip-wise multinomial statistic. Each row summarizes model adequacy for a given dataset (numbered as described in Table S.1.2). For each dataset, N is the number of sequences (*i.e.*, tips), and k is the number of biogeographic areas. Each column lists the posterior-predictive p-values for a given prior model (notation follows Table S.1.2) based on the tip-wise multinomial statistic. Red p-values indicate that the model is inadequate at the 95% level.

			$P(\mu)$ default				$P(\mu)$ alternative			
			$P(\Delta)$ default		$P(\Delta)$ alternative		$P(\Delta)$ default		$P(\Delta)$ alternative	
Dataset	N	k	Q_s	Q_a	Q_s	Q_a	Q_s	Q_a	Q_s	Q_a
1	62	23	**1.00**	**1.00**	**1.00**	**1.00**	0.46	0.96	0.44	0.62
2	209	8	**0.98**	0.97	0.96	0.96	0.70	0.80	0.59	0.61
3	183	7	0.91	0.96	0.88	0.93	0.18	0.93	0.21	0.89
4	96	7	**1.00**	**1.00**	**1.00**	**1.00**	0.92	0.97	0.89	0.94
5	792	8	**0.98**	0.92	0.97	0.93	0.87	0.75	0.81	0.58
6	927	10	**1.00**	**1.00**	**1.00**	**1.00**	**0.98**	**0.99**	0.96	0.93
7	466	8	**1.00**	**0.99**	**1.00**	**0.99**	0.96	0.96	0.96	0.95
8	1391	9	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
9	1240	9	**1.00**	**1.00**	**1.00**	**1.00**	0.39	**0.99**	0.40	**0.98**
10	141	18	**1.00**	**1.00**	**1.00**	**1.00**	0.97	**1.00**	0.92	0.96
11	1271	23	**1.00**	**1.00**	**1.00**	**1.00**	0.77	**0.99**	0.51	0.90
12	1908	22	0.97	0.95	0.96	0.96	0.42	0.72	0.19	0.19
13	1182	10	0.91	0.95	0.91	0.95	0.40	0.68	0.20	0.52
14	1182	22	0.89	0.94	0.84	0.89	0.06	0.22	**0.02**	0.09

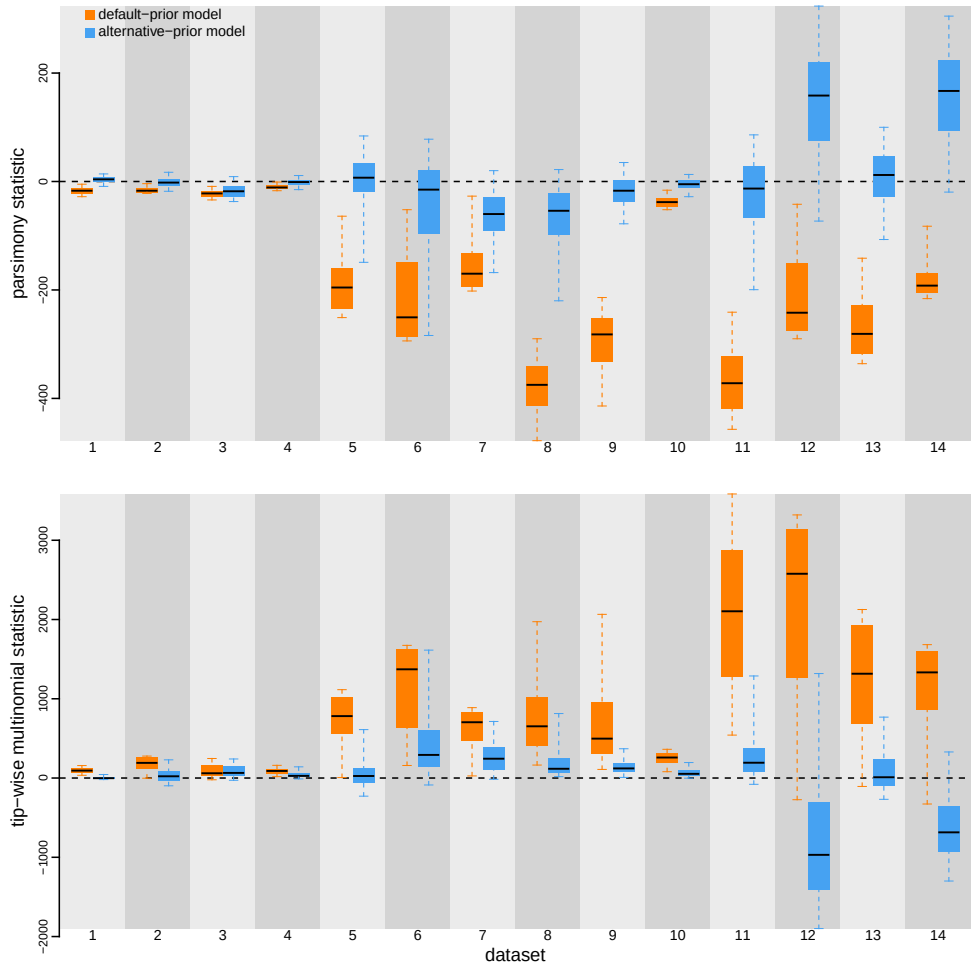

Figure S.1.2: Posterior-predictive distributions of the parsimony statistic (top panel) and the tip-wise multinomial statistic (bottom panel) under the preferred default- and alternative-prior models for all datasets. Each column depicts estimates for one of the 14 datasets (numbered as described in Table S.1.2). Within each column, the pair of boxplots depicts the posterior-predictive distributions of the summary statistic under the default (orange) and alternative (blue) prior models: the center of each box is the median predictive value of the summary statistic; the box and whiskers indicate the corresponding 50% and 95% posterior-predictive intervals, respectively. The horizontal dashed line indicates when the simulated and observed datasets produce identical value for the summary statistic. A model is judged to be inadequate (*i.e.*, incapable of generating geographic datasets that are similar to the observed data) if its 95% posterior-predictive interval does not overlap with the dashed line. Importantly, posterior-predictive simulation allows us to compare the absolute fit of the candidate models to the 14 geographic datasets: the preferred default prior models (orange) are always inadequate, whereas the preferred alternative prior models (blue) are almost always adequate.

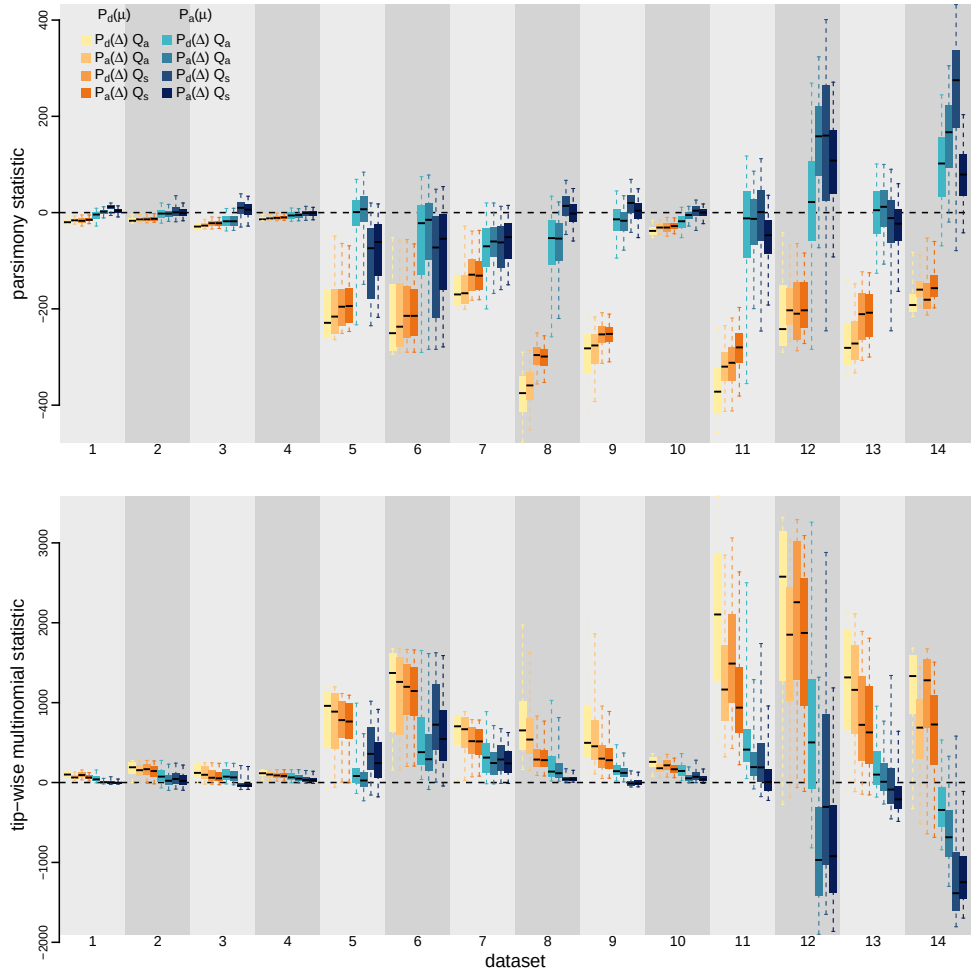

Figure S.1.3: Posterior-predictive distributions of the parsimony statistic (top panel) and the tip-wise multinomial statistic (bottom panel) under each of the eight prior models for all datasets. Each column depicts estimates for one of the 14 datasets (numbered as described in Table S.1.2). Within each column, the set of 8 boxplots depicts the posterior-predictive distributions of the summary statistic under each of the 8 prior models: the center of each box is the median predictive value of the summary statistic; the box and whiskers indicate the corresponding 50% and 95% posterior-predictive intervals, respectively. The horizontal dashed line indicates when the simulated and observed datasets produce identical value for the summary statistic. A model is judged to be inadequate (*i.e.*, incapable of generating geographic datasets that are similar to the observed data) if its 95% posterior-predictive interval does not overlap with the dashed line.

Joint and Sequential Bayesian Phylodynamic Inference

For simplicity, our description of the biogeographic model in the main text assumes that the phylogeny, Ψ, is known without error. However, empirical applications typically embed this biogeographic model in a larger "phylodynamic" model, which jointly models multiple aspects of epidemiological evolution, namely: (1) the *diversification model*, with parameters θ_Ψ that describe the branching process that generates the phylogeny; (2) the *substitution model*, with parameters θ_S that describe the process of molecular evolution over the branches of the tree, and; (3) the *geographic model*, with parameters $\theta_G = \{\boldsymbol{r}, \boldsymbol{\delta}, \mu\}$ that describe the dispersal of pathogens among areas. The resulting joint posterior density for the full phylodynamic model can be written as:

$$P(\Psi, \theta_\Psi, \theta_S, \theta_G \mid X, G) = \frac{P(X, G \mid \Psi, \theta_\Psi, \theta_S, \theta_G) P(\Psi, \theta_\Psi, \theta_S, \theta_G)}{P(X, G)}, \tag{S.1.1}$$

where X is an alignment of molecular sequence data and G is the geographic data. Conditional on the phylogeny, the processes of molecular and geographic evolution are assumed to be independent. Combined with an assumption that the parameters of the three model components are independent *a priori*, eq. S.1.1 can be written:

$$P(\Psi, \theta_\Psi, \theta_S, \theta_G \mid X, G) = \frac{P(X \mid \Psi, \theta_S) P(G \mid \Psi, \theta_G) P(\Psi \mid \theta_\Psi) P(\theta_\Psi) P(\theta_S) P(\theta_G)}{P(X, G)}. \tag{S.1.2}$$

In principle, this joint posterior density can be approximated using Markov chain Monte Carlo (MCMC). However, owing to the complexity of the joint model, these MCMC analyses may perform poorly in practice. To simplify the MCMC, it is possible to perform a "sequential" analysis consisting of two steps that together are equivalent to a joint analysis. The first step estimates the joint posterior density of phylogenies, diversification-model parameters, and substitution-model parameters:

$$P(\Psi, \theta_\Psi, \theta_S \mid X) = \frac{P(X \mid \Psi, \theta_S) P(\Psi \mid \theta_\Psi) P(\theta_\Psi) P(\theta_S)}{P(X)}; \tag{S.1.3}$$

this joint posterior density is approximated using MCMC. The second step uses the marginal posterior density of phylogenies from the first step as a prior to estimate the joint posterior distribution of the geographic-model parameters:

$$P(\Psi, \theta_G \mid G) = \frac{P(G \mid \Psi, \theta_G) P(\Psi) P(\theta_G)}{P(G)}, \tag{S.1.4}$$

where $P(\Psi)$ corresponds to the marginal posterior distribution of phylogenies from the first step, $P(\Psi \mid X)$. Again, this joint posterior density is estimated using MCMC. Proposals for the phylogeny are made by drawing a new phylogeny, Ψ', from the marginal prior distribution, $P(\Psi \mid X)$, and accepting the proposal with probability:

$$A = \min \left[1, \frac{P(G \mid \Psi', \theta_G) P(\Psi')}{P(G \mid \Psi, \theta_G) P(\Psi)} \times \frac{P(\Psi \mid X)}{P(\Psi' \mid X)} \right]. \tag{S.1.5}$$

Recognizing that each sample from the joint posterior distribution in second step is associated with a sample of Ψ from the first step, we can reconstitute the full joint posterior distribution (eq. S.1.1). That is, for the i^{th} sample from the posterior distribution of the second step with phylogeny Ψ^i, we can find the sample of the first step associated with Ψ^i, and "attach" the corresponding sample of parameters from the first step to the i^{th} sample of the second step. The resulting distribution of samples is theoretically equivalent to the joint posterior distribution of the full model (*i.e.*, the sequential analysis is theoretically equivalent to the joint analysis).

We note that `BEAST` provides two options for performing sequential analysis through `empiricalTreeDistributionModel`. The first option, evoked with the argument `MetropolisHastings = "true"` of the `empiricalTreeDistributionOperator`, uses the proposal mechanism described above and uses eq. S.1.5 to accept or reject proposals on the tree. The second option, evoked with the argument `MetropolisHastings = "false"`, proposes new trees by drawing them from the marginal prior density, and then accepts the proposed tree with probability 1. The second option does not result in an ergodic Markov chain with a stationary distribution equivalent to eq. S.1.4, and therefore is not equivalent to a full joint phylodynamic analysis (eq. S.1.1). For this reason, we use the argument `MetropolisHastings = "true"` for all empirical analyses described below.

Analyses of Empirical Datasets

General Analysis Protocol

To explore the empirical consequences arising from our theoretical concerns with the informative default priors, we collected 14 datasets from published empirical studies (see Table S.1.5), and reanalyzed each dataset using the sequential approach under a suite of biogeographic models, including all combinations of: (1) a symmetric and asymmetric rate matrix; (2) default and alternative priors on the number of dispersal routes; and (3) default and alternative priors on the average dispersal rate. We provide details of these two prior models in the Theoretical Concerns section of the main text and in Table S.1.1.

Table S.1.5: Empirical datasets information.

Study	Virus	Dataset	N	k
Dash et al. (2015)	Dengue	—	62	23
Wilfert et al. (2016)	DWV	lp	209	8
		rdrp	183	7
		vp3	96	7
Faria et al. (2014)	HIV	A	792	8
		B	927	10
		C	466	8
Bedford et al. (2015)	Influenza	H3	1391	9
		Yam	1240	9
Yao et al. (2015)	Rabies	—	141	18
Gao et al. (2022)	SARS-CoV-2	Global	1271	23
Alpert et al. (2021)	SARS-CoV-2	B.1.1.7 US	1908	22
Candido et al. (2020)	SARS-CoV-2	Brazil SchemeB	1182	10
		Brazil SchemeC	1182	22

Estimating the Marginal Posterior Distribution of Phylogenies from Molecular Sequence Data, $P(\Psi \mid X)$

The marginal posterior distribution of trees inferred in three of the empirical studies (Faria et al. 2014; Bedford et al. 2015; Candido et al. 2020) were available directly; we used these

posterior distributions of trees as the corresponding prior distributions for the second step of our sequential phylodynamic analyses. For the SARS-CoV-2 Global dataset (Gao et al. 2022), we conditioned the subsequent phylodynamic analyses on the maximum clade credibility (MCC) tree summarized from the marginal posterior distribution to ensure numerical stability of the analyses. We also conditioned on the MCC tree for the SARS-CoV-2 B.1.1.7 US dataset (Alpert et al. 2021) as original empirical study conditioned on that tree (instead of averaging over the marginal posterior distribution of trees).

The posterior distributions of trees were not published for the remainder of the empirical studies; accordingly, in these cases we first inferred the posterior distributions of trees from the corresponding sequence data, and then used the resulting posterior distributions of trees as the prior distributions in the second step of our sequential phylodynamic analyses. For each of these latter studies, we obtained the nucleotide sequences and sampling-time information from the original studies. When only the raw sequence data were available for a given study, we inferred the sequence alignment using `MUSCLE` version 3.8 (Edgar 2004). For each of the five (published or inferred) sequence alignments, we inferred the posterior probability density of trees under the identical diversification and substitution models as those used in the original studies, and then performed MCMC simulations to approximate the joint posterior distribution using `BEAST` version 1.8.2 (Drummond et al. 2012) (with `BEAGLE` version 3.1.2 [Ayres et al. 2019] enabled). Details of these analyses are available in the XML scripts included in our GitHub and Dryad repositories. For each dataset, we performed four replicate MCMC simulations; we set the length (100–200 million generations) and sampling frequency of each simulation to values that provided an adequate approximation of the posterior distribution of model parameters. We combined the posterior samples of trees from the four replicate simulations (after discarding burnin samples from each simulation) using `LogCombiner` version 1.8.2. We then subsampled the resulting composite posterior sample of trees to retain a total of 500–1000 trees (available in our GitHub and Dryad repositories); we used this posterior sample of trees as the prior distribution for the second step of the corresponding sequential analyses (detailed below).

Estimating the Joint Posterior Distribution of Geographic-Model Parameters, $P(\boldsymbol{r}, \boldsymbol{\delta}, \mu, \Psi \mid G)$

For each empirical dataset, we performed MCMC simulations to infer the joint posterior probability distribution under each prior model using `BEAST` version 1.8.2 (Drummond et al. 2012),

with BEAGLE version 3.1.2 (Ayres et al. 2019) enabled (except for the SARS-CoV-2 datasets, for which we used BEAST version 1.10.5 with BEAGLE version 3.2.0). For each candidate model, we performed four replicate MCMC simulations; we set the length (10–50 million generations) and sampling frequency of each MCMC simulation to values that provided an adequate approximation of the posterior distribution of the geographic-model parameters. Details of these analyses are available in the XML scripts included in our GitHub and Dryad repositories. In these repositories, we also provide R scripts that can be used to generate the XML scripts. We then discarded burnin samples drawn from the first 5–20% of each MCMC simulation, and then combined the remaining samples from the four replicate MCMC simulations using LogCombiner version 1.8.2. Finally, we generated the MCC tree from the composite posterior sample for each unique analysis using TreeAnnotator version 1.8.2.

Estimating the Posterior Distribution of Biogeographic History

We estimated the ancestral area at each internal node using the ancestral-state estimation algorithm (Yang 2014) implemented in BEAST. We calculated the expected number of dispersal events between each pair of areas using the "fast stochastic-mapping algorithm" developed by Minin and Suchard (2008a, see also Minin and Suchard 2008b, O'Brien et al. 2009) implemented in BEAST. The exceptions are SARS-CoV-2 datasets, where we inferred the number of dispersal events between each pair of areas by simulating the full biogeographic history using the stochastic-mapping algorithm Nielsen (2002); Hobolth and Stone (2009) implemented in BEAST. These two statistics were computed during the MCMC simulation used to infer the joint posterior distribution of geographic-model parameters, $P(\boldsymbol{r}, \boldsymbol{\delta}, \mu, \Psi \mid G)$ (*i.e.*, in the second step of our sequential analyses).

Estimating Marginal Likelihoods, $P(G)$

We used Bayes factors to evaluate the relative fit of each candidate prior model to each of the biogeographic datasets; to this end, we estimated the marginal likelihood for each prior model using both thermodynamic integration (Lartillot and Philippe 2006) and stepping-stone sampling (Xie et al. 2011; Baele et al. 2012). We ran four independent series of power-posterior simulations to estimate the marginal likelihood of each prior model. We set the chain length and sampling frequency of the power-posterior analysis at each stone, as well as the number of stones, to achieve stable marginal-likelihood estimates (both among the four replicates and

also between thermodynamic integration and stepping-stone sampling estimators). Details of these analyses are available in the XML scripts included in our GitHub and Dryad repositories.

Posterior-Predictive Simulations

We used posterior-predictive simulation to evaluate the absolute fit of each candidate prior model to each of the biogeographic datasets (Gelman et al. 1996; Bollback 2002). For each model and dataset combination, we combined the four MCMC replicates and simulated $m = 2500$ predictive datasets. For each predictive dataset, G_i^{sim}, we drew a vector of parameters, $\theta_i = \{\Psi_i, \boldsymbol{r_i}, \boldsymbol{\delta_i}, \mu_i\}$, at random from the combined MCMC samples, and simulated a dataset conditional on those parameters using the `sim.history()` function in the `R` package `phytools` (Revell 2012). We then calculated a difference statistic for the i^{th} simulated dataset as:

$$D_i = T(G_i^{\text{sim}} \mid \theta_i) - T(G^{\text{obs}} \mid \theta_i),$$

where G^{obs} is the observed biogeographic dataset, and $T(\cdot \mid \theta_i)$ is a summary statistic (detailed below). For the m predictive datasets for a given model and dataset combination, we calculated the posterior-predictive p-value as:

$$P = \left[\frac{1}{m}\sum_{i=1}^{m} D_i > 0\right] + \left[\frac{1}{2}\frac{1}{m}\sum_{i=1}^{m} D_i = 0\right],$$

where the first term measures the fraction of simulated statistics that are more extreme than the observed statistic, and the second term measures *half* the fraction of simulated statistics that are equal to the observed statisic (to accommodate discrete summary statistics, as described by Gelman et al. 2013). Posterior-predictive p-values between 0.025 and 0.975 indicating that the model is adequate and cannot be rejected (*i.e.*, the observed statistic is within the 95% posterior-predictive interval).

We used two summary statistics to assess model adequacy: (1) the *parsimony statistic*, and; (2) the *tip-wise multinomial statistic*. We calculated the posterior-predictive p-value for both of these statistics for each model and dataset combination. For the parsimony statistic, we simply calculated the parsimony score for the given simulated or observed dataset, conditional on the sampled tree, Ψ_i, using the `parsimony()` function in `R` package `phangorn` (Schliep 2010). The tip-wise multinomial statistic is similar to the multinomial statistic introduced by Goldman (1993) and used in posterior-predictive simulation by Bollback (2002), which treats the sites (columns) in a molecular alignment as outcomes of a multinomial trial. Our tip-wise statistic is

similar, but treats the states at the tips of the tree for the single geographic character (*i.e.*, site) as the outcomes of the multinomial trial. For the tip-wise multinomial statistic, we calculated:

$$T(G \mid \theta_i) = \sum_{i=1}^{k} n_i \ln(n_i/n),$$

where n is the number of tips, and n_i is the number of tips in state i. (Note that this statistic is also similar to the entropy statistic used to assess genetic variability along sequences; Shannon 1948; Schneider et al. 1986). Details about the computation of these two summary statistics are available in the `R` script included in our GitHub and Dryad repositories.

Data Cloning

We explored the use of a computational technique called *data cloning* to understand the sensitivity of posterior estimates to the choice of prior. Originally developed as a tool for using MCMC to perform maximum-likelihood inference (Robert 1993), and later used as a tool for understanding model identifiability for complex Bayesian models (Lele et al. 2007; Ponciano et al. 2009, 2012), data cloning involves performing a sequence of MCMC analyses with an increasing number of duplicates of the observed data. A particular MCMC in the sequence is defined by the number of duplicated datasets, $\beta_i \geq 1$, with the resulting posterior distribution being:

$$P(\theta \mid X)_{\beta_i} \propto P(X \mid \theta)^{\beta_i} P(\theta).$$

As $\beta_i \to \infty$ (assuming the model is identifiable), the joint posterior distribution converges to a point that corresponds to the joint maximum-likelihood estimate (MLE); if the joint posterior distribution does not converge to a point, then the model is non-identifiable (*i.e.*, the MLE may not be unique). When the model is identifiable, the rate at which the joint posterior distribution converges to the MLE is proportional to the amount of information available in the data relative to the strength of the prior, *i.e.*, when the prior is extremely (mis)informative, convergence to the MLE will be very slow.

We used data cloning to analyze each of the biogeographic datasets under the symmetric default- and alternative-prior models, with $\beta = \{1, 5, 10, 20\}$. This is achieved by duplicating the discrete-geography data in the `BEAST` XML scripts. These XML scripts are available in our GitHub and Dryad repositories. In these repositories, we also provide `R` scripts that can be used to generate the XML scripts.

To ensure good MCMC performance, we conducted the biogeographic inferences conditioning on the MCC tree inferred using the sequence data. In all cases, the inferred posterior distributions shrink as β increases. Under the alternative-prior model, the posterior-mean estimates remain mostly constant as β increases (*i.e.*, the posterior-mean estimates are almost identical to the MLEs). By contrast, the posterior-mean estimates under the default-prior model change drastically as β increases, converging to the MLEs very slowly (Figs. S.1.12 and S.1.13). These results indicate that the default-prior models exert much stronger influence on posterior estimates relative to the alternative-prior models.

MCMC Diagnosis

After initial inspection of the output log files using `Tracer` version 1.7.1 (Rambaut et al. 2018), we assessed MCMC performance using the `coda` package (Plummer et al. 2006) in `R` (R Core Team 2020). Specifically, we assess mixing and adequacy within each MCMC replicate by calculating the effective sample size (ESS) diagnostic for each continuous parameter (ensuring ESS values $\gg 100$) after discarding the first 5–20% of samples from each replicate simulation as the burn-in. We assessed convergence among replicate MCMC simulations by calculating the potential scale reduction factor (PSRF Gelman and Rubin 1992) diagnostic for each continuous parameter (ensuring R $\approx$ 1). We also assessed the convergence among replicates by calculating the ESS for each continuous parameter for each combined MCMC chain (independent replicates combined after discarding the burn-in), ensuring the ESS values $\gg 200$.

Parameter Summaries

As described in the Empirical Consequences section of the main text, for each dataset we summarized the following statistics: (1) marginal likelihood; (2) posterior-predictive summary statistics; (3) rate of dispersal between each pair of areas; (4) support for dispersal routes between each pair of areas; (5) average dispersal rate among all areas; (6) ancestral area at each internal node of the phylogeny; (7) total number of dispersal events among all areas, and; (8) number of dispersal events between each pair of areas. The `R` scripts used to summarize these statistics are available in our GitHub and Dryad repositories. We compared estimates of these statistics across all candidate models to assess the impact of the default and alternative priors. We report the meta summaries of these statistics across all the empirical datasets in this section, and provide these summaries for each dataset in detail in this section.

Expanded Meta Summaries of Empirical Analyses

In this section, we provide various summaries across all the empirical datasets. In the main text, we focus on comparisons between the preferred default-prior model and the preferred alternative-prior model; here, we provide results for pairwise comparisons between all the eight candidate prior models.

The Impact of Prior Choice on Pairwise Dispersal Rates

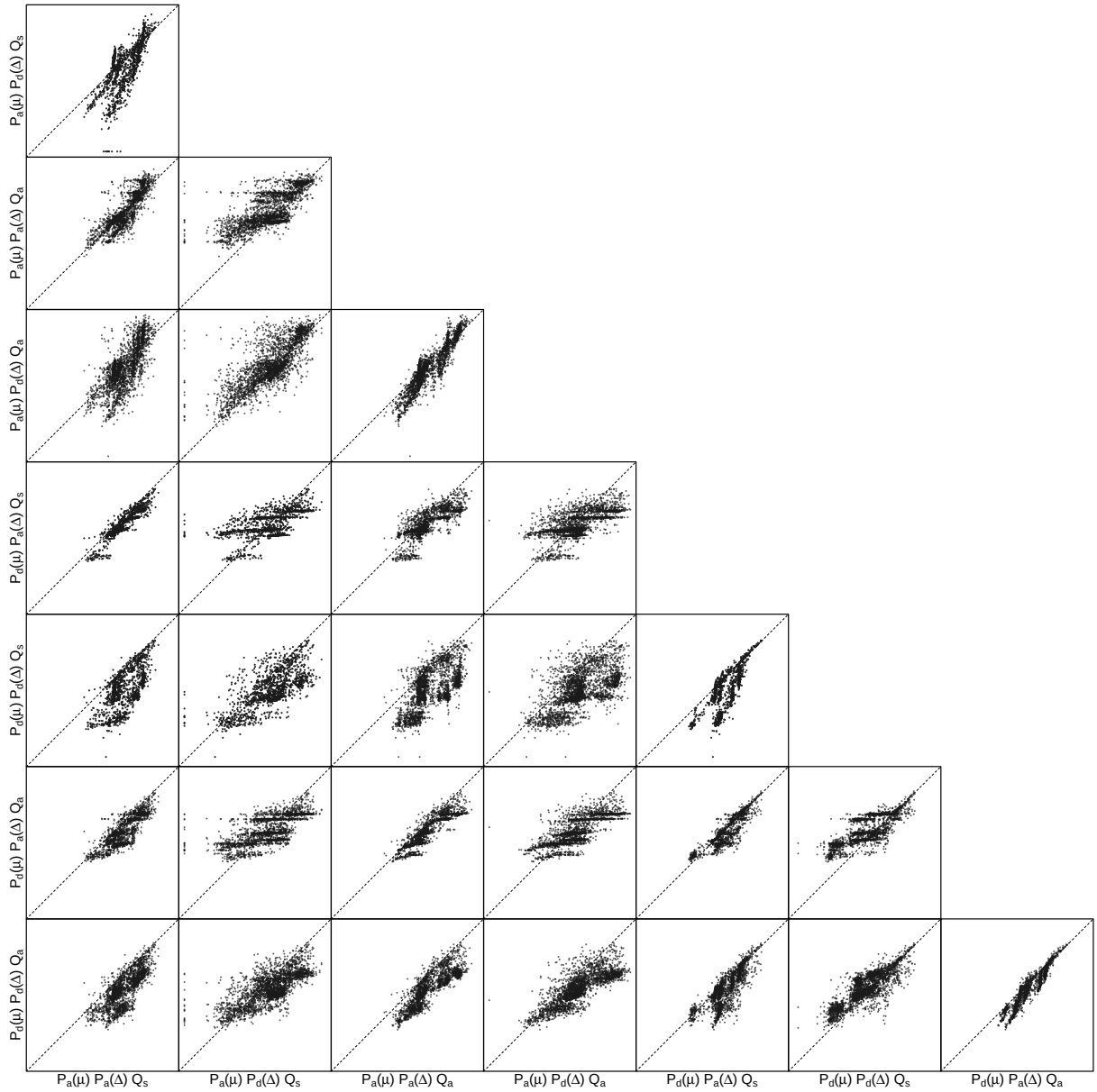

Figure S.1.4: The impact of prior choice on pairwise dispersal rates. Each cell of the plot compares posterior-mean estimate of the rate of dispersal between each pair of geographic areas, q_{ij}, between each pair of prior models, summarized across all datasets. Axis label notation for the prior models follows Table S.1.2.

The Impact of Prior Choice on Average Dispersal Rate

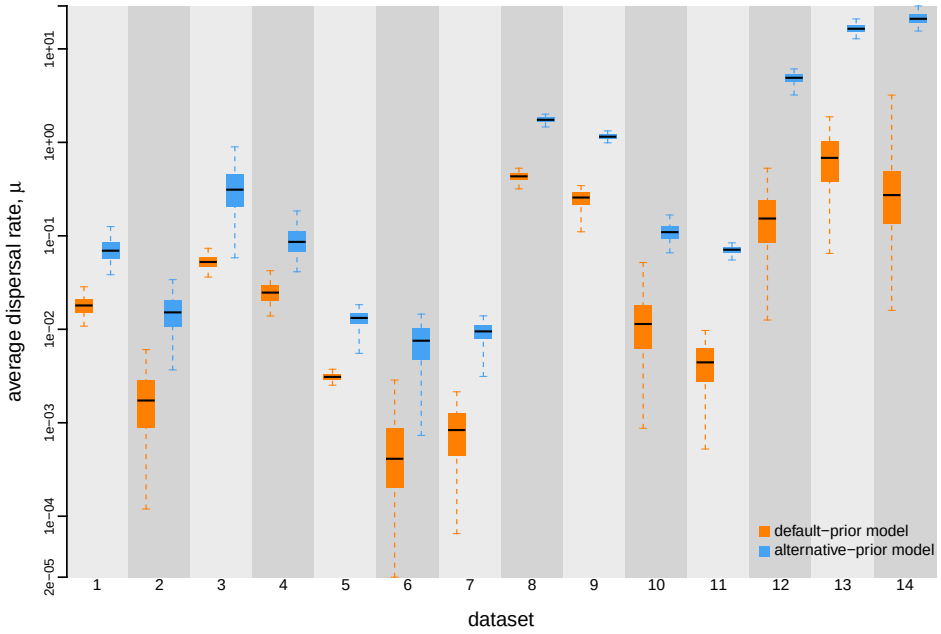

Figure S.1.5: The impact of prior choice on the average dispersal rate. Each column depicts estimates for one empirical dataset (see Table S.1.5 for the description of datasets). Within each column, each pair of boxplots depicts posterior estimates of the average dispersal rate, μ, under the default (orange) and alternative (blue) prior models: the center of each box indicates the posterior-median rate; the box and whiskers indicate the corresponding 50% and 95% credible intervals, respectively.

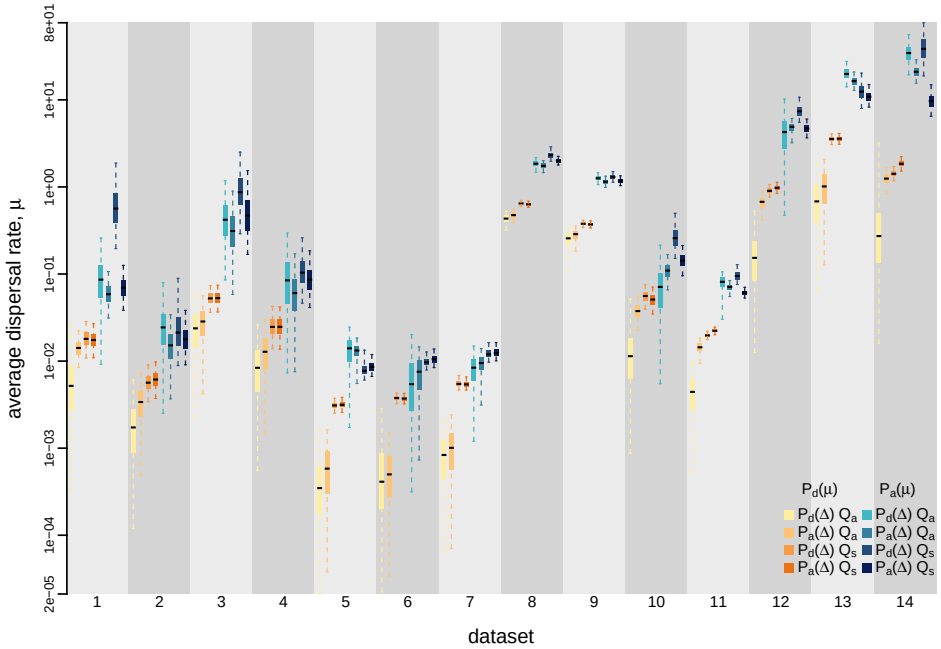

Figure S.1.6: The impact of prior choice on the average dispersal rate. Each column depicts estimates for one of the 14 datasets (description of datasets see Table S.1.5). Within each column, the set of eight boxplots depicts posterior estimates of the average dispersal rate, μ, under the prior models: the center of each box indicates the posterior-median rate; the box and whiskers indicate the corresponding 50% and 95% credible intervals.

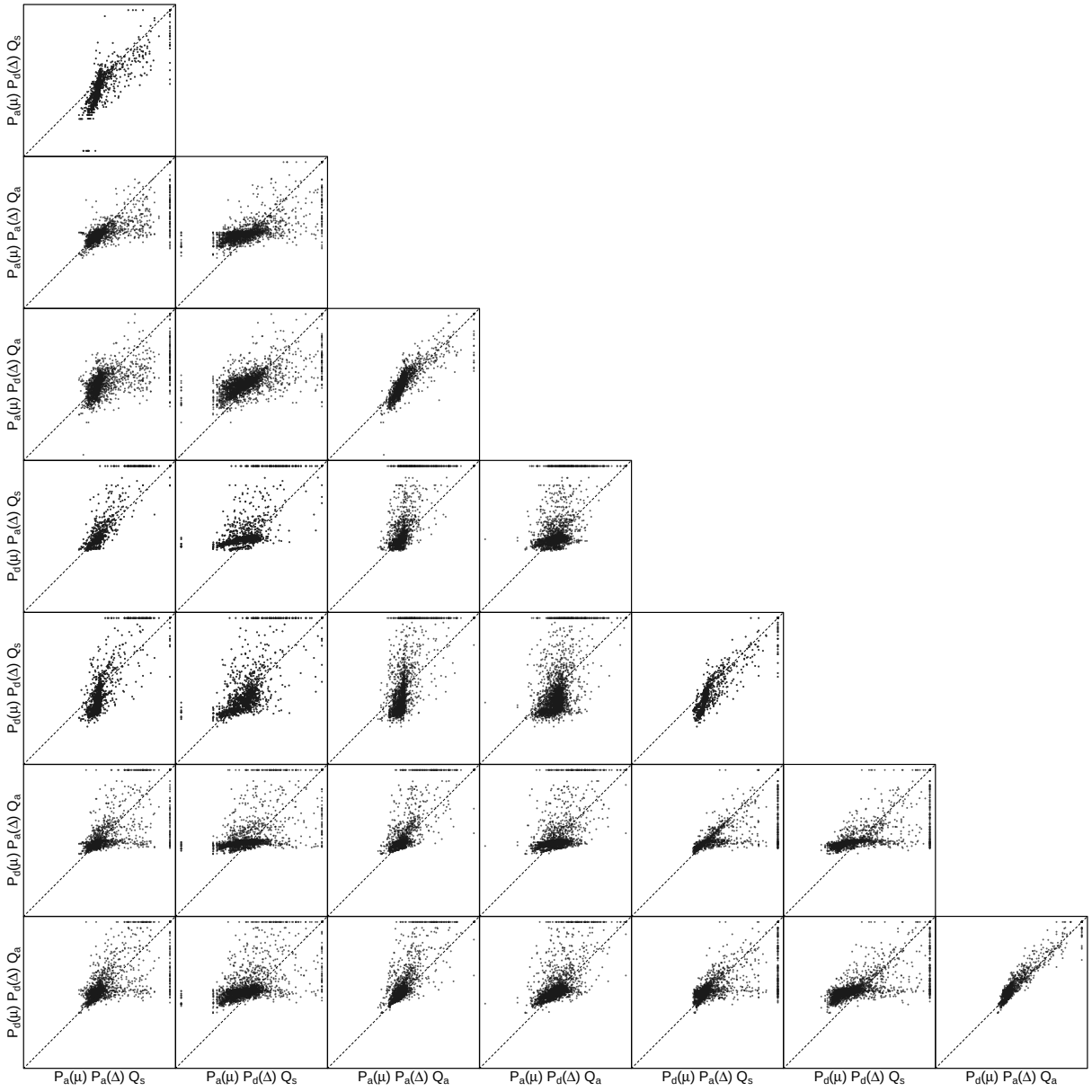

Figure S.1.7: The impact of prior choice on the inferred support for dispersal routes. Each cell of the plot compares the inferred support (2 ln BF) for pairwise dispersal routes between each pair of prior models, summarized across all datasets. Axis label notation for the prior models follows Table S.1.2.

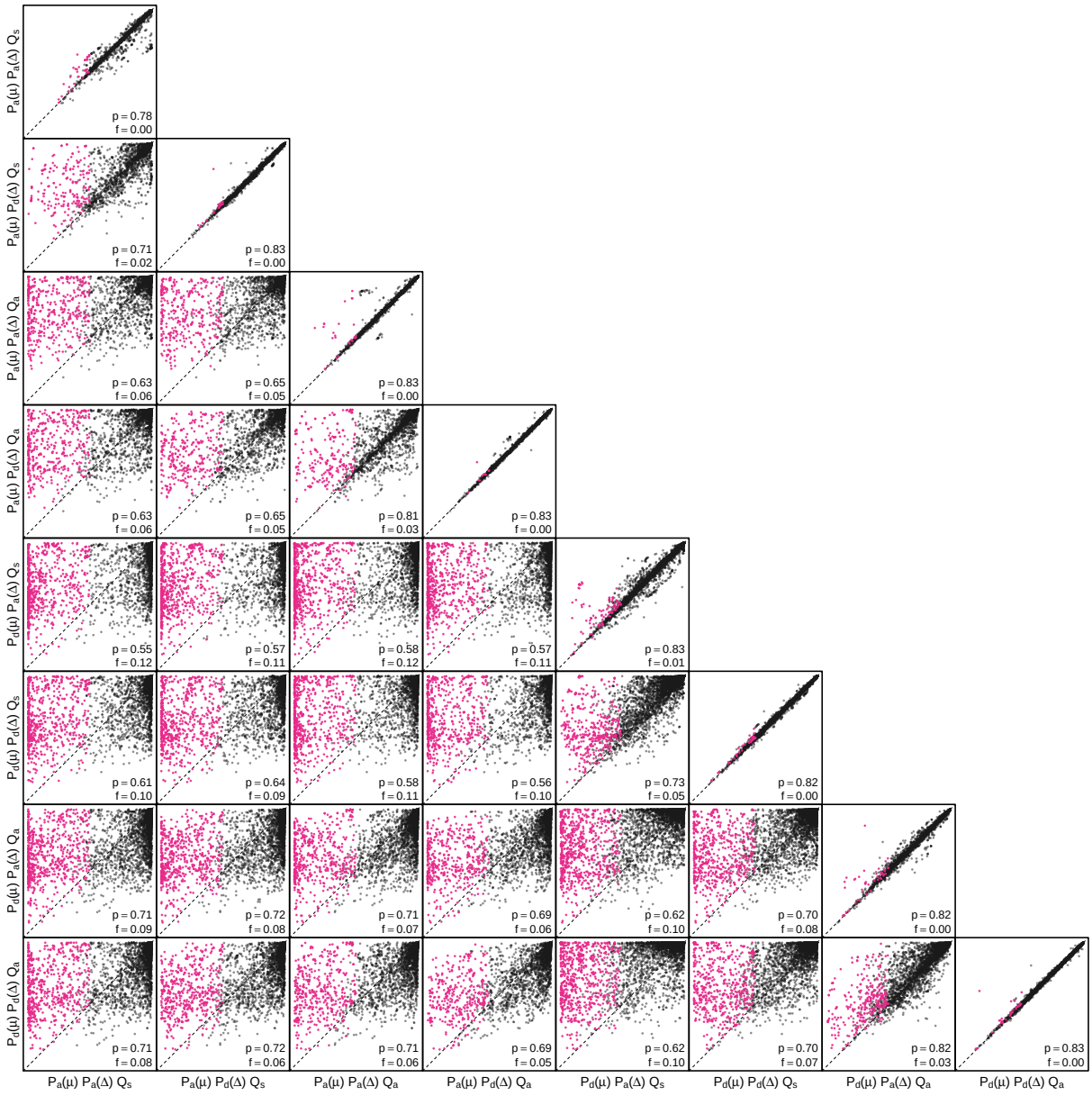

Figure S.1.8: The impact of prior choice on ancestral-area estimates. Each cell of the plot compares the estimated posterior probability of the maximum a posteriori (MAP) ancestral area between each pair of prior models, summarized across all datasets. Diagonal cells are comparisons between two replicates under the same prior model, assessing the convergence of MCMC simulations; off-diagonal cells are comparisons between different prior models, demonstrating the impact of the prior model on both the posterior probability and the identity of the MAP ancestral-area estimates at internal nodes. Pink dots represent the internal nodes where the MAP ancestral area inferred under the default-prior model differs from that inferred under the alternative-prior models. The statistic p denotes the fraction of internal nodes that are shared under the default- and alternative-prior models; f is the fraction of shared nodes where the MAP ancestral area differs under the default- and alternative-prior models. Axis label notation for the prior models follows Table S.1.2.

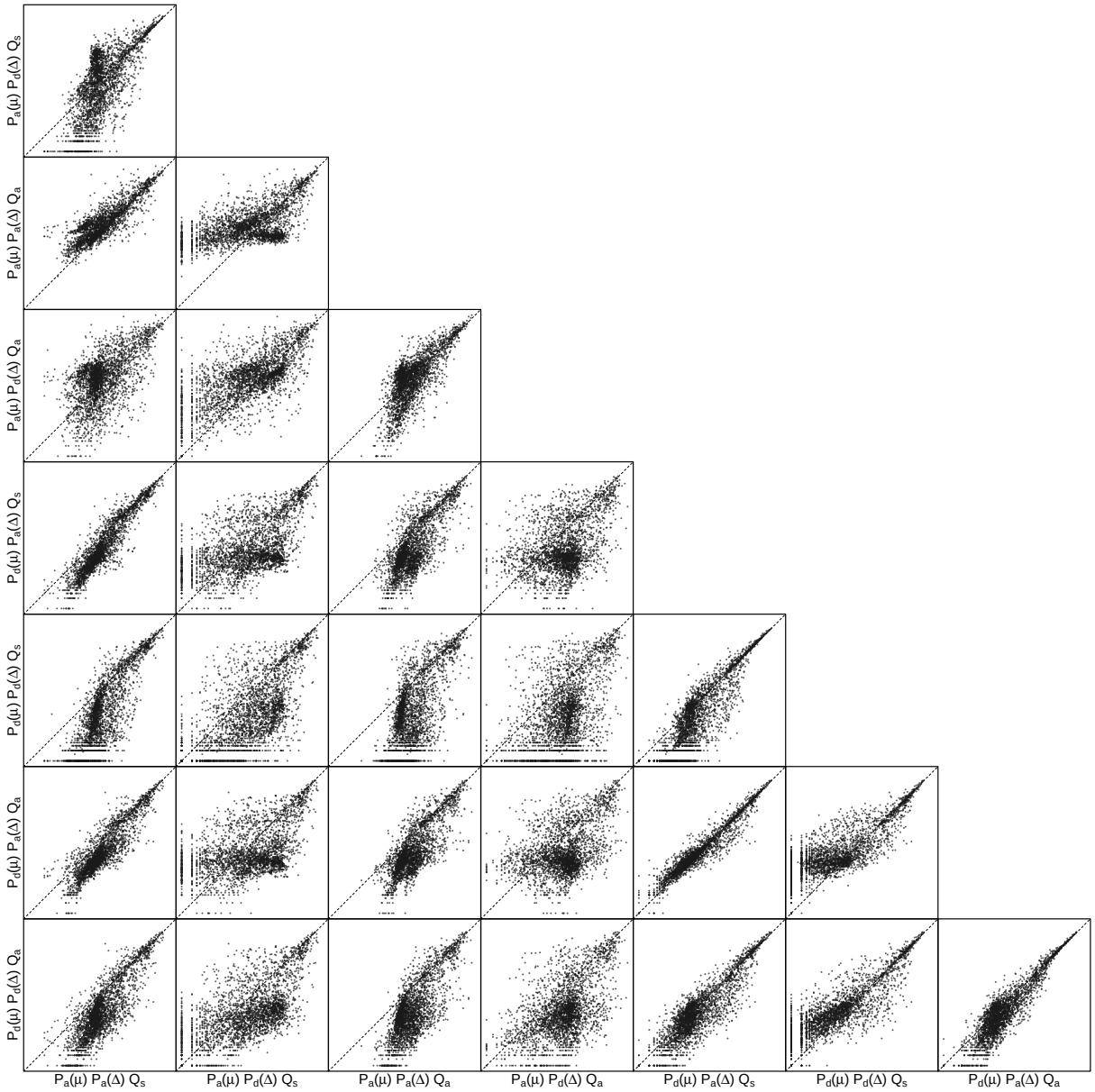

Figure S.1.9: The impact of prior choice on the inferred number of dispersal events between each pair of areas. Each cell of the plot compares the inferred number of pairwise dispersal events between each pair of prior models, summarized across all datasets. Model notation on the axis follows Table S.1.2.

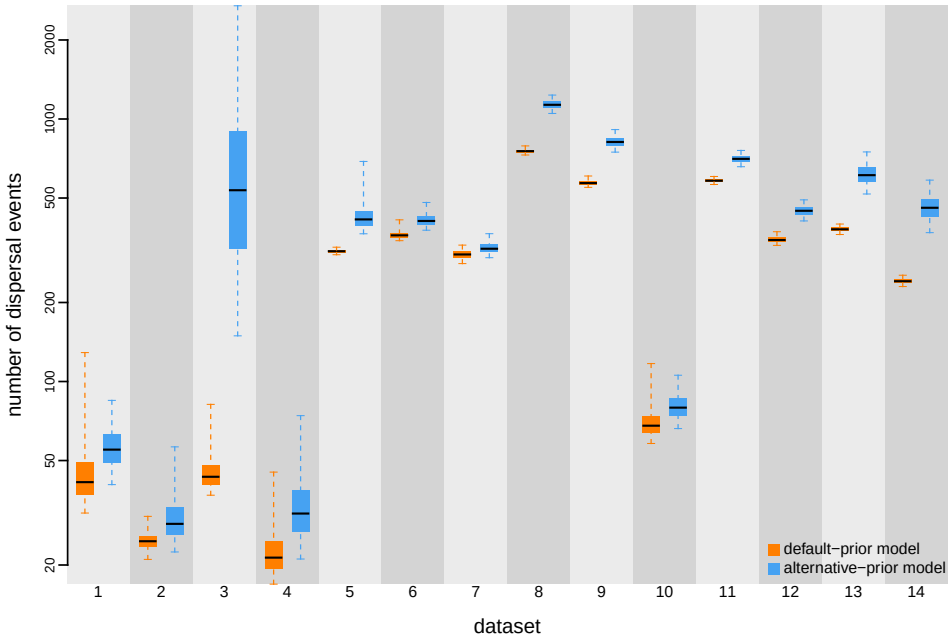

Figure S.1.10: The impact of prior choice on the inferred total number of dispersal events between all areas. Each column depicts estimates for one of the 14 datasets (description of datasets see Table S.1.5). Within each column, the pair of boxplots depicts posterior estimates of the total number of dispersal events under the default (orange) and alternative (blue) prior models: the center of each box indicates the posterior-median number of dispersal events; the box and whiskers indicate the corresponding 50% and 95% credible intervals, respectively.

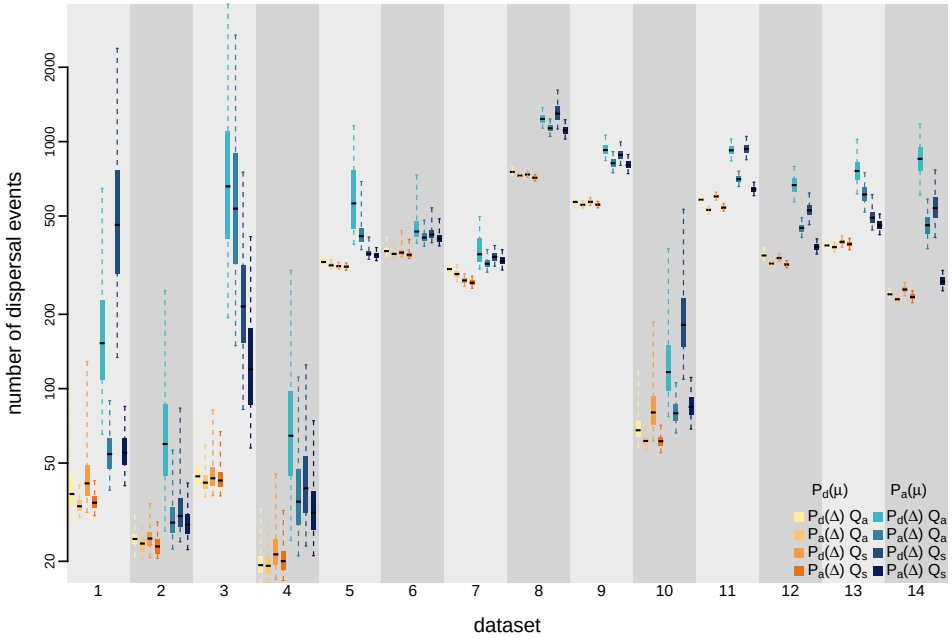

Figure S.1.11: The impact of prior choice on the inferred total number of dispersal events between all areas. Each column depicts estimates for one of the 14 datasets (description of datasets see Table S.1.5). Within each column, the set of eight boxplots depicts posterior estimates of the total number of dispersal events under each of the prior models: the center of each box indicates the posterior-median number of dispersal events; the box and whiskers indicate the corresponding 50% and 95% credible intervals, respectively.

Data Cloning

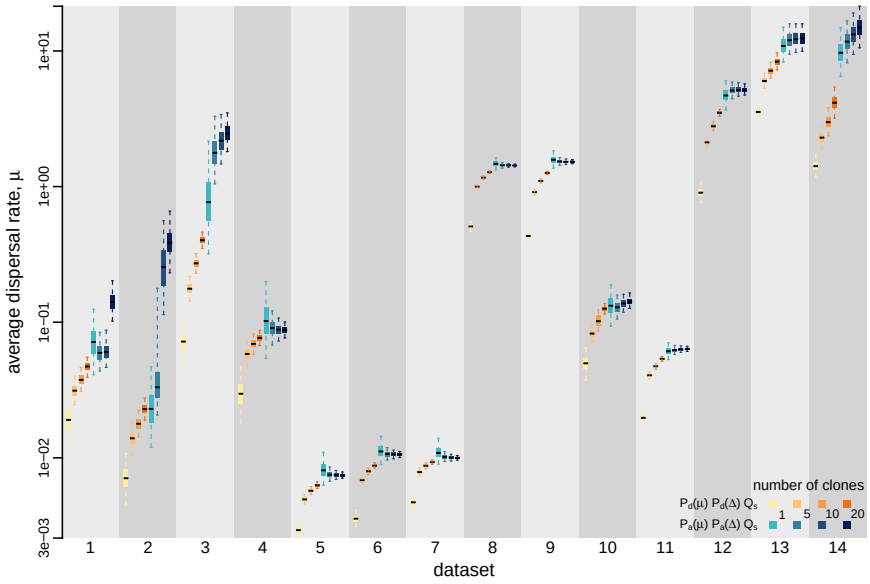

Figure S.1.12: Using data cloning to explore the impact of prior choice on posterior estimates of the average dispersal rate. Each column depicts estimates for one of the 14 datasets (description of datasets see Table S.1.5). Within each column, the set of eight boxplots depicts posterior estimates inferred from the associated dataset that has been cloned 1, 5, 10, 20 times: the center of each box indicates the posterior-median average dispersal rate; the box and whiskers indicate the corresponding 50% and 95% credible intervals, respectively.

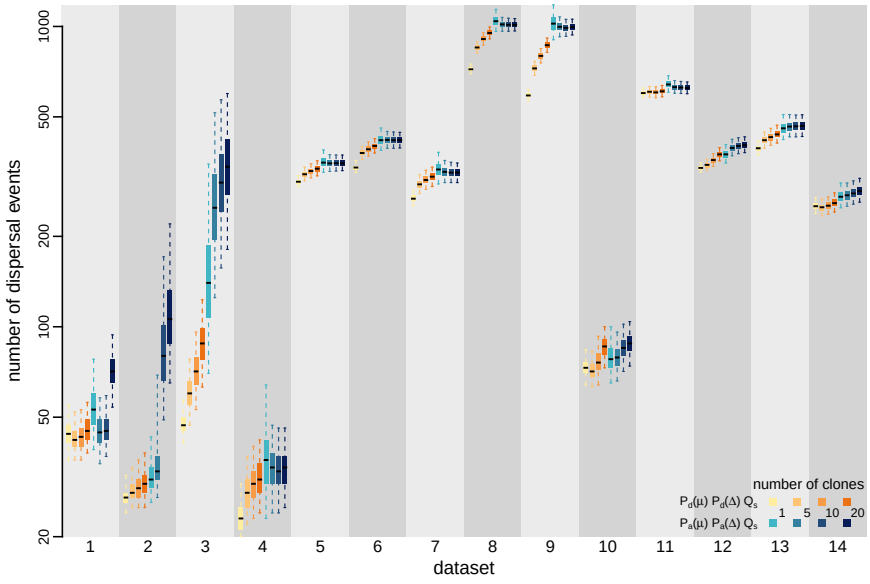

Figure S.1.13: Exploring the impact of prior choice on posterior estimates of the number of dispersal routes via data cloning. Each column depicts estimates for one of the 14 datasets (description of datasets see Table S.1.5). Within each column, the set of eight boxplots depicts posterior estimates inferred from the associated dataset that has been cloned 1, 5, 10, 20 times: the center of each box indicates the posterior-median number of dispersal events; the box and whiskers indicate the corresponding 50% and 95% credible intervals, respectively.

Expanded Dataset-Specific Summaries of Empirical Analyses

In this section, we provide dataset-specific summaries of the statistics we reported in the section above. These results reveal a highly consistent pattern across all the datasets, further demonstrating the widespread impact of prior misspecification in biogeographic inferences and the universality of this issue.

Dengue Virus

Dash et al. (2015) studied the geographic dynamics of Dengue virus type 1 (DENV-1) in India and inferred the history by which this virus dispersed throughout the world. This study contains a single dataset comprised of sequences of (part of) the envelope gene, sampled from across a large number of distant geographic areas over a protracted sampling interval (1956–2011). We acquired the sampling time and location data, as well as the GenBank accession numbers from the sequence names in the MCC tree figured in Dash et al. (2015, Fig. 3), and then obtained the nucleotide sequences from GenBank. This dataset has 62 sequences distributed among 23 defined geographic areas. We aligned the nucleotide sequences using `MUSCLE` version 3.8 (Edgar 2004). The files containing the GenBank accession numbers, the sequence alignment, and the sampling time and location data are available in our GitHub and Dryad repositories.

To infer the marginal posterior distribution of phylogenies given the sequence alignment, we specified a phylogenetic model with the following components: (1) the GTR+I+Γ_4 substitution model (Tavaré 1986; Yang 1994; Gu et al. 1995); (2) the uncorrelated lognormal (UCLN) branch-rate prior model (Drummond et al. 2006; Rannala and Yang 2007), and; (3) the Gaussian Markov Random Field (GMRF) Bayesian Skyride coalescent node-age model (Minin et al. 2008). Details of these analyses are available in the XML scripts included in our GitHub and Dryad repositories.

We ran four independent MCMC simulations in `BEAST` version 1.8.2 for 200 million generations each, sampling every 15000 generations. We first assessed the performance of each MCMC simulation using `Tracer` version 1.7.1 (Rambaut et al. 2018), removed the first 10% of samples from each chain as the burn-in, and then combined the remaining posterior samples of trees from the replicate simulations using `LogCombiner` version 1.8.2. This resulted in a posterior sample of 1200 trees (available in our GitHub and Dryad repositories), which we then used as the prior distribution of phylogenies for the second step of our sequential analyses.

The MCMC simulations used in the second step of our sequential analyses and of the analyses used to estimate marginal likelihoods under each prior model are described above in this section. Details of these analyses are available in the XML scripts included in our GitHub and Dryad repositories.

The Impact of Prior Choice on Biogeographic Model Fit

Table S.1.6: Marginal-likelihood estimates of the eight prior models for the Dengue virus dataset. Columns 2–5 list marginal likelihoods inferred from four replicate analyses; the last two columns list the mean and standard deviation of these marginal-likelihood estimates. Candidate models are listed in rows and include all possible combinations of: (1) instantaneous-rate matrices (symmetric, Q_s or asymmetric, Q_a); (2) priors on the average dispersal rate [default, $P_d(\mu)$ or alternative, $P_a(\mu)$], and; (3) priors on the number of dispersal routes [default, $P_d(\Delta)$ or alternative, $P_a(\Delta)$]. The preferred default- and alternative-prior models are indicated in bold text.

Model	replicate1	replicate2	replicate3	replicate4	mean	sd
$P_d(\mu)Q_aP_d(\Delta)$	-193.12	-192.95	-192.90	-192.38	-192.84	0.32
$P_d(\mu)Q_aP_a(\Delta)$	-170.72	-171.16	-170.55	-170.90	-170.83	0.26
$\boldsymbol{P_d(\mu)Q_sP_d(\Delta)}$	**-190.44**	**-190.62**	**-190.47**	**-190.39**	**-190.48**	**0.10**
$P_d(\mu)Q_sP_a(\Delta)$	-170.91	-171.57	-171.24	-171.15	-171.22	0.27
$P_a(\mu)Q_aP_d(\Delta)$	-158.16	-158.68	-158.82	-158.68	-158.58	0.29
$P_a(\mu)Q_aP_a(\Delta)$	-148.27	-148.10	-148.14	-148.31	-148.21	0.10
$P_a(\mu)Q_sP_d(\Delta)$	-152.76	-152.36	-152.54	-152.69	-152.59	0.18
$\boldsymbol{P_a(\mu)Q_sP_a(\Delta)}$	**-147.24**	**-147.46**	**-147.24**	**-147.33**	**-147.32**	**0.11**

The Impact of Prior Choice on Pairwise Dispersal Rates

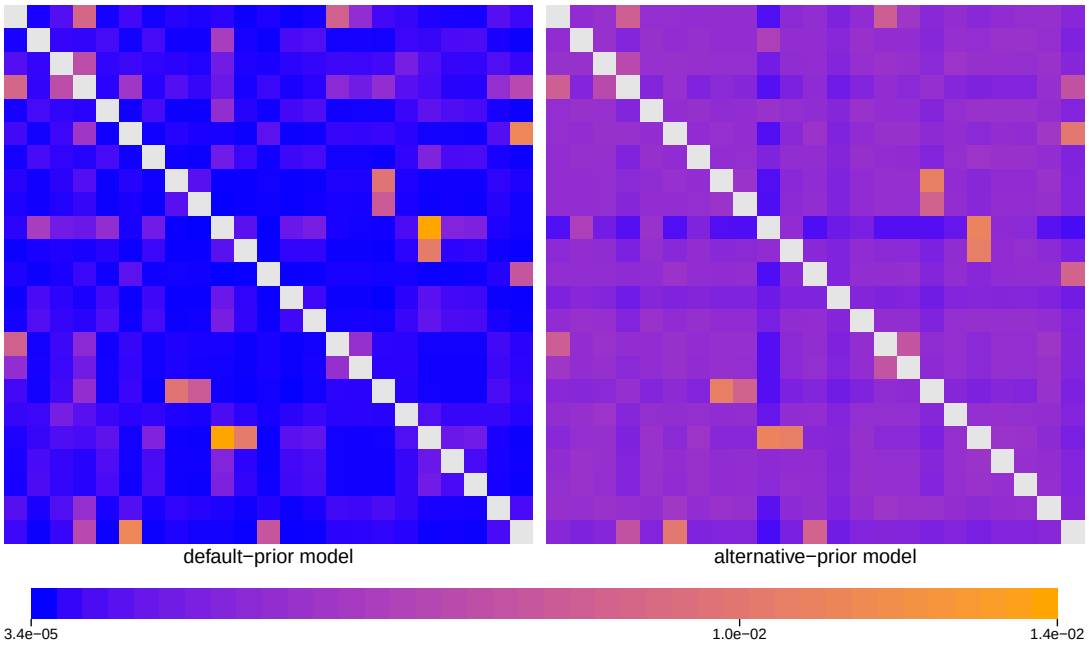

Figure S.1.14: The impact of prior choice on pairwise dispersal rates for the Dengue virus dataset. Heatmaps summarize posterior-mean estimates of the instantaneous rate of dispersal between each pair of geographic areas, q_{ij}, under the default (left) and alternative (right) prior models.

The Impact of Prior Choice on the Inferred Support for Dispersal Routes

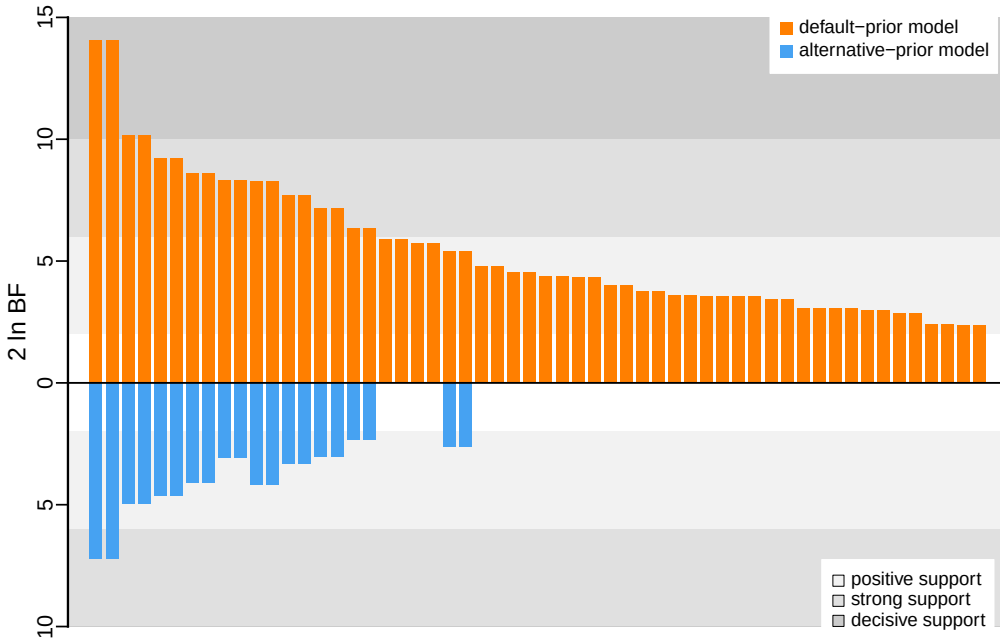

Figure S.1.15: The impact of prior choice on the inferred support for dispersal routes for the Dengue virus dataset. We compare the evidential support for each dispersal route for the Dengue virus dataset under the default (orange) and alternative (blue) prior models. Each bar indicates the $2\ln \mathrm{BF}$ for the corresponding dispersal route between two areas; only supported dispersal routes (*i.e.*, $2\ln \mathrm{BF} > 2$) are plotted.

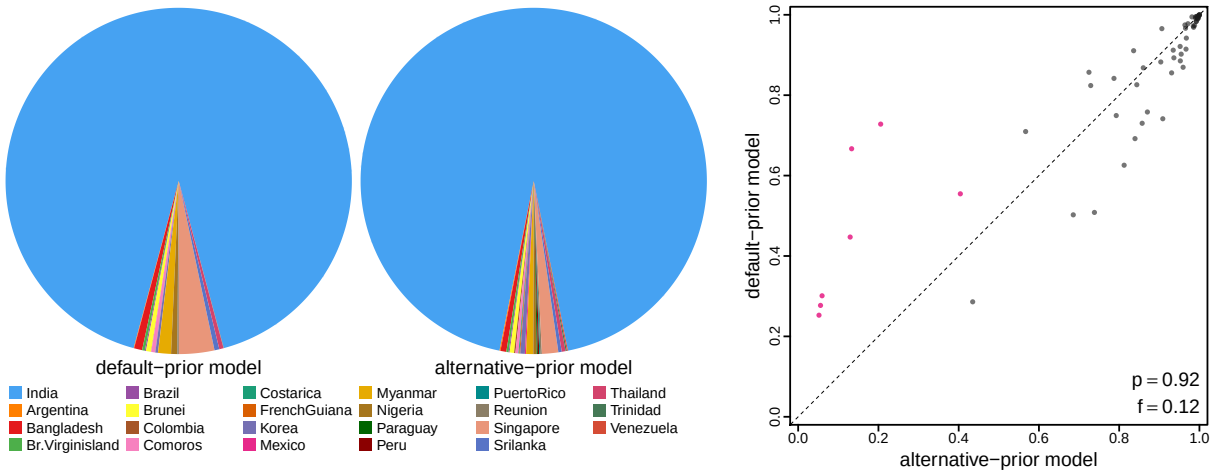

Figure S.1.16: The impact of prior choice ancestral-area estimates for the Dengue virus dataset. The left panel compares the posterior probability of each ancestral area at the root node under the default- and alternative-prior models for the Dengue virus dataset. The right panel plots the posterior probability of the most probable ancestral area under the default-prior model for each node in the MCC tree (y-axis) against the corresponding posterior probability of that area under the alternative-prior model (x-axis). Pink dots represent the internal nodes where the MAP ancestral area inferred under the default-prior model differs from that inferred under the alternative-prior models. The statistic p denotes the fraction of internal nodes that are shared under the default- and alternative-prior models; f, is the fraction of shared nodes where the MAP ancestral area differs under the default- and alternative-prior models. Note that the posterior probabilities of the MAP ancestral area under the default-prior model are generally higher than those under the alternative-prior model (*i.e.*, the default-prior model tends to mask uncertainty in the ancestral-area estimates).

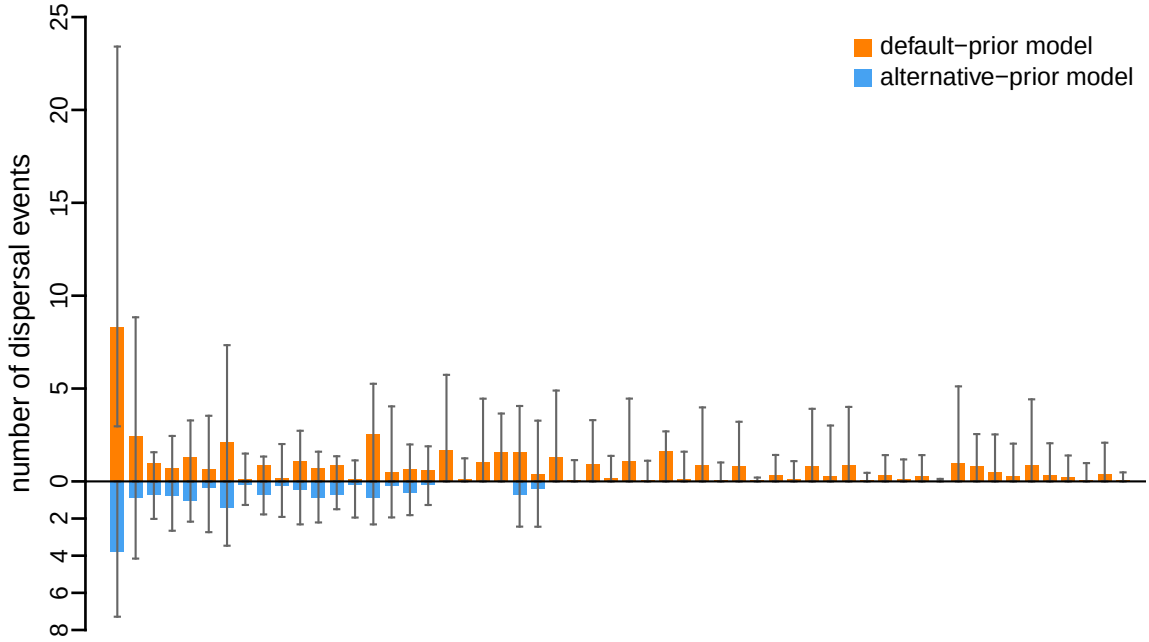

Figure S.1.17: The impact of prior choice on the inferred number of dispersal events between each pair of areas for the Dengue virus dataset. The reflected bar plot depicts the number of dispersal events inferred under the default (orange) and alternative (blue) prior models for the Dengue virus dataset. Each bar indicates the posterior-mean number of dispersal events between a pair of areas; whiskers indicate the 95% credible interval. Note that only the number of dispersal events over the "significant" dispersal routes (*i.e.*, $2 \ln \mathrm{BF} > 2$) are figured.

Deformed Wing Virus

Wilfert et al. (2016) explored the role of Varroa in the spread of DWV in honeybees (*i.e.*, whether they were the source of the virus or merely facilitated its spread among honeybees), and also identified significant dispersal routes of DWV between geographic populations. This study contains six datasets, including three molecular sequence alignments (lp, rdrp, and vp3) and two discrete-trait datasets (geographic areas and host species). We re-inferred the biogeographic history for each of the three molecular datasets (see Wilfert et al. (2016) for details about these datasets).

We acquired the `BEAST` XML scripts used in the original study—containing both the sequence alignment and sampling time and geographic location data—directly from the authors. For each gene region, we inferred the marginal posterior distribution of phylogenies under the phylogenetic models identical to those specified in Wilfert et al. (2016). Specifically, for each gene region we partitioned the alignments into two subsets (where the first subset included sites at the first and second codon positions, and the second subset included sites at the third codon position), and specified independent substitution models for each of these two partitions. Our phylogenetic models assume that the two data partitions share the same uncorrelated exponential (UCED) branch-rate model (Drummond et al. 2006; Rannala and Yang 2007), and the same exponential coalescent node-age model, but we specified a rate multiplier for each data subset to allow the average substitution rate to vary among data partitions. We constrained the mean of these rate multipliers to one so that these rates are identifiable under the uncorrelated branch-rate model. The phylogenetic models were identical for all three gene regions except for the substitution model specified for the partitions of each alignment. Following Wilfert et al. (2016), we specified independent TN93+I+Γ_4 model (Tamura and Nei 1993; Gu et al. 1995; Yang 1994) for the two partitions of the lp fragment, and specified independent HKY+Γ_4 model (Hasegawa et al. 1984, 1985) for the two partitions of the rdrp fragment, and specified independent HKY+I model for the two partitions of the vp3 fragment. Details of these analyses are available in the XML scripts included in our GitHub and Dryad repositories.

For each gene region, we ran four independent MCMC simulations in `BEAST` version 1.8.2 for 100 million generations each, sampling every 10000 generations (except for the rdrp fragment, where we it was necessary to run simulations for 200 million cycles and sample every 20000 generations to achieve adequate MCMC performance). We first assessed the performance

of each MCMC simulation, removed the first 10% of samples from each chain as the burn-in, and then combined the remaining posterior samples of trees from the replicate simulations using `LogCombiner` version 1.8.2. This resulted in a posterior sample of 360 trees (available in our GitHub and Dryad repositories), which we then used as the prior distribution of phylogenies.

The MCMC simulations used in the second step of our sequential analyses and of the analyses used to estimate marginal likelihoods under each prior model are described above in this section. Details of these analyses are available in the XML scripts included in our GitHub and Dryad repositories.

Lp fragment

The Impact of Prior Choice on Biogeographic Model Fit

Table S.1.7: Marginal-likelihood estimates of the eight prior models for the Deformed wing virus lp fragment dataset. Columns 2–5 list marginal likelihoods inferred from four replicate analyses; the last two columns list the mean and standard deviation of these marginal-likelihood estimates. Candidate models are listed in rows and include all possible combinations of: (1) instantaneous-rate matrices (symmetric, Q_s or asymmetric, Q_a); (2) priors on the average dispersal rate [default, $P_d(\mu)$ or alternative, $P_a(\mu)$], and; (3) priors on the number of dispersal routes [default, $P_d(\Delta)$ or alternative, $P_a(\Delta)$]. The preferred default- and alternative-prior models are indicated in bold text.

Model	replicate1	replicate2	replicate3	replicate4	mean	sd
$\boldsymbol{P_d(\mu)Q_aP_d(\Delta)}$	**-144.01**	**-144.10**	**-144.14**	**-144.08**	**-144.08**	**0.05**
$P_d(\mu)Q_aP_a(\Delta)$	-140.28	-140.33	-140.46	-140.68	-140.44	0.18
$P_d(\mu)Q_sP_d(\Delta)$	-145.12	-145.14	-145.03	-144.95	-145.06	0.09
$P_d(\mu)Q_sP_a(\Delta)$	-141.90	-141.88	-141.62	-141.63	-141.76	0.15
$P_a(\mu)Q_aP_d(\Delta)$	-130.94	-131.39	-130.97	-131.01	-131.08	0.21
$\boldsymbol{P_a(\mu)Q_aP_a(\Delta)}$	**-128.78**	**-128.90**	**-128.83**	**-128.69**	**-128.80**	**0.09**
$P_a(\mu)Q_sP_d(\Delta)$	-133.11	-133.43	-133.43	-133.48	-133.36	0.17
$P_a(\mu)Q_sP_a(\Delta)$	-129.85	-129.66	-129.76	-129.71	-129.74	0.08

The Impact of Prior Choice on Pairwise Dispersal Rates

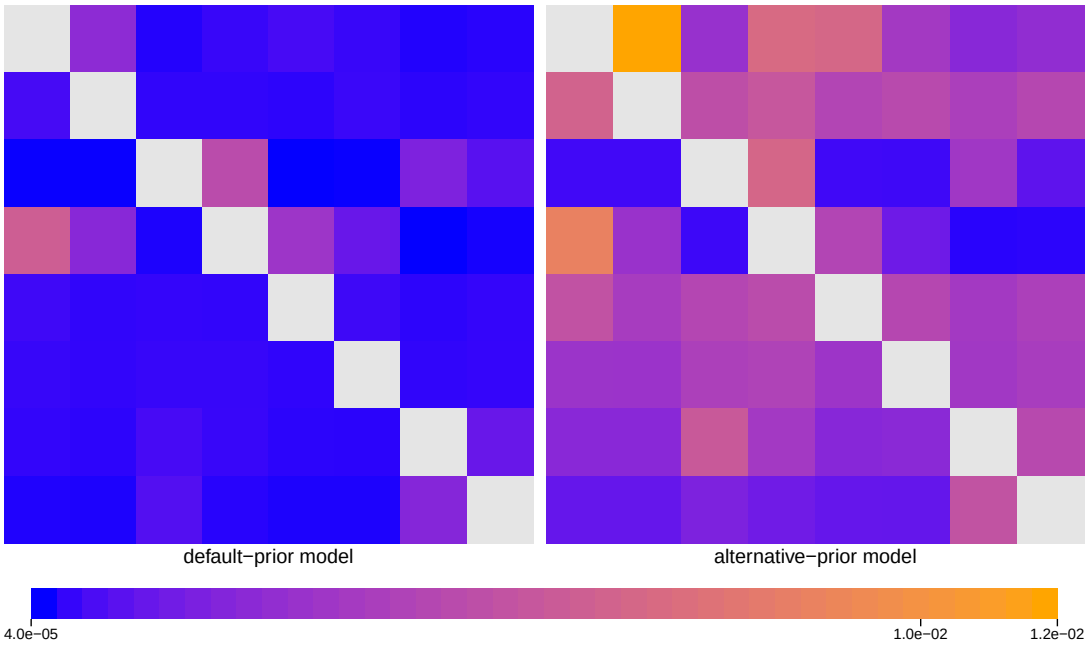

Figure S.1.18: The impact of prior choice on pairwise dispersal rates for the Deformed wing virus lp fragment dataset. Heatmaps summarize posterior-mean estimates of the instantaneous rate of dispersal between each pair of geographic areas, q_{ij}, under the default (left) and alternative (right) prior models.

The Impact of Prior Choice on the Inferred Support for Dispersal Routes

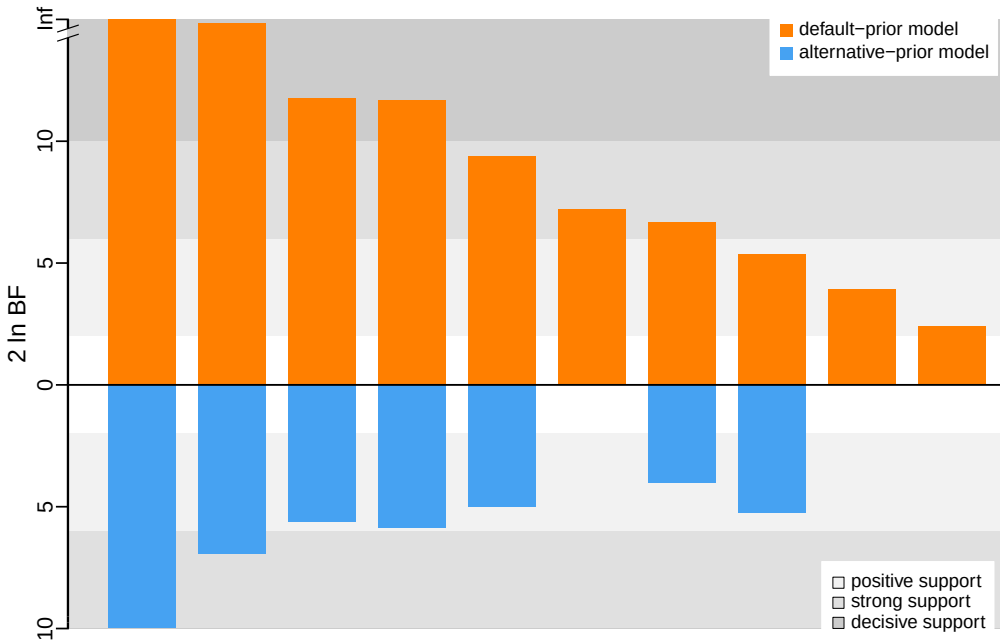

Figure S.1.19: The impact of prior choice on the inferred support for dispersal routes for the Deformed wing virus lp fragment dataset. We compare the evidential support for each dispersal route for the Deformed wing virus lp fragment dataset under the default (orange) and alternative (blue) prior models. Each bar indicates the $2\ln \mathrm{BF}$ for the corresponding dispersal route between two areas; only supported dispersal routes (*i.e.*, $2\ln \mathrm{BF} > 2$) are plotted.

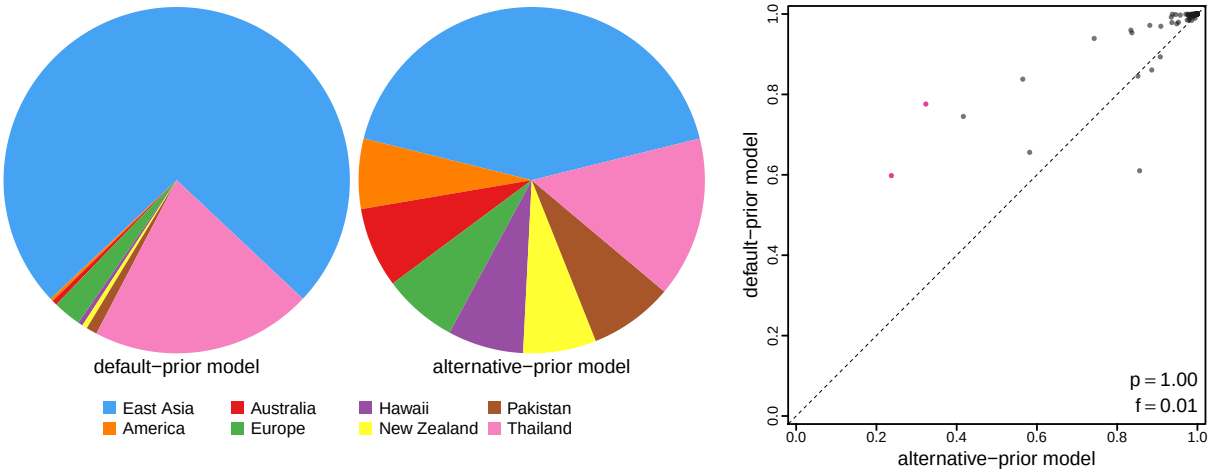

Figure S.1.20: The impact of prior choice ancestral-area estimates for the Deformed wing virus lp fragment dataset. The left panel compares the posterior probability of each ancestral area at the root node under the default- and alternative-prior models for the Deformed wing virus lp fragment dataset. The right panel plots the posterior probability of the most probable ancestral area under the default-prior model for each node in the MCC tree (y-axis) against the corresponding posterior probability of that area under the alternative-prior model (x-axis). Pink dots represent the internal nodes where the MAP ancestral area inferred under the default-prior model differs from that inferred under the alternative-prior models. The statistic p denotes the fraction of internal nodes that are shared under the default- and alternative-prior models; f, is the fraction of shared nodes where the MAP ancestral area differs under the default- and alternative-prior models. Note that the posterior probabilities of the MAP ancestral area under the default-prior model are generally higher than those under the alternative-prior model (*i.e.*, the default-prior model tends to mask uncertainty in the ancestral-area estimates).

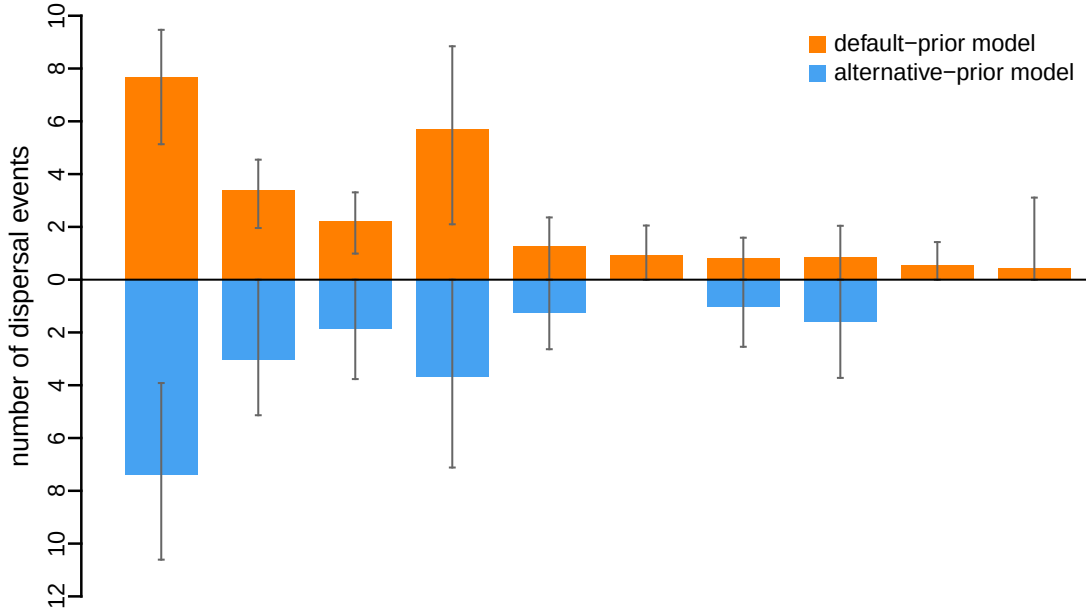

Figure S.1.21: The impact of prior choice on the inferred number of dispersal events between each pair of areas for the Deformed wing virus lp fragment dataset. The reflected bar plot depicts the number of dispersal events inferred under the default (orange) and alternative (blue) prior models for the Deformed wing virus lp fragment dataset. Each bar indicates the posterior-mean number of dispersal events between a pair of areas; whiskers indicate the 95% credible interval. Note that only the number of dispersal events over the "significant" dispersal routes (*i.e.*, $2 \ln \mathrm{BF} > 2$) are figured.

Rdrp fragment

The Impact of Prior Choice on Biogeographic Model Fit

Table S.1.8: Marginal-likelihood estimates of the eight prior models for the Deformed wing virus rdrp fragment dataset. Columns 2–5 list marginal likelihoods inferred from four replicate analyses; the last two columns list the mean and standard deviation of these marginal-likelihood estimates. Candidate models are listed in rows and include all possible combinations of: (1) instantaneous-rate matrices (symmetric, Q_s or asymmetric, Q_a); (2) priors on the average dispersal rate [default, $P_d(\mu)$ or alternative, $P_a(\mu)$], and; (3) priors on the number of dispersal routes [default, $P_d(\Delta)$ or alternative, $P_a(\Delta)$]. The preferred default- and alternative-prior models are indicated in bold text.

Model	replicate1	replicate2	replicate3	replicate4	mean	sd
$P_d(\mu)Q_aP_d(\Delta)$	-220.28	-220.43	-221.02	-220.70	-220.61	0.32
$P_d(\mu)Q_aP_a(\Delta)$	-217.03	-217.13	-217.16	-217.36	-217.17	0.14
$\boldsymbol{P_d(\mu)Q_sP_d(\Delta)}$	**-214.77**	**-215.04**	**-214.84**	**-214.93**	**-214.89**	**0.12**
$P_d(\mu)Q_sP_a(\Delta)$	-213.43	-213.19	-213.25	-213.16	-213.26	0.12
$P_a(\mu)Q_aP_d(\Delta)$	-172.54	-172.34	-172.70	-172.55	-172.53	0.14
$\boldsymbol{P_a(\mu)Q_aP_a(\Delta)}$	**-173.62**	**-173.82**	**-174.01**	**-174.16**	**-173.90**	**0.23**
$P_a(\mu)Q_sP_d(\Delta)$	-181.72	-181.87	-181.81	-181.56	-181.74	0.13
$P_a(\mu)Q_sP_a(\Delta)$	-181.86	-181.55	-181.81	-181.91	-181.78	0.16

The Impact of Prior Choice on Pairwise Dispersal Rates

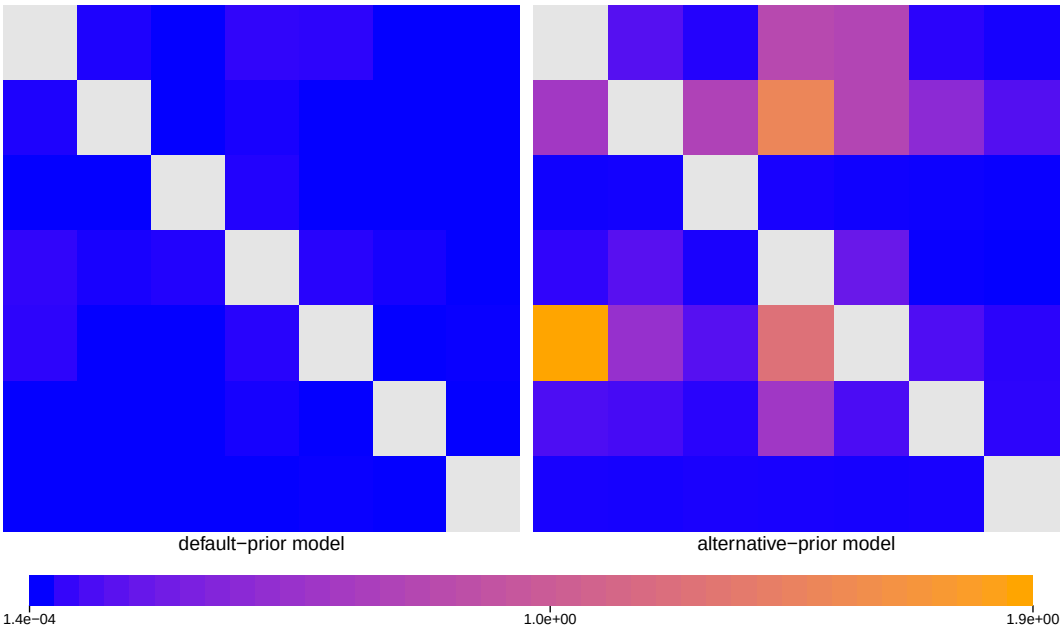

Figure S.1.22: The impact of prior choice on pairwise dispersal rates for the Deformed wing virus rdrp fragment dataset. Heatmaps summarize posterior-mean estimates of the instantaneous rate of dispersal between each pair of geographic areas, q_{ij}, under the default (left) and alternative (right) prior models.

The Impact of Prior Choice on the Inferred Support for Dispersal Routes

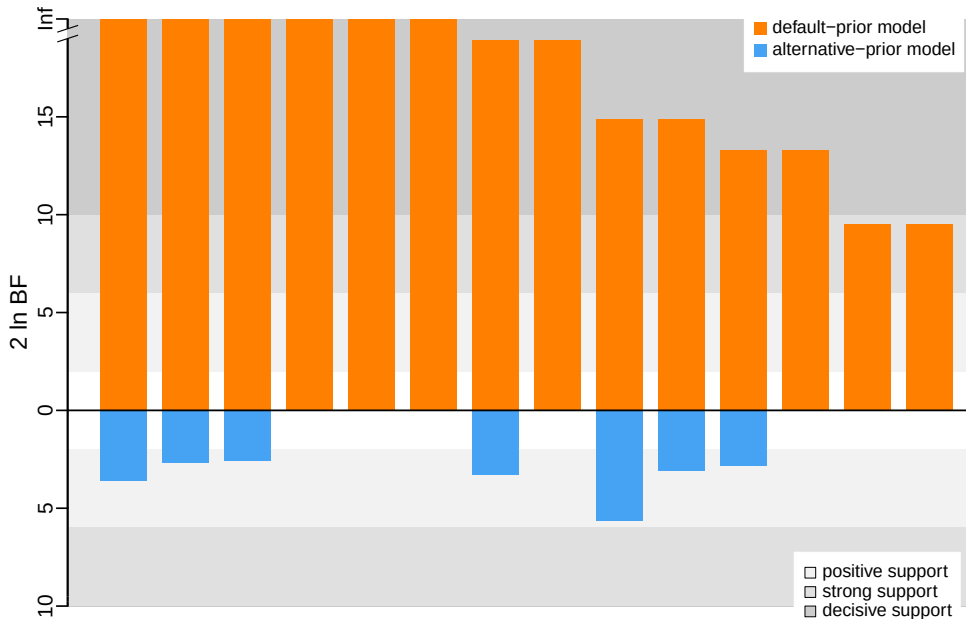

Figure S.1.23: The impact of prior choice on the inferred support for dispersal routes for the Deformed wing virus rdrp fragment dataset. We compare the evidential support for each dispersal route for the Deformed wing virus rdrp fragment dataset under the default (orange) and alternative (blue) prior models. Each bar indicates the $2\ln \mathrm{BF}$ for the corresponding dispersal route between two areas; only supported dispersal routes (*i.e.*, $2\ln \mathrm{BF} > 2$) are plotted.

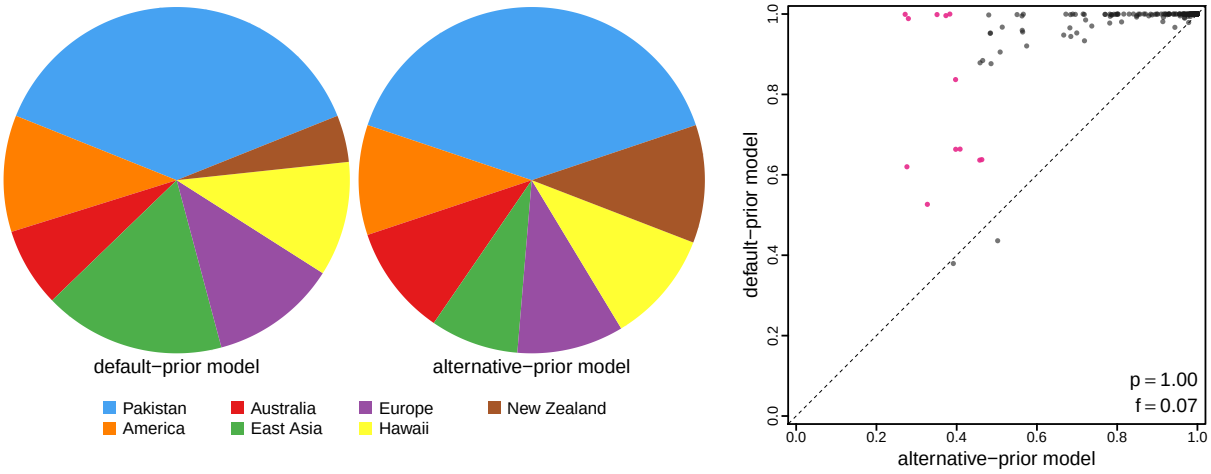

Figure S.1.24: The impact of prior choice ancestral-area estimates for the Deformed wing virus rdrp fragment dataset. The left panel compares the posterior probability of each ancestral area at the root node under the default- and alternative-prior models for the Deformed wing virus rdrp fragment dataset. The right panel plots the posterior probability of the most probable ancestral area under the default-prior model for each node in the MCC tree (y-axis) against the corresponding posterior probability of that area under the alternative-prior model (x-axis). Pink dots represent the internal nodes where the MAP ancestral area inferred under the default-prior model differs from that inferred under the alternative-prior models. The statistic p denotes the fraction of internal nodes that are shared under the default- and alternative-prior models; f, is the fraction of shared nodes where the MAP ancestral area differs under the default- and alternative-prior models. Note that the posterior probabilities of the MAP ancestral area under the default-prior model are generally higher than those under the alternative-prior model (*i.e.*, the default-prior model tends to mask uncertainty in the ancestral-area estimates).

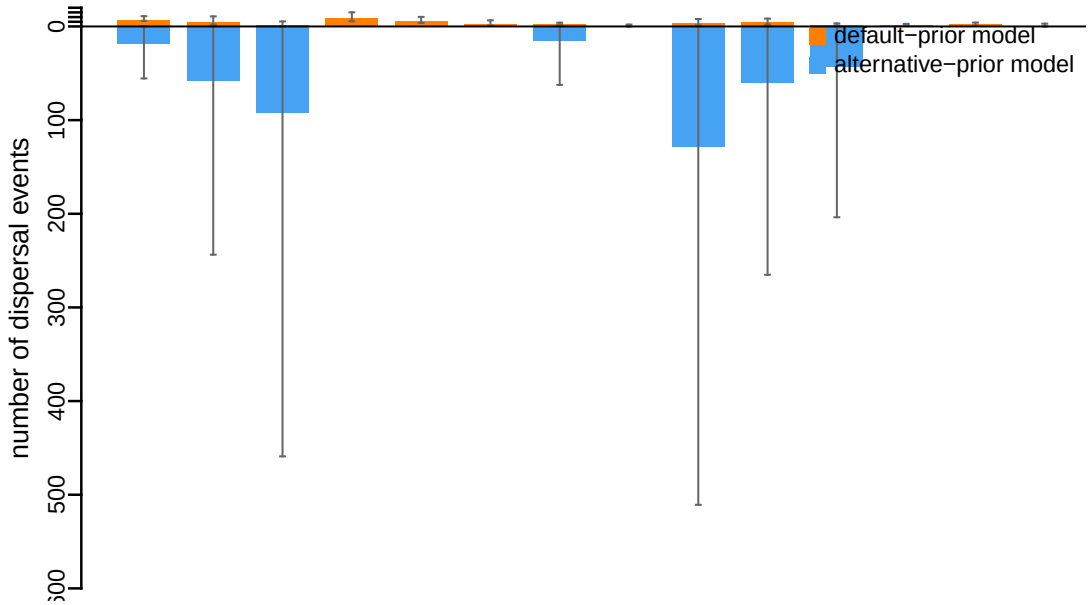

Figure S.1.25: The impact of prior choice on the inferred number of dispersal events between each pair of areas for the Deformed wing virus rdrp fragment dataset. The reflected bar plot depicts the number of dispersal events inferred under the default (orange) and alternative (blue) prior models for the Deformed wing virus rdrp fragment dataset. Each bar indicates the posterior-mean number of dispersal events between a pair of areas; whiskers indicate the 95% credible interval. Note that only the number of dispersal events over the "significant" dispersal routes (*i.e.*, $2 \ln \mathrm{BF} > 2$) are figured.

Vp3 fragment

The Impact of Prior Choice on Biogeographic Model Fit

Table S.1.9: Marginal-likelihood estimates of the eight prior models for the Deformed wing virus vp3 fragment dataset. Columns 2–5 list marginal likelihoods inferred from four replicate analyses; the last two columns list the mean and standard deviation of these marginal-likelihood estimates. Candidate models are listed in rows and include all possible combinations of: (1) instantaneous-rate matrices (symmetric, Q_s or asymmetric, Q_a); (2) priors on the average dispersal rate [default, $P_d(\mu)$ or alternative, $P_a(\mu)$], and; (3) priors on the number of dispersal routes [default, $P_d(\Delta)$ or alternative, $P_a(\Delta)$]. The preferred default- and alternative-prior models are indicated in bold text.

Model	replicate1	replicate2	replicate3	replicate4	mean	sd
$P_d(\mu)Q_aP_d(\Delta)$	-107.55	-107.90	-107.80	-107.65	-107.73	0.15
$P_d(\mu)Q_aP_a(\Delta)$	-104.57	-104.38	-104.47	-104.51	-104.48	0.08
$\boldsymbol{P_d(\mu)Q_sP_d(\Delta)}$	**-106.43**	**-106.45**	**-106.25**	**-106.35**	**-106.37**	**0.09**
$P_d(\mu)Q_sP_a(\Delta)$	-104.08	-104.15	-103.95	-104.11	-104.07	0.09
$P_a(\mu)Q_aP_d(\Delta)$	-93.58	-94.16	-93.99	-93.87	-93.90	0.25
$P_a(\mu)Q_aP_a(\Delta)$	-91.98	-92.00	-92.01	-92.06	-92.01	0.04
$P_a(\mu)Q_sP_d(\Delta)$	-93.00	-92.86	-92.76	-92.72	-92.84	0.12
$\boldsymbol{P_a(\mu)Q_sP_a(\Delta)}$	**-91.56**	**-91.30**	**-91.53**	**-91.51**	**-91.47**	**0.12**

The Impact of Prior Choice on Pairwise Dispersal Rates

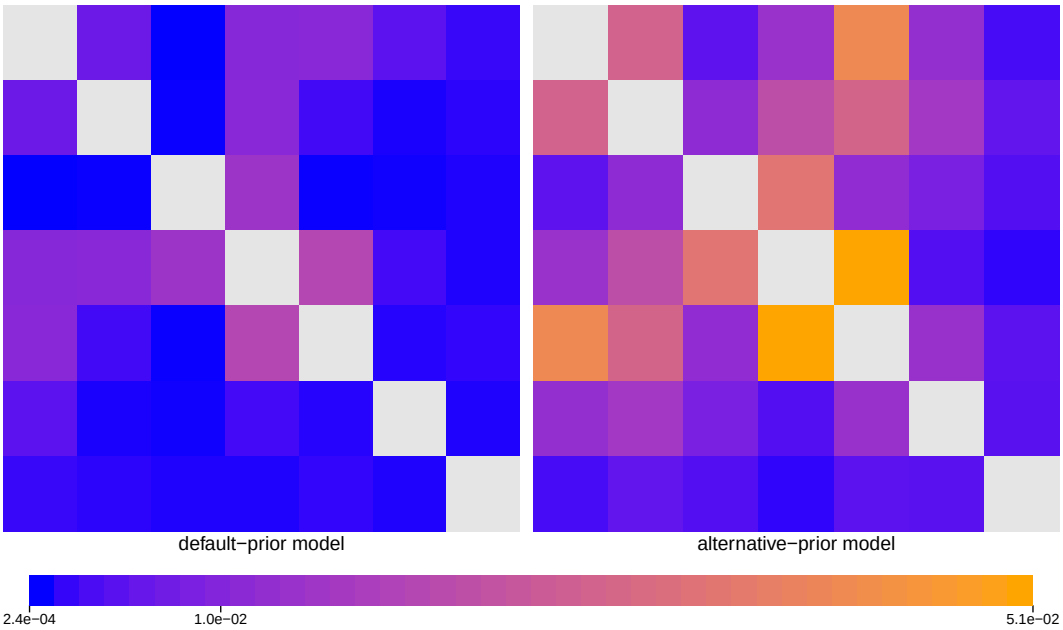

Figure S.1.26: The impact of prior choice on pairwise dispersal rates for the Deformed wing virus vp3 fragment dataset. Heatmaps summarize posterior-mean estimates of the instantaneous rate of dispersal between each pair of geographic areas, q_{ij}, under the default (left) and alternative (right) prior models.

The Impact of Prior Choice on the Inferred Support for Dispersal Routes

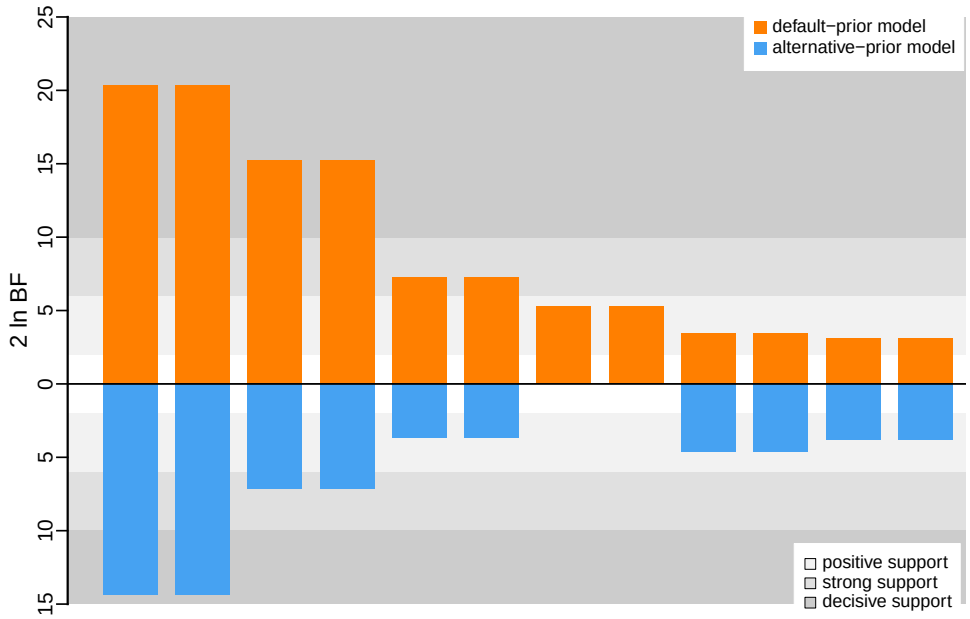

Figure S.1.27: The impact of prior choice on the inferred support for dispersal routes for the Deformed wing virus vp3 fragment dataset. We compare the evidential support for each dispersal route for the Deformed wing virus vp3 fragment dataset under the default (orange) and alternative (blue) prior models. Each bar indicates the $2\ln \mathrm{BF}$ for the corresponding dispersal route between two areas; only supported dispersal routes (*i.e.*, $2\ln \mathrm{BF} > 2$) are plotted.

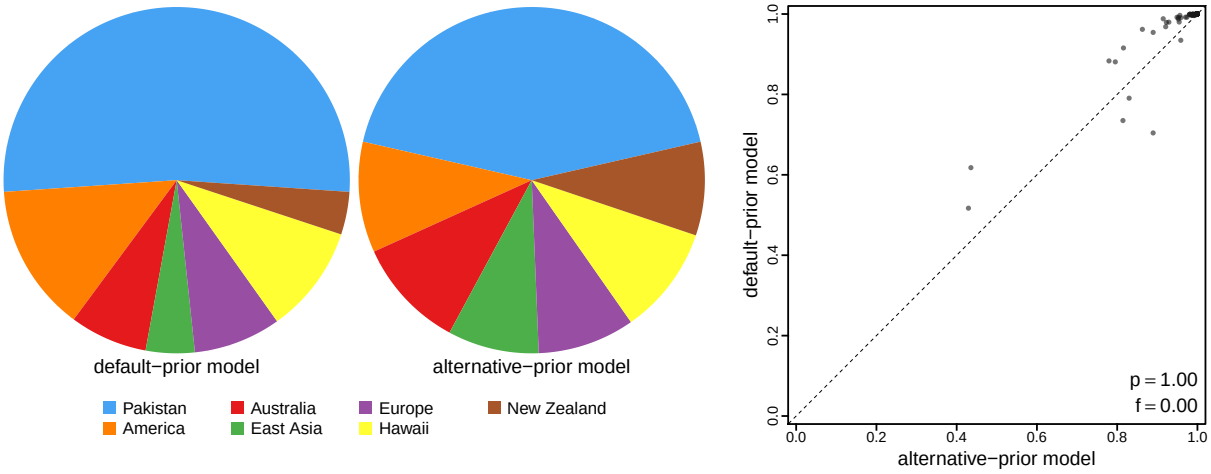

Figure S.1.28: The impact of prior choice ancestral-area estimates for the Deformed wing virus vp3 fragment dataset. The left panel compares the posterior probability of each ancestral area at the root node under the default- and alternative-prior models for the Deformed wing virus vp3 fragment dataset. The right panel plots the posterior probability of the most probable ancestral area under the default-prior model for each node in the MCC tree (y-axis) against the corresponding posterior probability of that area under the alternative-prior model (x-axis). Pink dots represent the internal nodes where the MAP ancestral area inferred under the default-prior model differs from that inferred under the alternative-prior models. The statistic p denotes the fraction of internal nodes that are shared under the default- and alternative-prior models; f, is the fraction of shared nodes where the MAP ancestral area differs under the default- and alternative-prior models. Note that the posterior probabilities of the MAP ancestral area under the default-prior model are generally higher than those under the alternative-prior model (*i.e.*, the default-prior model tends to mask uncertainty in the ancestral-area estimates).

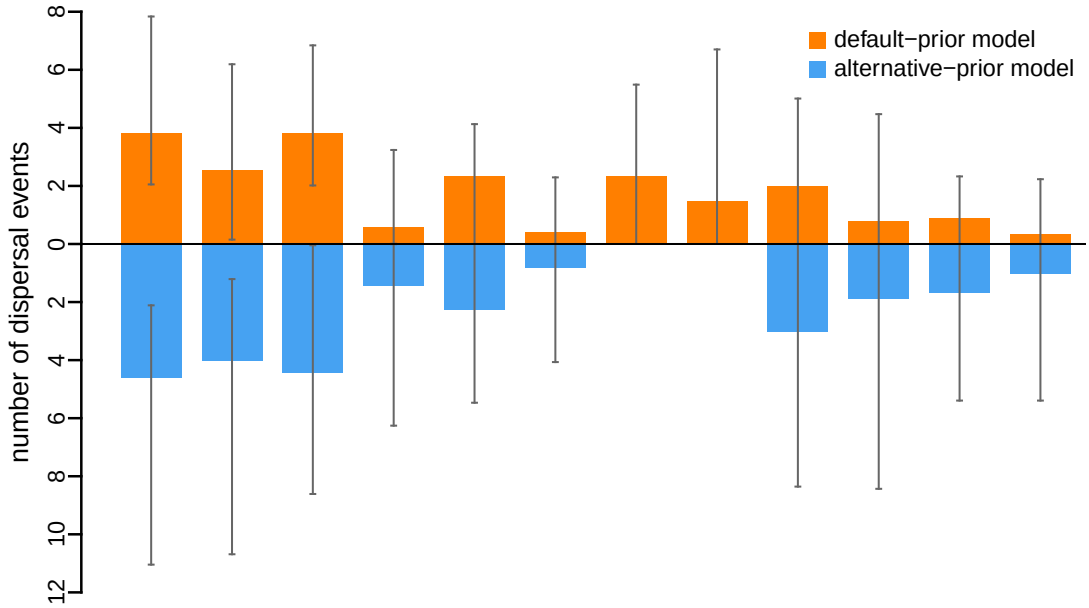

Figure S.1.29: The impact of prior choice on the inferred number of dispersal events between each pair of areas for the Deformed wing virus vp3 fragment dataset. The reflected bar plot depicts the number of dispersal events inferred under the default (orange) and alternative (blue) prior models for the Deformed wing virus vp3 fragment dataset. Each bar indicates the posterior-mean number of dispersal events between a pair of areas; whiskers indicate the 95% credible interval. Note that only the number of dispersal events over the "significant" dispersal routes (*i.e.*, $2 \ln \mathrm{BF} > 2$) are figured.

HIV

Faria et al. (2014) explored the origin and early spread of HIV-1 in human populations by analyzing sequences collected from central and southeast Africa and America between the 1980s and early 2000s. The authors used a down-sampling scheme of the complete dataset to verify the robustness of the main conclusions in the original study: this involved the creation of four data(sub)sets. Specifically, Dataset A includes 792 envelope C2V3 sequences collected between 1985–2004 from eight cities in the Democratic Republic of the Congo and the Republic of the Congo. Dataset B includes 927 sequences, with the addition of 67 subtype C sequences from southeast Africa (Zambia, Botswana, Tanzania, Kenya, Uganda, Burundi, Ethiopia and South Africa) sampled between 1986–2005, 67 sequences from the Americas (Haiti, Trinidad and Tobago and the USA) sampled between 1978–1997, and the ZR59 isolate obtained in 1959 from blood collected in Kinshasa. Dataset C includes 466 sequences that were down-sampled from Dataset A; Dataset D includes 601 sequences that were down-sampled from Dataset B; this down-sampling was motivated to decrease the representation of sequences sampled from Kinshasa (see (Faria et al. 2014) for details).

We acquired the sampling geographic location data of Datasets A, B, and C from the `BEAST` XML scripts provided by the original study; this sampling-area data are available in our GitHub and Dryad repositories.

The posterior distribution of phylogenies (used to perform sequential analyses in Faria et al. 2014) was obtained directly from the Dryad repository of the original study (also available in our GitHub and Dryad repositories).

We reanalyzed Datasets A, B, and C. The MCMC simulations used in the second step of our sequential analyses and of the analyses used to estimate marginal likelihoods under each prior model are described above in this section. Details of these analyses are available in the XML scripts included in our GitHub and Dryad repositories.

Dataset A

The Impact of Prior Choice on Biogeographic Model Fit

Table S.1.10: Marginal-likelihood estimates of the eight prior models for HIV dataset A. Columns 2–5 list marginal likelihoods inferred from four replicate analyses; the last two columns list the mean and standard deviation of these marginal-likelihood estimates. Candidate models are listed in rows and include all possible combinations of: (1) instantaneous-rate matrices (symmetric, Q_s or asymmetric, Q_a); (2) priors on the average dispersal rate [default, $P_d(\mu)$ or alternative, $P_a(\mu)$], and; (3) priors on the number of dispersal routes [default, $P_d(\Delta)$ or alternative, $P_a(\Delta)$]. The preferred default- and alternative-prior models are indicated in bold text.

Model	replicate1	replicate2	replicate3	replicate4	mean	sd
$P_d(\mu)Q_aP_d(\Delta)$	-1180.03	-1180.04	-1179.66	-1179.84	-1179.89	0.18
$P_d(\mu)Q_aP_a(\Delta)$	-1177.42	-1177.23	-1177.56	-1177.63	-1177.46	0.18
$\boldsymbol{P_d(\mu)Q_sP_d(\Delta)}$	**-1176.34**	**-1176.65**	**-1176.07**	**-1176.43**	**-1176.37**	**0.24**
$P_d(\mu)Q_sP_a(\Delta)$	-1175.19	-1175.21	-1175.62	-1175.43	-1175.36	0.20
$P_a(\mu)Q_aP_d(\Delta)$	-1041.67	-1042.14	-1042.10	-1042.41	-1042.08	0.30
$\boldsymbol{P_a(\mu)Q_aP_a(\Delta)}$	**-1037.31**	**-1037.94**	**-1037.65**	**-1037.51**	**-1037.60**	**0.26**
$P_a(\mu)Q_sP_d(\Delta)$	-1055.82	-1055.68	-1055.75	-1055.83	-1055.77	0.07
$P_a(\mu)Q_sP_a(\Delta)$	-1050.73	-1050.38	-1050.69	-1050.79	-1050.65	0.19

The Impact of Prior Choice on Pairwise Dispersal Rates

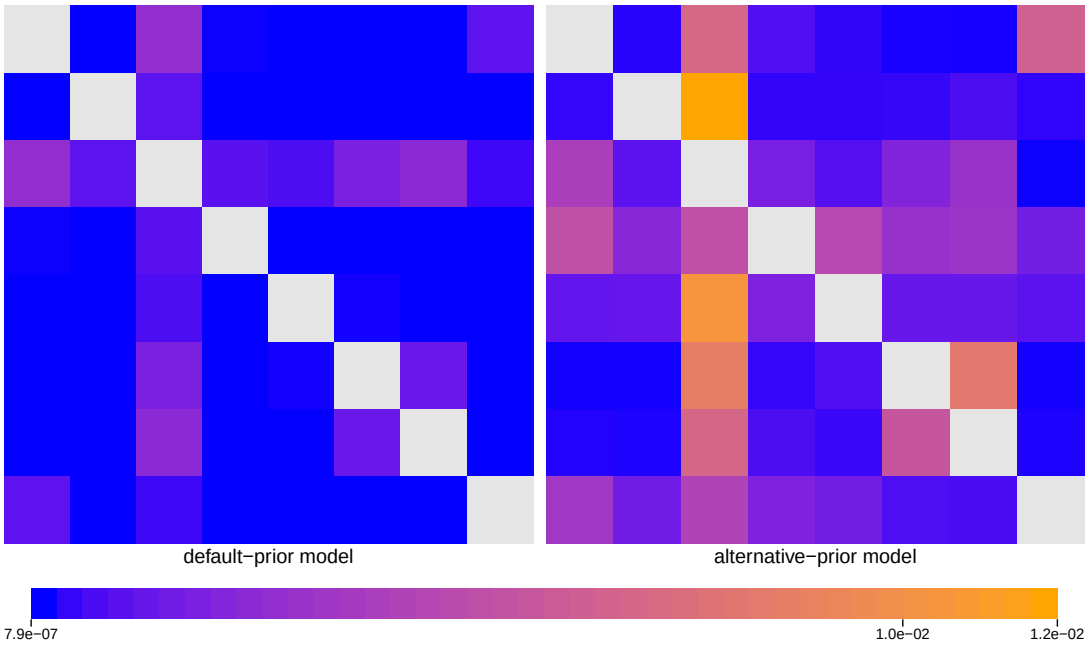

Figure S.1.30: The impact of prior choice on pairwise dispersal rates for HIV dataset A. Heatmaps summarize posterior-mean estimates of the instantaneous rate of dispersal between each pair of geographic areas, q_{ij}, under the default (left) and alternative (right) prior models.

The Impact of Prior Choice on the Inferred Support for Dispersal Routes

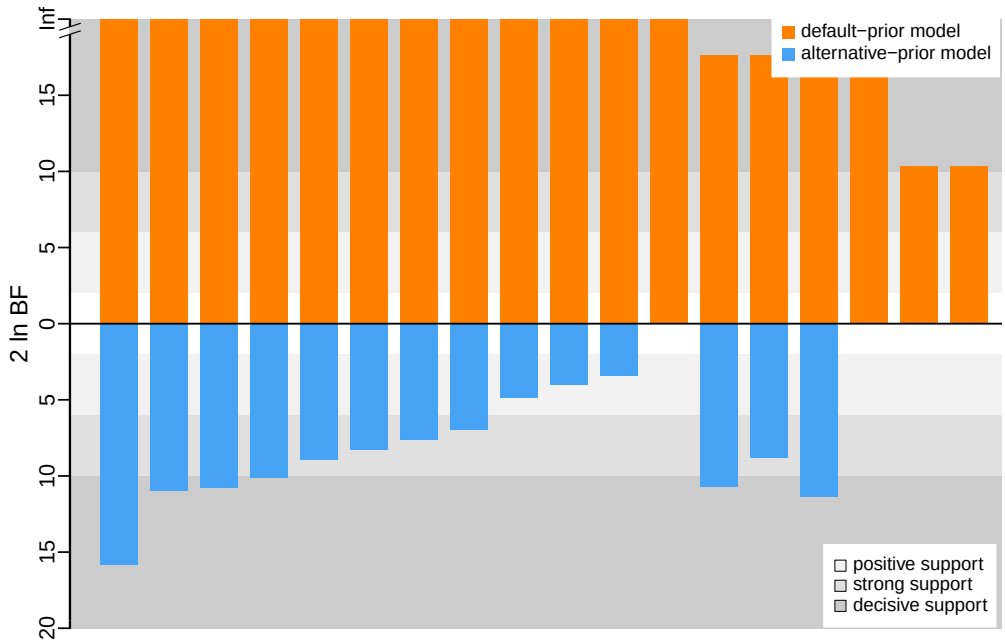

Figure S.1.31: The impact of prior choice on the inferred support for dispersal routes for HIV dataset A. We compare the evidential support for each dispersal route for HIV dataset A under the default (orange) and alternative (blue) prior models. Each bar indicates the $2\ln \mathrm{BF}$ for the corresponding dispersal route between two areas; only supported dispersal routes (*i.e.*, $2\ln \mathrm{BF} > 2$) are plotted.

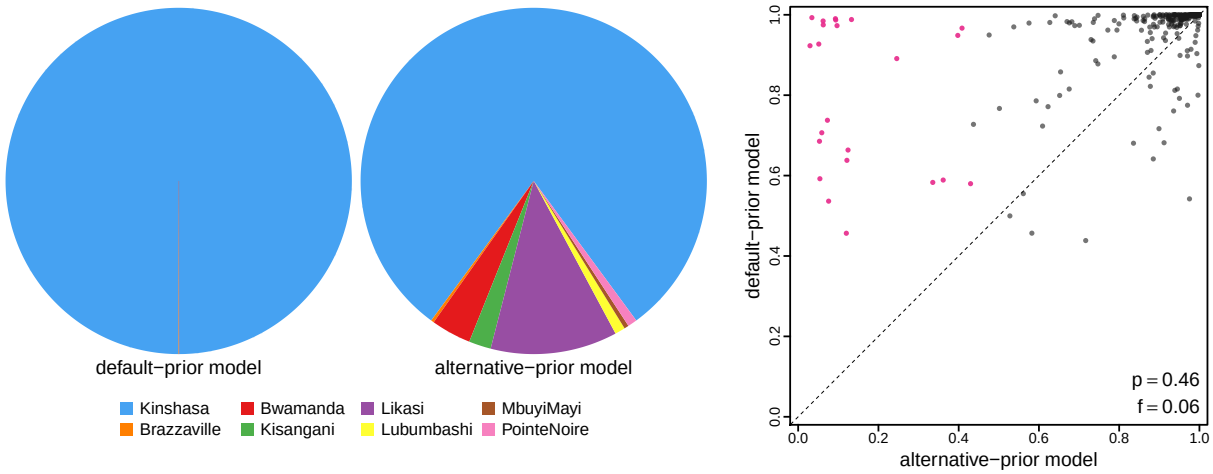

Figure S.1.32: The impact of prior choice ancestral-area estimates for HIV dataset A. The left panel compares the posterior probability of each ancestral area at the root node under the default- and alternative-prior models for HIV dataset A. The right panel plots the posterior probability of the most probable ancestral area under the default-prior model for each node in the MCC tree (y-axis) against the corresponding posterior probability of that area under the alternative-prior model (x-axis). Pink dots represent the internal nodes where the MAP ancestral area inferred under the default-prior model differs from that inferred under the alternative-prior models. The statistic p denotes the fraction of internal nodes that are shared under the default- and alternative-prior models; f, is the fraction of shared nodes where the MAP ancestral area differs under the default- and alternative-prior models. Note that the posterior probabilities of the MAP ancestral area under the default-prior model are generally higher than those under the alternative-prior model (*i.e.*, the default-prior model tends to mask uncertainty in the ancestral-area estimates).

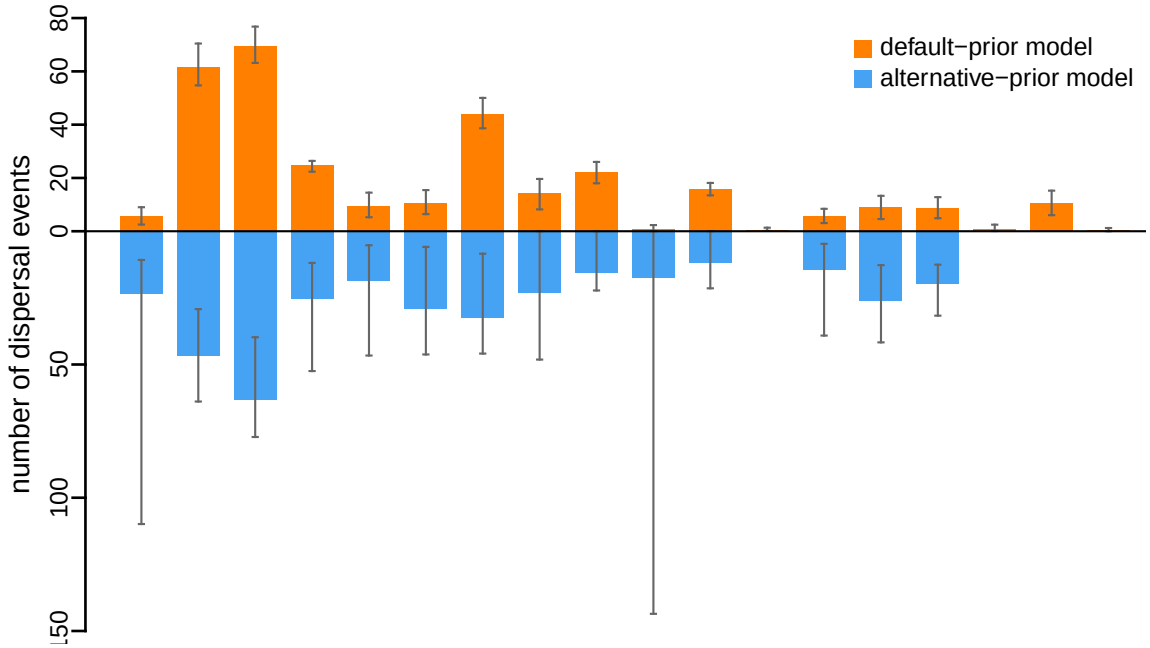

Figure S.1.33: The impact of prior choice on the inferred number of dispersal events between each pair of areas for HIV dataset A. The reflected bar plot depicts the number of dispersal events inferred under the default (orange) and alternative (blue) prior models for HIV dataset A. Each bar indicates the posterior-mean number of dispersal events between a pair of areas; whiskers indicate the 95% credible interval. Note that only the number of dispersal events over the "significant" dispersal routes (*i.e.*, $2 \ln \mathrm{BF} > 2$) are figured.

Dataset B

The Impact of Prior Choice on Biogeographic Model Fit

Table S.1.11: Marginal-likelihood estimates of the eight prior models for HIV dataset B. Columns 2–5 list marginal likelihoods inferred from four replicate analyses; the last two columns list the mean and standard deviation of these marginal-likelihood estimates. Candidate models are listed in rows and include all possible combinations of: (1) instantaneous-rate matrices (symmetric, Q_s or asymmetric, Q_a); (2) priors on the average dispersal rate [default, $P_d(\mu)$ or alternative, $P_a(\mu)$], and; (3) priors on the number of dispersal routes [default, $P_d(\Delta)$ or alternative, $P_a(\Delta)$]. The preferred default- and alternative-prior models are indicated in bold text.

Model	replicate1	replicate2	replicate3	replicate4	mean	sd
$\boldsymbol{P_d(\mu)Q_aP_d(\Delta)}$	**-1311.00**	**-1310.50**	**-1310.59**	**-1311.44**	**-1310.88**	**0.43**
$P_d(\mu)Q_aP_a(\Delta)$	-1309.74	-1310.36	-1310.37	-1310.16	-1310.15	0.29
$P_d(\mu)Q_sP_d(\Delta)$	-1335.72	-1336.59	-1336.31	-1337.06	-1336.42	0.56
$P_d(\mu)Q_sP_a(\Delta)$	-1335.18	-1334.56	-1334.92	-1334.93	-1334.90	0.26
$P_a(\mu)Q_aP_d(\Delta)$	-1168.93	-1168.28	-1169.26	-1168.98	-1168.86	0.41
$\boldsymbol{P_a(\mu)Q_aP_a(\Delta)}$	**-1164.71**	**-1164.34**	**-1164.92**	**-1164.55**	**-1164.63**	**0.24**
$P_a(\mu)Q_sP_d(\Delta)$	-1188.41	-1187.71	-1187.91	-1187.51	-1187.89	0.39
$P_a(\mu)Q_sP_a(\Delta)$	-1185.73	-1186.09	-1186.03	-1185.93	-1185.94	0.16

The Impact of Prior Choice on Pairwise Dispersal Rates

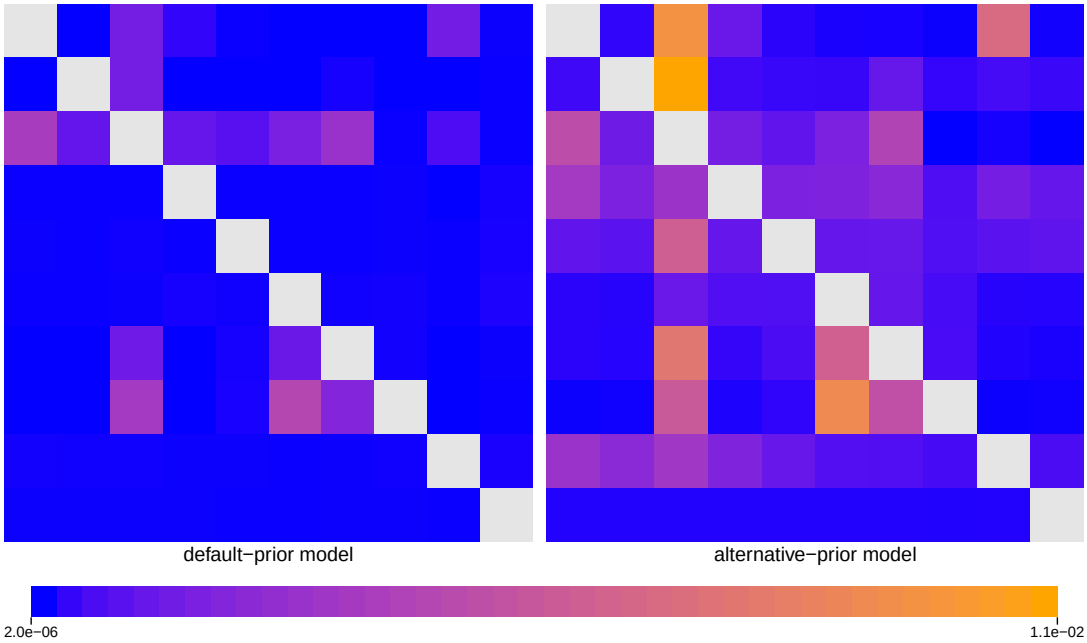

Figure S.1.34: The impact of prior choice on pairwise dispersal rates for HIV dataset B. Heatmaps summarize posterior-mean estimates of the instantaneous rate of dispersal between each pair of geographic areas, q_{ij}, under the default (left) and alternative (right) prior models.

The Impact of Prior Choice on the Inferred Support for Dispersal Routes

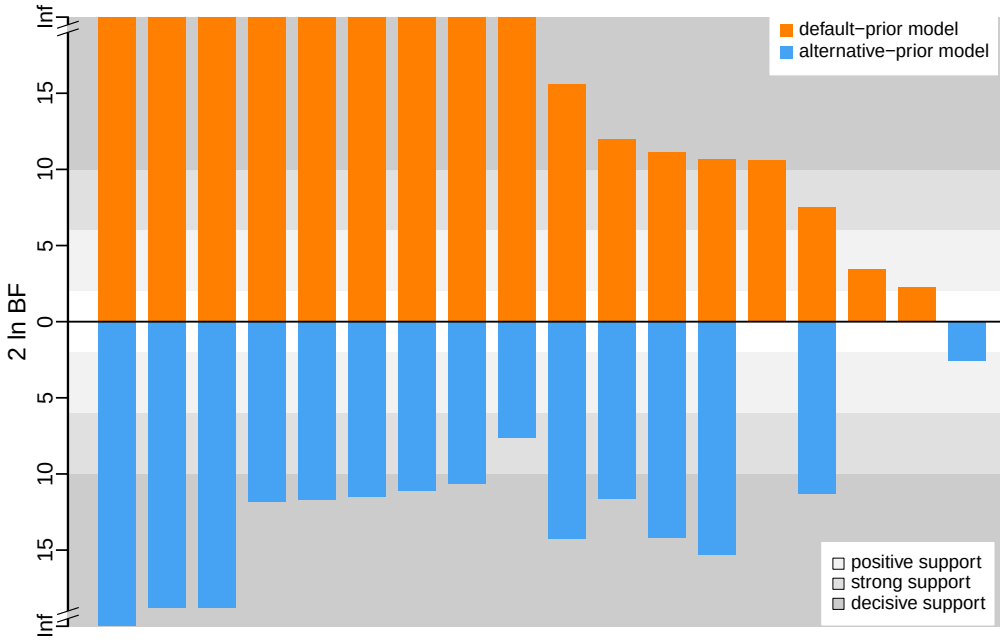

Figure S.1.35: The impact of prior choice on the inferred support for dispersal routes for HIV dataset B. We compare the evidential support for each dispersal route for HIV dataset B under the default (orange) and alternative (blue) prior models. Each bar indicates the $2\ln \mathrm{BF}$ for the corresponding dispersal route between two areas; only supported dispersal routes (*i.e.*, $2\ln \mathrm{BF} > 2$) are plotted.

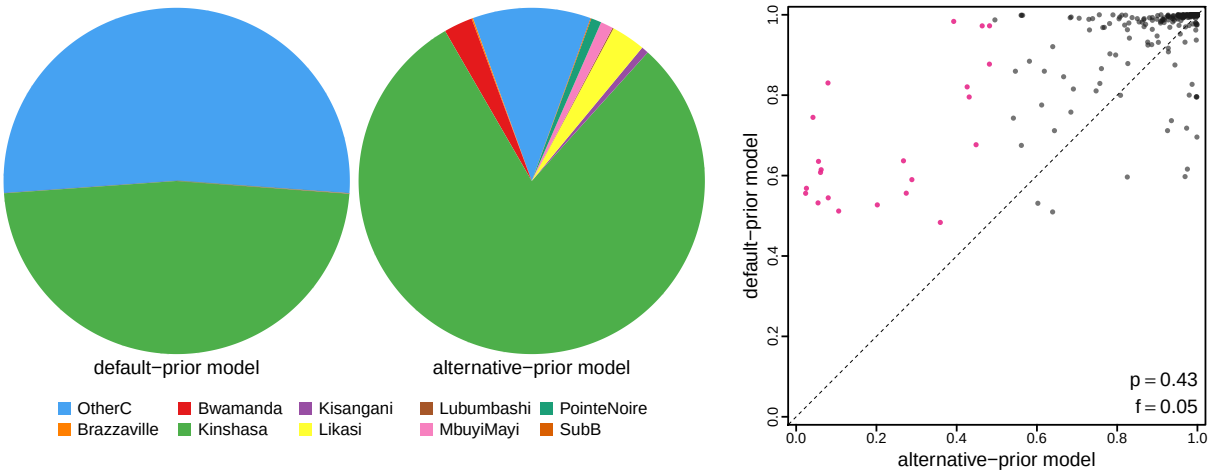

Figure S.1.36: The impact of prior choice ancestral-area estimates for HIV dataset B. The left panel compares the posterior probability of each ancestral area at the root node under the default- and alternative-prior models for HIV dataset B. The right panel plots the posterior probability of the most probable ancestral area under the default-prior model for each node in the MCC tree (y-axis) against the corresponding posterior probability of that area under the alternative-prior model (x-axis). Pink dots represent the internal nodes where the MAP ancestral area inferred under the default-prior model differs from that inferred under the alternative-prior models. The statistic p denotes the fraction of internal nodes that are shared under the default- and alternative-prior models; f, is the fraction of shared nodes where the MAP ancestral area differs under the default- and alternative-prior models. Note that the posterior probabilities of the MAP ancestral area under the default-prior model are generally higher than those under the alternative-prior model (*i.e.*, the default-prior model tends to mask uncertainty in the ancestral-area estimates).

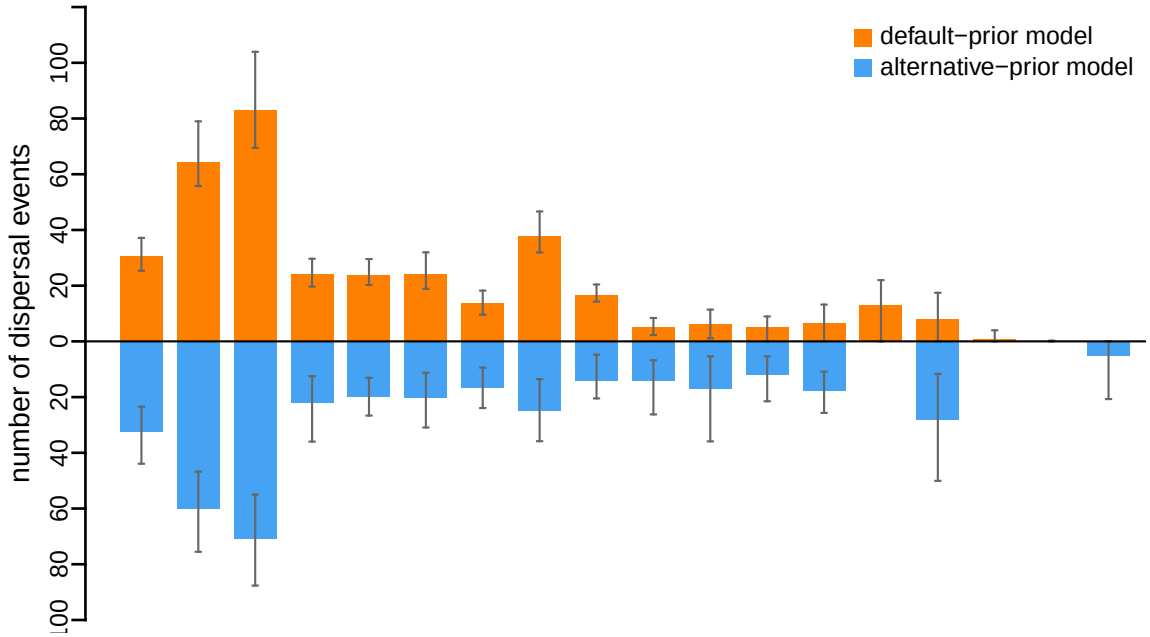

Figure S.1.37: The impact of prior choice on the inferred number of dispersal events between each pair of areas for HIV dataset B. The reflected bar plot depicts the number of dispersal events inferred under the default (orange) and alternative (blue) prior models for HIV dataset B. Each bar indicates the posterior-mean number of dispersal events between a pair of areas; whiskers indicate the 95% credible interval. Note that only the number of dispersal events over the "significant" dispersal routes (*i.e.*, $2 \ln \mathrm{BF} > 2$) are figured.

Dataset C

The Impact of Prior Choice on Biogeographic Model Fit

Table S.1.12: Marginal-likelihood estimates of the eight prior models for HIV dataset C. Columns 2–5 list marginal likelihoods inferred from four replicate analyses; the last two columns list the mean and standard deviation of these marginal-likelihood estimates. Candidate models are listed in rows and include all possible combinations of: (1) instantaneous-rate matrices (symmetric, Q_s or asymmetric, Q_a); (2) priors on the average dispersal rate [default, $P_d(\mu)$ or alternative, $P_a(\mu)$], and; (3) priors on the number of dispersal routes [default, $P_d(\Delta)$ or alternative, $P_a(\Delta)$]. The preferred default- and alternative-prior models are indicated in bold text.

Model	replicate1	replicate2	replicate3	replicate4	mean	sd
$\boldsymbol{P_d(\mu)Q_aP_d(\Delta)}$	**-837.74**	**-837.45**	**-838.03**	**-837.49**	**-837.68**	**0.27**
$P_d(\mu)Q_aP_a(\Delta)$	-835.21	-835.34	-835.25	-835.38	-835.29	0.08
$P_d(\mu)Q_sP_d(\Delta)$	-859.10	-859.03	-858.77	-858.98	-858.97	0.14
$P_d(\mu)Q_sP_a(\Delta)$	-858.01	-858.00	-858.05	-857.66	-857.93	0.18
$P_a(\mu)Q_aP_d(\Delta)$	-732.64	-732.74	-732.66	-732.95	-732.75	0.14
$\boldsymbol{P_a(\mu)Q_aP_a(\Delta)}$	**-726.94**	**-726.82**	**-726.55**	**-726.83**	**-726.79**	**0.17**
$P_a(\mu)Q_sP_d(\Delta)$	-730.16	-730.01	-730.04	-729.88	-730.02	0.11
$P_a(\mu)Q_sP_a(\Delta)$	-728.56	-728.34	-728.44	-728.38	-728.43	0.09

The Impact of Prior Choice on Pairwise Dispersal Rates

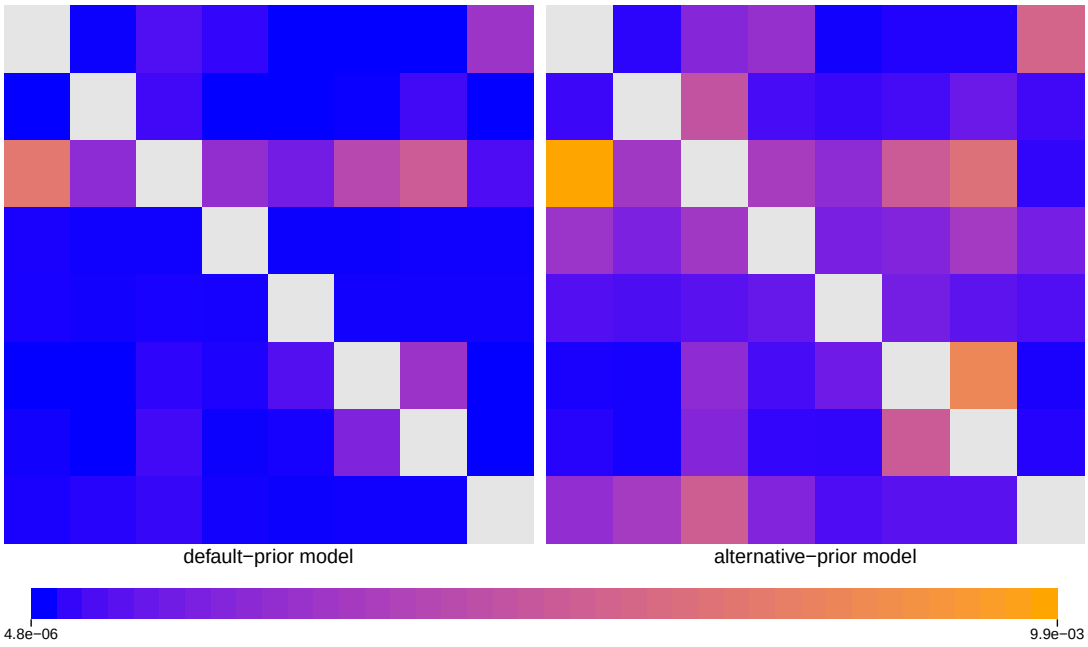

Figure S.1.38: The impact of prior choice on pairwise dispersal rates for HIV dataset C. Heatmaps summarize posterior-mean estimates of the instantaneous rate of dispersal between each pair of geographic areas, q_{ij}, under the default (left) and alternative (right) prior models.

The Impact of Prior Choice on the Inferred Support for Dispersal Routes

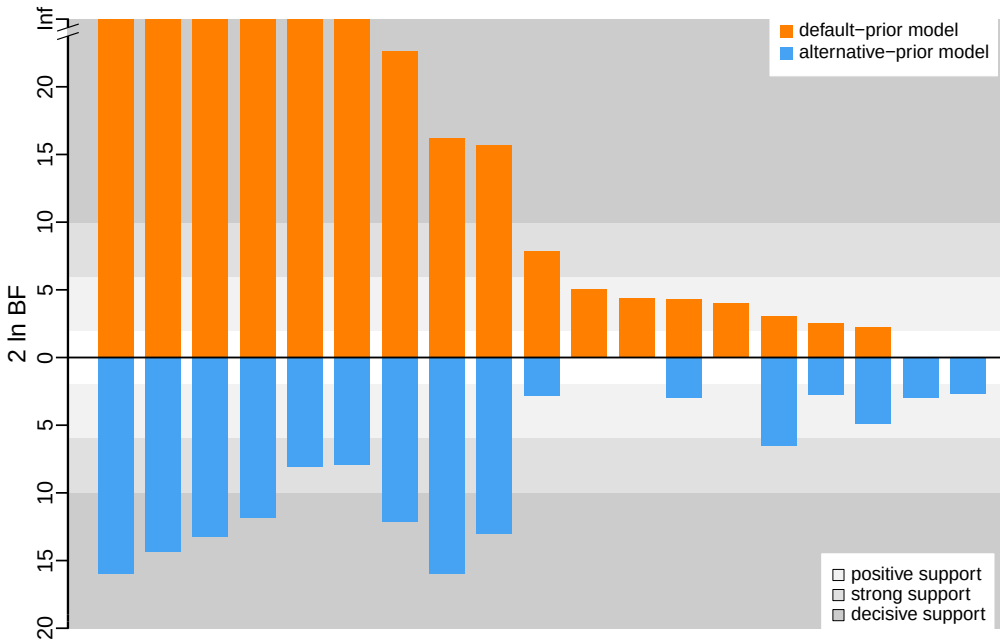

Figure S.1.39: The impact of prior choice on the inferred support for dispersal routes for HIV dataset C. We compare the evidential support for each dispersal route for HIV dataset C under the default (orange) and alternative (blue) prior models. Each bar indicates the $2\ln \mathrm{BF}$ for the corresponding dispersal route between two areas; only supported dispersal routes (*i.e.*, $2\ln \mathrm{BF} > 2$) are plotted.

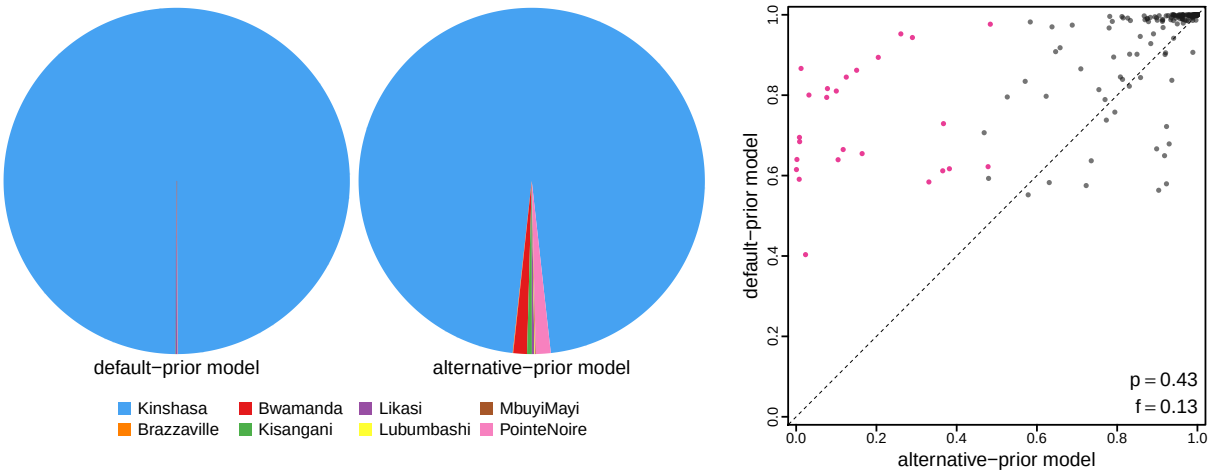

Figure S.1.40: The impact of prior choice ancestral-area estimates for HIV dataset C. The left panel compares the posterior probability of each ancestral area at the root node under the default- and alternative-prior models for HIV dataset C. The right panel plots the posterior probability of the most probable ancestral area under the default-prior model for each node in the MCC tree (y-axis) against the corresponding posterior probability of that area under the alternative-prior model (x-axis). Pink dots represent the internal nodes where the MAP ancestral area inferred under the default-prior model differs from that inferred under the alternative-prior models. The statistic p denotes the fraction of internal nodes that are shared under the default- and alternative-prior models; f, is the fraction of shared nodes where the MAP ancestral area differs under the default- and alternative-prior models. Note that the posterior probabilities of the MAP ancestral area under the default-prior model are generally higher than those under the alternative-prior model (*i.e.*, the default-prior model tends to mask uncertainty in the ancestral-area estimates).

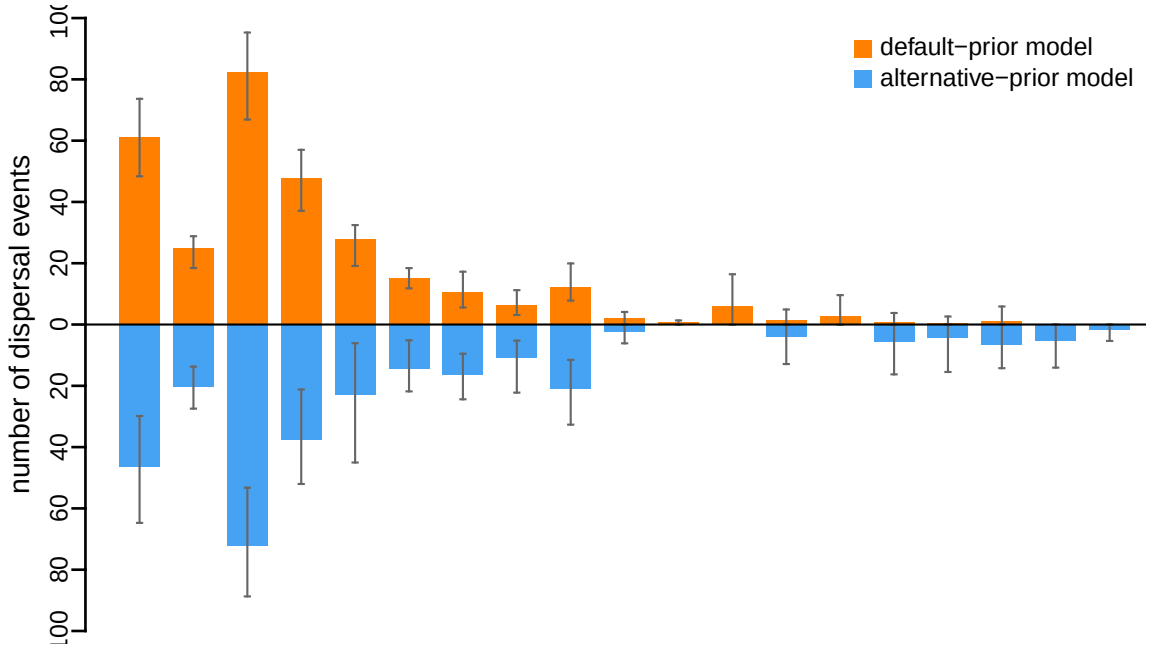

Figure S.1.41: The impact of prior choice on the inferred number of dispersal events between each pair of areas for HIV dataset C. The reflected bar plot depicts the number of dispersal events inferred under the default (orange) and alternative (blue) prior models for HIV dataset C. Each bar indicates the posterior-mean number of dispersal events between a pair of areas; whiskers indicate the 95% credible interval. Note that only the number of dispersal events over the "significant" dispersal routes (*i.e.*, $2 \ln \mathrm{BF} > 2$) are figured.

Influenza Virus

Bedford et al. (2015) inferred the geographic dynamics of the four most prevalent, globally circulating human seasonal influenza viruses (A/H3N2, A/H1N1, B/Victoria, and B/Yamagata). The authors collected complete sequences of the HA1 domain of the hemagglutinin (HA) gene for these influenza viruses, sampled across most of the major geographic areas between 2000–2012. They applied down-sampling schemes to the complete dataset to reduce the impact of possible surveillance biases, resulting into three data(sub)sets that were used to assess the robustness of the main conclusions to sampling effects. These includes: the "large" datasets (containing between 1999 to 4006 sequences for each virus); the "small" datasets (containing between 1240 to 1391 sequences for each virus), and the "alternative" datasets (containing between 1223 to 1967 sequences for each virus) (details see Bedford et al. 2015).

We explored the impact of prior choice on phylodynamic inferences for two of the datasets from their study; the "small" versions of the A/H3N2 and B/Yamagata datasets. Our choice of datasets was motivated by computational considerations (the comprehensive series of analyses in our study entails a very large computational burden even for the "small" datasets), coupled with our desire to include one virus from each of the two major types of human seasonal influenza virus (*i.e.*, A and B).

We acquired the marginal posterior probability distribution of phylogenies (inferred from the sequence data and used to perform sequential analyses in Bedford et al. 2015) directly from the Github repository of the original study. Each posterior distribution included 101 trees, which was then treated as the prior distribution of phylogenies in the second step of the sequential phylodynamic inference. The `trees` files containing these distributions are available in our GitHub and Dryad repositories. We acquired the sampling geographic location data from the XML scripts provided by the original study; this sampling-area data are available in our GitHub and Dryad repositories.

The MCMC simulations used in the second step of our sequential analyses and of the analyses used to estimate marginal likelihoods under each prior model are described above in this section. Details of these analyses are available in the XML scripts included in our GitHub and Dryad repositories.

A/H3N2 Dataset

The Impact of Prior Choice on Biogeographic Model Fit

Table S.1.13: Marginal-likelihood estimates of the eight prior models for Influenza A/H3N2 dataset. Columns 2–5 list marginal likelihoods inferred from four replicate analyses; the last two columns list the mean and standard deviation of these marginal-likelihood estimates. Candidate models are listed in rows and include all possible combinations of: (1) instantaneous-rate matrices (symmetric, Q_s or asymmetric, Q_a); (2) priors on the average dispersal rate [default, $P_d(\mu)$ or alternative, $P_a(\mu)$], and; (3) priors on the number of dispersal routes [default, $P_d(\Delta)$ or alternative, $P_a(\Delta)$]. The preferred default- and alternative-prior models are indicated in bold text.

Model	replicate1	replicate2	replicate3	replicate4	mean	sd
$\boldsymbol{P_d(\mu)Q_aP_d(\Delta)}$	**-2873.22**	**-2876.37**	**-2875.62**	**-2875.98**	**-2875.30**	**1.42**
$P_d(\mu)Q_aP_a(\Delta)$	-2825.59	-2827.47	-2826.41	-2826.32	-2826.45	0.77
$P_d(\mu)Q_sP_d(\Delta)$	-2883.60	-2887.96	-2889.36	-2887.22	-2887.03	2.46
$P_d(\mu)Q_sP_a(\Delta)$	-2838.53	-2837.93	-2838.49	-2838.33	-2838.32	0.27
$P_a(\mu)Q_aP_d(\Delta)$	-2303.17	-2303.72	-2303.91	-2303.04	-2303.46	0.42
$\boldsymbol{P_a(\mu)Q_aP_a(\Delta)}$	**-2275.33**	**-2276.09**	**-2275.23**	**-2275.43**	**-2275.52**	**0.39**
$P_a(\mu)Q_sP_d(\Delta)$	-2310.78	-2309.85	-2312.15	-2311.25	-2311.01	0.96
$P_a(\mu)Q_sP_a(\Delta)$	-2280.06	-2280.62	-2280.40	-2280.54	-2280.40	0.25

The Impact of Prior Choice on Pairwise Dispersal Rates

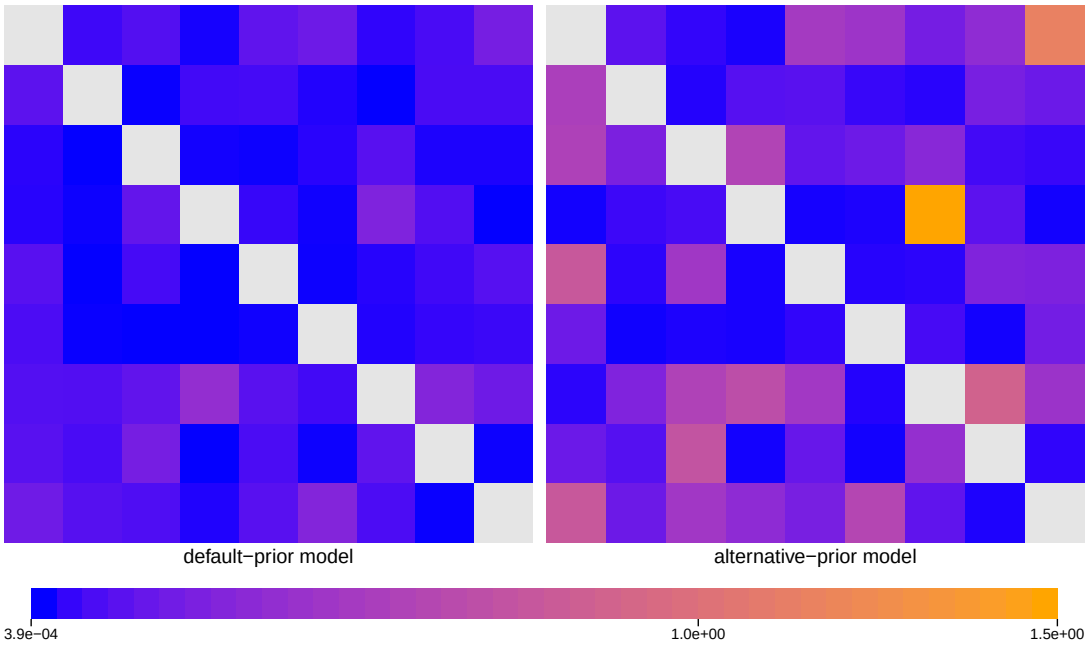

Figure S.1.42: The impact of prior choice on pairwise dispersal rates for Influenza A/H3N2 dataset. Heatmaps summarize posterior-mean estimates of the instantaneous rate of dispersal between each pair of geographic areas, q_{ij}, under the default (left) and alternative (right) prior models.

The Impact of Prior Choice on the Inferred Support for Dispersal Routes

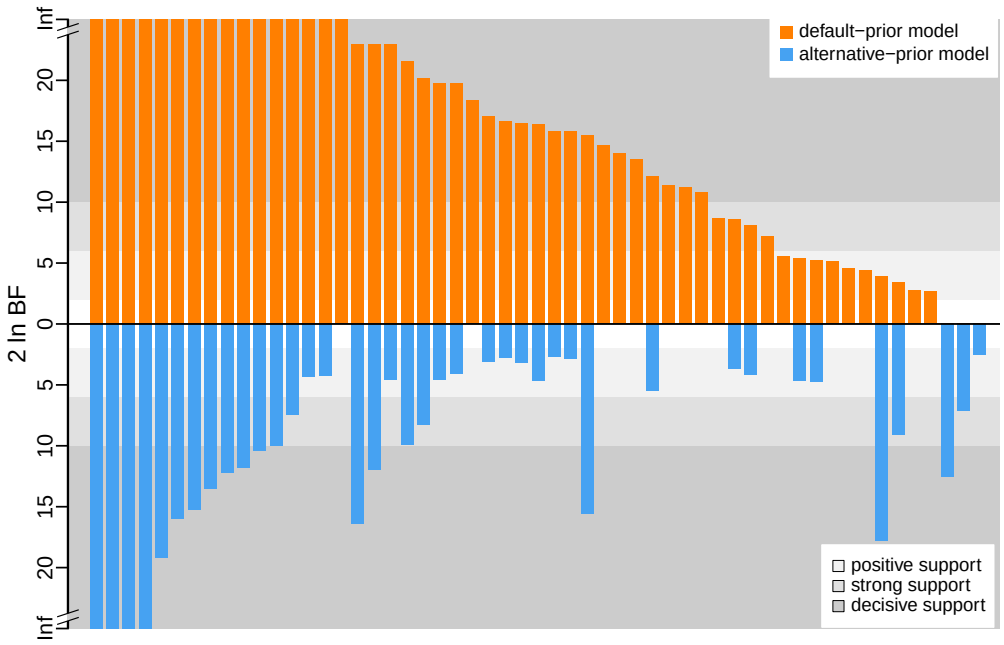

Figure S.1.43: The impact of prior choice on the inferred support for dispersal routes for Influenza A/H3N2 dataset. We compare the evidential support for each dispersal route for Influenza A/H3N2 dataset under the default (orange) and alternative (blue) prior models. Each bar indicates the $2\ln \mathrm{BF}$ for the corresponding dispersal route between two areas; only supported dispersal routes (*i.e.*, $2\ln \mathrm{BF} > 2$) are plotted.

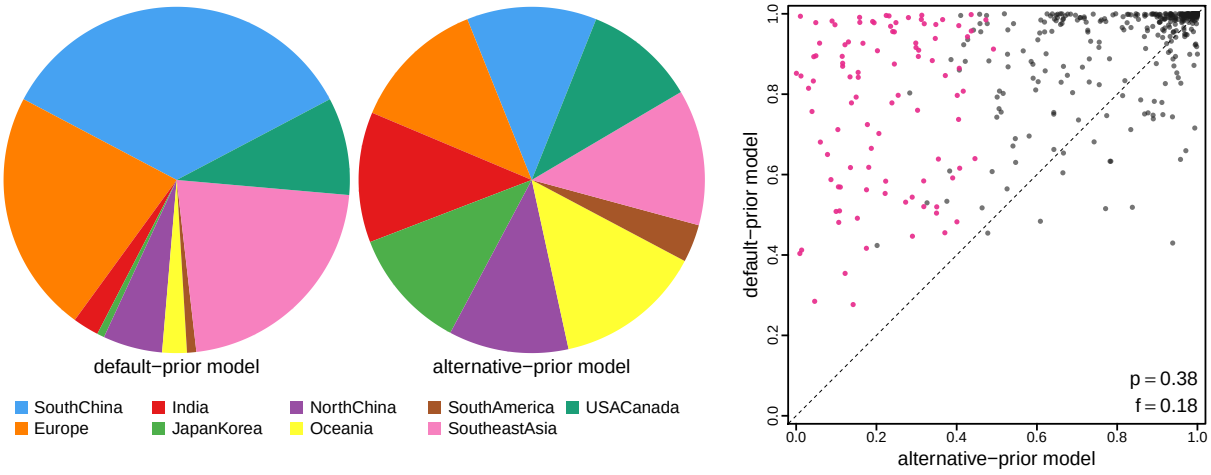

Figure S.1.44: The impact of prior choice ancestral-area estimates for Influenza A/H3N2 dataset. The left panel compares the posterior probability of each ancestral area at the root node under the default- and alternative-prior models for Influenza A/H3N2 dataset. The right panel plots the posterior probability of the most probable ancestral area under the default-prior model for each node in the MCC tree (y-axis) against the corresponding posterior probability of that area under the alternative-prior model (x-axis). Pink dots represent the internal nodes where the MAP ancestral area inferred under the default-prior model differs from that inferred under the alternative-prior models. The statistic p denotes the fraction of internal nodes that are shared under the default- and alternative-prior models; f, is the fraction of shared nodes where the MAP ancestral area differs under the default- and alternative-prior models. Note that the posterior probabilities of the MAP ancestral area under the default-prior model are generally higher than those under the alternative-prior model (*i.e.*, the default-prior model tends to mask uncertainty in the ancestral-area estimates).

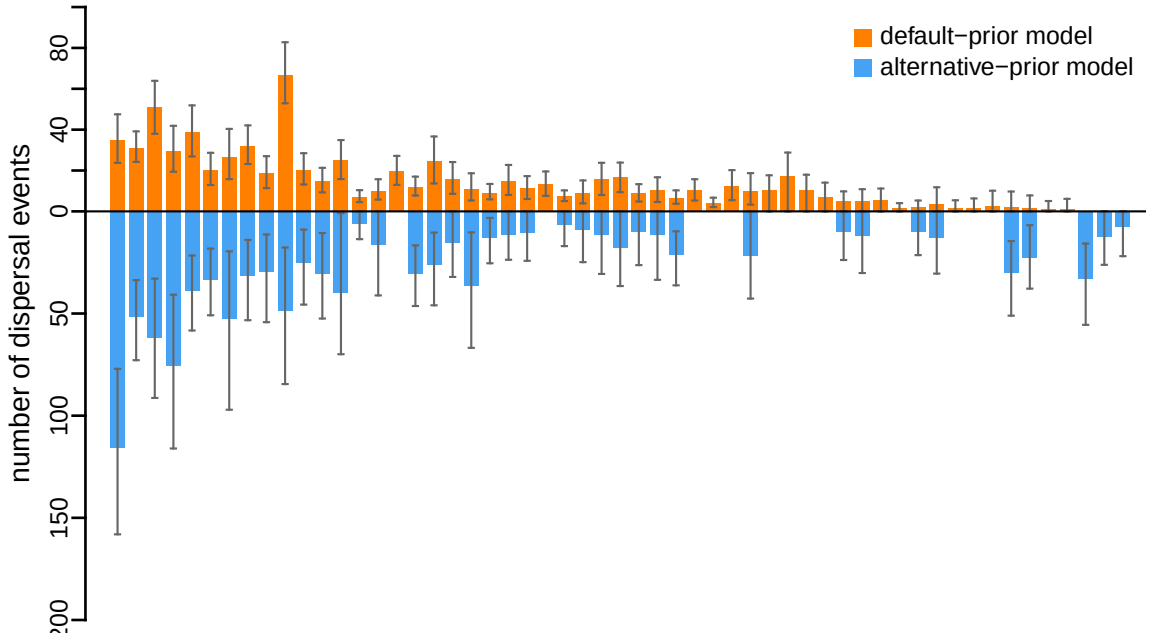

Figure S.1.45: The impact of prior choice on the inferred number of dispersal events between each pair of areas for Influenza A/H3N2 dataset. The reflected bar plot depicts the number of dispersal events inferred under the default (orange) and alternative (blue) prior models for Influenza A/H3N2 dataset. Each bar indicates the posterior-mean number of dispersal events between a pair of areas; whiskers indicate the 95% credible interval. Note that only the number of dispersal events over the "significant" dispersal routes (*i.e.*, $2 \ln \mathrm{BF} > 2$) are figured.

B/Yamagata Dataset

The Impact of Prior Choice on Biogeographic Model Fit

Table S.1.14: Marginal-likelihood estimates of the eight prior models for Influenza B/Yamagata dataset. Columns 2–5 list marginal likelihoods inferred from four replicate analyses; the last two columns list the mean and standard deviation of these marginal-likelihood estimates. Candidate models are listed in rows and include all possible combinations of: (1) instantaneous-rate matrices (symmetric, Q_s or asymmetric, Q_a); (2) priors on the average dispersal rate [default, $P_d(\mu)$ or alternative, $P_a(\mu)$], and; (3) priors on the number of dispersal routes [default, $P_d(\Delta)$ or alternative, $P_a(\Delta)$]. The preferred default- and alternative-prior models are indicated in bold text.

Model	replicate1	replicate2	replicate3	replicate4	mean	sd
$\boldsymbol{P_d(\mu)Q_aP_d(\Delta)}$	**-2335.56**	**-2333.27**	**-2334.64**	**-2334.57**	**-2334.51**	**0.94**
$P_d(\mu)Q_aP_a(\Delta)$	-2293.14	-2293.52	-2293.87	-2293.23	-2293.44	0.33
$P_d(\mu)Q_sP_d(\Delta)$	-2351.22	-2351.73	-2351.95	-2351.83	-2351.68	0.32
$P_d(\mu)Q_sP_a(\Delta)$	-2308.93	-2308.12	-2308.29	-2307.89	-2308.31	0.45
$P_a(\mu)Q_aP_d(\Delta)$	-1900.36	-1899.46	-1900.33	-1900.69	-1900.21	0.53
$\boldsymbol{P_a(\mu)Q_aP_a(\Delta)}$	**-1873.21**	**-1872.88**	**-1872.55**	**-1873.24**	**-1872.97**	**0.32**
$P_a(\mu)Q_sP_d(\Delta)$	-1929.77	-1930.08	-1930.90	-1930.34	-1930.27	0.48
$P_a(\mu)Q_sP_a(\Delta)$	-1898.25	-1898.70	-1898.84	-1898.62	-1898.60	0.25

The Impact of Prior Choice on Pairwise Dispersal Rates

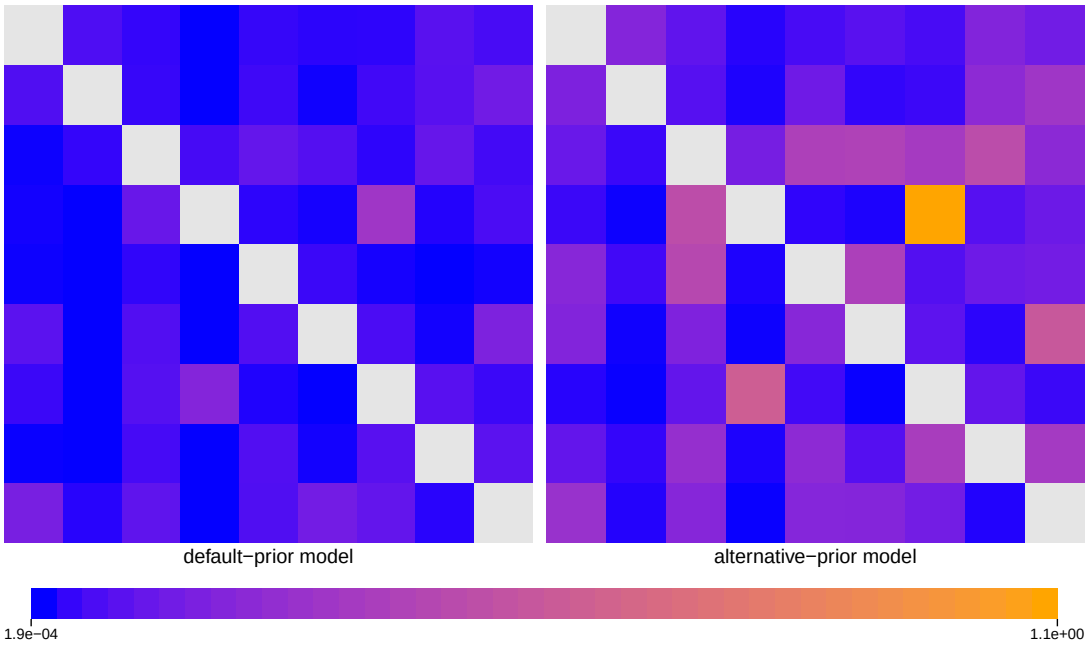

Figure S.1.46: The impact of prior choice on pairwise dispersal rates for Influenza B/Yamagata dataset. Heatmaps summarize posterior-mean estimates of the instantaneous rate of dispersal between each pair of geographic areas, q_{ij}, under the default (left) and alternative (right) prior models.

The Impact of Prior Choice on the Inferred Support for Dispersal Routes

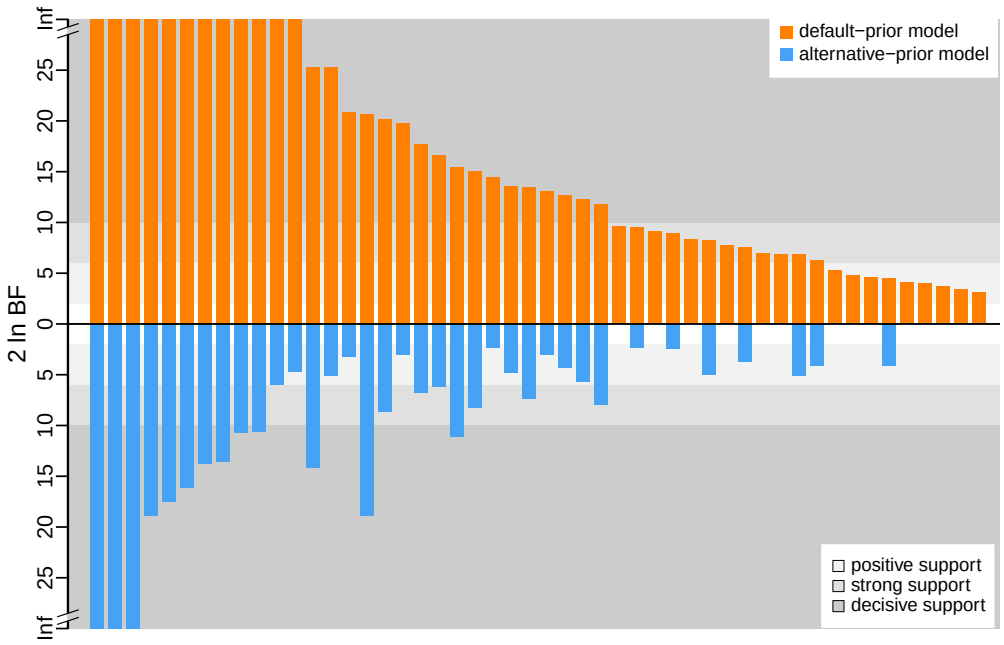

Figure S.1.47: The impact of prior choice on the inferred support for dispersal routes for Influenza B/Yamagata dataset. We compare the evidential support for each dispersal route for Influenza B/Yamagata dataset under the default (orange) and alternative (blue) prior models. Each bar indicates the $2\ln \mathrm{BF}$ for the corresponding dispersal route between two areas; only supported dispersal routes (*i.e.*, $2\ln \mathrm{BF} > 2$) are plotted.

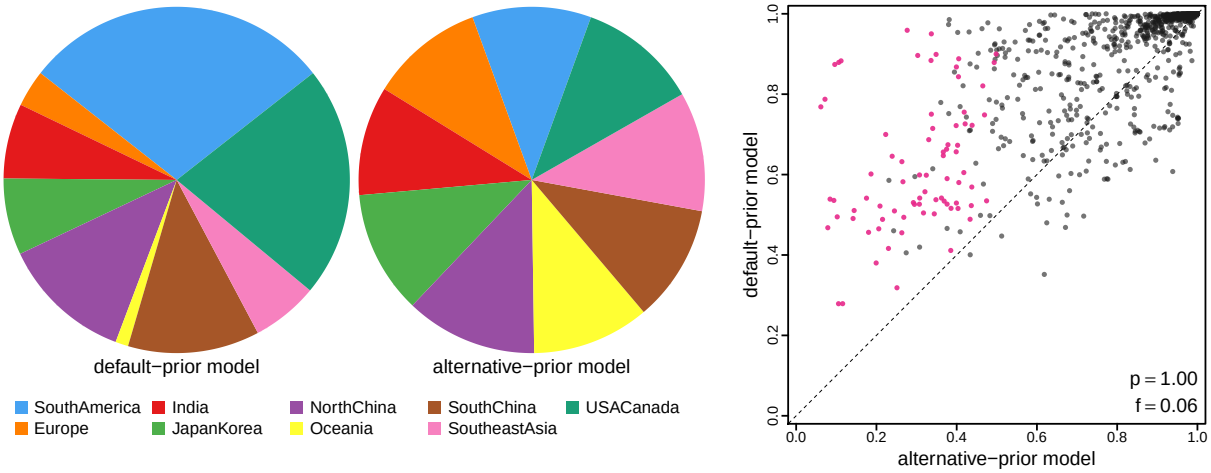

Figure S.1.48: The impact of prior choice ancestral-area estimates for Influenza B/Yamagata dataset. The left panel compares the posterior probability of each ancestral area at the root node under the default- and alternative-prior models for Influenza B/Yamagata dataset. The right panel plots the posterior probability of the most probable ancestral area under the default-prior model for each node in the MCC tree (y-axis) against the corresponding posterior probability of that area under the alternative-prior model (x-axis). Pink dots represent the internal nodes where the MAP ancestral area inferred under the default-prior model differs from that inferred under the alternative-prior models. The statistic p denotes the fraction of internal nodes that are shared under the default- and alternative-prior models; f, is the fraction of shared nodes where the MAP ancestral area differs under the default- and alternative-prior models. Note that the posterior probabilities of the MAP ancestral area under the default-prior model are generally higher than those under the alternative-prior model (*i.e.*, the default-prior model tends to mask uncertainty in the ancestral-area estimates).

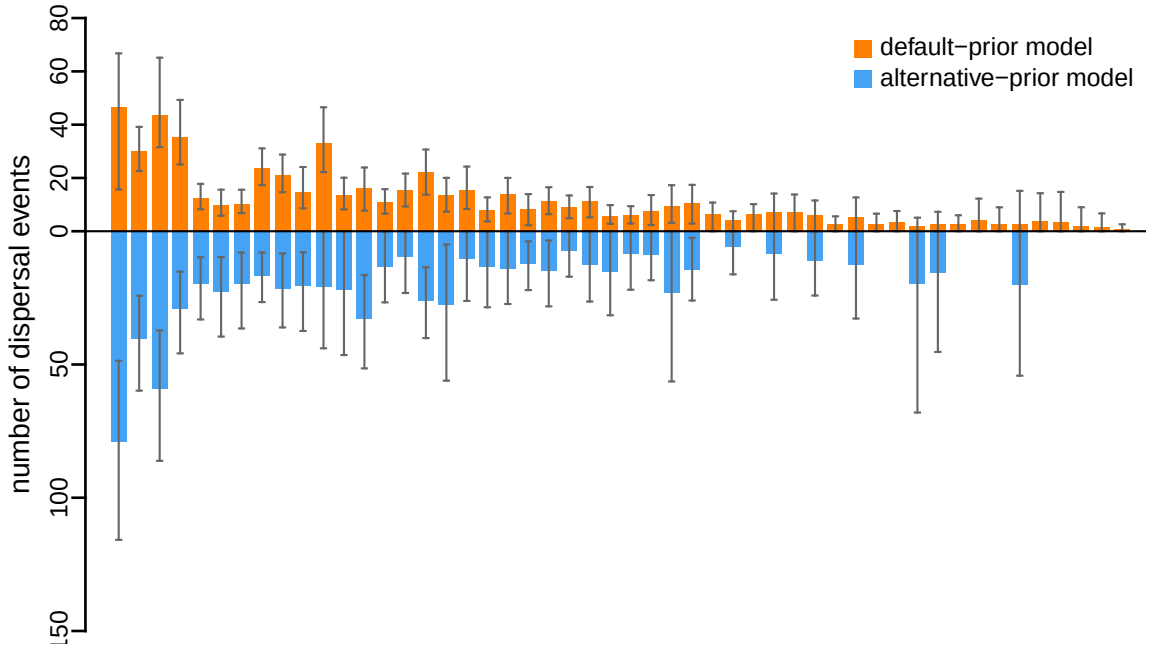

Figure S.1.49: The impact of prior choice on the inferred number of dispersal events between each pair of areas for Influenza B/Yamagata dataset. The reflected bar plot depicts the number of dispersal events inferred under the default (orange) and alternative (blue) prior models for Influenza B/Yamagata dataset. Each bar indicates the posterior-mean number of dispersal events between a pair of areas; whiskers indicate the 95% credible interval. Note that only the number of dispersal events over the "significant" dispersal routes (*i.e.*, $2 \ln \mathrm{BF} > 2$) are figured.

Rabies Virus

Yao et al. (2015) explored the geographic dynamics of the rabies virus in China based on sequences sampled across 19 provinces between 1986–2012. The authors performed separate phylodynamic analyses on two main lineages, Clade I and Clade II. Our reanalyses are based on the dataset defined by Clade I.

We acquired the sampling time and location data, as well as the GenBank accession numbers from Table S1 in Yao et al. (2015), and then obtained the nucleotide sequences from GenBank. This dataset has 141 sequences distributed among 18 geographic areas. We aligned the nucleotide sequences using `MUSCLE` version 3.8 (Edgar 2004). The files containing the GenBank accession numbers, the sequence alignment, and the sampling time and location data are available in our GitHub and Dryad repositories.

To infer the marginal posterior distribution of phylogenies given the sequence alignment, we specified a phylogenetic model with the following components: (1) the GTR+I+Γ_4 substitution model (Tavaré 1986; Yang 1994; Gu et al. 1995); (2) the uncorrelated lognormal (UCLN) branch-rate prior model (Drummond et al. 2006; Rannala and Yang 2007), and; (3) the Gaussian Markov Random Field (GMRF) Bayesian Skyride coalescent node-age model (Minin et al. 2008). Details of these analyses are available in the XML scripts included in our GitHub and Dryad repositories.

We ran four independent MCMC simulations in `BEAST` version 1.8.2 for 200 million generations each, sampling every 15000 generations. We first assessed the performance of each MCMC simulation using `Tracer` version 1.7.1 (Rambaut et al. 2018), removed the first 10% of samples from each chain as the burn-in, and then combined the remaining posterior samples of trees from the replicate simulations using `LogCombiner` version 1.8.2. This resulted in a posterior sample of 1200 trees (available in our GitHub and Dryad repositories), which we then used as the prior distribution of phylogenies for the second step of our sequential analyses.

The MCMC simulations used in the second step of our sequential analyses and of the analyses used to estimate marginal likelihoods under each prior model are described above in this section. Details of these analyses are available in the XML scripts included in our GitHub and Dryad repositories.

Table S.1.15: Marginal-likelihood estimates of the eight prior models for the Rabies virus dataset. Columns 2–5 list marginal likelihoods inferred from four replicate analyses; the last two columns list the mean and standard deviation of these marginal-likelihood estimates. Candidate models are listed in rows and include all possible combinations of: (1) instantaneous-rate matrices (symmetric, Q_s or asymmetric, Q_a); (2) priors on the average dispersal rate [default, $P_d(\mu)$ or alternative, $P_a(\mu)$], and; (3) priors on the number of dispersal routes [default, $P_d(\Delta)$ or alternative, $P_a(\Delta)$]. The preferred default- and alternative-prior models are indicated in bold text.

Model	replicate1	replicate2	replicate3	replicate4	mean	sd
$\boldsymbol{P_d(\mu)Q_aP_d(\Delta)}$	**-310.66**	**-310.07**	**-309.44**	**-309.42**	**-309.90**	**0.59**
$P_d(\mu)Q_aP_a(\Delta)$	-297.53	-297.26	-296.90	-297.23	-297.23	0.26
$P_d(\mu)Q_sP_d(\Delta)$	-319.33	-319.90	-319.12	-319.13	-319.37	0.37
$P_d(\mu)Q_sP_a(\Delta)$	-299.37	-299.85	-299.47	-299.63	-299.58	0.21
$P_a(\mu)Q_aP_d(\Delta)$	-266.85	-266.86	-266.05	-266.78	-266.64	0.39
$\boldsymbol{P_a(\mu)Q_aP_a(\Delta)}$	**-258.36**	**-257.80**	**-258.00**	**-258.39**	**-258.14**	**0.29**
$P_a(\mu)Q_sP_d(\Delta)$	-269.21	-269.26	-269.49	-269.30	-269.32	0.12
$P_a(\mu)Q_sP_a(\Delta)$	-258.67	-258.73	-258.40	-258.49	-258.57	0.15

The Impact of Prior Choice on Pairwise Dispersal Rates

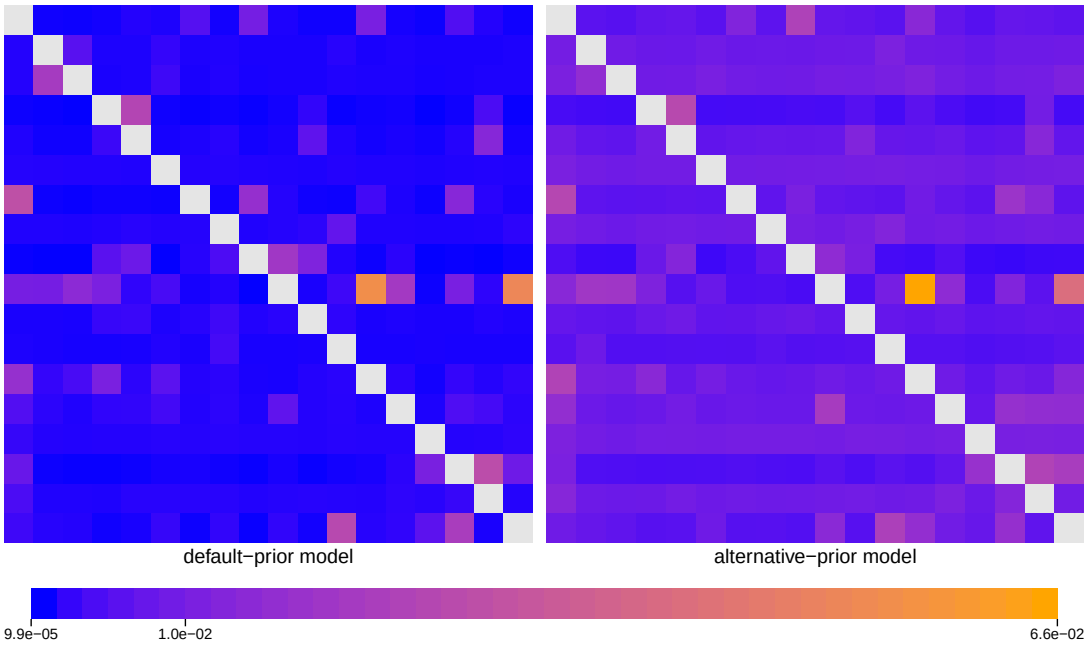

Figure S.1.50: The impact of prior choice on pairwise dispersal rates for the Rabies virus dataset. Heatmaps summarize posterior-mean estimates of the instantaneous rate of dispersal between each pair of geographic areas, q_{ij}, under the default (left) and alternative (right) prior models.

The Impact of Prior Choice on the Inferred Support for Dispersal Routes

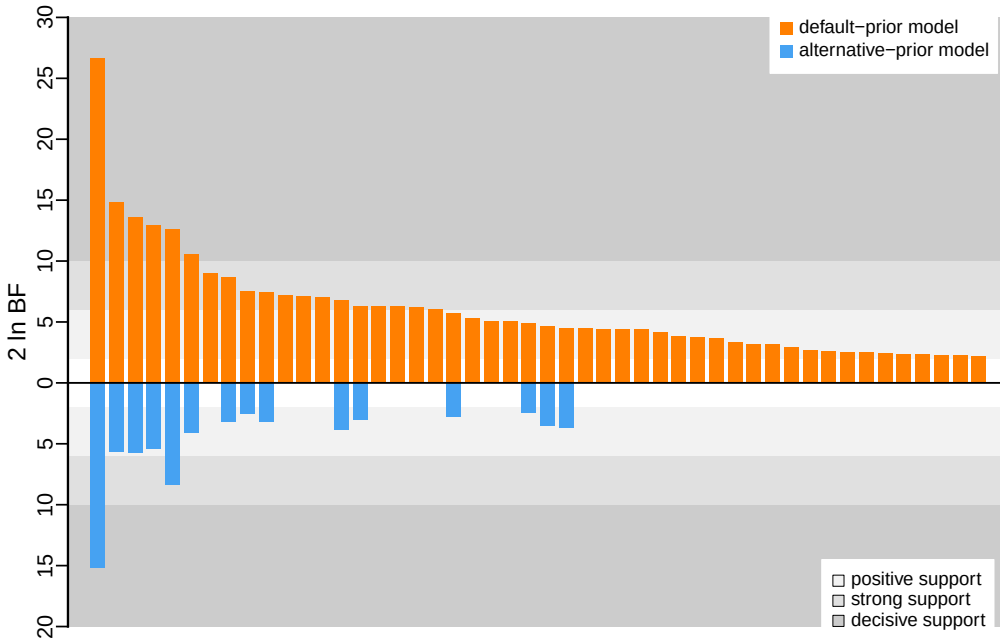

Figure S.1.51: The impact of prior choice on the inferred support for dispersal routes for the Rabies virus dataset. We compare the evidential support for each dispersal route for the Rabies virus dataset under the default (orange) and alternative (blue) prior models. Each bar indicates the $2\ln \mathrm{BF}$ for the corresponding dispersal route between two areas; only supported dispersal routes (*i.e.*, $2\ln \mathrm{BF} > 2$) are plotted.

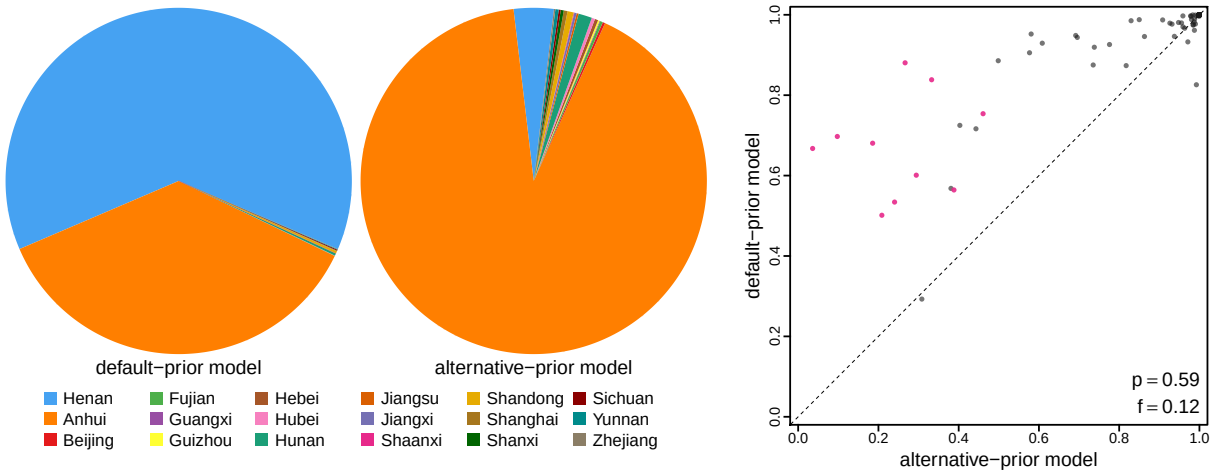

Figure S.1.52: The impact of prior choice ancestral-area estimates for the Rabies virus dataset. The left panel compares the posterior probability of each ancestral area at the root node under the default- and alternative-prior models for the Rabies virus dataset. The right panel plots the posterior probability of the most probable ancestral area under the default-prior model for each node in the MCC tree (y-axis) against the corresponding posterior probability of that area under the alternative-prior model (x-axis). Pink dots represent the internal nodes where the MAP ancestral area inferred under the default-prior model differs from that inferred under the alternative-prior models. The statistic p denotes the fraction of internal nodes that are shared under the default- and alternative-prior models; f, is the fraction of shared nodes where the MAP ancestral area differs under the default- and alternative-prior models. Note that the posterior probabilities of the MAP ancestral area under the default-prior model are generally higher than those under the alternative-prior model (*i.e.*, the default-prior model tends to mask uncertainty in the ancestral-area estimates).

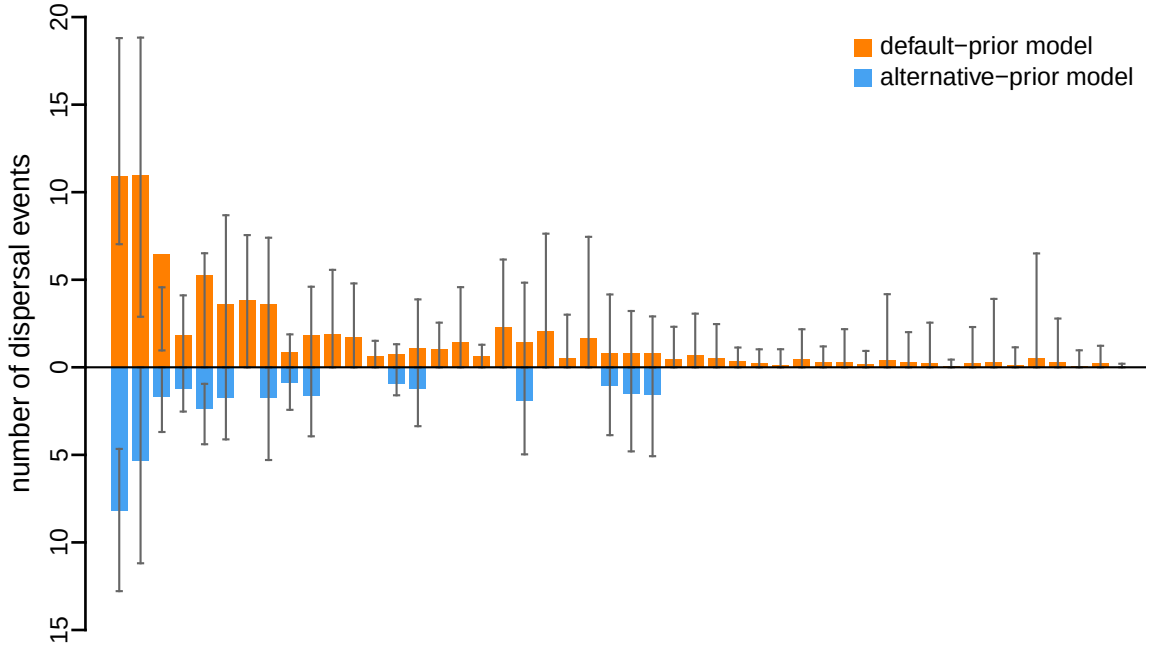

Figure S.1.53: The impact of prior choice on the inferred number of dispersal events between each pair of areas for the Rabies virus dataset. The reflected bar plot depicts the number of dispersal events inferred under the default (orange) and alternative (blue) prior models for the Rabies virus dataset. Each bar indicates the posterior-mean number of dispersal events between a pair of areas; whiskers indicate the 95% credible interval. Note that only the number of dispersal events over the "significant" dispersal routes (*i.e.*, $2 \ln \mathrm{BF} > 2$) are figured.

SARS-CoV-2 Global

Gao et al. (2022) explored the early spread of the COVID-19 pandemic and the efficacy of mitigation measures on limiting the spread using publicly available SARS-CoV-2 genomic sequences collected during the early phase of the pandemic. Here we used one of the datasets produced in that study to assess the impact of prior misspecification on phylodynamic inference of biogeographic history. This dataset contains 1271 SARS-CoV-2 genomic sequences that were originally obtained from the Global Initiative on Sharing All Influenza Data (GISAID, Shu and McCauley 2017). Details about the data curation process can be found in Gao et al. (2022). An alignment were then inferred using `MUSCLE` version 3.8 (Edgar 2004) with the curated dataset. The sampling time and geographic location associated with each sequence of the alignment were also acquired from GISAID.

Different from the other datasets, where we performed the second step of the sequential phylodynamic inferences by marginalizing over the posterior distribution of trees inferred using sequence data and the associated sampling times (but not sampling locations), here we conditioned on the maximum clade credibility (MCC) tree summarized from the posterior distribution to ensure numerical stability of the analyses. This MCC tree was obtained directly from Gao et al. (2022); details (including model and prior specification, as well as the `BEAST` analyses settings) about the analyses that estimated the tree can be found in that study. Following Gao et al. (2022), we discretized the globe into 23 geographic areas to; see Gao et al. (2022) for detailed description of the geographic-area delineation. The MCC tree and the file containing the associated sampling time and location of each sequence are available in our GitHub and Dryad repositories.

The MCMC simulations used in the second step of our sequential analyses and of the analyses used to estimate marginal likelihoods under each prior model are described above in this section. Details of these analyses are available in the XML scripts included in our GitHub and Dryad repositories.

Table S.1.16: Marginal-likelihood estimates of the eight prior models for the SARS-CoV-2 Global dataset. Columns 2–5 list marginal likelihoods inferred from four replicate analyses; the last two columns list the mean and standard deviation of these marginal-likelihood estimates. Candidate models are listed in rows and include all possible combinations of: (1) instantaneous-rate matrices (symmetric, Q_s or asymmetric, Q_a); (2) priors on the average dispersal rate [default, $P_d(\mu)$ or alternative, $P_a(\mu)$], and; (3) priors on the number of dispersal routes [default, $P_d(\Delta)$ or alternative, $P_a(\Delta)$]. The preferred default- and alternative-prior models are indicated in bold text.

Model	replicate1	replicate2	replicate3	replicate4	mean	sd
$\boldsymbol{P_d(\mu)Q_aP_d(\Delta)}$	**-2527.09**	**-2526.18**	**-2525.69**	**-2524.43**	**-2525.85**	**1.11**
$P_d(\mu)Q_aP_a(\Delta)$	-2503.84	-2504.09	-2505.79	-2504.27	-2504.50	0.88
$P_d(\mu)Q_sP_d(\Delta)$	-2559.29	-2560.11	-2561.17	-2560.25	-2560.21	0.77
$P_d(\mu)Q_sP_a(\Delta)$	-2504.00	-2504.09	-2502.86	-2503.08	-2503.51	0.63
$P_a(\mu)Q_aP_d(\Delta)$	-2166.40	-2164.48	-2162.31	-2165.59	-2164.69	1.78
$\boldsymbol{P_a(\mu)Q_aP_a(\Delta)}$	**-2161.51**	**-2160.05**	**-2160.49**	**-2160.21**	**-2160.57**	**0.65**
$P_a(\mu)Q_sP_d(\Delta)$	-2255.39	-2255.53	-2260.00	-2255.15	-2256.52	2.33
$P_a(\mu)Q_sP_a(\Delta)$	-2200.05	-2200.87	-2200.58	-2200.18	-2200.42	0.37

The Impact of Prior Choice on Pairwise Dispersal Rates

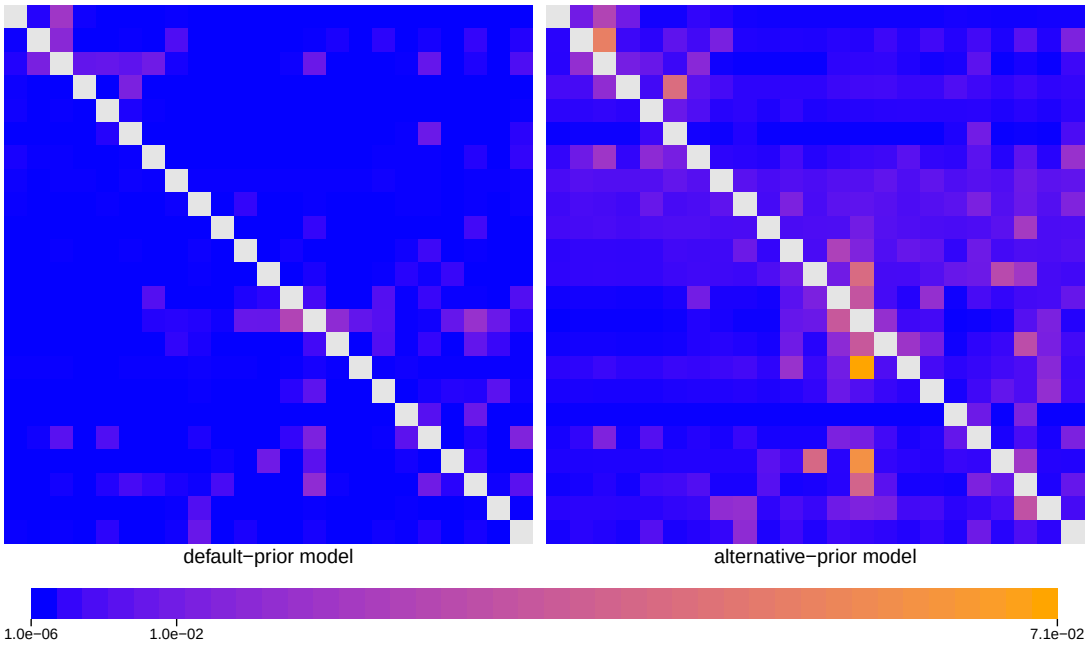

Figure S.1.54: The impact of prior choice on pairwise dispersal rates for the SARS-CoV-2 Global dataset. Heatmaps summarize posterior-mean estimates of the instantaneous rate of dispersal between each pair of geographic areas, q_{ij}, under the default (left) and alternative (right) prior models.

The Impact of Prior Choice on the Inferred Support for Dispersal Routes

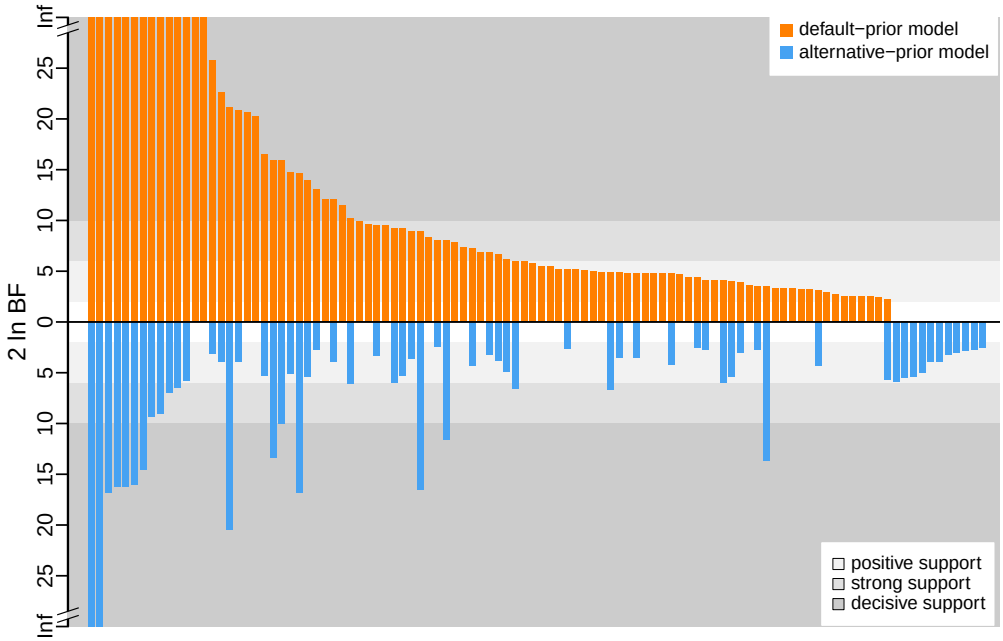

Figure S.1.55: The impact of prior choice on the inferred support for dispersal routes for the SARS-CoV-2 Global dataset. We compare the evidential support for each dispersal route for the SARS-CoV-2 Global dataset under the default (orange) and alternative (blue) prior models. Each bar indicates the $2\ln \mathrm{BF}$ for the corresponding dispersal route between two areas; only supported dispersal routes (*i.e.*, $2\ln \mathrm{BF} > 2$) are plotted.

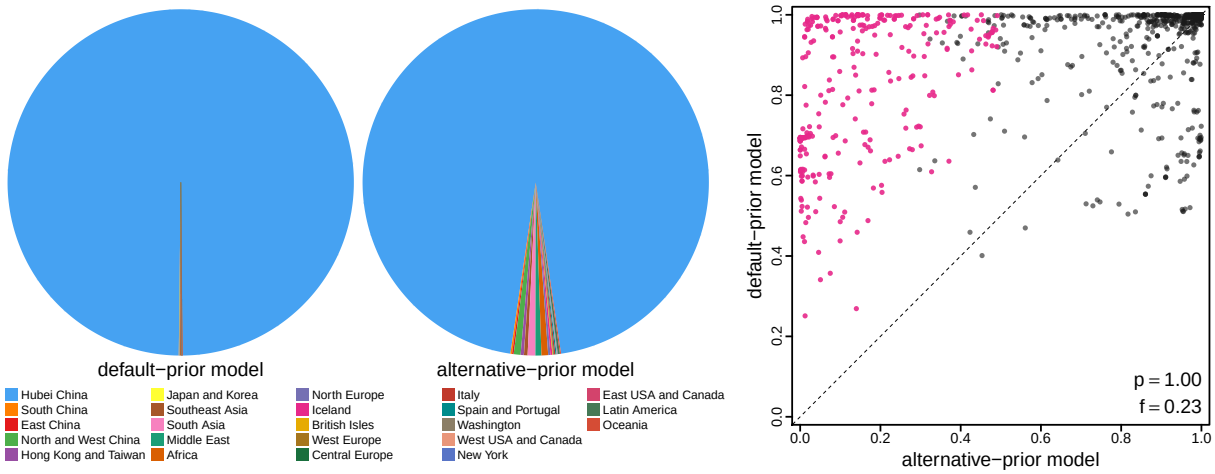

Figure S.1.56: The impact of prior choice ancestral-area estimates for the SARS-CoV-2 Global dataset. The left panel compares the posterior probability of each ancestral area at the root node under the default- and alternative-prior models for the SARS-CoV-2 Global dataset. The right panel plots the posterior probability of the most probable ancestral area under the default-prior model for each node in the MCC tree (y-axis) against the corresponding posterior probability of that area under the alternative-prior model (x-axis). Pink dots represent the internal nodes where the MAP ancestral area inferred under the default-prior model differs from that inferred under the alternative-prior models. The statistic p denotes the fraction of internal nodes that are shared under the default- and alternative-prior models; f, is the fraction of shared nodes where the MAP ancestral area differs under the default- and alternative-prior models. Note that the posterior probabilities of the MAP ancestral area under the default-prior model are generally higher than those under the alternative-prior model (*i.e.*, the default-prior model tends to mask uncertainty in the ancestral-area estimates).

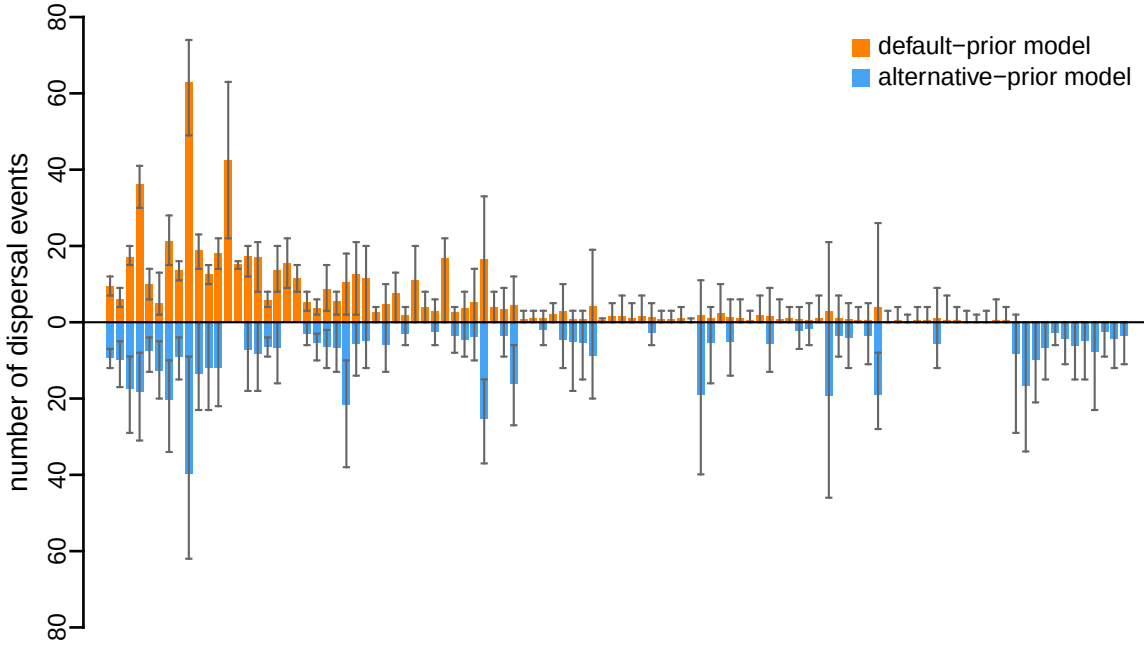

Figure S.1.57: The impact of prior choice on the inferred number of dispersal events between each pair of areas for the SARS-CoV-2 Global dataset. The reflected bar plot depicts the number of dispersal events inferred under the default (orange) and alternative (blue) prior models for the SARS-CoV-2 Global dataset. Each bar indicates the posterior-mean number of dispersal events between a pair of areas; whiskers indicate the 95% credible interval. Note that only the number of dispersal events over the "significant" dispersal routes (*i.e.*, $2 \ln \mathrm{BF} > 2$) are figured.

SARS-CoV-2 B.1.1.7 US

Alpert et al. (2021) explored the introduction, establishment, and geographic dispersal dynamics of the B.1.1.7 variant of SARS-CoV-2 in the United States. The authors produced a dataset containing 1908 SARS-CoV-2 genomic sequences, subsampled from all the B.1.1.7 variant genomes available on GISAID (Shu and McCauley 2017) as of February 26, 2021, focussing on the samples from the US. They discretized the geographic space by states (for the US samples) or by Europe or not Europe (for the international samples), resulting in 22 (20 US and 2 international) geographic areas. Details about the data and the curation procedures can be found in Alpert et al. (2021).

Here we explored the impact of prior choice on phylodynamic inference of biogeographic history using this dataset. Following the original study, we conditioned on the summary tree (obtained directly from the Github repository of Alpert et al. 2021) inferred without the geographic data (*i.e.*, using the sequence data and the sampling time for each sequence) to perform the biogeographic inference. This summary tree and the file containing the associated sampling time and location of each sequence are available in our GitHub and Dryad repositories.

The MCMC simulations used in the second step of our sequential analyses and of the analyses used to estimate marginal likelihoods under each prior model are described above in this section. Details of these analyses are available in the XML scripts included in our GitHub and Dryad repositories.

Table S.1.17: Marginal-likelihood estimates of the eight prior models for the SARS-CoV-2 B.1.1.7 US dataset. Columns 2–5 list marginal likelihoods inferred from four replicate analyses; the last two columns list the mean and standard deviation of these marginal-likelihood estimates. Candidate models are listed in rows and include all possible combinations of: (1) instantaneous-rate matrices (symmetric, Q_s or asymmetric, Q_a); (2) priors on the average dispersal rate [default, $P_d(\mu)$ or alternative, $P_a(\mu)$], and; (3) priors on the number of dispersal routes [default, $P_d(\Delta)$ or alternative, $P_a(\Delta)$]. The preferred default- and alternative-prior models are indicated in bold text.

Model	replicate1	replicate2	replicate3	replicate4	mean	sd
$\boldsymbol{P_d(\mu)Q_aP_d(\Delta)}$	**-1992.03**	**-1988.56**	**-1986.68**	**-1989.42**	**-1989.18**	**2.22**
$P_d(\mu)Q_aP_a(\Delta)$	-1968.56	-1967.08	-1967.14	-1968.09	-1967.72	0.73
$P_d(\mu)Q_sP_d(\Delta)$	-2052.70	-2049.29	-2051.74	-2049.58	-2050.83	1.66
$P_d(\mu)Q_sP_a(\Delta)$	-1980.12	-1979.06	-1977.57	-1979.64	-1979.10	1.11
$P_a(\mu)Q_aP_d(\Delta)$	-1733.59	-1734.20	-1731.54	-1729.36	-1732.17	2.19
$\boldsymbol{P_a(\mu)Q_aP_a(\Delta)}$	**-1720.18**	**-1722.60**	**-1721.11**	**-1720.78**	**-1721.17**	**1.03**
$P_a(\mu)Q_sP_d(\Delta)$	-1848.77	-1845.14	-1847.95	-1846.93	-1847.20	1.56
$P_a(\mu)Q_sP_a(\Delta)$	-1792.84	-1790.98	-1793.38	-1791.87	-1792.27	1.06

The Impact of Prior Choice on Pairwise Dispersal Rates

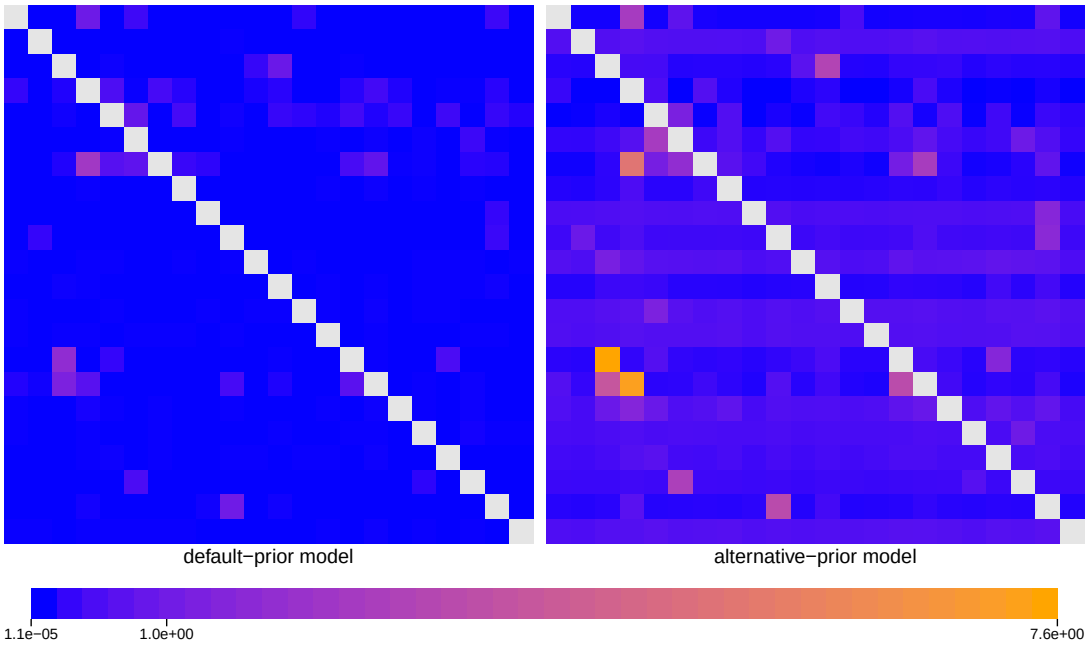

Figure S.1.58: The impact of prior choice on pairwise dispersal rates for the SARS-CoV-2 B.1.1.7 US dataset. Heatmaps summarize posterior-mean estimates of the instantaneous rate of dispersal between each pair of geographic areas, q_{ij}, under the default (left) and alternative (right) prior models.

The Impact of Prior Choice on the Inferred Support for Dispersal Routes

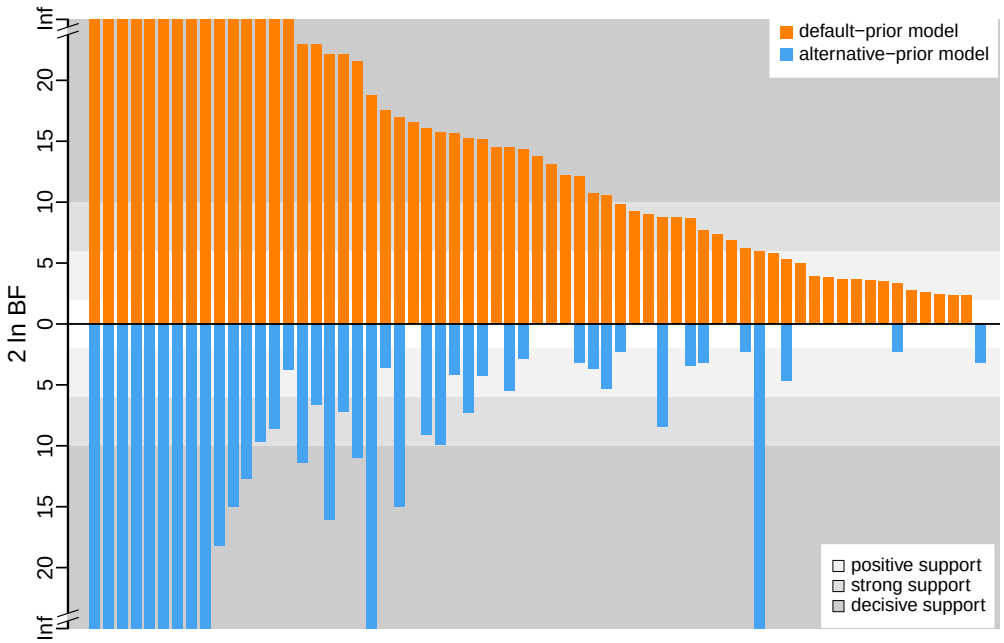

Figure S.1.59: The impact of prior choice on the inferred support for dispersal routes for the SARS-CoV-2 B.1.1.7 US dataset. We compare the evidential support for each dispersal route for the SARS-CoV-2 B.1.1.7 US dataset under the default (orange) and alternative (blue) prior models. Each bar indicates the $2\ln\mathrm{BF}$ for the corresponding dispersal route between two areas; only supported dispersal routes (*i.e.*, $2\ln\mathrm{BF} > 2$) are plotted.

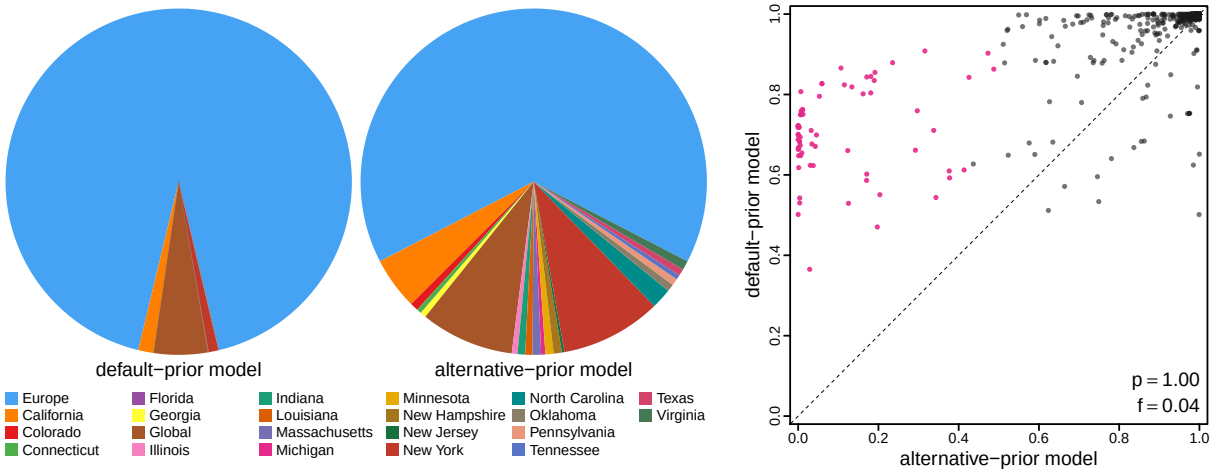

Figure S.1.60: The impact of prior choice ancestral-area estimates for the SARS-CoV-2 B.1.1.7 US dataset. The left panel compares the posterior probability of each ancestral area at the root node under the default- and alternative-prior models for the SARS-CoV-2 B.1.1.7 US dataset. The right panel plots the posterior probability of the most probable ancestral area under the default-prior model for each node in the MCC tree (y-axis) against the corresponding posterior probability of that area under the alternative-prior model (x-axis). Pink dots represent the internal nodes where the MAP ancestral area inferred under the default-prior model differs from that inferred under the alternative-prior models. The statistic p denotes the fraction of internal nodes that are shared under the default- and alternative-prior models; f, is the fraction of shared nodes where the MAP ancestral area differs under the default- and alternative-prior models. Note that the posterior probabilities of the MAP ancestral area under the default-prior model are generally higher than those under the alternative-prior model (*i.e.*, the default-prior model tends to mask uncertainty in the ancestral-area estimates).

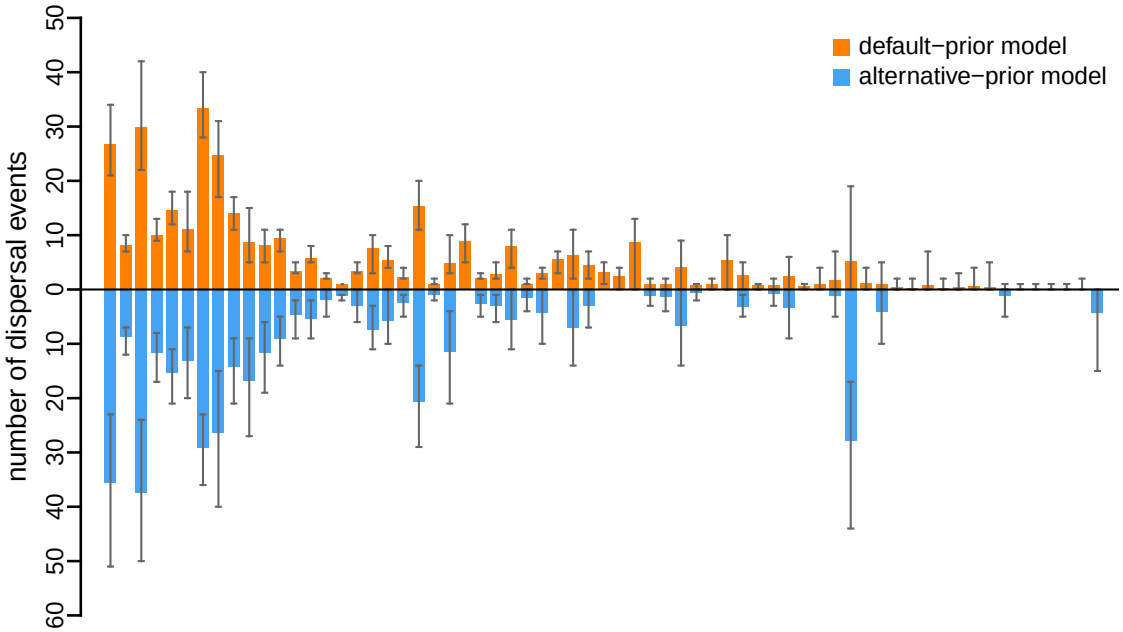

Figure S.1.61: The impact of prior choice on the inferred number of dispersal events between each pair of areas for the SARS-CoV-2 B.1.1.7 US dataset. The reflected bar plot depicts the number of dispersal events inferred under the default (orange) and alternative (blue) prior models for the SARS-CoV-2 B.1.1.7 US dataset. Each bar indicates the posterior-mean number of dispersal events between a pair of areas; whiskers indicate the 95% credible interval. Note that only the number of dispersal events over the "significant" dispersal routes (*i.e.*, $2 \ln \mathrm{BF} > 2$) are figured.

SARS-CoV-2 Brazil

Candido et al. (2020) investigated the early spread of the SARS-CoV-2 epidemic in Brazil and the efficacy of mitigation measures on limiting that spread. The authors combined newly sequenced SARS-CoV-2 genomes sampled from Brazil with the genomes available on GISAID (Shu and McCauley 2017) as April 24, 2020 to produce a SARS-CoV-2 sequence dataset focussing on the epidemic in Brazil. This dataset contains 1182 SARS-CoV-2 genomes, including 490 sampled from Brazil and 692 subsampled from the sequences collected outside of Brazil. The authors then discretized the geographic space with three different ways, generating three geographic datasets, including: (1) Brazil or not Brazil (totaling two areas, scheme A); (2) five Brazilian regions (“Southeast‘‘, “Northeast”, “North”, “Centre-West”, and “South”) and five international regions (North America, Europe, Asia, Oceania, and Africa) (totaling 10 areas, scheme B), and; (3) 21 Brazil states and one other area representing the sampling location for all the international sequences (totaling 22 areas, scheme C). Details about the data and the curation procedures can be found in Candido et al. (2020).

Here we explored the impact of prior choice on phylodynamic inference of biogeographic history using the second (SchemeB) and third (SchemeC) geographic datasets. We acquired the marginal posterior probability distribution of phylogenies (inferred from the sequence data and used to perform sequential analyses in Candido et al. 2020) directly from the Dryad repository of the original study. This posterior distribution included 1000 trees, which was then treated as the prior distribution of phylogenies in the second step of the sequential phylodynamic inference. The `trees` file containing this distribution is available in our GitHub and Dryad repositories. We acquired the sampling geographic location data from the XML scripts provided by the original study; this sampling-area data are available in our GitHub and Dryad repositories.

The MCMC simulations used in the second step of our sequential analyses and of the analyses used to estimate marginal likelihoods under each prior model are described above in this section. Details of these analyses are available in the XML scripts included in our GitHub and Dryad repositories.

SchemeB Dataset

The Impact of Prior Choice on Biogeographic Model Fit

Table S.1.18: Marginal-likelihood estimates of the eight prior models for SARS-CoV-2 Brazil SchemeB dataset. Columns 2–5 list marginal likelihoods inferred from four replicate analyses; the last two columns list the mean and standard deviation of these marginal-likelihood estimates. Candidate models are listed in rows and include all possible combinations of: (1) instantaneous-rate matrices (symmetric, Q_s or asymmetric, Q_a); (2) priors on the average dispersal rate [default, $P_d(\mu)$ or alternative, $P_a(\mu)$], and; (3) priors on the number of dispersal routes [default, $P_d(\Delta)$ or alternative, $P_a(\Delta)$]. The preferred default- and alternative-prior models are indicated in bold text.

Model	replicate1	replicate2	replicate3	replicate4	mean	sd
$\boldsymbol{P_d(\mu)Q_aP_d(\Delta)}$	**-1745.64**	**-1745.52**	**-1748.64**	**-1748.59**	**-1747.10**	**1.75**
$P_d(\mu)Q_aP_a(\Delta)$	-1736.13	-1734.77	-1738.46	-1738.56	-1736.98	1.85
$P_d(\mu)Q_sP_d(\Delta)$	-1772.70	-1772.92	-1776.57	-1776.07	-1774.57	2.04
$P_d(\mu)Q_sP_a(\Delta)$	-1758.22	-1757.65	-1761.34	-1761.31	-1759.63	1.97
$P_a(\mu)Q_aP_d(\Delta)$	-1546.46	-1547.06	-1549.53	-1549.10	-1548.04	1.51
$\boldsymbol{P_a(\mu)Q_aP_a(\Delta)}$	**-1530.22**	**-1530.37**	**-1532.87**	**-1532.54**	**-1531.50**	**1.40**
$P_a(\mu)Q_sP_d(\Delta)$	-1592.98	-1593.10	-1596.21	-1595.96	-1594.56	1.76
$P_a(\mu)Q_sP_a(\Delta)$	-1570.65	-1569.97	-1572.74	-1573.07	-1571.61	1.53

The Impact of Prior Choice on Pairwise Dispersal Rates

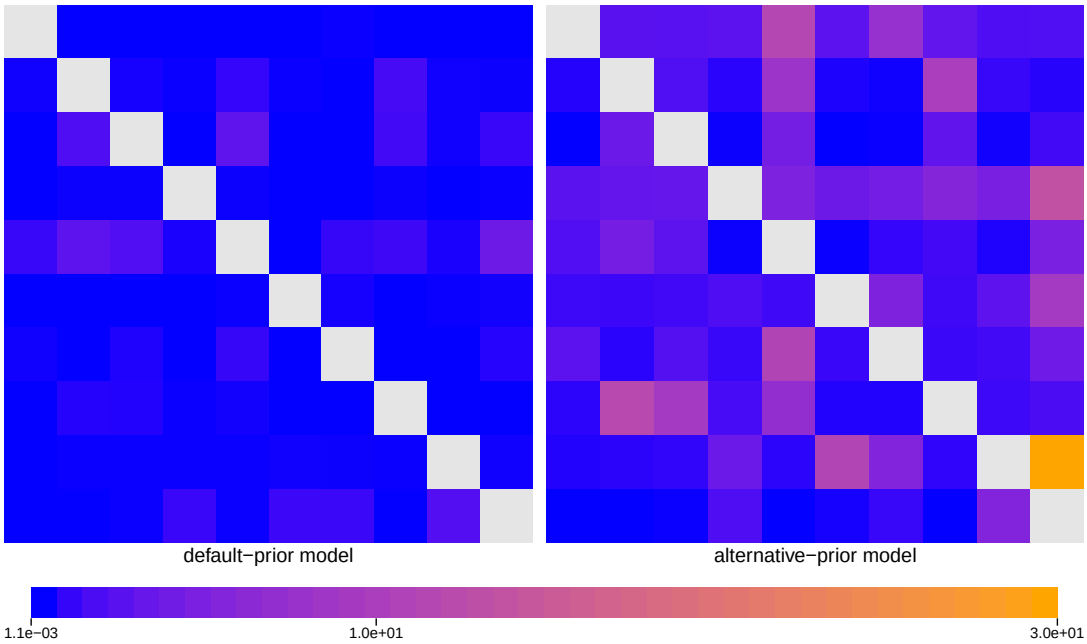

Figure S.1.62: The impact of prior choice on pairwise dispersal rates for SARS-CoV-2 Brazil SchemeB dataset. Heatmaps summarize posterior-mean estimates of the instantaneous rate of dispersal between each pair of geographic areas, q_{ij}, under the default (left) and alternative (right) prior models.

The Impact of Prior Choice on the Inferred Support for Dispersal Routes

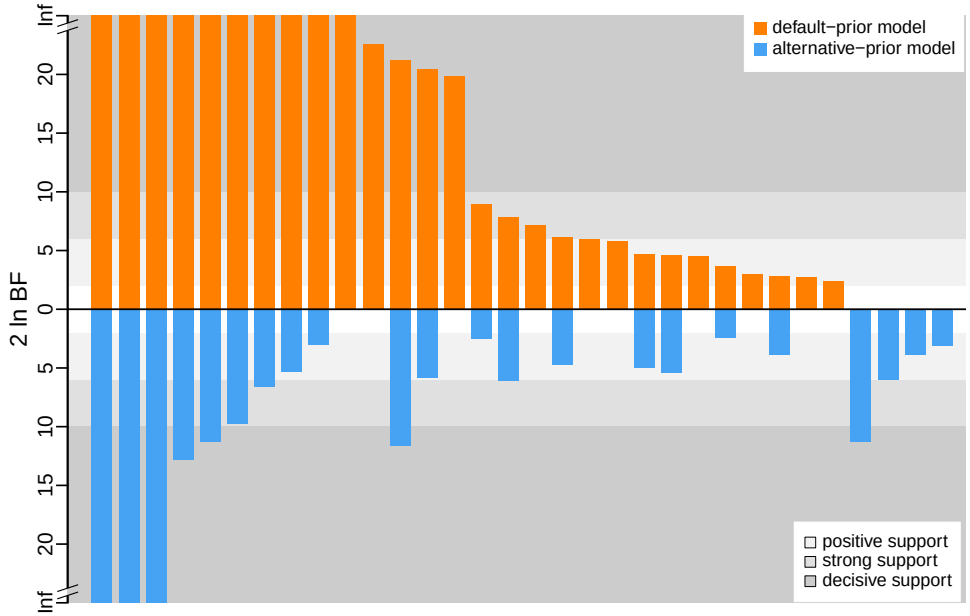

Figure S.1.63: The impact of prior choice on the inferred support for dispersal routes for SARS-CoV-2 Brazil SchemeB dataset. We compare the evidential support for each dispersal route for SARS-CoV-2 Brazil SchemeB dataset under the default (orange) and alternative (blue) prior models. Each bar indicates the $2\ln \mathrm{BF}$ for the corresponding dispersal route between two areas; only supported dispersal routes (*i.e.*, $2\ln \mathrm{BF} > 2$) are plotted.

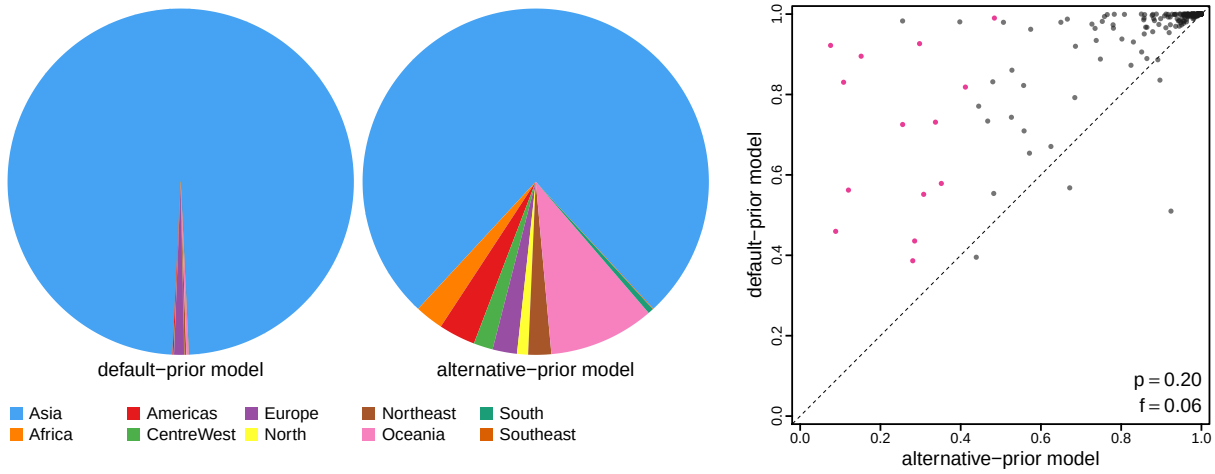

Figure S.1.64: The impact of prior choice ancestral-area estimates for SARS-CoV-2 Brazil SchemeB dataset. The left panel compares the posterior probability of each ancestral area at the root node under the default- and alternative-prior models for SARS-CoV-2 Brazil SchemeB dataset. The right panel plots the posterior probability of the most probable ancestral area under the default-prior model for each node in the MCC tree (y-axis) against the corresponding posterior probability of that area under the alternative-prior model (x-axis). Pink dots represent the internal nodes where the MAP ancestral area inferred under the default-prior model differs from that inferred under the alternative-prior models. The statistic p denotes the fraction of internal nodes that are shared under the default- and alternative-prior models; f, is the fraction of shared nodes where the MAP ancestral area differs under the default- and alternative-prior models. Note that the posterior probabilities of the MAP ancestral area under the default-prior model are generally higher than those under the alternative-prior model (*i.e.*, the default-prior model tends to mask uncertainty in the ancestral-area estimates).

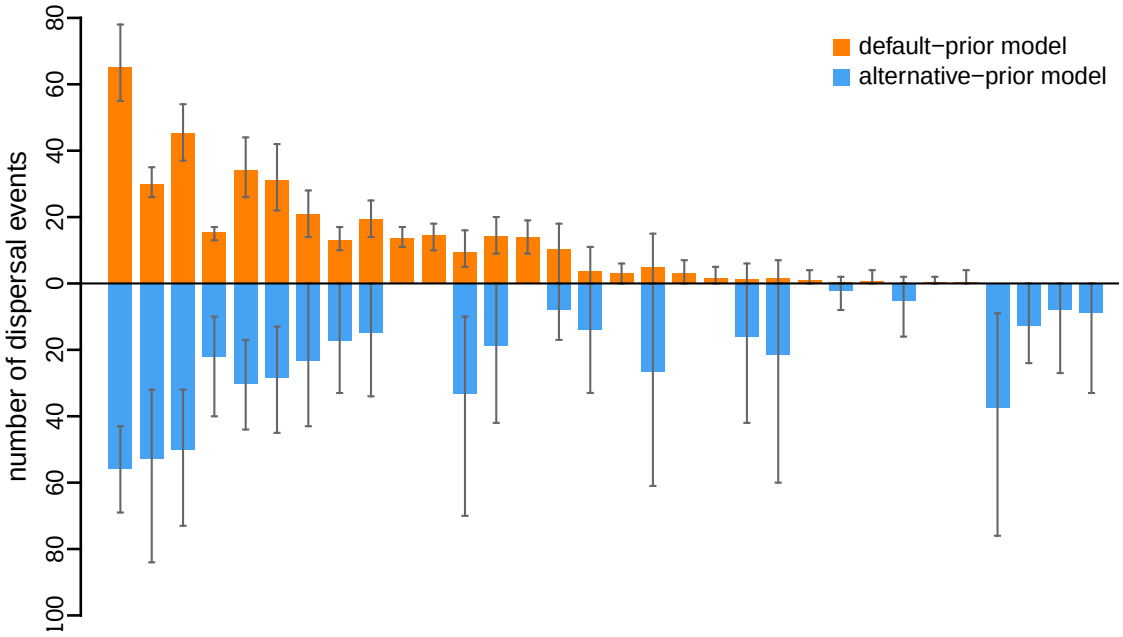

Figure S.1.65: The impact of prior choice on the inferred number of dispersal events between each pair of areas for SARS-CoV-2 Brazil SchemeB dataset. The reflected bar plot depicts the number of dispersal events inferred under the default (orange) and alternative (blue) prior models for SARS-CoV-2 Brazil SchemeB dataset. Each bar indicates the posterior-mean number of dispersal events between a pair of areas; whiskers indicate the 95% credible interval. Note that only the number of dispersal events over the "significant" dispersal routes (*i.e.*, $2 \ln \mathrm{BF} > 2$) are figured.

SchemeC Dataset

The Impact of Prior Choice on Biogeographic Model Fit

Table S.1.19: Marginal-likelihood estimates of the eight prior models for SARS-CoV-2 Brazil SchemeC dataset. Columns 2–5 list marginal likelihoods inferred from four replicate analyses; the last two columns list the mean and standard deviation of these marginal-likelihood estimates. Candidate models are listed in rows and include all possible combinations of: (1) instantaneous-rate matrices (symmetric, Q_s or asymmetric, Q_a); (2) priors on the average dispersal rate [default, $P_d(\mu)$ or alternative, $P_a(\mu)$], and; (3) priors on the number of dispersal routes [default, $P_d(\Delta)$ or alternative, $P_a(\Delta)$]. The preferred default- and alternative-prior models are indicated in bold text.

Model	replicate1	replicate2	replicate3	replicate4	mean	sd
$\boldsymbol{P_d(\mu)Q_aP_d(\Delta)}$	**-1366.00**	**-1366.66**	**-1370.88**	**-1371.08**	**-1368.66**	**2.70**
$P_d(\mu)Q_aP_a(\Delta)$	-1351.90	-1352.41	-1355.62	-1355.61	-1353.89	2.01
$P_d(\mu)Q_sP_d(\Delta)$	-1374.38	-1373.54	-1376.97	-1377.22	-1375.53	1.85
$P_d(\mu)Q_sP_a(\Delta)$	-1353.55	-1353.73	-1355.72	-1357.09	-1355.02	1.69
$P_a(\mu)Q_aP_d(\Delta)$	-1226.37	-1225.32	-1226.78	-1224.41	-1225.72	1.07
$\boldsymbol{P_a(\mu)Q_aP_a(\Delta)}$	**-1220.44**	**-1221.27**	**-1222.42**	**-1221.61**	**-1221.43**	**0.82**
$P_a(\mu)Q_sP_d(\Delta)$	-1290.69	-1288.45	-1291.19	-1291.38	-1290.43	1.35
$P_a(\mu)Q_sP_a(\Delta)$	-1242.86	-1242.87	-1244.99	-1245.57	-1244.07	1.42

The Impact of Prior Choice on Pairwise Dispersal Rates

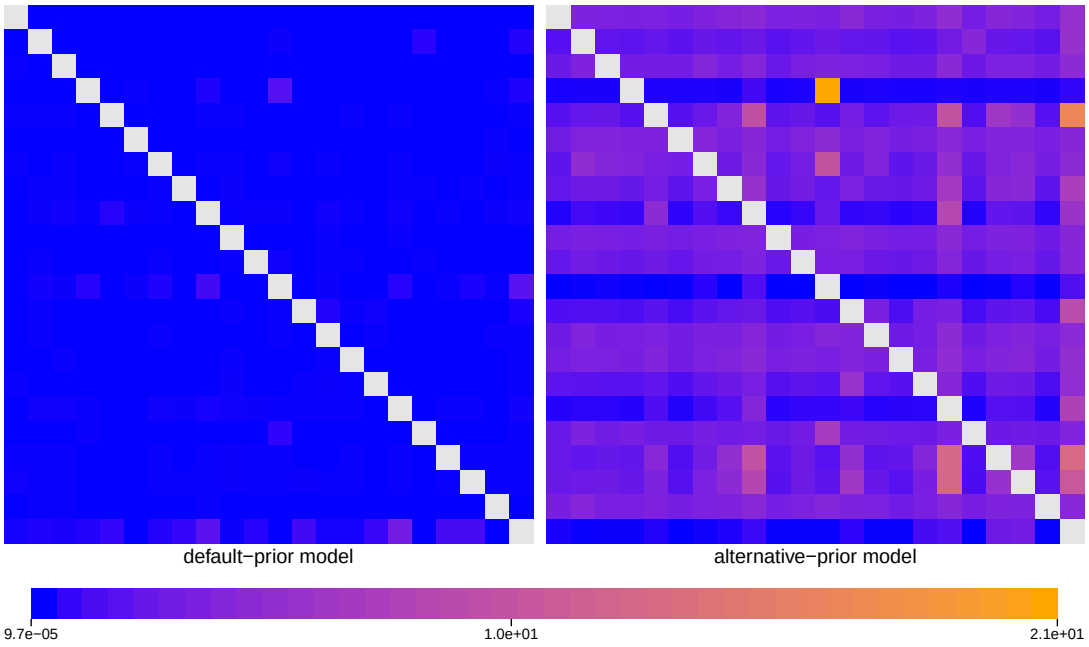

Figure S.1.66: The impact of prior choice on pairwise dispersal rates for SARS-CoV-2 Brazil SchemeC dataset. Heatmaps summarize posterior-mean estimates of the instantaneous rate of dispersal between each pair of geographic areas, q_{ij}, under the default (left) and alternative (right) prior models.

The Impact of Prior Choice on the Inferred Support for Dispersal Routes

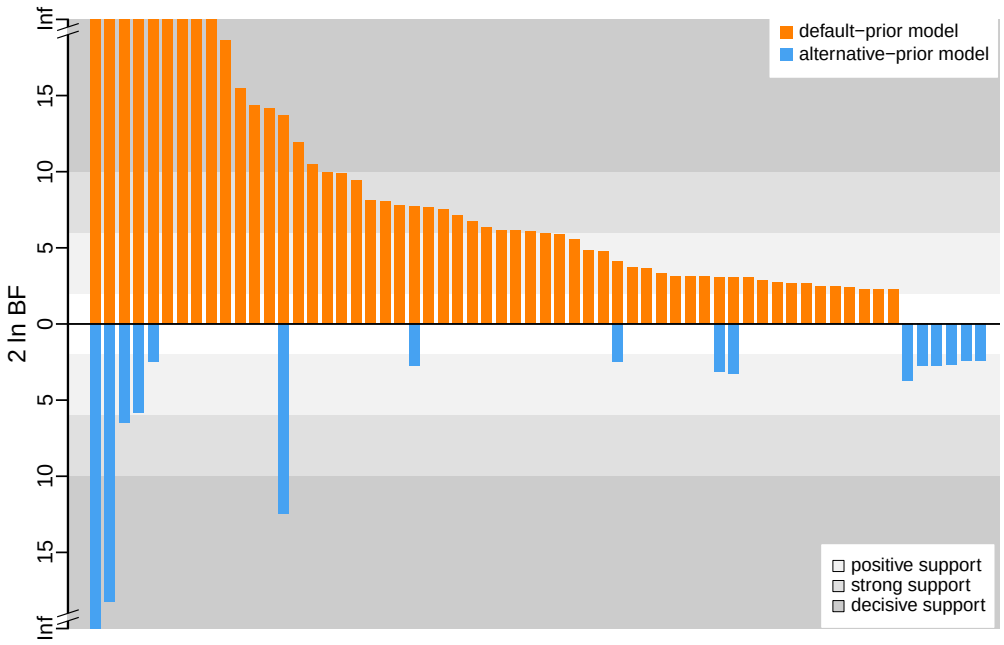

Figure S.1.67: The impact of prior choice on the inferred support for dispersal routes for SARS-CoV-2 Brazil SchemeC dataset. We compare the evidential support for each dispersal route for SARS-CoV-2 Brazil SchemeC dataset under the default (orange) and alternative (blue) prior models. Each bar indicates the $2\ln \mathrm{BF}$ for the corresponding dispersal route between two areas; only supported dispersal routes (*i.e.*, $2\ln \mathrm{BF} > 2$) are plotted.

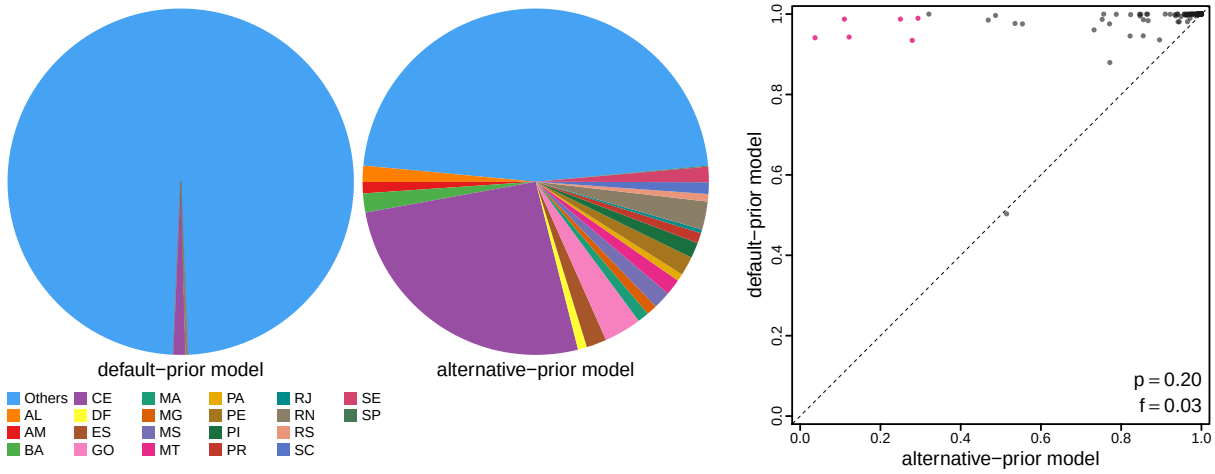

Figure S.1.68: The impact of prior choice ancestral-area estimates for SARS-CoV-2 Brazil SchemeC dataset. The left panel compares the posterior probability of each ancestral area at the root node under the default- and alternative-prior models for SARS-CoV-2 Brazil SchemeC dataset. The right panel plots the posterior probability of the most probable ancestral area under the default-prior model for each node in the MCC tree (y-axis) against the corresponding posterior probability of that area under the alternative-prior model (x-axis). Pink dots represent the internal nodes where the MAP ancestral area inferred under the default-prior model differs from that inferred under the alternative-prior models. The statistic p denotes the fraction of internal nodes that are shared under the default- and alternative-prior models; f, is the fraction of shared nodes where the MAP ancestral area differs under the default- and alternative-prior models. Note that the posterior probabilities of the MAP ancestral area under the default-prior model are generally higher than those under the alternative-prior model (*i.e.*, the default-prior model tends to mask uncertainty in the ancestral-area estimates).

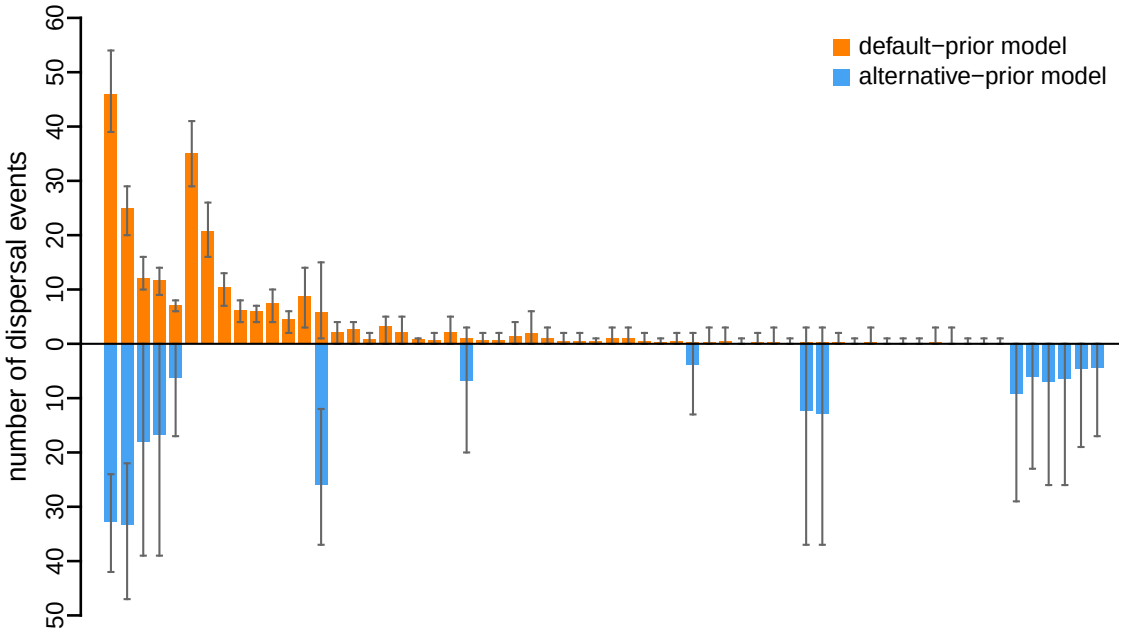

Figure S.1.69: The impact of prior choice on the inferred number of dispersal events between each pair of areas for SARS-CoV-2 Brazil SchemeC dataset. The reflected bar plot depicts the number of dispersal events inferred under the default (orange) and alternative (blue) prior models for SARS-CoV-2 Brazil SchemeC dataset. Each bar indicates the posterior-mean number of dispersal events between a pair of areas; whiskers indicate the 95% credible interval. Note that only the number of dispersal events over the "significant" dispersal routes (*i.e.*, $2 \ln \mathrm{BF} > 2$) are figured.

Chapter 2

PrioriTree: an Interactive Web Utility for Specifying Priors and Assessing Their Impacts in BEAST Biogeographic Analysis

Abstract.—Phylodynamic methods are central to studies of the geographic and demographic history of pathogen outbreaks. Inference under discrete-geographic phylodynamic models—which involve many parameters that must be inferred from minimal information—may be sensitive to our prior beliefs about the model parameters. We present an interactive utility, `PrioriTree`, to help researchers identify and accommodate prior sensitivity in discrete-geographic inferences. Specifically, `PrioriTree` provides a suite of functions to generate input files for—and summarize output from—`BEAST` analyses for performing robust Bayesian inference, data-cloning analyses, and assessing the relative and absolute fit of candidate discrete-geographic (prior) models to empirical datasets. `PrioriTree` is distributed as an R package available at https://github.com/jsigao/prioritree, with a comprehensive user manual provided at https://bookdown.org/jsigao/prioritree_manual/.

Introduction

Phylogenies are increasingly used to study the dispersal dynamics and history of spread of pathogens. The discrete-geographic phylodynamic method developed by Lemey *et al.* (Lemey et al. 2009; Edwards et al. 2011)—implemented in the popular `BEAST` software package (Drummond et al. 2012; Suchard et al. 2018)—is now the standard approach used to infer key aspects of the biogeographic history of pathogen epidemics, including: (1) the area in which an epidemic first originated; (2) the dispersal routes by which the pathogen spread among geographic areas, and; (3) the number of dispersal events between areas. However, these discrete-

geographic models contain many parameters that must be inferred from minimal information (the single geographic area in which each pathogen occurs); inferences under this approach are therefore inherently sensitive to the assumed priors on the model parameters. Unfortunately, the priors implemented as the defaults in BEAST—and used in the vast majority of published studies—are strongly informative and extremely unrealistic; these misinformative priors distort inferences of biogeographic history (see Chapter 1 for details).

Motivated by these considerations, here we present PrioriTree, an interactive web utility developed to help researchers set up and summarize BEAST biogeographic analysis. Specifically, PrioriTree is designed to help researchers: (1) interactively (and graphically) specify priors to generate input files for phylodynamic analyses using BEAST; (2) specify input files for (and generate summaries from) BEAST analyses to assess the prior sensitivity biogeographic inferences, and; (3) specify input files for (and generate summaries from) BEAST analyses to assess the relative and absolute fit of discrete-geographic (prior) models.

FEATURES

There are several non-mutually exclusive strategies to deal with prior sensitivity: (1) specify biologically-motivated/informed priors (*i.e.*, where the prior probability is focussed on 'reasonable' values); (2) specify diffuse/uninformative priors (*i.e.*, where the prior probability is spread (virtually) evenly over a wide range of 'plausible' values); (3) assess whether the posterior estimates (especially of the focal parameters) are sensitive to the prior choice (*e.g.*, perform replicate inferences under a range of candidate priors), and; (4) assess whether the specified prior model adequately describes the underlying data-generating process or compare the relative fit of competing prior models to the data. PrioriTree provides a suite of functions to help users pursue these strategies.

Interactively Set up BEAST Discrete-Geographic Phylodynamic Analyses with Visualized Priors

Before performing the analysis, users may have some intuitions (*e.g.*, knowledge learned from previous analyses) about the parameters that they want to express in their priors. PrioriTree allows users to specify these biologically motivated priors in an interactive manner; it provides the flexibility to specify a range of (hyper)priors and dynamically renders the resulting prior distribution according to the specification in real time (Fig. 2.1).

Alternatively, when users have no prior knowledge about the parameter or prefer not to express such knowledge, they may choose to specify a generic uninformative prior probability distribution that is flexible enough to be updated by the data. `PrioriTree` provides multiple such distributions to serve as the candidate choice, where the default choice of each prior has been identified to perform well in many empirical analyses (see Chapter 1). `PrioriTree` displays the distribution of the selected prior on discrete-geographic model parameter—including the number of dispersal routes, Δ, the average dispersal rate, μ, and the resulting prior distribution on the expected number of dispersal events—with the associated prior mean and 95% credible interval listed alongside (Fig. 2.1).

Specify `BEAST` Analyses and Summarize the Results to Assess Prior Sensitivity in Biogeographic Inference

In either of these scenarios, users may wish to ensure that the impact of the specified prior on the posterior is minimal, and the posterior estimates—which they will draw their biological conclusions upon—are rather robust to the prior choice. A simple but effective way to identify prior sensitivity is to compare the prior to the posterior for each parameter: if the inferred pos-

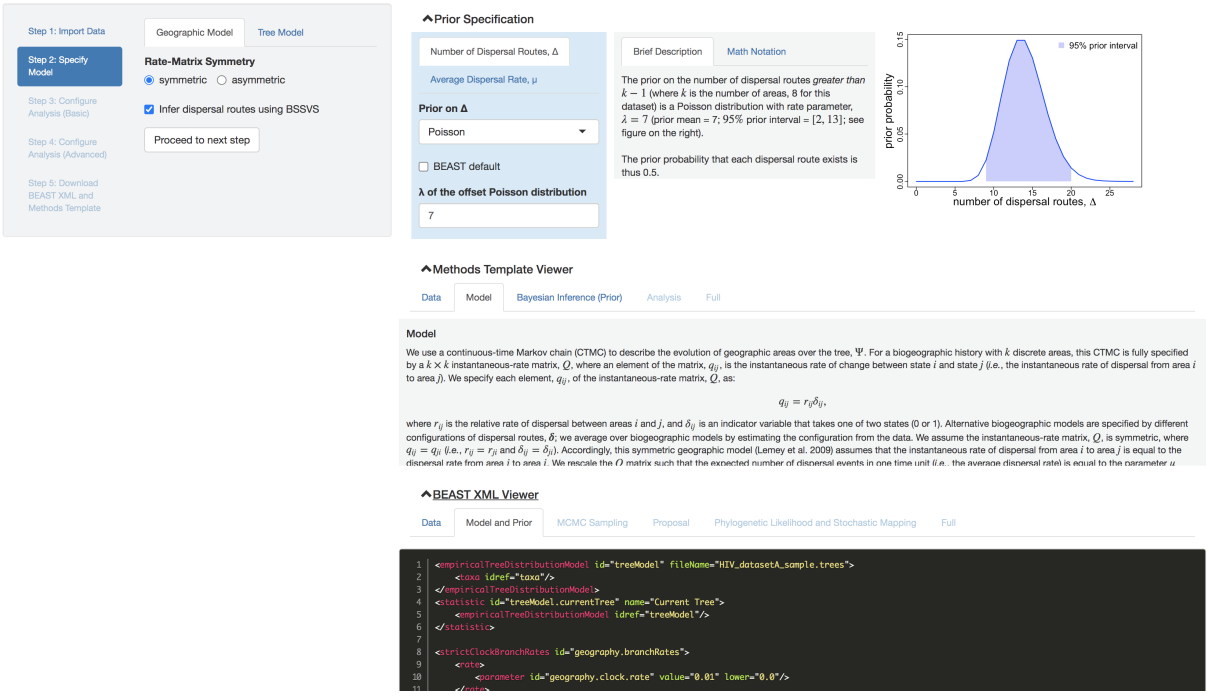

Figure 2.1: `PrioriTree` **main interface.** The main interface is divided into two panels: user input in the left panel, and the right panel dynamically renders the corresponding prior distributions, methods text, and `BEAST` XML script.

terior distribution appears to be (virtually) identical to the specified prior distribution, users should be concerned and motivated to further examine the potential prior-sensitivity issue. Although `PrioriTree` allows users to visualize the prior distribution, the induced prior may deviate substantially from the specified prior due to parameter interactions. It may thus be safer to estimate the joint prior probability distribution of our model parameters using MCMC, and then compare the inferred marginal prior distribution for each parameter to its corresponding inferred marginal posterior distribution. `PrioriTree` provides functions for setting up and summarizing `BEAST` analyses that infer the joint prior distribution (Fig. 2.2, top panel, green).

If the inferred posterior appears to be substantially different from the specified prior, it is still plausible that the prior has exerted a stronger impact than the user would like. To address this concern, we can perform a series of MCMC analyses—of the same dataset under the same inference model—where we iteratively change one (or more) (hyper)priors of our inference model for each separate analysis. We then compare the resulting series of marginal posterior probability distributions for a given parameter to assess whether (or how much) our estimates change under different priors. If the marginal posterior probability distributions vary substantially (especially if they resemble their corresponding marginal prior probability distributions), then we conclude that this parameter exhibits prior sensitivity. This approach, called robust Bayesian inference, is especially applicable when users have multiple candidate priors in mind—either from competing prior hypotheses or alternative flexible probability distributions—when configuring the analysis. `PrioriTree` allows users to set up and summarize `BEAST` robust-Bayesian analyses to examine whether the posterior estimates of a given parameter are robust to the choice of prior (Fig. 2.2, top panel, purple).

Complementary to robust Bayesian inference, another approach called data cloning can also be used to assess prior sensitivity of the biogeographic inference (Robert 1993; Lele et al. 2007; Ponciano et al. 2009, 2012). Under this approach, we also perform a series of MCMC analyses—but under identical priors—where we iteratively increment the number of copies ("clones") of our original dataset used in each separate analysis. We then explore the resulting series of marginal posterior probability distributions for a given parameter to assess how our estimates change as the level of information in the data increases (*i.e.*, as we increment the number of data clones). `PrioriTree` allows users to set up and summarize `BEAST` data-cloning analyses (Fig. 2.2, top panel, gray).

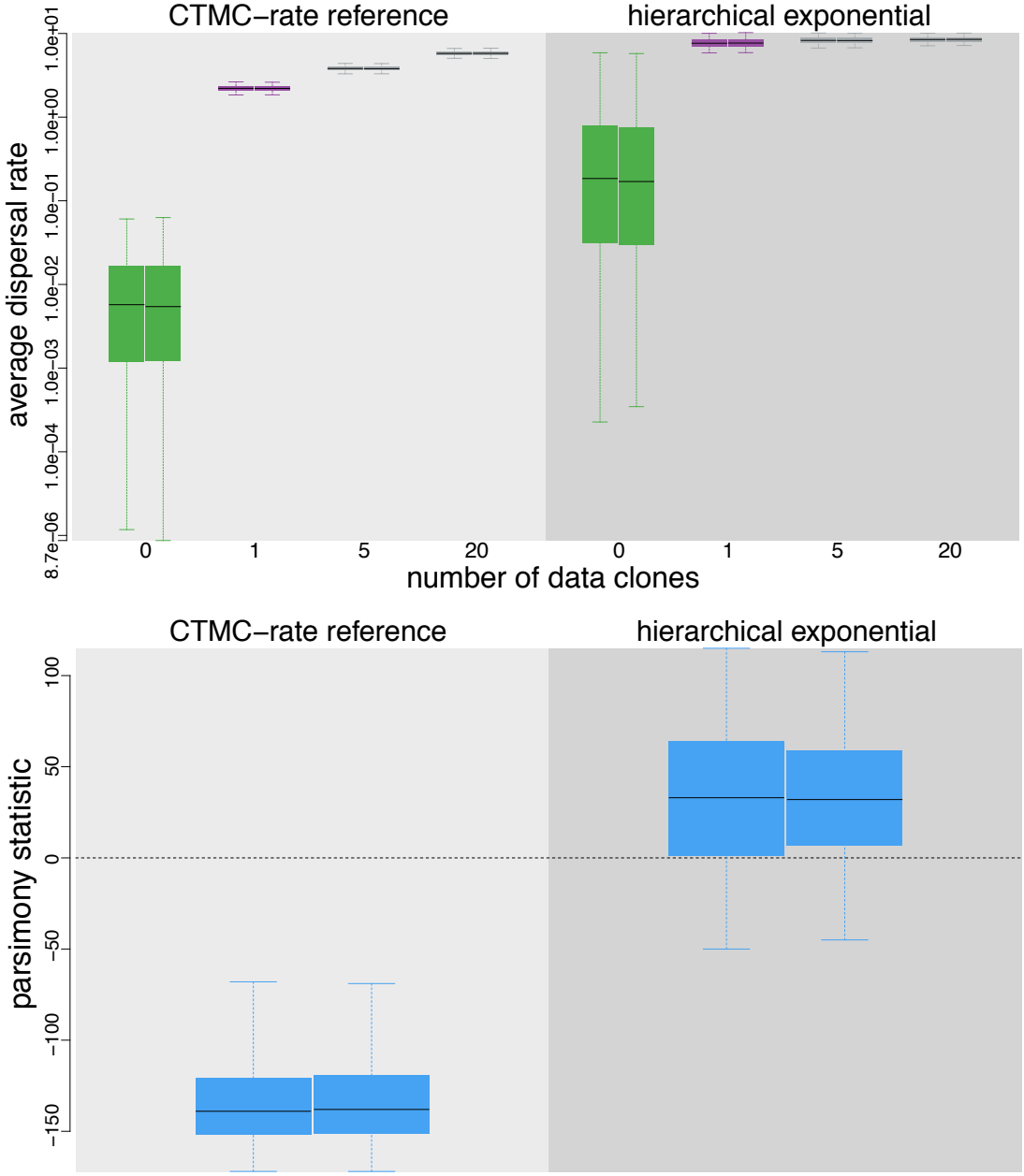

Figure 2.2: Example figures produced by `PrioriTree`. `PrioriTree` allows users to summarize `BEAST` biogeographic analyses for assessing prior sensitivity and model fit. We present the result figures using a SARS-CoV-2 dataset as an example. Top) Combination of inferred distributions of the average dispersal rate, μ, to show the prior (*i.e.*, inferred with no data; green), the posterior (*i.e.* inferred with the data; purple), and the data-cloned posterior (*i.e.* inferred with various numbers of clones of the data; gray). Each box plot corresponds to a single `BEAST` MCMC simulation; the pair of box plots under each type of inference indicates effectively identical distributions, confirming the convergence between MCMC replicates. The left subpanel shows the resulting distributions under the `BEAST` default prior on μ, and the right subpanel shows the corresponding distributions under an alternative diffuse prior. Bottom) Posterior-predictive distribution of the parsimony statistic under the `BEAST` default prior (left subpanel) and the alternative diffuse prior (right subpanel). The statistic computed for the observed data (dash horizontal line) overlaps with the posterior-predictive distributions under the alternative prior while falls outside of the 95% distributions under the `BEAST` default prior, indicating that the default prior is inadequate while the alternative is adequate.

Specify BEAST Analyses and Summarize the Results to Assess Relative or Absolute Model Fit

To assess the impacts of the priors in the discrete-geographic phylodynamic model, PrioriTree also implements functions to help users perform posterior-predictive checking (Gelman et al. 1996; Bollback 2002) to assess the adequacy (*i.e.*, absolute fit) of the specified (prior)model in describing the underlying dispersal process. Posterior-predictive checking assesses how close our inferred dispersal process is to the true process by simulating datasets under the assumed model and posterior estimates, and then comparing them with the observed data. When the simulated data resemble the observed data closely, we consider that the assumed model provides an adequate fit to the data in an absolute sense (*i.e.*, not comparing to other competing (prior)models). PrioriTree allows users to perform posterior-predictive simulations using the output of BEAST discrete-geographic analyses, and then compute and visualize the test statistics to assess the model adequacy (Fig. 2.2, bottom panel). PrioriTree also provides functions to set up BEAST power-posterior analyses to estimate the marginal likelihood (Lartillot and Philippe 2006; Xie et al. 2011; Baele et al. 2012) for comparing the relative fit of competing (prior)models to the geographic data.

Additional Features of PrioriTree

PrioriTree assumes that the phylogeny and biogeographic history are inferred sequentially in the discrete-geographic phylodynamic analysis. Under this sequential-inference approach, the phylogeny of the study group has typically been estimated from an alignment of sequence data using BEAST. PrioriTree therefore requires users to provide an input file containing a previously inferred tree or distribution of trees. If the input file contains a posterior distribution of trees, PrioriTree allows users to specify how to marginalize over the the distribution to accommodate phylogenetic uncertainty in the discrete-geographic inference.

Users can also set up other BEAST discrete-geographic inferences (*e.g.*, inferring the number of dispersal events between each pair of geographic areas) in PrioriTree. Apart from visualizing the prior distributions, PrioriTree also displays the BEAST XML script and the methods description text in separate panels, which also change dynamically according to the input and specification. Finally, users can download the BEAST analysis script, the figures summarizing the analysis, and the associated text description—that can serve as a template for the methods section of their study—produced by PrioriTree.

AVAILABILITY AND IMPLEMENTATION

`PrioriTree` is open-source and freely available both as a web utility (https://jsigao.shinyapps.io/prioritree/) and as an `R` package. The source code for `PrioriTree` is available at https://github.com/jsigao/prioritree. A comprehensive user manual is presented as the supplementary material of this chapter, and actively maintained at https://bookdown.org/jsigao/prioritree_manual/.

SUPPLEMENTARY MATERIAL: PRIORITREE MANUAL

Overview

PrioriTree is an interactive web utility designed to help researchers specify input files for—and process output files from—analyses of biogeographic history performed using the BEAST software package (Drummond et al. 2012; Suchard et al. 2018). The discrete-geographic models implemented in BEAST (Lemey et al. 2009; Edwards et al. 2011) contain many parameters that must be inferred from minimal information (the single geographic area in which each pathogen occurs); inferences under this approach are therefore inherently sensitive to the assumed priors on the model parameters. We recently demonstrated that the priors implemented as the defaults in BEAST—and used in the vast majority of published studies—are strongly informative and extremely unrealistic; these misinformative priors distort inferences of biogeographic history (see Chapter 1 for details).

These considerations motivated our development of PrioriTree to help researchers: (1) interactively set up BEAST discrete-geographic phylodynamic analyses with visualized priors (see this section); (2) specify BEAST analyses and summarize the results for assessing the sensitivity of biogeographic inference to the specified priors (see this section), and; (3) specify BEAST analyses and summarize the results for assessing the adequacy of the specified geographic (prior) model in describing the dispersal process and comparing competing models (see this section).

In each section below, we start with a theoretical-background subsection to explain the functionalities provided by PrioriTree and the related theory, and then provide details regarding specific operations needed in PrioriTree to set up or summarize the corresponding BEAST analysis. Example input files can be found in this downloadable folder. More example files with real (larger) datasets can be found in the supplementary repository for Chapter 1.

Interactively Configuration of BEAST Discrete-Geographic Phylodynamic Analyses Using Visualized Priors

When we have no prior knowledge about the parameter, we may want to specify a generic prior distribution that is flexible enough to be updated by the data. PrioriTree provides multiple such distributions to serve as the candidate choice, where the default choice of each prior has been identified to perform well in many empirical analyses (Chapter 1). PrioriTree will display the distribution of the selected prior, with the associated prior mean and 95% credible interval listed alongside.

Alternatively, we may have some intuitions (*e.g.*, knowledge learned from previous analyses) about the parameters that we want to express in our priors. PrioriTree allows users to specify these biologically motivated priors in an interactive manner; it provides the flexibility to specify a range of (hyper)priors and dynamically renders the resulting prior distribution according to the specification in real time.

You can also configure other settings (*e.g.*, inferring number of dispersal events between each pair of geographic areas) of the BEAST analysis in PrioriTree as well. According to the input, the changes to the BEAST XML script and methods description are viewable on the fly. At the end, PrioriTree generates a readily runnable BEAST XML script (as well as the associated methods template) to perform the analysis that you conceive.

Below we first provide an in-depth introduction to discrete-geographic phylodynamic inferences, including the discrete-geographic model (see this subsection), the prior-sensitivity nature of biogeographic inference and the prior choices provided by PrioriTree (see this subsection), dispersal-history inference (see this subsection), the tree model (see this subsection), and MCMC (see this subsection). Go directly to the quickstart section for a short tutorial focusing on how to set up a BEAST discrete-geographic phylodynamic analysis in PrioriTree.

Theoretical Background and PrioriTree Introduction: Discrete-geographic model

The process of geographic dispersal over the tree, Ψ, is described as a continuous-time Markov chain (CTMC). For a biogeographic history with k discrete-geographic areas, this stochastic process is fully specified by a $k \times k$ instantaneous-rate matrix, $\mathbf{Q}$, where an element of the matrix, q_{ij}, specifies the instantaneous rate of change between state i and state j, *i.e.*, the instantaneous rate of dispersal from area i to area j. By convention, we rescale the $\mathbf{Q}$ matrix such that the

expected (average) number of dispersal events in one time unit is equal to the parameter μ: this is the average rate of dispersal among all k discrete geographic areas.

Bayesian stochastic search variable selection (BSSVS)

For most inference problems, the number of discrete geographic areas, k, is large, such that the discrete-geographic model includes many parameters, while the data are limited to a single observation (the geographic area occupied by each tip in the tree). Accordingly, inference under these discrete-geographic models raises concerns about our ability to estimate each parameter in the matrix. (For details, see the prior-sensitivity subsection below.) This concern motivated Lemey and colleagues (Lemey et al. 2009) to develop an approach to reduce the complexity of the discrete-geographic model called Bayesian stochastic search variable selection (BSSVS).

This approach involves specifying each element, q_{ij}, of the instantaneous-rate matrix, $\mathbf{Q}$, as:

$$q_{ij} = r_{ij}\delta_{ij},$$

where r_{ij} is the relative rate of dispersal between areas i and j, and δ_{ij} is an indicator variable that takes one of two states (0 or 1). When $\delta_{ij} = 1$, the instantaneous dispersal rate for the corresponding element, q_{ij}, is simply $q_{ij} = r_{ij}$.

Conversely, when $\delta_{ij} = 0$, the instantaneous dispersal rate for the corresponding element, q_{ij}, is zero, effectively removing that parameter from the discrete-geographic model. The idea here is to exclude superfluous elements of the $\mathbf{Q}$ matrix in order to reduce the number of parameters that must be inferred from the (inherently minimal amount of) data.

A given $\mathbf{Q}$ matrix therefore entails a vector of δ_{ij} (*i.e.*, $\boldsymbol{\delta}$) and a vector of r_{ij} (*i.e.*, $\boldsymbol{r}$). Each unique $\boldsymbol{\delta}$ vector—a string of zeros and ones for each of the possible pairwise dispersal routes between the k geographic areas—corresponds to a unique discrete-geographic model.

By default, `BEAST` uses BSSVS to average over discrete-geographic models with different degrees of complexity. If BSSVS is not toggled in `PrioriTree`, the dispersal-route indicator vector, $\boldsymbol{\delta}$, will be removed from the model and thus the rate matrix is simply $\boldsymbol{r}$.

Form of the instantaneous-rate matrix

Alternative discrete-geographic models may be specified based on the symmetry of the instantaneous-rate matrix. The discrete-geographic model described by Lemey et al. (2009) assumes that the rate matrix, $\mathbf{Q}$, is symmetric, where $q_{ij} = q_{ji}$ (*i.e.*, $r_{ij} = r_{ji}$ and $\delta_{ij} = \delta_{ji}$). Accordingly, this model assumes that the instantaneous rate of dispersal from area i to area j is

equal to the dispersal rate from area j to area i. For a dataset with k areas, the symmetric model has $\binom{k}{2}$ dispersal-route indicators and up to $\binom{k}{2}$ relative-rate parameters.

A subsequent extension (Edwards et al. 2011) allows the $\mathbf{Q}$ matrix to be asymmetric, where q_{ij} and q_{ji} are not constrained to be equal. Accordingly, this model allows the rate of dispersal from area i to area j to be different from the rate of dispersal from area j to area i. For a dataset with k areas, the asymmetric model has $k \times (k-1)$ dispersal-route indicators and up to $k \times (k-1)$ relative-rate parameters.

Inherent Prior Sensitivity of Biogeographic Inference

We estimate the parameters of our discrete-geographic models—including the vector of instantaneous dispersal rates between each pair of areas, $\mathbf{Q}$, and the average dispersal rate across all geographic areas, μ—from our observations (including the geographic data, G, and the previously inferred phylogeny, Ψ). Specifically, we estimate the joint posterior probability distribution of our discrete-geographic model parameters conditional on the data using Bayes theorem:

$$\overbrace{P(\mathbf{Q}, \mu \mid G, \Psi)}^{\text{posterior distribution}} = \frac{\overbrace{P(G \mid \mathbf{Q}, \mu, \Psi)}^{\text{likelihood}} \overbrace{P(\mathbf{Q})P(\mu)}^{\text{prior distribution}}}{\underbrace{P(G \mid \Psi)}_{\text{marginal likelihood}}}.$$

The joint posterior probability distribution, $P(\mathbf{Q}, \mu \mid G, \Psi)$, reflects our beliefs about the parameter values after evaluating our data. The posterior is an updated version of our joint prior probability distribution, $P(\mathbf{Q})P(\mu)$, which reflects our beliefs about the parameter values before evaluating our data. Our prior is updated by the information in our data via the likelihood function, $P(G \mid \mathbf{Q}, \mu, \Psi)$, which is the probability of observing our geographic data under the discrete-geographic model.

In many cases, Bayesian inference is robust to the choice of prior: posterior estimates are dominated by the information in the data, allowing us to safely ignore the issue of prior choice. In other cases, however, posterior estimates will be strongly influenced by our choice of prior: specifically, when our data contain limited information about a parameter in our inference model, the posterior probability distribution inferred for that parameter will closely resemble the assumed prior probability distribution. This phenomenon—referred to as prior sensitivity—is an inherent feature of inference under discrete-geographic models.

To illustrate this issue, contrast typical inferences under discrete-geographic models and

substitution models. Both models describe the evolution of discrete states (geographic areas or nucleotide bases) from the root, along the branches, to the tips of our study phylogeny as a continuous-time Markov chain (CTMC). For a process with k discrete states, the process is completely described by a $k \times k$ matrix of instantaneous rates, $\mathbf{Q}$. Each element of this matrix, q_{ij}, describes the instantaneous rate of change between two states, i and j. The most complex time-reversible substitution model, the GTR model, has $k = 4$ states ($\{A, C, G, T\}$), and six instantaneous-rate parameters that are typically inferred from a sequence alignment with hundreds or thousands of sites.

By contrast, discrete-geographic models typically have many more parameters and much less data. For example, an inference problem with $k = 10$ discrete geographic areas, the symmetric discrete-geographic model has $\binom{k}{2} = 45$ instantaneous-rate parameters, and the asymmetric discrete-geographic model has $k \times (k - 1) = 90$ instantaneous-rate parameters. Moreover, the parameters of these discrete-geographic models must be estimated from an 'alignment' with a single 'site'; *i.e.*, the geographic dataset includes a single observation (the area occupied by each tip in the tree).

There are several strategies to deal with prior sensitivity:

- specify biologically-motivated/informed priors (*i.e.*, where the prior probability is focussed on 'reasonable' values);
- specify diffuse/uninformative priors (*i.e.*, where the prior probability is spread evenly over a wide range of 'plausible' values), or;
- assess prior sensitivity (*e.g.*, perform replicate inferences under a range of candidate priors, and/or assess the relative or absolute fit of alternative prior models to our data).

Unfortunately, the priors on discrete-geographic model parameters implemented as the defaults in `BEAST` are highly (mis)informative: *i.e.*, the default priors reflect extremely strong and biologically unrealistic assumptions about the underlying dispersal process (Chapter 1). Worse still, these default priors have been used in the vast majority ($\sim$ 93%) of published studies that have used `BEAST` to infer biogeographic history, and these default priors have been shown to strongly (and adversely) distort central conclusions of biogeographic studies (Chapter 1).

Prior on the number of dispersal routes

Recall that—when using BSSVS in BEAST—each element of the instantaneous-rate matrix, **Q**, is specified as:

$$q_{ij} = r_{ij}\delta_{ij},$$

where r_{ij} is the relative rate of dispersal between areas i and j, and δ_{ij} is an indicator variable that takes one of two states (0 or 1). Each vector, $\boldsymbol{\delta}$, specifies a unique configuration of dispersal routes, which corresponds to a unique discrete-geographic model. The total number of dispersal routes for a given discrete-geographic model is denoted Δ. For a given value of Δ, there may be multiple distinct discrete-geographic models. For example, a dataset with $k = 3$ geographic areas has the vector of relative rates $\boldsymbol{r} = \{r_{12}, r_{13}, r_{23}\}$, such the space of symmetric discrete-geographic models includes: three models with $\Delta = 1$ dispersal route, $\boldsymbol{\delta} = \{(1,0,0),(0,1,0),(0,0,1)\}$; three models with $\Delta = 2$ dispersal routes, $\boldsymbol{\delta} = \{(1,1,0),(0,1,1),(1,0,1)\}$, and; a single model with $\Delta = 3$ dispersal routes, $\boldsymbol{\delta} = \{(1,1,1)\}$. Lemey and colleagues (Lemey et al. 2009) impose a prior on the space of discrete-geographic models by:

1. placing a prior on the total number of dispersal routes, Δ, and;
2. assuming that all discrete-geographic models with a given value of Δ are equiprobable.

Together, these assumptions induce a prior probability that a given dispersal route between areas i and j exists, *i.e.*, the probability that $\delta_{ij} = 1$.

PrioriTree allows you to specify three alternative prior probability distributions on the number of dispersal routes, Δ.

Poisson prior

The default prior on Δ implemented in BEAST is a Poisson probability distribution. The specific parameterization of the Poisson prior depends on the a/symmetry of the discrete-geographic model. For symmetric discrete-geographic models, the default prior on Δ is an offset Poisson distribution. Specifically, for a dataset with k discrete geographic areas, the Poisson prior on Δ is offset by $(k-1)$, where all discrete-geographic models for which $\Delta < (k-1)$ have zero prior probability (*i.e.*, such models are disallowed a priori). The motivation for this offset is mathematical: a dataset with k geographic areas cannot be realized under a CTMC with fewer than $(k-1)$ non-zero q_{ij} values (*i.e.*, dispersal routes). [The real constraint on the geographic

model is that it must be irreducible; *i.e.*, it must be possible to reach each area from every other area either directly or indirectly. A model with fewer than $(k-1)$ dispersal routes cannot be irreducible; however, a model with at least $(k-1)$ dispersal routes is not guaranteed to be irreducible.] The prior probability of models for which $\Delta \geq (k-1)$ is Poisson with rate parameter, $\lambda = ln(2)$. As specified, this Poisson prior places approximately 50 percent of the prior probability on discrete-geographic models with the absolute minimum number of dispersal routes.

For asymmetric discrete-geographic models, `BEAST` specifies a default prior on Δ with rate parameter $\lambda = (k-1)$. (Note that this prior does not enforce a minimum number of dispersal routes, *i.e.*, the Poisson prior distribution under the asymmetric discrete-geographic model is not offset.)

The default Poisson prior on Δ in `BEAST` reflects a very strong preference (*i.e.*, a very informative prior) for discrete-geographic models with the minimal number of dispersal routes. When the default Poisson prior on Δ is specified using `BEAUti`, it is not possible for users to adjust the parameterization (*i.e.*, to make the prior less informative by changing the value of the Poisson-rate parameter, λ). By contrast, when the Poisson prior on Δ is specified using `PrioriTree`, users are able to adjust the parameterization. By default, `PrioriTree` specifies a more diffuse Poisson prior on Δ, where the expected number of dispersal routes is about half the maximum number. [As in `BEAST`, the Poisson prior for symmetric discrete-geographic models is also offset by $(k-1)$.] `PrioriTree` allows users to adjust the Poisson-rate parameter, λ, to specify the desired shape for the Poisson prior on Δ. Both the mean and the variance of the Poisson prior distribution scale linearly with λ.

Beta-Binomial prior

We can specify an alternative prior on the number of dispersal routes, Δ, by treating each dispersal-route indicator, δ_{ij}, as a Bernoulli random variable, such that the total number of dispersal routes follows a Binomial distribution. The Binomial probability distribution has two parameters: n is the number of trials (equal to the maximum number of dispersal routes for a dataset with k areas), and p is the success probability (equal to the probability that each dispersal route exists; *i.e.*, that $\delta_{ij} = 1$). We treat p as a random variable to be estimated from the data. Specifically, we specify a Beta prior probability distribution on p. The Beta prior has two hyperparamters—the shape parameters α and β—that can be interactively modified by the user

so that the resulting Beta-Binomial prior on Δ has the desired mean and 95% prior interval. The expected (mean) number of dispersal routes increases as α/β increases.

Uniform prior

Finally, we can specify an alternative prior on the number of dispersal routes, Δ, by assuming that all possible values (*i.e.*, between zero and the maximum number of dispersal routes) are equiprobable. This uniform prior on Δ is a special case of the Beta-Binomial distribution described above, which is specified when the values of both shape parameters of the Beta prior—α and β—are set to one. Under the uniform prior on Δ, the prior probability that each dispersal route exists is uniformly distributed between 0 and 1.

Prior on the average dispersal rate

Recall that the rate matrix, $\mathbf{Q}$, is rescaled such that the average rate of dispersal between all areas is μ. For a tree of length T (*i.e.*, the sum of all branch durations in the dated phylogeny), the expected number of dispersal events is $\mu \times T$. Therefore, the prior on the average dispersal rate, μ, represents our prior belief about the number of dispersal events over the tree.

`PrioriTree` allows you to specify three alternative prior probability distributions on the average dispersal rate, μ.

CTMC-rate reference prior

The default prior on μ implemented in `BEAST` is a Gamma probability distribution with shape parameter $\alpha = 0.5$ and rate parameter $\beta = T$. (Note that this Gamma prior is labelled as a 'CTMC-rate reference prior' in the `BEAST` utility, `BEAUTi`.) The Gamma distribution has a mean of α/β; therefore, this prior expresses the belief that the average rate of dispersal is $0.5/T$.

The default Gamma prior on μ in `BEAST` reflects a very strong preference (*i.e.*, a very informative prior) for biogeographic histories with an implausibly small number of dispersal events. A dataset with k geographic areas requires a biogeographic history with at least $(k-1)$ dispersal events. The expected number of dispersal events is $\mu \times T$. Accordingly, the number of dispersal events expected a priori under the default Gamma prior on μ is 0.5, independent of the duration of the entire biogeographic history (*i.e.*, the tree length, T), or the number of areas, k, involved in the geographic history. Similarly, the prior distribution on the number of dispersal events is independent of T and k: the 95% prior interval is $[0, 3]$ dispersal events, which implies that we would be very surprised if a biogeographic history of any duration with

any number of areas involved more than three dispersal events.

Hierarchical-exponential prior

`PrioriTree` allows users to specify an exponential prior on the average dispersal rate, μ; this exponential prior has a single hyperparameter, the exponential-rate parameter, λ. The mean of this exponential prior is $1/\lambda$, which we treat as a random variable to be estimated from the data. Specifically, we specify a Gamma hyperprior on λ. The Gamma distribution has two parameters—α (shape) and β (rate)—where the mean of the Gamma hyperprior is α/β and the variance is α/β^2. This Gamma hyperprior on the exponential-rate parameter, λ, is constrained such that $\alpha = \beta$: this constraint ensures that the resulting hierarchical-exponential prior is proper (*i.e.*, that it integrates to one, and so obeys the law of total probability). This hierarchical prior distribution is known as the K distribution (Jakeman and Pusey 1978).

Under this prior on the average dispersal rate, the prior distribution on the number of dispersal events (sensibly) scales with T; *i.e.*, the expected number of dispersal events increases with the duration of the biogeographic history. `PrioriTree` allows users to simultaneously modify the shape/rate parameter. Note that the mean of this hierarchical prior is always one (*i.e.*, independent of the shape/rate parameter), but its variance scales inversely with the shape/rate parameter.

Empirical-exponential prior

Finally, we can specify an alternative exponential prior on the average dispersal rate, μ, that adopts an 'empirical-Bayesian' approach for specifying the exponential-rate parameter, λ. In Bayesian inference, we specify a prior distribution for a given parameter that reflects our beliefs about its parameter values before we evaluate our study data. Empirical-Bayesian inference, by contrast, effectively entails some 'double dipping' of the study data; that is, we first estimate the (hyper)parameters of our inference model from our study data, and then use those (hyper)parameter estimates to specify one or more of the corresponding (hyper)priors of our inference model.

Our empirical-Bayesian approach for specifying the prior on the average dispersal rate, μ, involves computing the parsimony score (the minimum number of dispersal events) required to explain our observed geographic data (*i.e.*, the distribution of areas across the tips of our study tree). This parsimony score represents a minimum bound on the true number of dis-

persal events in the biogeographic history that gave rise to our observations. We can therefore leverage these parsimony scores to inform our prior on μ. Specifically, we specify a value for the exponential-rate parameter, λ, such that the resulting prior on the number of dispersal events is focused on values greater than the parsimony score.

By default, `PrioriTree` sets the parsimony score at the lower quartile (*i.e.*, 25% quantile) of the prior distribution on the number of dispersal events. You can adjust the mean of this resulting prior distribution (while using the parsimony score as a reference) according to their biological beliefs about the dispersal intensity.

Theoretical Background and `PrioriTree` Introduction: Inferring Biogeographic History

Empirical biogeographic studies often report summaries that are based on the conditional probability distribution of biogeographic histories over the tree. The distribution of histories depends on—*i.e.*, is conditioned on—the instantaneous-rate matrix, $\mathbf{Q}$, the biogeographic data, G, and the phylogeny, Ψ. Conceptually, for a given tree and rate matrix, we imagine simulating a geographic history over the tree from the root to its tips, where the rate matrix specifies the waiting times between dispersal events. We can construct the conditional distribution of biogeographic histories by simulating a large number of individual histories, and retaining only those histories that realize the observed geographic areas at the tips, G. This conditional distribution contains all of the information required to compute two commonly reported summaries: the ancestral areas at internal nodes of the tree, and the number of dispersal events between geographic areas.

We can infer the total number of dispersal events among all k discrete geographic areas, and/or we can infer the number of dispersal events between each pair of geographic areas. `BEAST` implements two algorithms for computing the number of dispersal events: the first option—referred to as "fast stochastic mapping" in `PrioriTree`—relies on analytical integration (Minin and Suchard 2008a,b; O'Brien et al. 2009). As the name suggests, this is a more computationally efficient option for inferring biogeographic histories; however, this method only computes the expected (average) number of dispersal events on each branch, but does not allow us to infer all the details of the biogeographic history (such as the exact timing of dispersal events over the tree).

The second option for inferring biogeographic histories—referred to as "stochastic map-

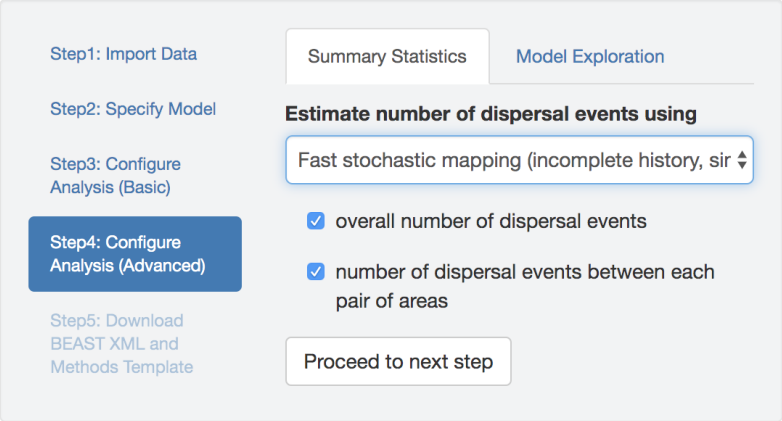

Figure S.2.1: Options for inferring biogeographic history. You can specify either the integration-based or simulation-based methods for stochastic mapping of biogeographic histories, and for either method, you can choose to estimate the total and/or pairwise number of dispersal events.

ping" in `PrioriTree`—relies on a simulation-based algorithm (Nielsen 2002; Rodrigue et al. 2007; Hobolth and Stone 2009). This method for mapping biogeographic histories is more computationally intensive, but also allows us to infer additional details (such as the exact timing of dispersal events over the tree).

When inferring biogeographic histories with either stochastic-mapping option, `PrioriTree` allows you to specify the type of dispersal events—*i.e.*, the total number of dispersal events among all areas and/or the pairwise number of dispersal events between each pair of areas—in your `BEAST` analysis (Fig. S.2.1). If you do not specify either type of dispersal event (total or pairwise number) and you are using the "fast stochastic mapping" algorithm (the first option), `PrioriTree` will not infer biogeographic histories (*i.e.*, it will not write the part of XML script that instructs `BEAST` to perform computations for the expected number of dispersal events). Conversely, if you do not specify either type of dispersal event (total or pairwise number) and you are using the "stochastic mapping" algorithm (the second option), `PrioriTree` will still write the part of the XML script that instructs `BEAST` to infer the full biogeographic history (*i.e.*, the number of dispersal events will not be written to the parameter log file, but they can still be retrieved from the tree log file).

Theoretical Background and `PrioriTree` Introduction: Accommodating Phylogenetic Uncertainty

The discrete-geographic model describes the process of dispersal over the phylogeny of our study group. Typically, the phylogeny and biogeographic history are inferred sequentially. Under this sequential-inference scenario, for example, we might first estimate the phylogeny for our study group from an alignment of sequence data using `BEAST`. The resulting phylogenetic estimate is then used to infer the biogeographic history of our study group.

`PrioriTree` requires that you provide an input file containing a previously inferred study tree. If the input file includes a single tree (*e.g.*, an MCC summary tree in a `.tre` file), the discrete-geographic inference will treat the phylogeny as a fixed variable (*i.e.*, this effectively assumes that the phylogeny is 'known'). Alternatively, if the input file contains multiple trees (*e.g.*, a posterior distribution of trees in a `.trees` file), the discrete-geographic inference will be marginalized (*i.e.*, 'averaged') over those trees to accommodate phylogenetic uncertainty.

`BEAST` provides two options for accommodating phylogenetic uncertainty in sequential analyses. The first option—evoked with the argument `MetropolisHastings = "true"`—treats the posterior sample of trees as a prior distribution while inferring the joint posterior distribution of the discrete-geographic model parameters. Under this option, trees are proposed and accepted/rejected using the standard Metropolis–Hastings MCMC algorithm. Specifically, the MCMC proposes a move to a new tree at a frequency specified by the corresponding proposal weight, and when a new tree is being proposed, it is randomly drawn from the posterior sample of trees, and the proposed tree is then accepted or rejected according to the computed accep-

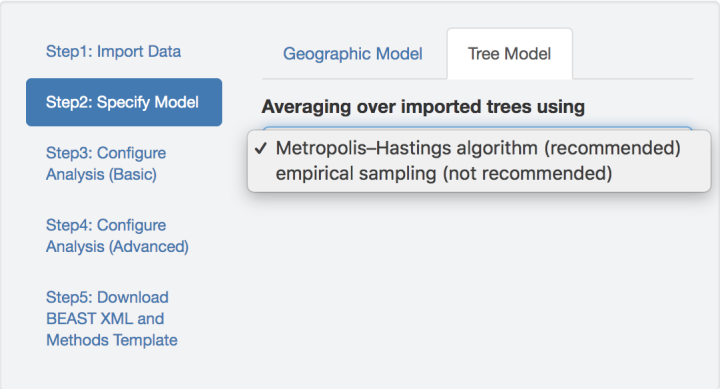

Figure S.2.2: Specify the tree model.

tance probability. Accordingly, sequential inference under this option is (theoretically) equivalent to performing joint inference of the phylogeny and biogeographic history (*i.e.*, where the phylogeny and biogeographic history are simultaneously inferred from a single, combined dataset that includes both the sequence and geographic data).

The second option for accommodating phylogenetic uncertainty—evoked with the argument `MetropolisHastings = "false"`—averages over the posterior sample of trees using an ad hoc MCMC algorithm. Similar to the previous option, the MCMC proposes a move to a new tree randomly drawn from the posterior sample of trees. In contrast to the correct Metropolis–Hastings algorithm, however, proposed trees are always accepted (*i.e.*, disregarding the acceptance probability of the proposed tree). This procedure therefore ignores the fact that our geographic data (*i.e.*, the geographic area for each tip in the tree) will have different probabilities of being observed for different trees. Accordingly, sequential inference under this option effectively assumes that the probability of observing the geographic data is independent of the underlying phylogeny. We caution that this approach for averaging biogeographic inferences over a posterior sample of trees—which is the default option implemented in `BEAST`—will not correctly sample the joint posterior probability distribution of the phylogeny and biogeographic history.

Theoretical Background and `PrioriTree` Introduction: MCMC

The Metropolis–Hastings algorithm

Recall that Bayesian inference is focused on the joint posterior probability distribution of model parameters. The posterior probability cannot be solved analytically, so we must resort to numerical methods to approximate the joint posterior probability. Here we briefly describe the numerical method used to estimate the joint posterior probability distribution of discrete-geographic model parameters in `BEAST`: the Metropolis–Hastings Markov chain Monte Carlo (MCMC) algorithm.

The Metropolis–Hastings algorithm (Metropolis et al. 1953; Hastings 1970) entails simulating a Markov chain that has a stationary distribution that is the joint posterior probability distribution of the discrete-geographic model parameters. The 'state' of the chain, θ, is a fully specified model, *i.e.*, a specific phylogeny with divergence times, Ψ, and a specific set of values for each parameter of the discrete-geographic model, $\{\boldsymbol{r}, \delta, \mu\}$. The Metropolis–Hastings

MCMC algorithm involves six main steps:

1. Initialize the chain with specific values for all parameters, $\theta = \{\Psi, \boldsymbol{r}, \boldsymbol{\delta}, \mu\}$. The initial parameter values might be specified arbitrarily, or might be drawn from the corresponding prior probability distribution for each parameter.

2. Select a single parameter according to its proposal weight. For example, if we assigned a proposal weight of 10 to the average dispersal rate parameter, μ, and assigned a total proposal weight of 100 to all parameters, then the probability of selecting μ is $10 \div 100 = 0.1$.

3. Propose a new value for the selected parameter. Each parameter in the model will have one or more stochastic proposal mechanisms: in general, a proposal mechanism is simply a probability distribution that is centered on the current parameter value, from which we randomly draw a new parameter value. By changing one of the parameter values, we have proposed a new possible state of the chain, θ'.

4. Calculate the probability of accepting the proposed change, R:

$$R = \min\left[1,\ \underbrace{\frac{f(G \mid \theta')}{f(G \mid \theta)}}_{\text{likelihood ratio}} \cdot \underbrace{\frac{f(\theta')}{f(\theta)}}_{\text{prior ratio}} \cdot \underbrace{\frac{f(\theta \mid \theta')}{f(\theta' \mid \theta)}}_{\text{proposal ratio}}\right]$$

   The acceptance probability, R, is the lesser of two values: *i.e.*, it is either equal to one or the product of three ratios:

   - Likelihood ratio: the likelihood ratio is simply the probability of our observed data given the proposed state of the chain, θ', divided by the probability of our observed data given the current state of the chain, θ. We calculate the likelihood for any given parameterization of the discrete-geographic model (*i.e.*, either θ' or θ) using the Felsenstein pruning algorithm.
   - Prior ratio: is simply the prior probability of the proposed state, θ', divided by the prior probability of the current state, θ. In Bayesian inference, each parameter is a random variable, and so is described by a prior probability density. Accordingly, we can simply 'look up' the prior probability of any specific parameter value.
   - Proposal ratio: the proposal (aka Hastings) ratio ensures that Markov chain is ergodic—*i.e.*, that the probability of proposing a move from state θ to θ' is equal

to the probability of proposing a move from state θ' to θ—which ensures that the samples provide a valid approximation of the target (joint posterior probability) distribution.

5. Generate a uniform random number between zero and one, $U[0,1]$. If $U < R$, accept the proposed change as the next state of the chain (*i.e.*, $\theta' \rightarrow \theta$); otherwise, the current state of the chain becomes the next state of the chain (*i.e.*, $\theta \rightarrow \theta$).

6. Repeat steps 2–5 an 'adequate' number of times.

A chain following the simple rules outlined above will sample parameter values with a frequency that is proportional to their posterior probability. That is, the proportion of time that the chain spends in any particular state is a valid approximation of the posterior probability of that state. To help understand why this is true, let's have a closer look at how we compute the acceptance probability, R. To this end, we will ignore the third term, the proposal ratio (as it only ensures that acceptance probabilities are based on the product of the first two terms); the simplified equation

$$R \propto \left[\frac{f(G \mid \theta') \cdot f(\theta')}{f(G \mid \theta) \cdot f(\theta)} \right] = \underbrace{\frac{f(\theta' \mid G)}{f(\theta \mid G)}}_{\text{posterior ratio}}$$

makes it clear that the MCMC simulation will visit states (parameter values) proportional to their relative posterior probability. [Recall that the posterior probability, $P(\theta \mid G)$, is proportional to the product of the likelihood function, $P(G \mid \theta)$, and the prior probability, $P(\theta)$.]

At specified intervals, the state of the chain is written to a log file (the `.trees` or `.log` files that are output by `BEAST`). Each row of parameter values in these log files represents a sample of the joint posterior probability distribution. We can query the joint posterior sample to make inferences on any parameter of interest: *e.g.*, we might infer the marginal posterior probability density for the average dispersal rate parameter, μ, by constructing a histogram (frequency distribution) of sampled values from the corresponding column in our log file.

`PrioriTree` allows you to specify a number of settings to control the MCMC simulation, which are grouped under two panels: `MCMC Sampling` and `Proposal Weights`.

MCMC sampling

MCMC simulation length

Given that we are numerically approximating the joint posterior probability distribution by collecting MCMC samples, a key issue is how long we need to run the MCMC simulation to adequately approximate the posterior probability. The length of the MCMC simulation refers to the total number of MCMC cycles or generations, where a cycle in `BEAST` entails a single proposal to a single parameter (*i.e.*, one iteration through steps 2–5 in the M–H algorithm, described above). A number of factors will impact the length of the simulation required to adequately approximate the joint posterior probability distribution, including dataset size, model complexity (including the number of parameters and the degree/nature of interactions among parameters), the specified (hyper)priors, and the efficiency of the MCMC simulation (including the choice of proposal mechanisms, the tuning parameters of those proposals, the weights assigned to each proposal mechanism, and the specified sampling frequency). Accordingly, it is impossible to know the required length of the MCMC simulation a priori; instead we must determine the minimal simulation length experimentally by iteratively running an MCMC simulation, diagnosing MCMC performance, adjusting MCMC settings, and re-running the MCMC simulation.

MCMC sampling frequency

The samples collected during an MCMC simulation are highly autocorrelated. That is, two successive states of the MCMC simulation differ at most by a single parameter value (if the proposed state was accepted) or will be identical (if the proposed state was rejected). Accordingly, we commonly 'thin' the samples collected by the MCMC simulation by writing the parameters to a log file at a specified frequency; this is referred to as the MCMC sampling frequency. This convention is largely practical, as it reduces the size of the log files. The number of MCMC samples written to our log files is therefore equal to the chain length divided by the sampling frequency. In setting the sampling frequency, we are trying to strike a balance between collecting as many samples as possible for a given chain length (sample at a high frequency), while simultaneously reducing both the file size and the degree of autocorrelation among sampled parameter values (sample at a low frequency). Again, determining the optimal sampling frequency entails a trial-and-error approach, which is typically facilitated by computing MCMC diagnostics, such as the effective sample size (ESS), described below.

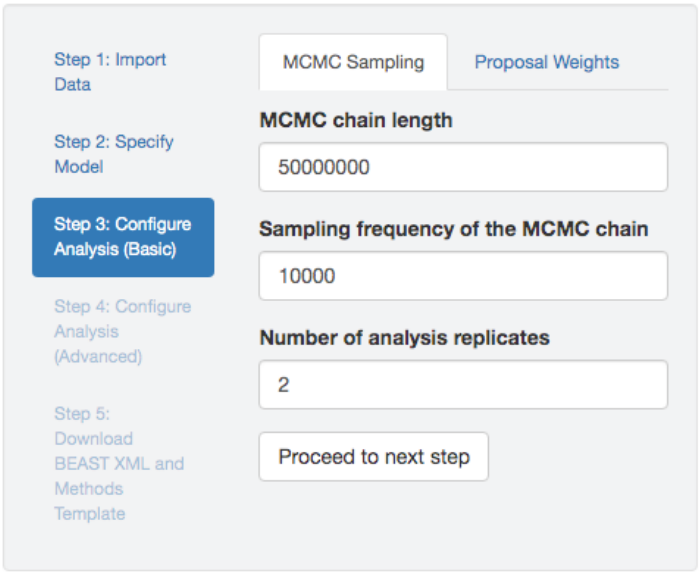

Figure S.2.3: MCMC settings. You can specify the simulation length, sampling frequency, and number of replicate simulations.

Number of replicate MCMC simulations

Because we are using numerical methods to approximate the joint posterior probability distribution, it is critical to assess the performance of our MCMC simulations. Many diagnostics have been developed to assess MCMC performance, but the most powerful rely on comparing aspects of replicate MCMC simulations. Two or more replicate MCMC simulations are identical—with identical data, model, and priors—except for the random-number seed. Our ability to detect MCMC pathologies increases with the number of replicate simulations that we perform: as a rule of thumb, we recommend performing (at least) four replicate MCMC simulations. However, as with most aspects of MCMC simulation, the actual number of replicate simulations required to rigorously diagnose MCMC performance for a given inference problem (model and dataset) is determined by trial and error. Note replicate MCMC simulations are useful beyond helping us diagnose MCMC performance; that is, we can combine the post-burnin samples from our replicate MCMC simulations to construct a 'composite' posterior sample (*i.e.*, a composite log file), and we can use this composite posterior sample as the basis for our parameter estimates.

Proposal weights

Before discussing proposal weights, it may be helpful to first describe the goal that we are trying to achieve by adjusting proposal weights, and the diagnostics we use to assess our proximity to that goal. As described above, the MCMC simulation length is the total number of MCMC cycles (or generations), the sampling frequency is interval (in number of cycles) at which we write the state of the chain to our log files, such that the number of samples in our log file is the chain length divided by the sampling frequency. Imagine, for example, that we run an MCMC simulation for a length of 1,000,000 cycles with a sampling frequency of 100. We will have sampled 10,000 parameter values, however, the autocorrelation of MCMC states means that we have fewer than 10,000 independent samples.

To estimate the number of effectively independent samples in our log file, we need to compute the effective sample size (ESS) diagnostic. This summary statistic, in turn, relies on a second MCMC summary statistic, the autocorrelation time (ACT) diagnostic. The ACT statistic indicates the number of successive MCMC samples over which the values for a given parameter are correlated. Accordingly, the ESS for a given parameter is simply the total number of MCMC samples divided by the ACT for that parameter. For example, imagine that the average dispersal rate parameter, μ, in our hypothetical MCMC scenario has an autocorrelation time of 100, then the ESS for this parameter is $10,000 \div 100 = 100$.

The objective is to achieve a sufficiently large ESS value for every (hyper)parameter in our model. The threshold for the minimal ESS value is somewhat arbitrary; by convention, an ESS $\geq$ 200 is considered adequate. [As an aside, it may be useful to augment this arbitrary ESS threshold by visually inspecting the distribution of sampled parameter values. Our objective is to collect an adequate number of samples for a given parameter in order to estimate its marginal posterior probability distribution. Accordingly, if we have collected an adequate number of samples for a given parameter, a histogram of those samples (*e.g.*, plotted in Tracer) should resemble a probability distribution.]

We might be tempted to focus largely/exclusively on the ESS values of the 'focal' parameters of our model (while ignoring those for the 'nuisance' parameters in our model). However, this would be unwise. Our assignment of a parameter to focal or nuisance status is entirely subjective: we are free to make inferences about (focus on) any parameter in our model by summarizing the samples in the corresponding column of our log file. Nevertheless, the re-

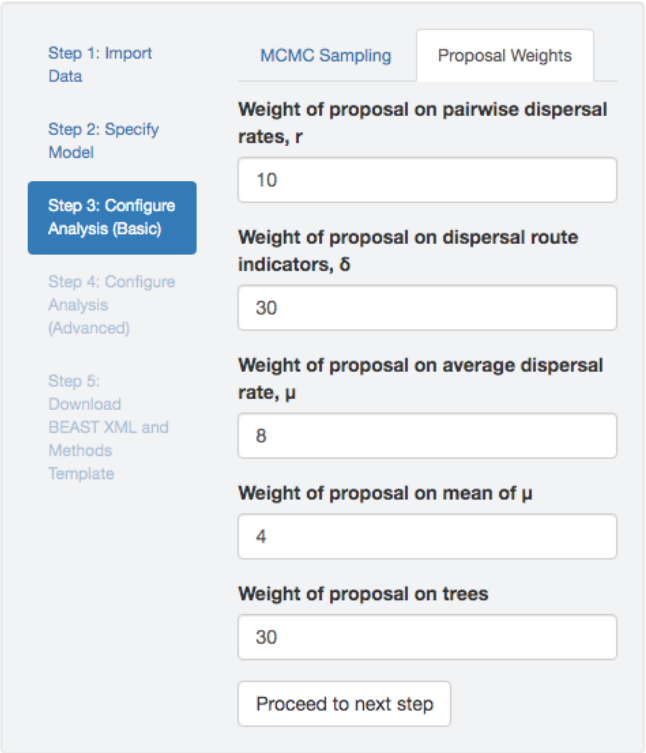

Figure S.2.4: Specify proposal weights.

liability of these marginal posterior probability distributions (based on a column of our log file) depends on an adequate approximation of the corresponding joint posterior probability distribution (all of the columns of our log file). In other words, all of the parameters of our model collectively (jointly) describe the process that gave rise to our observations, so a reliable estimate of any parameter requires that we adequately approximate all model parameters.

Virtually any MCMC simulation—if run long enough—will eventually achieve adequate ESS values for all identifiable parameters. It is common for ESS values to vary substantially across parameters; *i.e.*, where the ESS values for most parameters are extremely large by the time we achieve a minimal ESS value for one or two 'straggler' parameters. The objective is to achieve adequate ESS values for all parameters from the shortest possible MCMC simulation; this requires that the ESS values for all parameters are approximately equal (*i.e.*, such that all parameters reach adequate ESS values at the same point in the MCMC simulation). We can control the uniformity of the ESS values—and thereby optimize the efficiency of our MCMC simulation—by adjusting the proposal weights.

The proposal weights in `BEAST` are relative. For example, an MCMC simulation with two

proposal mechanisms—α with proposal weight 1, and β with proposal weight 2—is identical to the setting with proposal weights $\alpha = 10$ and $\beta = 20$ (*i.e.*, under both settings, the probability that we propose a change to parameter α is $1 \div (1 + 2) = 0.33 \equiv 10 \div (10 + 20) = 0.33$). It is common for the ESS values of some parameters to increase more slowly than others. These difficult parameters have longer autocorrelation times (ACT), owing to low acceptance rates for proposed changes to these parameters and/or because of correlations involving these parameters. The general idea is to increase the proposal weight for these difficult parameters so that their ESS values increase at approximately the same rate as those of other parameters.

Our experience with a given model and proposal mechanisms may inform our choice of initial proposal weights. (In fact, our experience with discrete-geographic models and proposals informed our choice of default proposal weights specified in `PrioriTree`, Fig. S.2.4) However, the optimal proposal weights will vary from analysis to analysis; therefore identifying the optimal set of proposal weights is yet again a trial-and-error process.

We suggest the following iterative procedure for optimizing proposal weights (and MCMC efficiency). First, perform a relatively short, preliminary ('shakedown') MCMC simulation using the default `PrioriTree` proposal weights. Next, examine the resulting ESS values for all of the parameters. If the ESS values are strongly uneven—*i.e.*, where most parameter values have similar ESS values, but one or a few have very low ESS values—increase the proposal weights for the parameters with relatively low ESS values (and/or decrease the proposal weights for the parameters with relatively high ESS values). Iterate this process—run a shakedown MCMC simulation, assess ESS values, and adjust proposal weights—until the ESS values for all parameters are approximately uniform. Then set up your final MCMC simulations using the optimized proposal weights, running each replicate MCMC simulation until the ESS value for each parameter is ≥ 200.

A final note about the size of log files. If the length of the MCMC simulation required to achieve adequate ESS values for all parameters results in MCMC log files becoming prohibitively large, you may wish to decrease the sampling frequency.

Quickstart: Set up a BEAST Discrete-Geographic Phylodynamic Analysis

Here, we walk through how to use PrioriTree to set up a basic BEAST discrete-geographic phylodynamic analysis with a focus on prior specification. This basic analysis infers the discrete-geographic model parameters and the ancestral areas at internal nodes of the phylogeny. The functionality described in this section can be found in the Analysis Setup main panel of the program. See the theoretical-background subsections above for more detailed explanations of how to specify alternative models and priors, and how to set up further analyses (*e.g.*, inferring the number of pathogen dispersal events between epidemic areas).

Step 1: Prepare input

The first step is to prepare your data for input into PrioriTree. PrioriTree requires two input files: one that contains the discrete-geographic data, and another that contains the phylogeny (either a single summary tree or distribution of trees inferred from a previous analysis). Note that other panels (including the Methods Template Viewer panel and the BEAST XML Viewer panel) will be enabled once both of the input files are uploaded (and pass the validity checks).

Discrete-geographic data File

The geographic data file needs to be either a .csv or .tsv file which contains two (or more) columns. The header (first row) of the file contains the names of the columns. By default, PrioriTree assumes that the first column contains the taxon names, and the second column contains the geographic area that each taxon was sampled from. If the columns in your data file

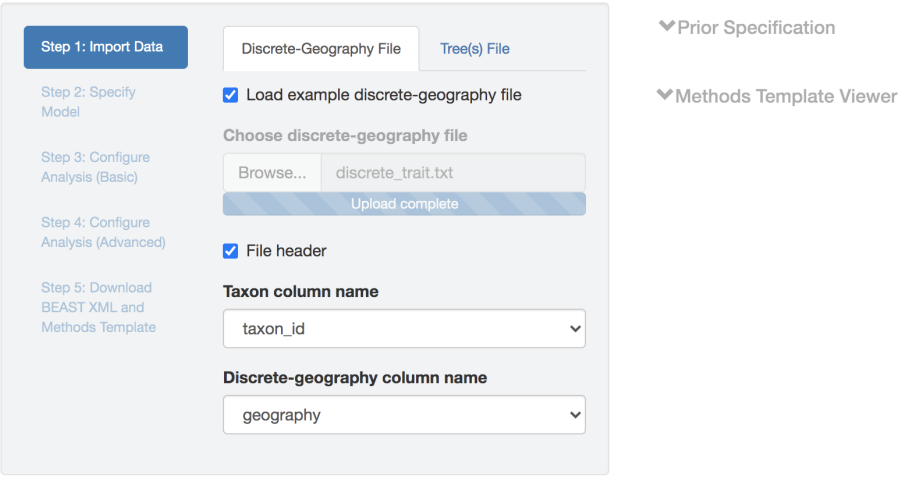

Figure S.2.5: Import data panel (discrete-geographic data file).

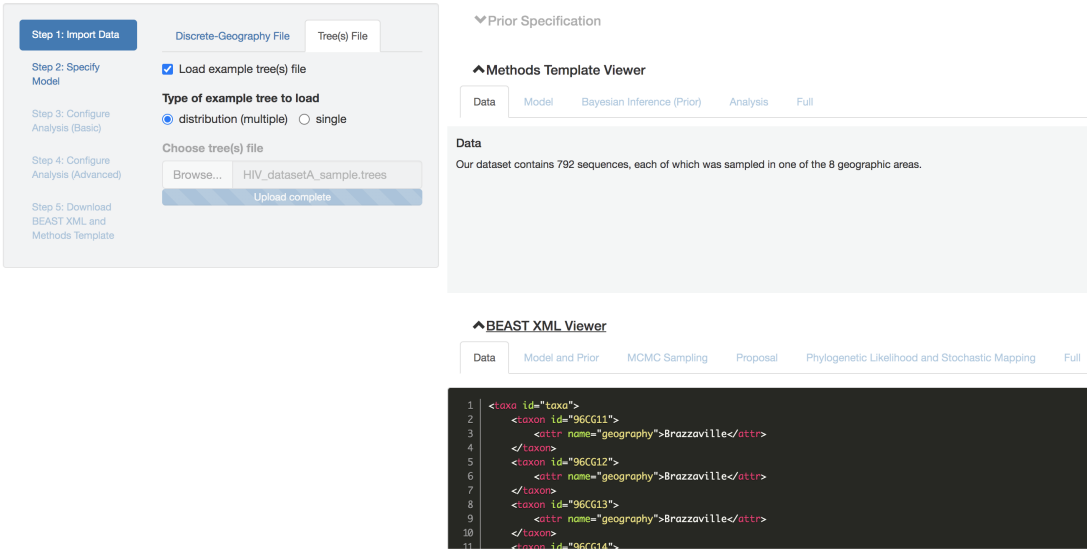

Figure S.2.6: Import data panel (tree file).

are in a different order, you can select the columns containing the taxon name and geographic data from the drop-down menu after uploading the geographic-data file (other columns will then be ignored).

To help you get a quick sense of the program, an example geographic-data file is available with the program; load it by checking the `Load example discrete-geography file` box. The example geographic data file is also available here in the supplementary repository for Chapter 1).

Tree file

The second input file contains the phylogeny to condition on (for a single tree) or marginalize over (a distribution of trees) in the discrete-geographic phylodynamic analysis. This phylogeny is supposed to be inferred by previous analyses without using the geographic data. If the tree file contains more than one tree, then `PrioriTree` will specify your `BEAST` analysis to average over this distribution of trees during the MCMC.

See this for an example file that contains a single summary phylogeny, and this for an example file that contains a distribution of phylogenies. Check the `Load example tree(s) file` box to load one of these example tree files to `PrioriTree`.

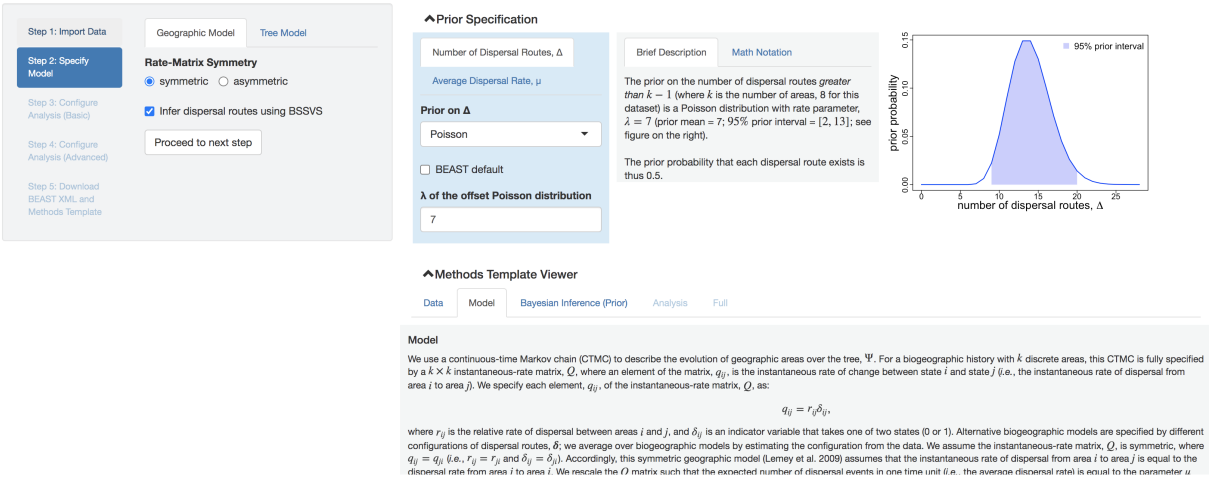

Figure S.2.7: Model and prior specification panels.

Step 2: Specify model and configure basic analysis

Now that input files have been read in, it is time to specify the discrete-geographic model. Click the tab labelled `Step 2: Specify Model` (which won't be clickable until both data files are uploaded). Given that model and prior specification can have a large impact on discrete-geographic phylodynamic inferences, we strongly recommend going through this panel carefully to take advantage of the interactive and visual feature of `PrioriTree` for specifying model and priors.

Specify model and prior

Model and prior specification is the core of `PrioriTree`. Below we briefly describe the basic steps for specifying the discrete-geographic model and priors; for more details, see the corresponding theoretical-background subsections above and Chapter 1.

Specify model.—Specify the discrete-geographic model in the left panel and choose the associated priors in the right panel (Fig. S.2.7). Two tabs in this model-specification panel correspond to the geographic model and tree model, respectively.

In general, the tree-model part is not something to worry about, as `PrioriTree` assumes that your focus is on the discrete-geographic inference. If the imported tree file contains only a single summary tree, it will be treated as a fixed variable in the inference; while if the imported tree file contains a distribution of phylogenies, the default tree-model configuration should be appropriate.

Now let's focus on the discrete-geographic model. In this panel, there are two input fields:

1. `Rate-Matrix Symmetry`: whether the forward and backward dispersal rates between a given pair of geographic areas are assumed to be identical (symmetric) or allowed to be different (asymmetric).
2. `BSSVS`: whether to use Bayesian Stochastic Search Variable Selection (BSSVS) to estimate the number of dispersal routes (see Lemey et al. 2009 and Chapter 1) for detailed explanations).

Specify prior.—Priors may qualitatively affect the main biological conclusions drawn from discrete-geographic phylodynamic inferences (see Chapter 1). In particular, the priors on the average dispersal rate and the number of dispersal routes can be strongly impactful. In the `Prior Specification` panel, we provide a tool for specifying priors based on your biological knowledge before performing the current analysis. The left subpanel allows you to choose the prior distribution on each model parameter, as well as adjust the parameters of the prior distribution, the middle subpanel displays a brief summary of the currently selected prior, and the right subpanel renders the distribution of the currently selected prior.

You can switch between the two parameters using the tabs in the left subpanel. For the prior on the average dispersal rate, `PrioriTree` plots both the average dispersal rate itself (top figure, right subpanel), as well as the resulting prior distribution on the number of dispersal events across the entire dispersal history (bottom figure, right subpanel). The vertical dashed

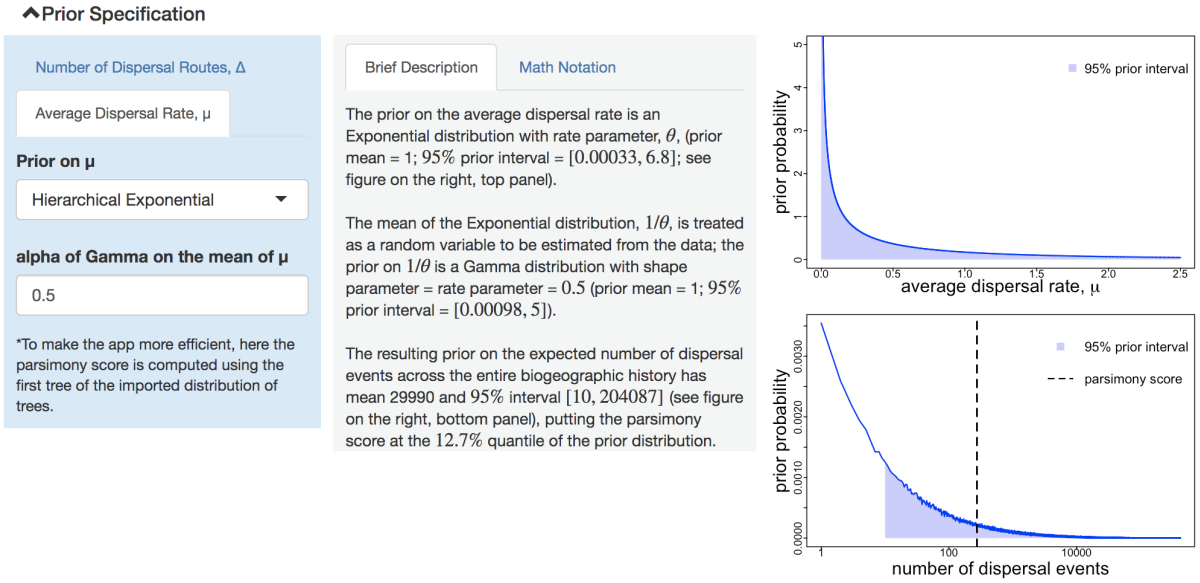

Figure S.2.8: Specify prior on average dispersal rate.

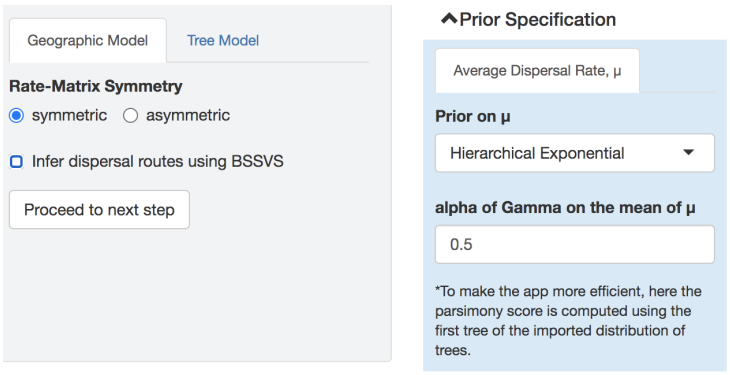

Figure S.2.9: Prior specification panel when not using BSSVS.

line in the bottom figure indicates the parsimony score; you may use it as a reference to assess how biologically realistic the prior is.

The dispersal-route indicators and their sum, the number of dispersal routes, will be part of the model when you choose to average over geographic models using BSSVS; otherwise (when they are not part of the model) the prior on the number of dispersal routes naturally disappear (Fig. S.2.9).

Click the `Proceed to next step` button once you are done with the model and prior specification step. If none of the default settings has been changed, a warning message will pop up as a reminder (Fig. S.2.10). Click the `Yes` button to proceed if you do intend to stay with the default settings. This behavior applies to all the remaining major steps in `PrioriTree`.

Step 3: Configure basic analysis settings

After specifying the model and priors, configure the MCMC settings of the `BEAST` analysis. The default settings of `PrioriTree` are likely to be appropriate for most discrete-geographic analyses. If the MCMC fails, you may want to increase the total number of generations (MCMC chain length) and/or adjust the proposal weights; see the MCMC theoretical-background subsection above for detailed instructions.

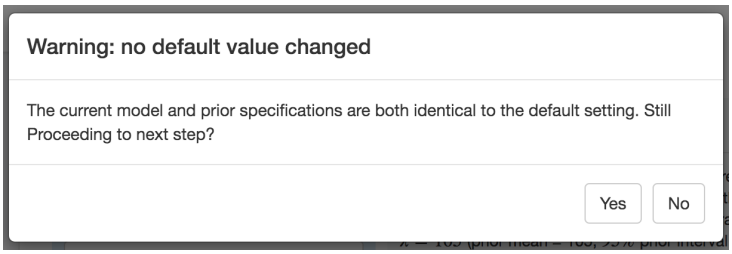

Figure S.2.10: Warning message when no default settings has been changed.

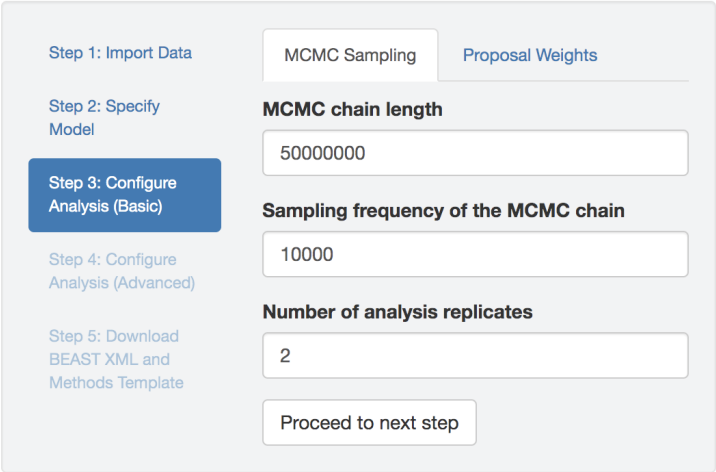

Figure S.2.11: Basic analysis-setting configuration panel.

Brief explanation of the input fields:

- MCMC chain length: the total number of generations for the analysis to run;
- MCMC sampling frequency: how frequent (every how many generations) an MCMC sample will be written to the log files, and;
- Proposal weight: the frequency of performing each proposal.

In addition, you may also choose the number of replicate XML scripts to produce using the third field of this panel. As MCMC is a numerical algorithm approximating the posterior distribution, it is necessary to run multiple MCMCs targeting the same distribution to ensure the results converge; the default number of replicates is thus set to 2 (although 4–8 replicates are generally preferable provided sufficient computational resources).

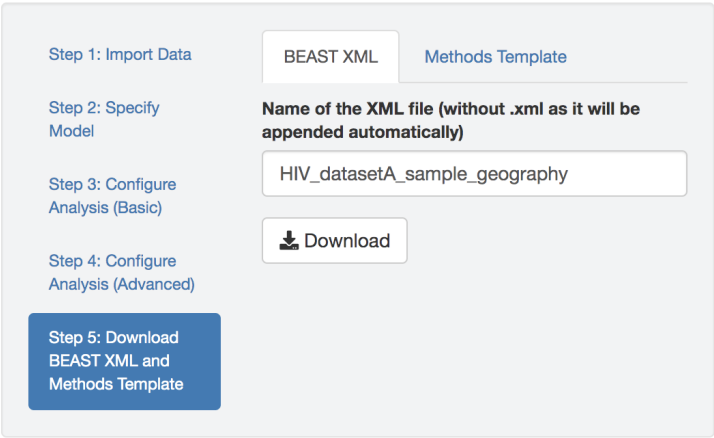

Figure S.2.12: Download the `BEAST` XML script.

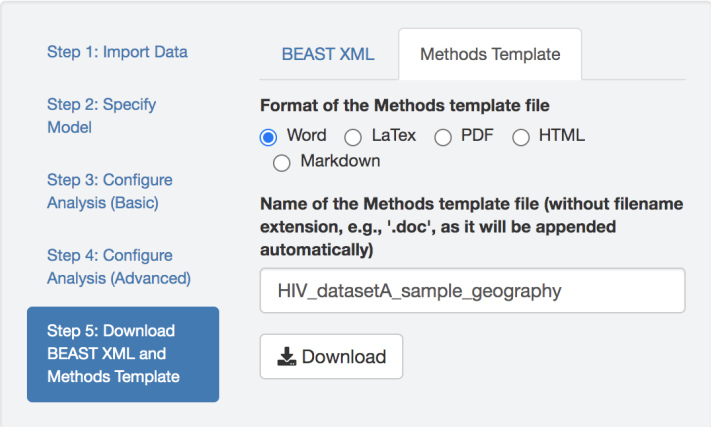

Figure S.2.13: Download the methods-description text file.

Step 4: Save output

At the end, when all the settings are complete, download the readily runnable `BEAST` XML script produced by `PrioriTree`. The default name of the XML script is generated by `PrioriTree` to indicate the type of analysis specified. If the inference involves averaging over a distribution of phylogenies (*i.e.*, the imported tree file contains a distribution of trees), the tree file needs to be put in the same directory as the XML script for it to run. An example XML-script zipped folder generated by `PrioriTree` can be found here.

You may also download a text document that describes the model and prior specification, as well as the analysis configuration. An example methods-description text file generated by `PrioriTree` is available here.

Set up and Summarize `BEAST` Analysis for Assessing Prior Sensitivity in Biogeographic Inference

Recall that inference under discrete-geographic models—where many parameters must be inferred from minimal information—is inherently prior sensitive; *i.e.*, the posterior probability distributions of the discrete-geographic model parameters that we infer from our geographic data are apt to be influenced by the prior probability distributions that we assume for those parameters. In this section, we describe some of the features implemented in `PrioriTree` that are intended to help you identify prior sensitivity in your biogeographic analyses, including joint prior distribution estimation (see this subsection), robust Bayesian inference (see this subsection), and data cloning (see this subsection). In each subsection below, we start with the related theoretical background and then go into details to show the specific operations you need to do in `PrioriTree` to set up and summarize the corresponding `BEAST` analysis.

Estimating the Prior

Theoretical background

A simple but effective way to identify prior sensitivity is to compare the (specified) prior to the (inferred) posterior probability distributions for each parameter: if a parameter is prior sensitive, its inferred posterior probability distribution will be (virtually) identical to whatever prior probability distribution we specified for that parameter. `PrioriTree` allows users to visualize the prior distributions for the geographic model parameters—including the number of dispersal routes, Δ, the average dispersal rate, μ, and the resulting prior distribution on the expected number of dispersal events—which we can then compare to their corresponding posterior distributions to assess prior sensitivity.

However, a possible limitation of this approach is related to the induced priors caused by parameter interactions. Imagine, for example, that we specify (and visualize) a uniform prior for a hypothetical parameter, θ, in `PrioriTree`, but (unforeseen) parameter interactions induce an exponential prior for θ. That is, the independent uniform prior that we initially specified for θ—when marginalized over the joint prior probability distribution of all model parameters—is marginally exponential. After performing our MCMC simulation, we observe that the inferred marginal posterior for θ resembles an exponential distribution, which departs strongly from the uniform prior distribution that we specified for θ, leading us to incorrectly conclude that

this parameter is unlikely to be prior sensitive. Accordingly, it may be safer to estimate the joint prior probability distribution of our model parameters using MCMC, and then compare the inferred marginal prior probability distribution for each parameter to its corresponding inferred marginal posterior probability distribution.

To understand how we estimate the joint prior probability distribution using MCMC, first recall how the M–H algorithm estimates the joint posterior probability distribution. Central to the M–H algorithm is the acceptance probability, R—the probability that we accept a move to a proposed state (set of parameter values)—which is essentially based on the ratio of the posterior probabilities of the proposed (θ') and current (θ) states:

$$R \propto \left[\frac{f(G \mid \theta') \cdot f(\theta')}{f(G \mid \theta) \cdot f(\theta)} \right] = \underbrace{\frac{f(\theta' \mid G)}{f(\theta \mid G)}}_{\text{posterior ratio}}.$$

Because we have replaced all of our geographic data, G, with `"?"`, the likelihood of any parameter value will be identical, such that the first term of the acceptance probability (the likelihood ratio of the proposed and current states) cancels out:

$$R \propto \left[\underbrace{\frac{f(G \mid \theta')}{f(G \mid \theta)}}_{\text{likelihood ratio}} \cdot \underbrace{\frac{f(\theta')}{f(\theta)}}_{\text{prior ratio}} \right] = \underbrace{\frac{f(\theta')}{f(\theta)}}_{\text{prior ratio}},$$

which makes it clear that the MCMC simulation will visit states (parameter values) proportional to their relative prior probability. We can then query the joint prior sample from the

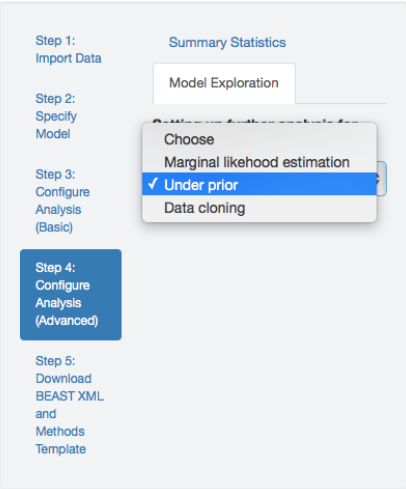

Figure S.2.14: Set up an MCMC simulation to target the joint prior probability distribution.

MCMC simulation to summarize the marginal prior probability distribution for any parameter: *e.g.*, we might infer the marginal prior probability density for the average dispersal rate parameter, μ, by constructing a histogram (frequency distribution) of sampled values from the corresponding column in our log file. These inferred marginal prior probability distributions can then be compared to their corresponding marginal posterior probability distributions to assess prior sensitivity.

Quickstart

To use `PrioriTree` to setup a `BEAST` MCMC simulation targeting the joint prior distribution, select the `Under prior` option in the dropdown menu in the `Model Exploration` panel (Fig. S.2.14). When this option is selected, `PrioriTree` will replace the observed data (*i.e.*, the geographic area where each species was sampled) with `"?"` in the XML script.

Robust Bayesian Inference

Theoretical background

We can assess the prior sensitivity of our biogeographic inferences using an approach called robust Bayesian inference. The fancy name belies the simplicity of this approach; we perform a series of MCMC analyses—of the same dataset under the same inference model—where we iteratively change one (or more) (hyper)priors of our inference model for each separate analysis. We then compare the resulting series of marginal posterior probability distributions for a given parameter to assess whether (or how much) our estimates change under different priors. We usually make this comparison visually, by plotting distributions for a given parameter under the range of candidate priors that we explored.

If the inferred marginal posterior probability distributions are (more or less) identical under a range of corresponding priors, we can safely conclude that our estimates of this parameter are robust to the choice of prior. Conversely, if the marginal posterior probability distributions vary substantially (and resemble their corresponding marginal prior probability distributions), then we would conclude that this parameter exhibits prior sensitivity (*i.e.*, that there is little information in our study data to estimate this parameter). The latter scenario indicates that we need to take further steps; for example, by removing this parameter from our inference model (if possible), or (if not) by making an effort to objectively choose among alternative priors (*e.g.*, by assessing the relative and/or absolute fit of the data to alternative priors).

Quickstart

We have implemented functions in `PrioriTree` to help you perform robust Bayesian inference by generating graphical summaries of parameter estimates under different priors. `PrioriTree` assumes that you have performed `BEAST` analyses under identical model but with different priors. To examine if the posterior estimates are robust to these alternative priors, `PrioriTree` take `BEAST` output files to generate plots that show the inferred posterior distributions under different priors.

Step 1: Import `BEAST` log files

You can upload one or multiple (analysis replicates) `BEAST` log files (that contain parameter estimates under the a given model and prior combination) to an input field; different input fields correspond to different priors (Fig. S.2.15). These parameter log files can include samples from

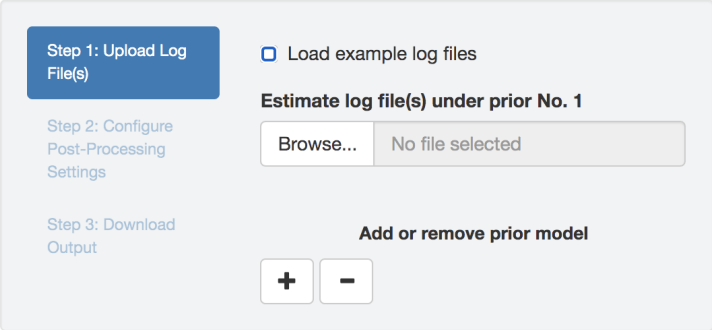

Figure S.2.15: Initial panel: uploading log files.

the joint posterior distribution (*i.e.*, estimates with data) and/or samples from the joint prior distribution (*i.e.*, estimates without data; see the the subsection above for details about how to set up this type of `BEAST` analyses using `PrioriTree`). `PrioriTree` assumes that the uploaded log files are produced using the `BEAST` XML scripts generated by itself; *i.e.*, for summarizing robust Bayesian analysis, only the log files with `_underprior` or `_posterior` as part of their name strings will be included (other uploaded files are ignored). Check the `Load example log files` box to load example log files to `PrioriTree`.

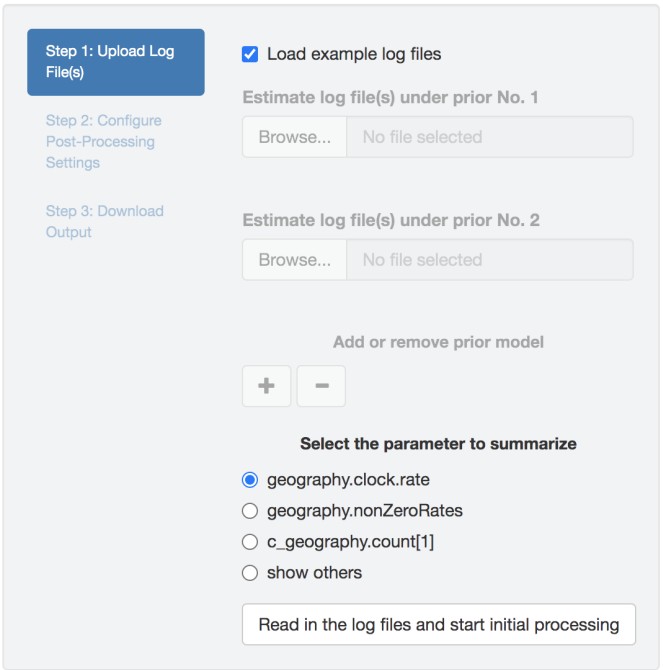

Figure S.2.16: Select the parameter to examine.

Select the parameter to examine.—Once the log files are successfully uploaded, `PrioriTree` finds the intersection of parameters (column names), and then displays them in the way shown in Fig. S.2.16: the focal parameters (*i.e.*, the parameters commonly focused by empirical studies), if exist, will be listed as radio buttons, while all the remaining parameters will only be presented (as a dropdown menu) when the `show others` option is chosen. Simultaneously, when the parameter selection panel appears, the start-processing button will also be enabled; clicking this button triggers the computationally demanding log-parsing action. Note here only a single parameter can be selected (either one of the radio buttons or one item from the drop-down menu); if later you select another parameter to examine, the start-processing button will be enabled again and the log-parsing step needs to be re-executed.

Step 2: Configure post-processing settings

Once `PrioriTree` finishes parsing the log files, two main panels will be enabled: the processing-setting panel (on the left) and the result-visualization panel (on the right; see Fig. S.2.17). All the operations you may perform under this section should be computationally inexpensive so that the changes to the figure and/or table should be seen immediately.

Output processing settings.—Within the processing-setting panel, a separate scrollable collapsible subpanel is displayed under each prior model; each repeated chunk within that subpanel corresponds to the settings you can adjust for each log file (led by the log file name). With

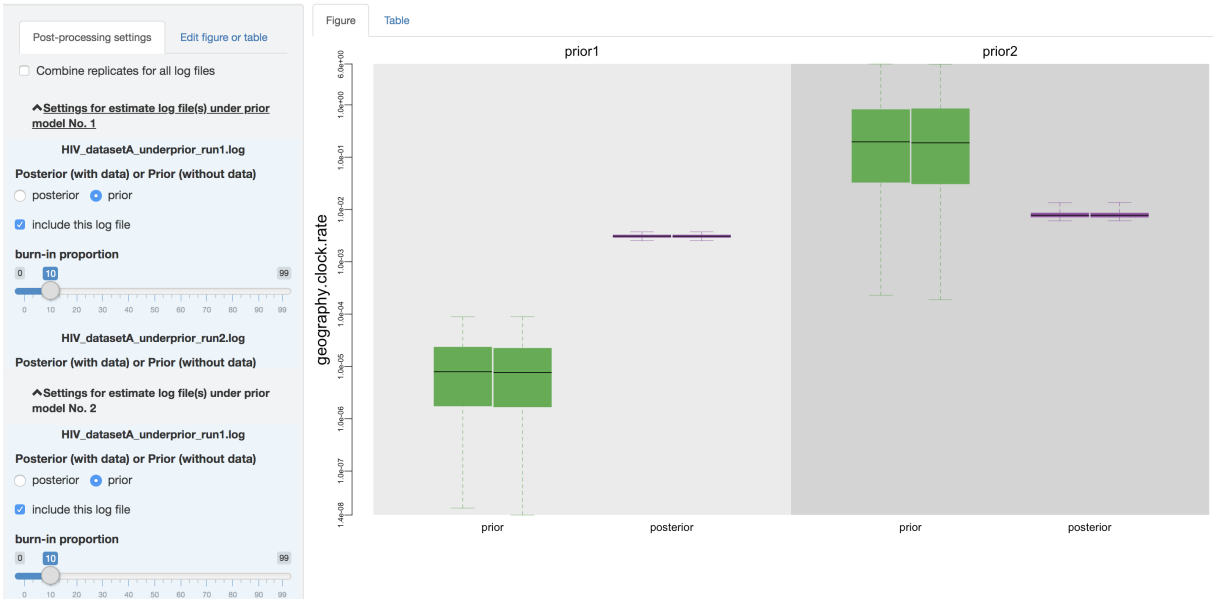

Figure S.2.17: Configure post-processing settings.

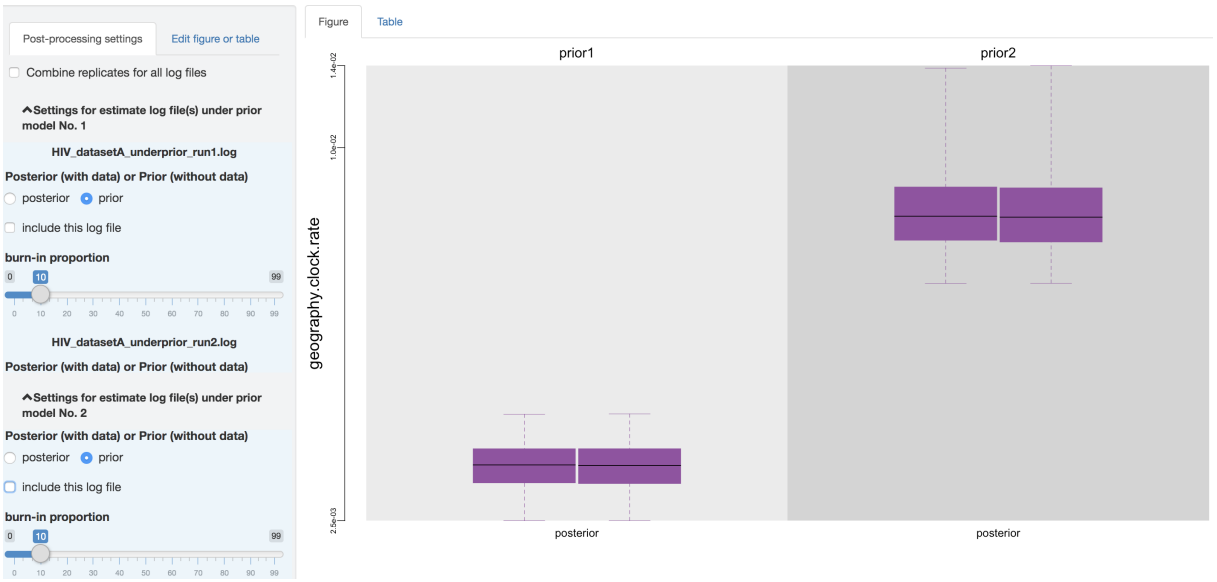

Figure S.2.18: Exclude some log files.

the first item (as radio buttons), you may adjust (if it has not been guessed incorrectly by `PrioriTree`) whether the given parameter log file was sampled from the joint posterior distribution or the joint prior distribution.

The second item allows you to exclude some log files without having to re-execute the log-parsing step. For instance, you may desire to compare the estimated posterior and prior distributions under each prior model in the first place, which may indicate the sensitivity of

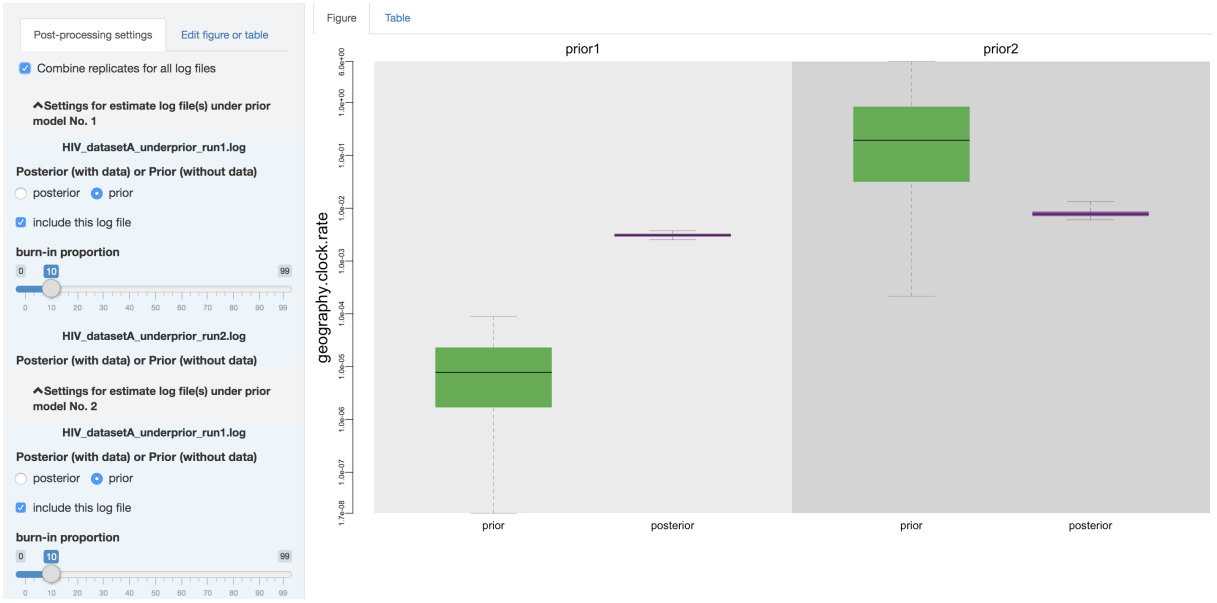

Figure S.2.19: Combine analysis replicates.

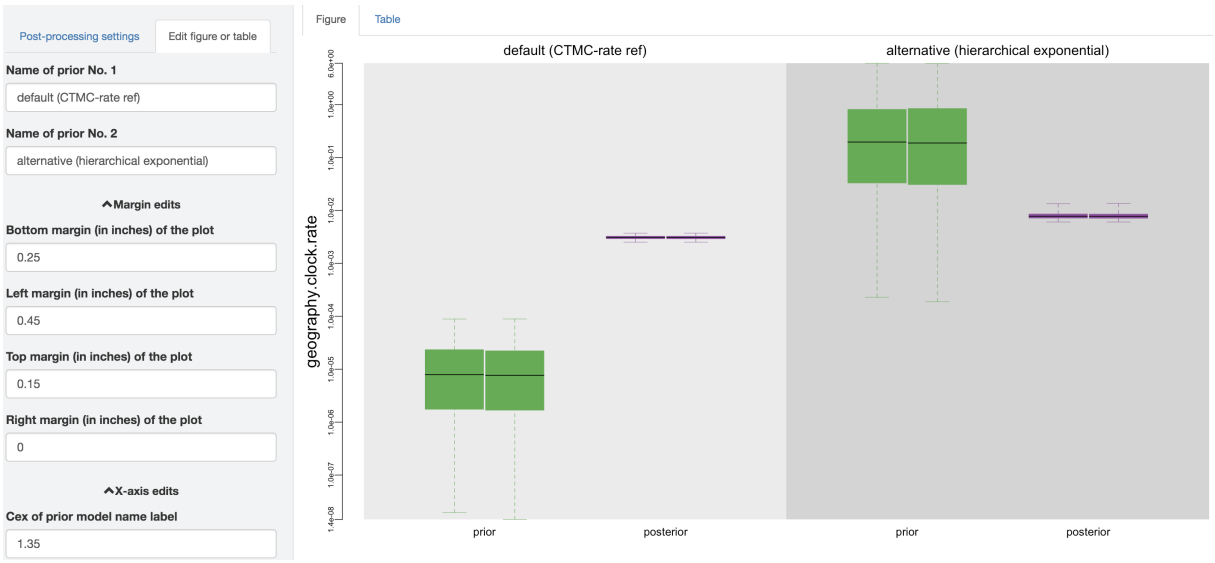

Figure S.2.20: Edit figure.

posterior estimates to the specified prior model. Then if you only want to present the comparison between the estimated posterior distribution, you can uncheck the log files for the estimated prior distribution.

Thirdly, you can adjust the burnin proportion of each log file independently using the slider object. Also, you may choose to combine all the replicate log files (*i.e.*, estimates under identical model and prior specification but sampled from independent MCMC chains) using the check box on the top of the post-processing panel, once confirming that the replicate MCMCs have converged.

Figure Table

Show 10 entries Search:

	Prior model	Posterior or prior	Replicate	Mean	Lower 95% CI	Upper 95% CI
1	default (CTMC-rate ref)	prior	1	1.77537779748323e-05	1.95705876309928e-08	8.88731237618221e-05
2	default (CTMC-rate ref)	prior	2	1.68748219415303e-05	1.44329779643237e-08	8.91538637494517e-05
3	default (CTMC-rate ref)	posterior	1	0.00309874550373551	0.00251940318606008	0.00373475490008824
4	default (CTMC-rate ref)	posterior	2	0.0030950058032378	0.00252008704022875	0.00373805920904519
5	alternative (hierarchical exponential)	prior	1	0.867378480678831	0.000229746356254162	6.04476984738492
6	alternative (hierarchical exponential)	prior	2	0.886740425365977	0.000189137091555126	5.95170773310767
7	alternative (hierarchical exponential)	posterior	1	0.00818278092594316	0.00605615343577873	0.0134105327848021
8	alternative (hierarchical exponential)	posterior	2	0.00816616791464084	0.00605892654265039	0.0135355014843589

Showing 1 to 8 of 8 entries Previous 1 Next

Figure S.2.21: Distribution-summary table.

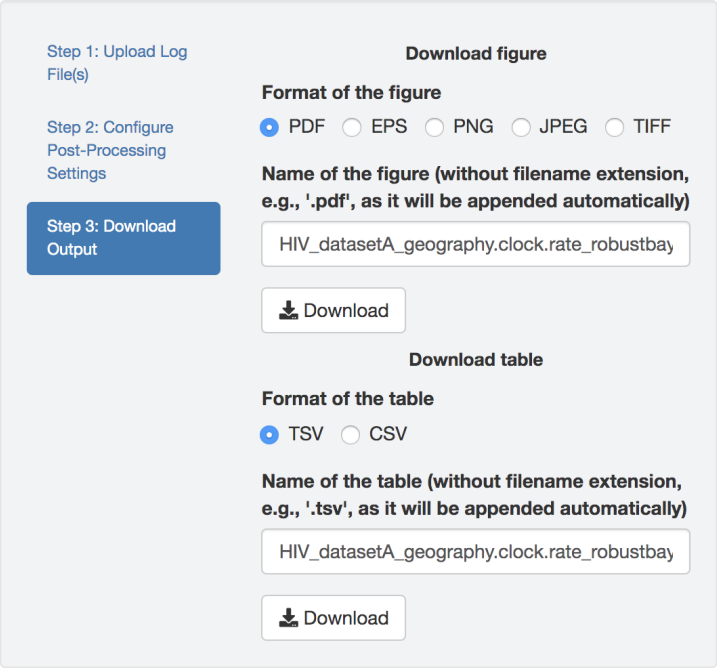

Figure S.2.22: Download figure and/or table.

Figure edits.—You may perform further cosmetic edits to the figure and/or table before saving them. The fields under the figure-edit panel provide flexibility in modifying the appearance of the figure. You can view the exact values of the mean and 95% credible interval of each log file under the table tab.

Step 3: Save output

At the end, when all the settings are complete, you can download the figure and/or the table under the desired format. The default figure/table name generated by `PrioriTree` indicates the type of analysis and the examined parameter.

Data Cloning

Theoretical background

We can also assess the prior sensitivity of our biogeographic inferences using an approach called data cloning (Robert 1993; Lele et al. 2007; Ponciano et al. 2009, 2012). Under this approach, we perform a series of MCMC analyses—under the same inference model with identical priors—where we iteratively increment the number of copies ("clones") of our original dataset used in each separate analysis. We then explore the resulting series of marginal posterior probability distributions for a given parameter to assess how our estimates change as the level of information in the data increases (*i.e.*, as we increment the number of data clones).

We might think of data cloning as the inverse of robust Bayesian inference; as described above, robust Bayesian inference involves a series of analyses where we hold the inference model and data constant, but iteratively change the prior probability distribution for a parameter to explore how the choice of prior impacts the corresponding marginal posterior probability distribution. By contrast, data cloning involves a series of analyses where we hold the inference model and prior constant, but iteratively change the number of copies of the original data to explore how the level of information in the data impacts the inferred marginal posterior probability distribution.

A particular MCMC in a sequence of data clones is defined by the number of replicate copies of our original data, $\beta_i \geq 1$, with the resulting posterior distribution being:

$$P(\theta \mid X)_{\beta_i} \propto P(X \mid \theta)^{\beta_i} P(\theta).$$

If we were to set $\beta_i = 0$, we would be targeting the joint prior probability distribution (*i.e.*, we would be running the MCMC without data), when $\beta_i = 1$, we are targeting the joint posterior probability distribution (*i.e.*, we would be running the MCMC using our original dataset). As $\beta_i \to \infty$, the marginal posterior distribution for the parameter under consideration will converge to a point value that is identical to the maximum-likelihood estimate (MLE) for that parameter (if the parameter is identifiable).

Of interest here is the relative rate at which the marginal posterior probability distribution for the parameter under scrutiny—given the prior specified for that parameter—converges to the MLE as we increase the clone number. If the prior is very informative (*i.e.*, focused on a narrow range of parameter values) and the prior mean is far from the MLE value, the rate of

convergence will be slow. Conversely, if a prior is more diffuse (*i.e.*, spread over a relatively wide range of parameter values) and the prior mean is rather close to the MLE value, the rate of convergence will be relatively fast. When the information in the data is limited, we would generally prefer a prior that has a faster convergence rate. We usually assess the convergence rate visually, by plotting posterior distributions for a given parameter under the range of β_i values that we explored.

Quickstart

`PrioriTree` provides functions that allow you to generate XML files with a series of clone numbers (that you can subsequently analyze using `BEAST`), and also includes functions to allow you to visualize the results of your data-cloning experiments.

Set up `BEAST` data-cloning analysis

You can generate XML files using `PrioriTree` that specify the number of data clones (Fig. S.2.15). Effectively, `PrioriTree` duplicates your original geographic data (*i.e.*, the area in which each tip was sampled). So, specifying k clones will generate an alignment with k copies of the original "site"; *e.g.*, setting $k = 5$ will generate an alignment with 5 sites that are identical copies of the original, single site (*i.e.*, with an identical distribution of areas across tips).

After generating XML files that specify a series of data-cloning analyses (*e.g.*, with 5, 10, 20 copies of the original data), you would then analyze each XML file in `BEAST`. You may set up a series of data clones for a single choice of prior (*e.g.*, to assess the proximity of the mean of that prior to the MLE for the parameter under scrutiny), or you might set up a series of data clones

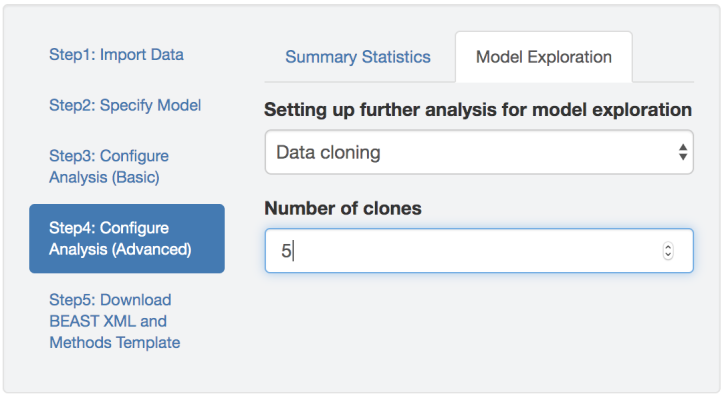

Figure S.2.23: Set up data-cloning analysis.

for two or more candidate priors (*e.g.*, to assess the relative rate of convergence to the MLE under different priors for the parameter under scrutiny). `PrioriTree` provides features to help you explore the results of your data-cloning experiments—by generating plots of the marginal posterior probability distributions under different clone numbers and/or priors—which we describe below.

Note that posterior probability density will converge to the MLE as we increase the amount of information in the data if: (1) our inference model is identifiable (*i.e.*, that each unique set of parameter values has a corresponding unique likelihood value), and; (2) that we specify priors with soft bounds for all of our model parameters (*i.e.*, that one or more of our priors does not assign zero prior probability [does not "box out"] the corresponding maximum-likelihood estimate for that parameter.)

Summarize data-cloning analysis

`PrioriTree` assumes that you have performed a sequence of `BEAST` analyses with increasing number of copies of the data using identical model (under one or multiple priors). `PrioriTree` take the resulting `BEAST` output files as input to generate plots that show the inferred distributions under various numbers of data clones (as well as under different priors if they exist).

Step 1: Import log files.—You can upload multiple (from analysis replicates and/or under different number of clones) `BEAST` log files (that contain parameter estimates under the a given model and prior combination) to a input field; different input fields correspond to different

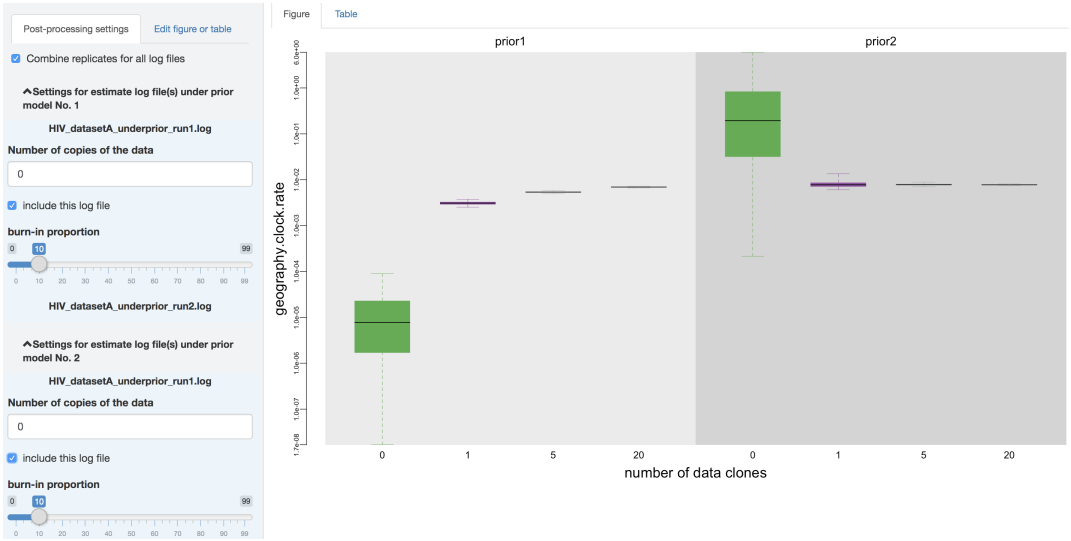

Figure S.2.24: Combine analysis replicates.

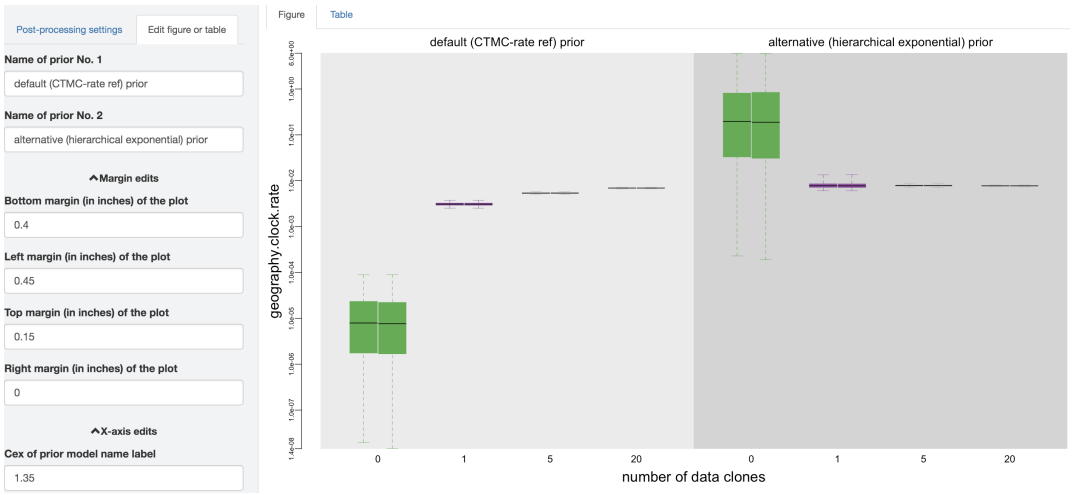

Figure S.2.25: Edit figure.

priors (Fig. S.2.15). These log files may include the estimates sampled from data-cloned distributions (with various numbers of copies of the data), the joint posterior distribution (one copy of the data), or the joint prior distribution (zero copy of the data). `PrioriTree` assumes that the uploaded log files are produced using `BEAST` XML scripts generated by itself; *i.e.*, for summarizing data-cloning analysis, upload log files with `_datacloning`, `_posterior`, or `_underprior` as part of their name strings.

Steps 2–5.—The following workflow of summarizing data-cloning analysis is effectively identical to the robust Bayesian analysis (see the robust Bayesian quickstart subsection), so we only present Figs. S.2.24–S.2.26 to show the different output figures.

Figure | Table

Show 10 entries Search:

	Prior model	Number of data clones	Replicate	Mean	Lower 95% CI	Upper 95% CI
1	default (CTMC-rate ref) prior	0	1	1.77537779748323e-05	1.95705876309928e-08	8.88731237618221e-05
2	default (CTMC-rate ref) prior	0	2	1.68748219415303e-05	1.44329779643237e-08	8.91538637494517e-05
3	default (CTMC-rate ref) prior	1	1	0.00309874550373551	0.00251940318606008	0.00373475490008824
4	default (CTMC-rate ref) prior	1	2	0.0030950058032378	0.00252008704022875	0.00373805920904519
5	default (CTMC-rate ref) prior	5	1	0.00537380032563668	0.00496518395717419	0.00578555633854775
6	default (CTMC-rate ref) prior	5	2	0.00537943783023824	0.0049654199183492	0.0058234244044098
7	default (CTMC-rate ref) prior	20	1	0.00692432714972432	0.00661358016178822	0.00723175309143433
8	default (CTMC-rate ref) prior	20	2	0.00692617781140897	0.00663502221476068	0.00723296017225248
9	alternative (hierarchical exponential) prior	0	1	0.867378480678831	0.000229746356254162	6.04476984738492
10	alternative (hierarchical exponential) prior	0	2	0.886740425365977	0.00018937091555126	5.95170773310767

Showing 1 to 10 of 16 entries Previous 1 2 Next

Figure S.2.26: Distribution-summary table.

Set up and Summarize `BEAST` Analysis for Assessing Relative or Absolute Model Fit

For a given phylodynamic study, we typically wish (or need) to consider several candidate discrete-geographic models (where alternative models might specify a/symmetric rate matrices, different priors on the average dispersal rate and/or number of dispersal routes, etc.). Comparing the fit of competing phylodynamic models to our study data offers two important benefits. First, model-based inference—including phylodynamic inference—assumes that our inference model provides a reasonable description of the process that generated our study data; otherwise, our inferences—including estimates of relative and/or average dispersal rates and any summaries based on those parameter estimates (ancestral areas, dispersal histories, number of dispersal events, etc.)—are apt to be unreliable. Additionally, comparing alternative discrete-geographic models provides a means to objectively test hypotheses regarding the history of dispersal (*i.e.*, by assessing the relative fit of our data to competing models that are specified to include/exclude a parameter relevant to the hypothesis under consideration).

To this end, `PrioriTree` implements functions to help you assess both the relative (see this subsection) and absolute fit of discrete-geographic models to an empirical dataset (see this subsection). In each subsection below, we start with the related theoretical background and then go into details to show the specific operations you need to do in `PrioriTree` to set up and summarize the corresponding `BEAST` analysis.

Compare the Relative Fit of Competing Models

Theoretical background

We assess the relative fit of two or more candidate discrete-geographic models to a given dataset by computing Bayes factors, which is based on comparing the average fit (*i.e.*, the 'marginal likelihood') of competing models to that dataset. The apparent simplicity of Bayes factors belies some rather challenging conceptual and computational issues. Here, we begin by describing a relevant probability concept (marginal likelihood), then detail the numerical methods that we use to estimate marginal likelihoods (stepping-stone simulation), and then describe how to compute (and interpret the results of) Bayes factors. Finally, we describe how to assess (and improve) the reliability of our marginal-likelihood estimates using `PrioriTree` and `BEAST`.

Marginal likelihood

Marginal likelihoods can be a challenging concept. In the simplest terms, it is the average fit of a model to a dataset. More precisely, the marginal likelihood is the probability of observing the data (*i.e.*, the likelihood) averaged over all values for every parameter in the model, weighted by the prior probability of those parameter values (*i.e.*, it is the likelihood averaged over the joint prior probability distribution of the model parameters).

Recall that we've previously encountered this probability term (lurking in the denominator of Bayes theorem):

$$\overbrace{P(\mathbf{Q}, \mu \mid G, \Psi)}^{\text{posterior distribution}} = \frac{\overbrace{P(G \mid \mathbf{Q}, \mu, \Psi)}^{\text{likelihood}} \overbrace{P(\mathbf{Q})P(\mu)}^{\text{prior distribution}}}{\underbrace{P(G \mid \Psi)}_{\textbf{marginal likelihood}}}.$$

[Note that the marginal likelihood for model M_i is conditional on the model, *i.e.*, $P(G \mid M_i)$. By convention, however, this dependence is often suppressed or ignored, as in the equation above.]

You may also recall that the MCMC algorithms that we use to approximate the joint posterior probability density, $P(\mathbf{Q}, \mu \mid G, \Psi)$, involves simulating a Markov chain that samples states—where each state, θ, is a fully specified model $\theta = \{\Psi, \mathbf{Q}, \mu\}$—based on their relative posterior probabilities:

$$R \propto \left[\underbrace{\frac{f(G \mid \theta')}{f(G \mid \theta)}}_{\text{likelihood ratio}} \cdot \underbrace{\frac{f(\theta')}{f(\theta)}}_{\text{prior ratio}} \cdot \underbrace{\frac{f(\theta \mid \theta')}{f(\theta' \mid \theta)}}_{\text{proposal ratio}} \right] = \underbrace{\frac{f(\theta' \mid G)}{f(\theta \mid G)}}_{\text{posterior ratio}}.$$

In other words, the MCMC algorithms that we use to estimate posterior probabilities of discrete-geographic model parameters from our data completely (and deliberately) avoid calculating the denominator of Bayes theorem; *i.e.*, the marginal likelihood that we need to compute Bayes factors!

Estimating marginal likelihoods

In order to estimate marginal likelihoods, we must resort to alternative numerical methods. These methods are variously referred to as "stepping-stone" or "power-posterior" sampling algorithms. These algorithms essentially involve running a series of MCMC simulations over a sequence of "stones" that allow us to step from the joint posterior probability distribution to

the joint prior probability distribution. For each stone, i, we raise the likelihood by a power, β_i, such that MCMC for this stone is estimating the distribution:

$$P(\theta \mid G, \beta_i) = \frac{P(G \mid \theta)^{\beta_i} P(\theta)}{P(G)}.$$

When $\beta = 1$ the MCMC samples from the joint posterior probability distribution, and when $\beta = 0$ the MCMC samples from the joint prior probability probability distribution. For intermediate values of β in $1 \rightarrow 0$, the MCMC samples from increasingly distorted ("heated") versions of the posterior distribution.

To perform a stepping-stone simulation, we first need to specify the number of stones, k, that we will use to span the posterior and prior distributions (we usually specify a relatively large number, *e.g.*, $k \geq 32$). Next, we need to decide how to space our k stones. The most common approach spaces the stones as k quantiles of a Beta probability distribution (where the quantiles divide the distribution into $k - 1$ intervals with equal probability). The Beta distribution has two shape parameters, where the second one is set to one by convention (as we only need one degree of freedom for this distribution). We might, for example, set the first shape parameters α also to one, which specifies a uniform probability distribution (as a special case of the Beta), such that the k quantiles in this case would be uniformly distributed between the posterior and prior. However, as we move from the posterior to the prior, while the difference between β_{i+1} and β_i stays constant, the overlap between consecutive power-posterior distributions ($P(\theta \mid G, \beta_i)$ and $P(\theta \mid G, \beta_{i+1})$) becomes increasingly small. The approximation works poorly when the overlap becomes too small. Following the `BEAST` default, `PrioriTree` specifies the sequence of β values following evenly-spaced quantiles of a Beta$(0.3, 1.0)$ distribution (*i.e.*, $alpha = 0.3$), so that more values of β are put near 0 than near 1 (originally recommended by Xie et al. 2011).

Computing Bayes factors

Often, we compare two competing models—models M_0 and M_1, for example—by computing the Bayes factor:

$$\mathrm{BF}_{01} = \frac{P(G \mid M_0)}{P(G \mid M_1)}$$

Bayes factors greater than 1 reflect positive support for the model in the numerator, whereas Bayes factors less than 1 reflect positive support for the model in the denominator. Bayes factors

near 1 indicates that both models perform relatively the same. When comparing more than two models, we simply compute the Bayes factor between each pair of models and rank the models accordingly. Since we compute log-marginal-likelihoods, it's convenient to express the Bayes factors as:

$$2 \ln \mathrm{BF}_{01} = 2 \left(\ln P(G \mid M_0) - \ln P(G \mid M_1) \right),$$

where the factor of two is simply conventional.

Interpreting Bayes factors

Kass and Raftery (1995) provide rough guidelines for interpreting the strength of support indicated by Bayes factors:

BF_{01}	$2 \ln \mathrm{BF}_{01}$	Support for model M_0
1 to 3	0 to 2	Equivocal
3 to 20	2 to 6	Positive
20 to 150	6 to 10	Strong
> 150	> 10	Decisive

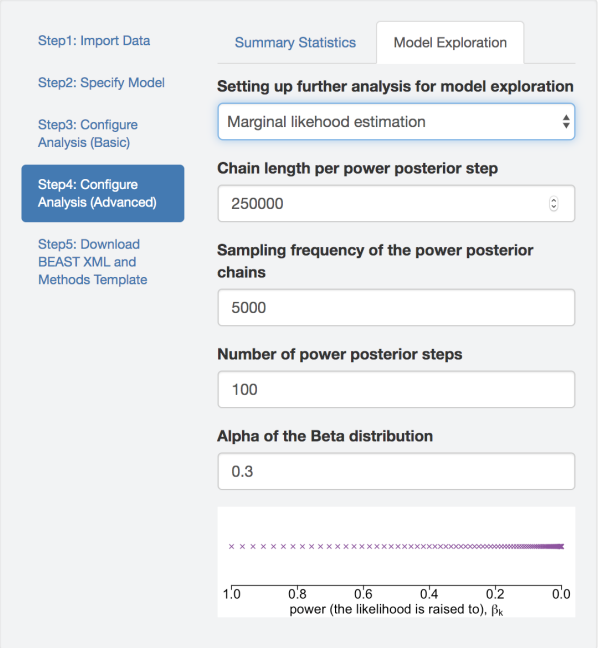

Figure S.2.27: Marginal likelihood estimation analyses.

Quickstart

`PrioriTree` sets up the marginal-likelihood `BEAST` analysis by appending a `marginalLikelihoodEstimator` section in the XML, after the analysis configuration section of the MCMC that approximates the joint posterior distribution. This will allow us to estimate marginal likelihood through both thermodynamic integration (Lartillot and Philippe 2006) and stepping-stone sampling (Xie et al. 2011; Baele et al. 2012).

The number of powers and how many MCMC generations under each power may also strongly impact the accuracy of the estimates; default values are likely to be sufficient for most empirical datasets and models. However, the most straightforward way to check the convergence of marginal likelihood estimates is to run multiple replicates of the analyses to see if we get stable estimates across replicates. If the estimates differ significantly among replicates (say greater than a few log-likelihood units, especially if it is on the same scale as the difference between the log marginal-likelihood estimates under competing models), consider increasing the number of powers and/or the MCMC chain length under each power.

Posterior-Predictive Checking

Theoretical background

We may wish to assess how close our inferred process (*i.e.*, the assumed the model and the posterior parameter estimates under the model) is to the true process that gave rise to the observed data. One way to achieve this is to simulate datasets under the assumed model and posterior estimates, and then compare them with the observed data. When the simulated data resemble the observed data closely, we consider that the assumed model provides an adequate fit to the data in an absolute sense (*i.e.*, not comparing to any other competing models). This model-adequacy assessment approach is referred to as posterior-predictive checking (Gelman et al. 1996; Bollback 2002).

Each individual posterior-predictive simulation is performed by drawing a vector of parameters, $\theta_i = \{\Psi_i, \boldsymbol{r_i}, \boldsymbol{\delta_i}, \mu_i\}$, at random from the MCMC samples approximating the joint posterior distribution, and then simulating a predictive dataset, G_i^{sim}, conditional on those parameters. Repeating this simulation procedure m times, we obtain m predictive datasets.

A difference statistic can then be calculated for the i^{th} simulated dataset as:

$$D_i = T(G_i^{\text{sim}} \mid \theta_i) - T(G^{\text{obs}} \mid \theta_i),$$

where G^{obs} is the observed biogeographic dataset, and $T(\cdot \mid \theta_i)$ is a summary statistic (detailed below).

For the m predictive datasets, the posterior-predictive p-value is calculated as:

$$P = \frac{1}{m} \sum_{i=1}^{m} D_i \geq 0,$$

with values between 0.025 and 0.975 indicating that the model is adequate and cannot be rejected (*i.e.*, the observed statistic is within the 95% posterior-predictive interval).

Two summary statistics can be used to assess model adequacy: (1) the parsimony statistic, and; (2) the tip-wise multinomial statistic. For the parsimony statistic, we simply calculated the parsimony score for the given simulated or observed dataset, conditional on the sampled tree, Ψ_i (achieved in `PrioriTree` by calling the `parsimony()` function in `R` package `phangorn` Schliep 2010).

The tip-wise multinomial statistic treats the states at the tips of the tree for the single geographic character (*i.e.*, site) as the outcomes of the multinomial trial. [This is similar to the

multinomial statistic introduced by Goldman (1993) and used in posterior-predictive simulation by Bollback (2002), which treats the sites (columns) in a molecular alignment as outcomes of a multinomial trial.] For the tip-wise multinomial statistic, we calculated:

$$T(G \mid \theta_i) = \sum_{i=1}^{k} n_i \ln(n_i/n),$$

where n is the number of tips, and n_i is the number of tips in state i.

Quickstart

To perform posterior-predictive checking, `PrioriTree` requires you to provide the observed data as well as the estimates (as log and tree files produced by `BEAST`) inferred from the data, assuming these inference outputs are generated by `BEAST` using the XML scripts produced by `PrioriTree` (only when this is the case, `PrioriTree` can reliably parse the log file and figure out the exact discrete-geographic model used in the inference, so that it can simulate data under that model). Once all the required input files are provided, you can start the simulation in `PrioriTree`, and then `PrioriTree` will generate plots to show the posterior-predictive distributions for each replicate analysis (as well as under different priors if they exist).

Step 1: import files

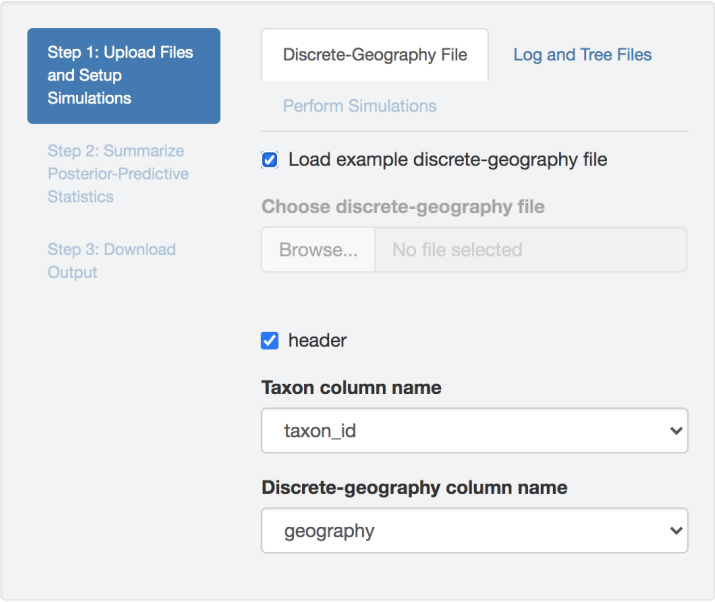

Figure S.2.28: Import discrete-geographic data.

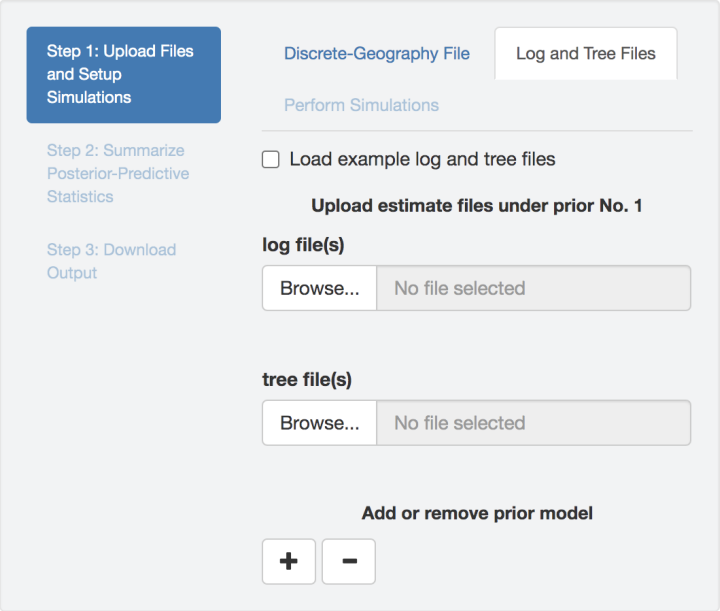

Figure S.2.29: Upload `BEAST` output files.

Discrete-geographic data file.—The discrete-geographic data file needs to be either a `.csv` or `.tsv` file which contains two (or more) columns. The header (first row) of the file contains the names of the columns. By default, `PrioriTree` assumes that the first column contains the taxon names, and the second column contains the geographic area for that taxon. If the columns in your data file are in a different order, you can select the columns containing the taxon name and geographic data from the drop-down menu after uploading the discrete-geographic data file into `PrioriTree` (the other columns are ignored). Check the `Load example discrete-geography file` box to read in an example discrete-geographic data file.

`BEAST` analysis output log and tree files.—You can upload one or multiple (analysis replicates) `BEAST` log/tree files (that contain parameter estimates under the a given model and prior combination) to an input field; different input fields correspond to different priors (Fig. S.2.30). Note here, not only the log file(s), you also need to upload the associated tree file(s). We need to know the tree sampled simultaneously with each parameter-estimate sample to simulate the dataset, as well as to compute the parsimony statistic.

`PrioriTree` assumes each sample in the log file and in its associated tree file match each other; *i.e.*, they should have been sampled from the same iteration of a `BEAST` analysis (the default behavior when the XML scripts generated by `PrioriTree` were used). If you have combined replicate analyses or have thinned their estimates files, identical operations need to

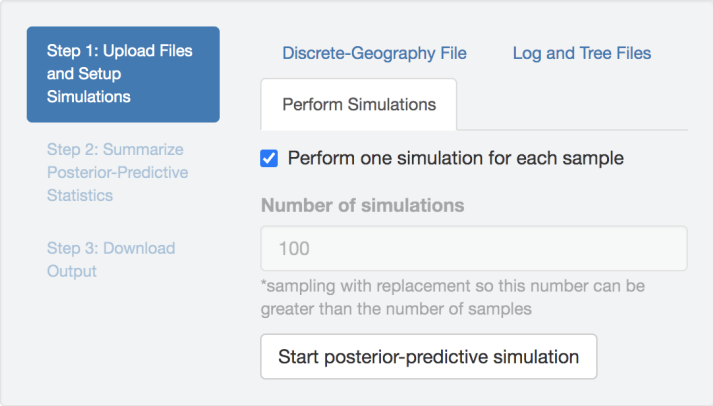

Figure S.2.30: Start posterior-predictive simulations.

be applied to both the log file and the associated tree file. Check the `Load example log and tree files` box to read in an example sets of log and tree files.

Perform posterior-predictive simulations.—Once all the input files (including the geographic-data file, parameter-estimates log file(s) and the associated tree file(s)) are uploaded and they are valid (both in terms of their own format and the match between them), the `Perform Simulations` tab will be enabled. In the `Perform Simulations` panel, you can either choose to simulate a dataset using each sample in the uploaded distribution by checking the `Perform one simulation for each sample` box, or specify the desired number of simulated datasets by unchecking the box first and then editing the simulation-number field. Finally, once the simulation configuration is done, click the `Start posterior-predictive simulation` button to initiate the computationally demanding log-parsing and forward-simulation actions. This simulation step may take a noticeable amount of time to complete, which scales with the number of sequences and number of geographic areas of the dataset, as well as the number of simulations specified. If later you change any of the uploaded files or the number of simulations to perform, click the `Start posterior-predictive simulation` button to re-execute the simulation step.

Step 2: Configure post-processing settings

Once `PrioriTree` finishes the posterior-predictive simulations, two panels will be enabled automatically: the processing-setting panel (on the left) and the result-visualization panel (on the right; Fig. S.2.31). All the operations you may perform under this section should be computationally inexpensive so that the changes to the figure and/or table should be seen immediately.

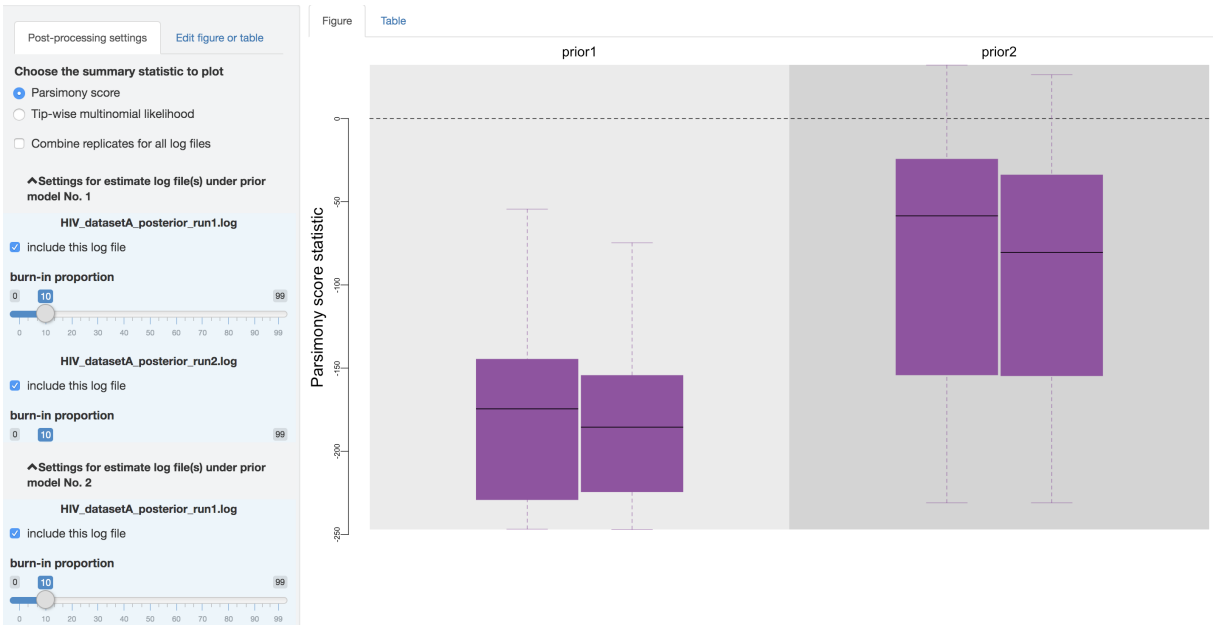

Figure S.2.31: Configure post-processing settings.

Output processing settings.—First, you can choose to visualize the posterior-predictive distributions under one of the two available statistics: 1) parsimony statistic and 2) tip-wise multinomial statistic (see the theoretical-background subsection above for details), and switch between them using the radio buttons on the top of the `post-processing settings` panel.

Below it, there is a checkbox that you can click to combine all the replicate analysis (*i.e.*, estimates under identical model and prior specification), once confirming that the replicate

Figure S.2.32: Posterior-predictive distributions of the tip-wise multinomial likelihood statistic.

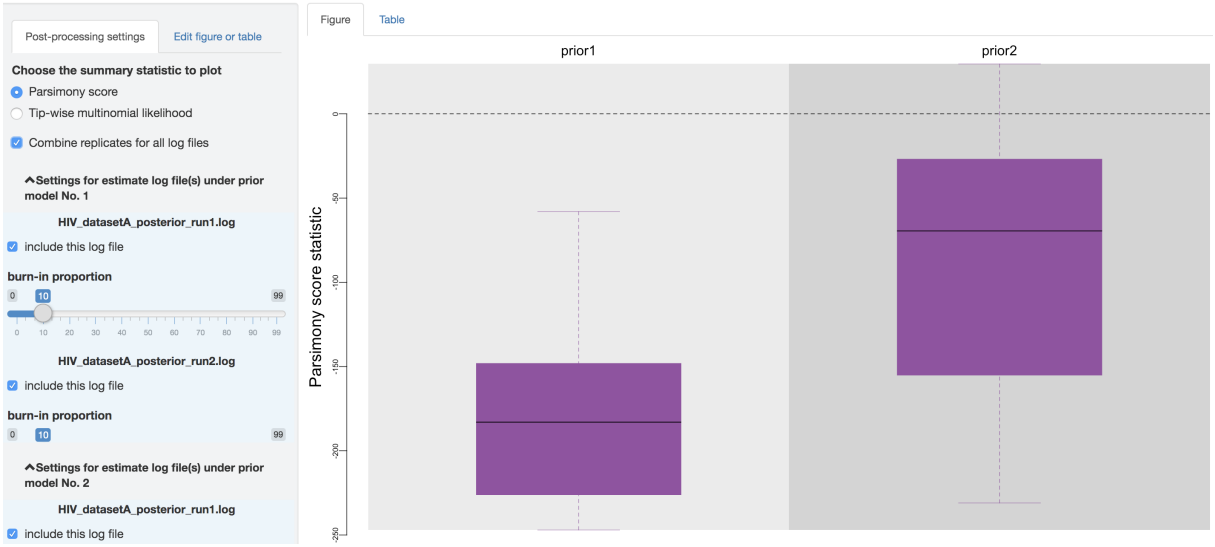

Figure S.2.33: Combine replicates.

MCMCs have converged.

Below the checkbox, a separate scrollable collapsible subpanel is displayed under each prior model; each repeated chunk within that subpanel contains the settings you can adjust for each log file (led by the log file name). The first item allows you to exclude some log files without having to re-execute the log parsing step. With the second item, you can adjust the burnin proportion of each analysis independently using the slider object.

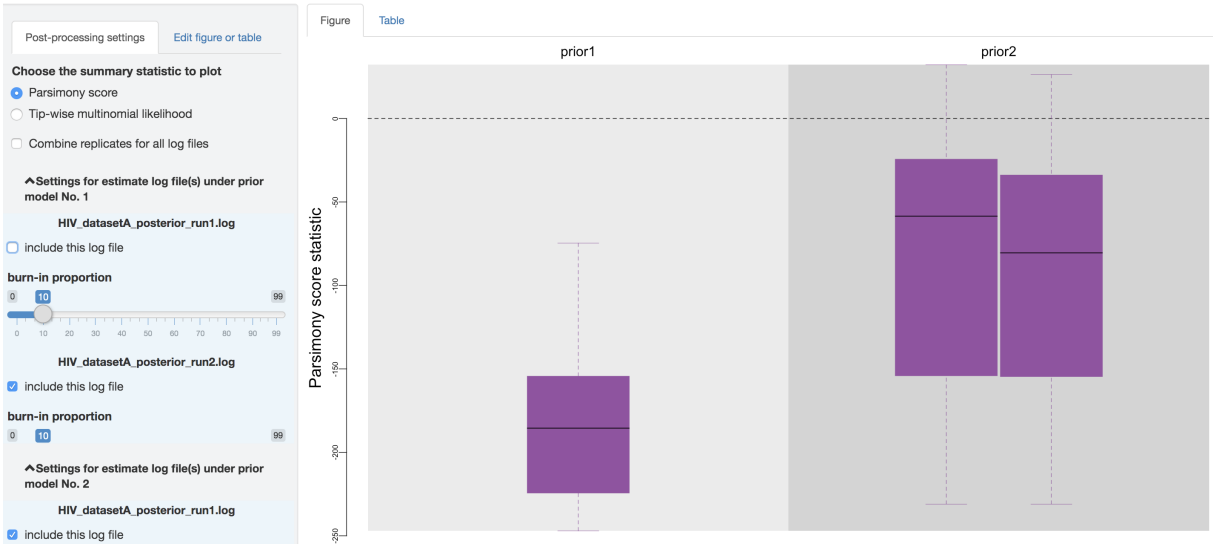

Figure S.2.34: Exclude some log files.

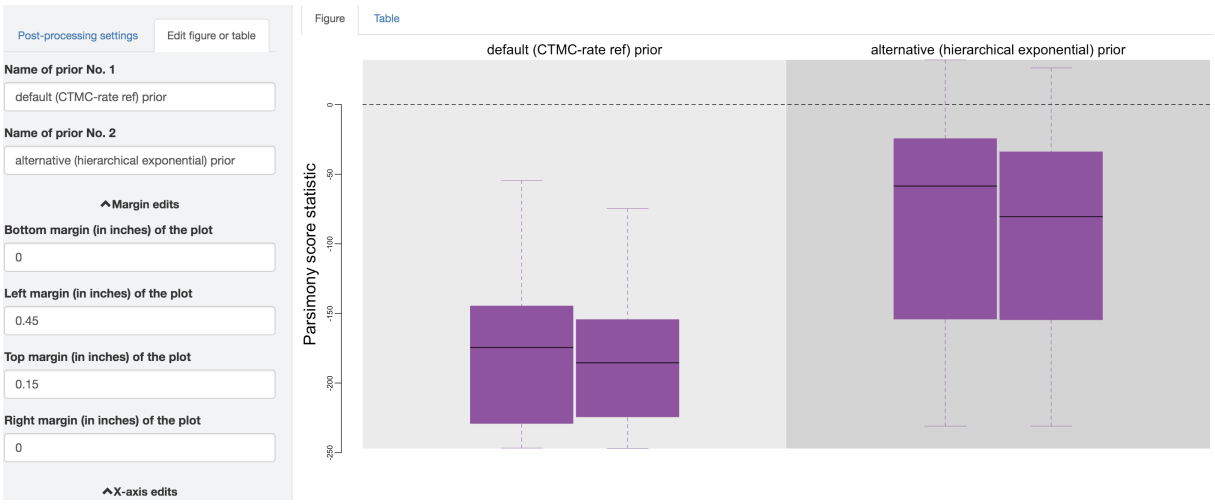

Figure S.2.35: Edit figure.

Figure edits.—Users may perform further cosmetic edits to the figure and/or table before they go to next step to save the output. The fields under the figure-edit panel should give you flexibility in modifying the appearance of the figure. Users can also view the exact posterior-predictive p-values of each analysis for both statistics together under the table tab.

Step 3: save output

Save figure or table output.—At the end, when all the settings are complete, download the figure and/or the table under the desired format. The default figure/table name generated by `PrioriTree` indicates the type analysis and the selected posterior-predictive statistic (only for the figure as the table contains both statistics).

Figure Table

Show 10 entries Search:

	Prior model	Replicate	Posterior-predictive p-value for parsimony score statistic	Posterior-predictive p-value for tip-wise multinomial likelihood statistic
1	prior1	1	0	0.977777777777778
2	prior1	2	0	0.988888888888889
3	prior2	1	0.0666666666666667	0.888888888888889
4	prior2	2	0.1	0.866666666666667

Showing 1 to 4 of 4 entries Previous 1 Next

Figure S.2.36: Posterior-predictive p-values table.

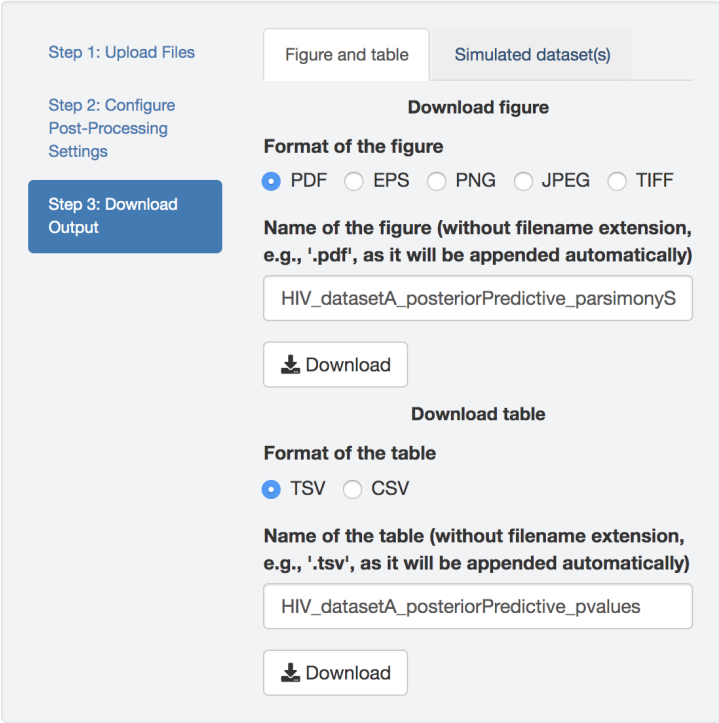

Figure S.2.37: Download figure and/or table.

Save the simulated datasets

You can also download the posterior-predictive simulated datasets and summarize them in other ways (*e.g.*, using alternative summary statistics other than the two provided by `PrioriTree`). For each analysis, `PrioriTree` simulates datasets (each of which contains state of every tip in the tree) and then writes them out as a single `.tsv` file, where each column indicates a tip (first row contains tip names as column names) while each row contains a simulated dataset (so the number of rows, after the first header row, is identical to the number of post-

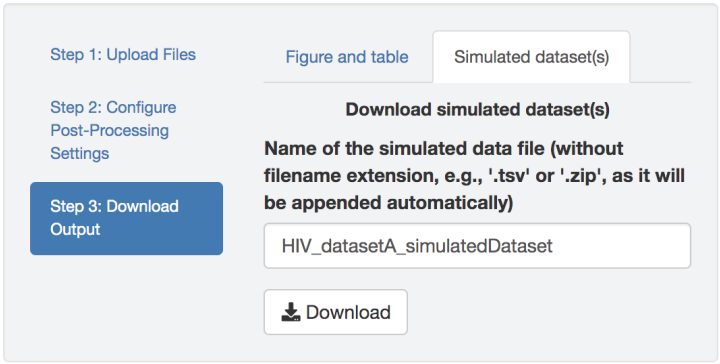

Figure S.2.38: Download the simulated datasets.

burnin samples of the corresponding analysis). When there are multiple analyses (replicates and/or under different prior models), `PrioriTree` will produce a zipped folder that contains all the `.tsv` files (where each `.tsv` file contains the simulated datasets for the corresponding analysis). The name of each `.tsv` file contains the prior model name as well as the replicate id as part of its string, so that you can match them to the uploaded analysis files. An example zipped folder that contains the simulated datasets is available here.

Chapter 3

New Phylogenetic Models Incorporating Interval-Specific Dispersal Dynamics Improve Inference of Disease Spread

Abstract.—Phylodynamic methods reveal the spatial and temporal dynamics of viral geographic spread, and have featured prominently in studies of the COVID-19 pandemic. Virtually all such studies are based on phylodynamic models that assume—despite direct and compelling evidence to the contrary—that rates of viral geographic dispersal are constant through time. Here, we: (1) extend phylodynamic models to allow both the average and relative rates of viral dispersal to vary independently between pre-specified time intervals; (2) implement methods to infer the number and timing of viral dispersal events between areas; and (3) develop statistics to assess the absolute fit of phylodynamic models to empirical datasets. We first validate our new methods using simulations, and then apply them to a SARS-CoV-2 dataset from the early phase of the COVID-19 pandemic. We show that: (1) under simulation, failure to accommodate interval-specific variation in the study data will severely bias parameter estimates; (2) in practice, our interval-specific phylodynamic models can significantly improve the relative and absolute fit to empirical data; and (3) the increased realism of our interval-specific phylodynamic models provides qualitatively different inferences regarding key aspects of the COVID-19 pandemic—revealing significant temporal variation in global viral dispersal rates, viral dispersal routes, and the number of viral dispersal events between areas—and alters interpretations regarding the efficacy of intervention measures to mitigate the pandemic.

INTRODUCTION

Phylodynamic methods encompass a suite of models for inferring various aspects of pathogen biology, including: (1) patterns of variation in demography through time (Drummond et al. 2005; Minin et al. 2008; Gill et al. 2013, 2016); (2) the history of geographic spread either over continuous space (Lemey et al. 2010; Pybus et al. 2012; Gill et al. 2017) or among a set of discrete-geographic areas Edwards et al. (2011); Lemey et al. (2009), and; (3) the interaction between demography and geographic history (Kühnert et al. 2016; De Maio et al. 2015; Müller et al. 2017, 2019). Our focus here is on discrete-geographic phylodynamic models. These phylodynamic methods have been used extensively to understand the spatial and temporal spread of disease outbreaks and have played a central role for inferring key aspects of the COVID-19 pandemic, such as the geographic location and time of origin of the disease, the rates and geographic routes by which it spread, and the efficacy of various mitigation measures to limit its geographic expansion (Worobey et al. 2020; Candido et al. 2020; Dellicour et al. 2021; Douglas et al. 2021; Lemey et al. 2021; Kraemer et al. 2021; Alpert et al. 2021; Nadeau et al. 2021; Washington et al. 2021; Müller et al. 2021; Bedford et al. 2020; Wilkinson et al. 2021; Davies et al. 2021; Tegally et al. 2021; Fauver et al. 2020; du Plessis et al. 2021).

These phylodynamic methods adopt an explicitly probabilistic approach that model the process of viral dispersal among a set of discrete-geographic areas (Baele et al. 2017). The observations include the times and locations of viral sampling, and the genomic sequences of the sampled viruses. These data are used to estimate the parameters of phylodynamic models, which include a dated phylogeny of the viral samples, the global dispersal rate (the average rate of dispersal among all geographic areas), and the relative dispersal rates (the dispersal rate between each pair of geographic areas).

The vast majority (651 of 666, 97.7%; Fig. S.3.1) of discrete-geographic phylodynamic studies are based on the earliest models (Lemey et al. 2009; Edwards et al. 2011), which assume that viral dispersal dynamics—including the average and relative rates of viral dispersal—remain constant over time. However, real-world observations indicate that the average and/or relative rates of viral dispersal inevitably vary during disease outbreaks. For example, relative rates of viral dispersal typically change as a disease is introduced to (and becomes prevalent in) new areas, and begins dispersing from those areas to other areas. Dispersal dynamics are also generally impacted by the initiation (or alteration or cessation) of area-specific mitigation measures

(*e.g.*, domestic shelter-in-place policies) that change the rate of viral transmission within an area and the relative rate of dispersal to other areas. Similarly, average rates of viral dispersal may change in response to the initiation (or alteration or cessation) of more widespread intervention efforts—*e.g.*, multiple area-specific mitigation measures, international-travel bans—that collectively impact the overall viral dispersal rate.

In this paper, we: (1) extend discrete-geographic phylodynamic models to allow both the average and relative dispersal rates to vary independently across pre-specified time intervals; (2) enable stochastic mapping under these interval-specific phylodynamic models to estimate the number and timing of viral dispersal events between areas, and; (3) develop statistics to assess the absolute fit of phylodynamic models to empirical datasets. We first validate the theory and implementation of our new phylodynamic methods using analyses of simulated data, and then provide an empirical demonstration of these methods with analyses of a SARS-CoV-2 dataset from the early phase of the COVID-19 pandemic.

Extending Phylodynamic Models

Anatomy of interval-specific phylodynamic models

Phylodynamic models of dispersal include two main components (Fig. 3.1): a *phylogenetic model* that allows us to estimate a dated phylogeny for the sampled viruses, Ψ, and a *biogeographic model* that describes the history of viral dispersal over the tree as a continuous-time Markov chain. For a geographic history with k discrete areas, this stochastic process is fully specified by a $k \times k$ instantaneous-rate matrix, $\mathbf{Q}$, where an element of the matrix, q_{ij}, is the instantaneous rate of change between state i and state j (*i.e.*, the instantaneous rate of dispersal from area i to area j). We rescale the $\mathbf{Q}$ matrix such that the average rate of dispersal between all areas is μ; this represents the average rate of viral dispersal among all areas (Yang 2014).

We could specify alternative biogeographic models based on the assumed constancy of the dispersal process. For example, the simplest possible model assumes that the average dispersal rate, μ, and the relative dispersal rates, $\mathbf{Q}$, remain constant over the entire history of the viral outbreak. Typically, viral outbreaks are punctuated by events that are likely to impact the average rate of viral dispersal (*e.g.*, the onset of an international-travel ban) and/or the relative rates of viral dispersal between pairs of areas (*e.g.*, the initiation of localized mitigation measures). We can incorporate information on such events into our phylodynamic inference by

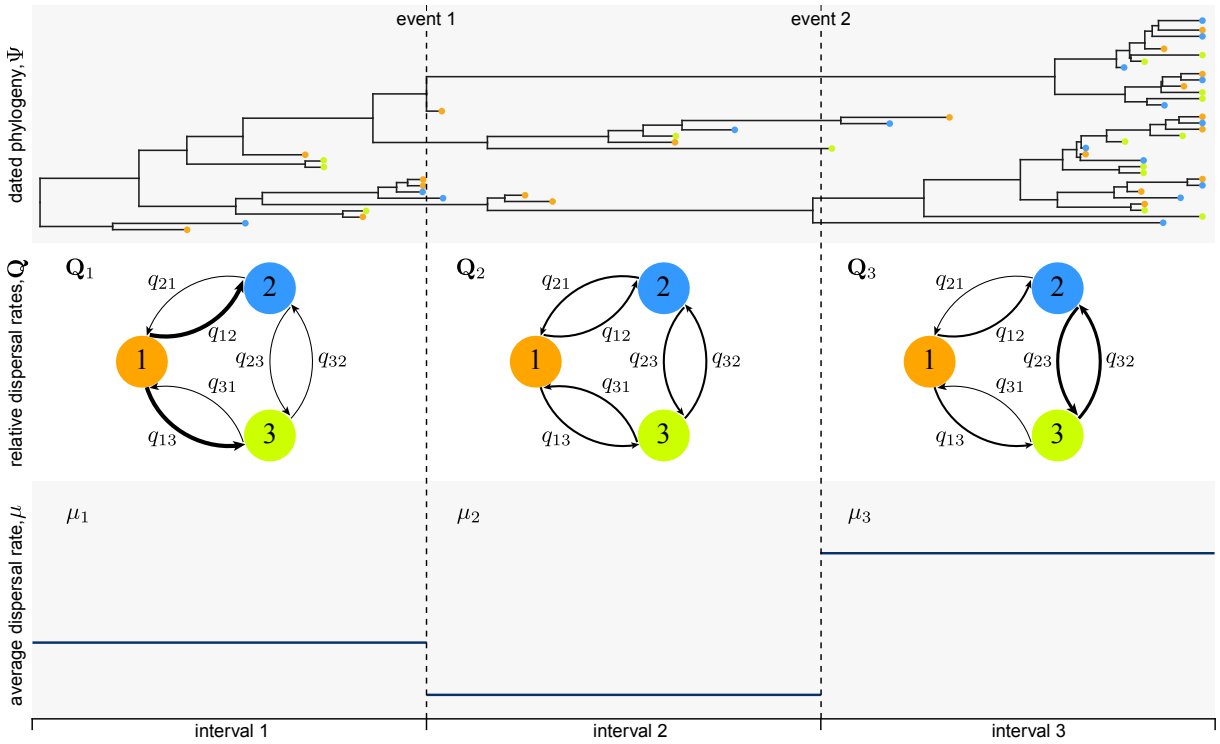

Figure 3.1: Interval-specific phylodynamic models accommodate variation in the process of viral dispersal. Phylodynamic models include two main components: a *phylogenetic model* that specifies the relationships and divergence times of the sampled viruses, Ψ (top panel), and a *biogeographic model* that describes the history of viral dispersal among a set of discrete-geographic areas—here, areas 1 (orange), 2 (blue), and 3 (green)—from the root to the tips of the dated viral tree. Parameters of the biogeographic model include an instantaneous-rate matrix, **Q**, that specifies relative rates of viral dispersal between each pair of areas (here, each element of the matrix, q_{ij}, is represented as an arrow that indicates the direction and relative dispersal rate from area i to area j; middle panel), and a parameter that specifies the average rate of viral dispersal between all areas, μ (lower panel). Although most phylodynamic studies assume that the process of viral dispersal is constant through time, disease outbreaks are typically punctuated by events that impact the average and/or relative rates of viral dispersal among areas. Here, for example, the history involves two events (*e.g.*, mitigation measures) that define three intervals, where both **Q** and μ are impacted by each of these events, such that the interval-specific parameters are $(\mathbf{Q}_1,\mathbf{Q}_2,\mathbf{Q}_3)$ and (μ_1,μ_2,μ_3). Our framework allows investigators to specify phylodynamic models with two or more intervals, where each interval has independent relative and/or average dispersal rates, which are then estimated from the data.

specifying interval-specific models. That is, the investigator specifies the number of intervals, the boundaries between each interval, and the parameters that are specific to each interval according to the presumed changes in the history of viral dispersal. For example, we might specify an interval-specific model (Membrebe et al. 2019) that assumes that the average rate of viral dispersal varies among two or more intervals (while assuming that the relative rates of viral dispersal remain constant across intervals). Conversely, an interval-specific model (Bielejec et al. 2014) might allow the relative rates of viral dispersal to vary among two or more time intervals (while assuming that the average rate of viral dispersal remains constant across intervals). Alternatively, a more complex interval-specific model might allow both the average rate

of viral dispersal and the relative rates of viral dispersal to vary among two or more intervals. We extend interval-specific phylodynamic models to allow *both* the relative *and* average dispersal rates to vary independently across two or more pre-defined intervals. Here, we describe how to compute transition probabilities, perform inference, simulate histories, and assess the absolute fit of interval-specific phylodynamic models.

Computing Transition Probabilities

The transition-probability matrix, $\mathbf{P}$, describes the probability of transitioning from state i to state j (*i.e.*, dispersing from area i to area j) along a branch with a finite duration; importantly, a branch may span two or more intervals with different relative and/or absolute dispersal rates.

Allowing average dispersal rates to vary across intervals.

Under a constant phylodynamic model, the transition-probability matrix for a branch is $\mathbf{P} = \exp(\mathbf{Q}v)$, where $v = \mu t$ represents the expected number of dispersal events on a branch of duration t with an average dispersal rate μ. However, under a phylodynamic model with interval-specific average dispersal rates (Membrebe et al. 2019)—which allows the average dispersal rate to vary among intervals, but assumes that relative dispersal rates are constant across all intervals—a given branch in a phylogeny may span two or more intervals with different average dispersal rates ("average-rate intervals"). The transition-probability matrix for the branch is then computed as the matrix exponential:

$$\mathbf{P} = \exp(\mathbf{Q} \sum_{l=1}^{n} v_l), \tag{3.1}$$

where $\mathbf{Q}$ is the instantaneous-rate matrix, n is the number of average-rate intervals spanned by the branch, and v_l is the expected number of dispersal events on the branch in average-rate interval l. Recall that $v_l = \mu_l t_l$, where μ_l is the average dispersal rate during interval l and t_l is the time spent in interval l.

Allowing relative dispersal rates to vary across intervals

Under a phylodynamic model with interval-specific relative dispersal rates (Bielejec et al. 2014)—which allows the instantaneous-rate of dispersal between each pair of areas to vary among intervals, but assumes that the average dispersal rate is constant across all intervals—a given branch may span two or more intervals with different $\mathbf{Q}$ matrices ("relative-rate intervals"). In this case, the transition-probability matrix for each relative-rate interval l, $\mathbf{P}_l$, is

computed as:

$$\mathbf{P}_l = \exp(\mathbf{Q}_l v_l), \tag{3.2}$$

where $\mathbf{Q}_l$ is the instantaneous-rate matrix in relative-rate interval l, and $v_l = \mu t_l$ is the average dispersal rate multiplied by the time spent in interval l. The transition-probability matrix for the entire branch is then computed as the matrix product of interval-specific transition-probability matrices:

$$\mathbf{P} = \prod_{l=1}^{m} \mathbf{P}_l, \tag{3.3}$$

where m is the number of relative-rate intervals spanned by the branch.

Allowing average and relative dispersal rates to vary across intervals

We combine the two approaches described above to compute transition-probability matrices under an interval-specific model that allows both the average dispersal rate and the relative dispersal rates to vary independently among intervals. Let a given branch span m relative-rate intervals. The expected number of dispersal events in each such interval l, v_l, is computed as:

$$v_l = \sum_{p=1}^{n} \mu_p t_p, \tag{3.4}$$

where n is the number of average-rate intervals spanned by interval l, μ_p is the dispersal rate in average-rate interval p, and t_p is the time spent in average-rate interval p. We then substitute equation (3.4) into equation (3.2), and apply equation (3.3) as normal to compute the transition-probability matrix for the entire branch. An example computation is illustrated in Fig. 3.2 for a scenario in which a branch spans two different relative-rate intervals and three different average-rate intervals.

We modified `BEAST` source code to implement the above equation for computing $\mathbf{P}$ under our interval-specific phylodynamic models that allow both μ and $\mathbf{Q}$ to vary independently among two or more pre-specified intervals.

Inference under interval-specific phylodynamic models

We estimate parameters of the interval-specific phylodynamic models within a Bayesian statistical framework. Specifically, we use numerical algorithms—Markov chain Monte Carlo (MCMC) simulation—to approximate the joint posterior probability distribution of the phylodynamic model parameters—the dated phylogeny, Ψ, the set of relative dispersal rates, $\mathbf{Q}$,

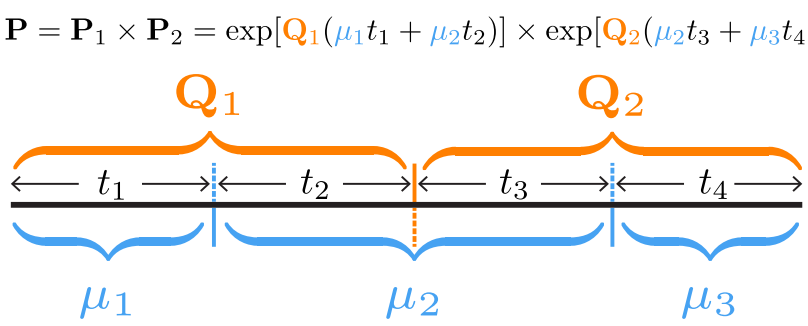

Figure 3.2: Computing the transition-probability matrix for a branch spanning intervals where both the average and relative dispersal rates vary. An example illustrating the transition-probability matrix computation for a branch spanning two relative-rate intervals ($\mathbf{Q}_1$ and $\mathbf{Q}_2$) and three average-rate intervals (μ_1, μ_2, μ_3).

and the average dispersal rates, μ—from the study data (*i.e.*, the location and times of viral sampling, and the genomic sequences of the sampled viruses).

Simulating dispersal histories under interval-specific phylodynamic models

We have also implemented numerical algorithms—stochastic mapping—to simulate histories of viral dispersal under interval-specific phylodynamic models; these methods allow us to estimate the number of dispersal events between a specific pair of areas, the number of dispersal events from one area to a set of two or more areas, and the total number dispersal events among all areas. Stochastic mapping—initially proposed by Nielsen (2002; see also Huelsenbeck et al. 2003; Bollback 2006; Minin and Suchard 2008b; Rodrigue et al. 2007; Hobolth and Stone 2009)—is commonly used to sample dispersal histories over branches of a phylogeny. Here, we extend this approach to sample dispersal histories under our interval-specific models.

Let a given branch start at time T_0 with state i and end at time T_m with state k. Further, let the dispersal process change (either by changing the average or relative dispersal rates) $m-1$ times on the branch at times $\{T_1, \ldots, T_{m-1}\}$, resulting in m intervals. For interval l, denote the average dispersal rate as μ_l, the instantaneous-rate matrix as $\mathbf{Q}_l$, and the duration as t_l. We simulate a dispersal history along this branch using a two-step procedure: (1) we first sample the state at each of the $m-1$ time points, and; (2) we then simulate the history between each time point, conditional on the states sampled in the first step.

To simulate the states at each time point, we first compute a transition-probability matrix for each interval:

$$\mathbf{P}_l = \exp\left(\mathbf{Q}_l \mu_l t_l\right).$$

We then calculate the probability of state j at the first time point, T_1, given that the branch

begins in state i and ends in state k, as:

$$P(j \mid i,k) \propto \mathbf{P}_{ij,1} \times \left[\prod_{l=2}^{m} \mathbf{P}_l\right]_{jk},$$

where the first term is the probability of transitioning from state i (the state at the beginning of the branch) to state j at the first time point, and the second term is the probability of transitioning from state j to state k (the state at the end of the branch) over the remaining time intervals. We compute this for each state j, and sample the state in proportion to these probabilities. We then repeat this process for each remaining time point, recursively conditioning on the state sampled at the previous time point and the state at the end of the branch.

Second, we simulate histories within each interval. For a given time interval, we simulate histories conditional on the start and end states generated in the first step using the uniformization algorithm described by Rodrigue *et al.* (2007; see also Fearnhead and Sherlock 2006; Hobolth and Stone 2009).

Assessing the absolute fit of interval-specific phylodynamic models

For a given phylodynamic study, we might wish to consider several candidate interval-specific models (where each candidate model specifies a unique number of intervals, set of interval boundaries, and/or interval-specific parameters). Comparing the fit of these competing phylodynamic models to the data offers two benefits: (1) confirming that our inference model adequately describes the process that gave rise to data will improve the accuracy of the corresponding inferences (*i.e.*, estimates of relative and/or average dispersal-rate parameters and viral dispersal histories), and; (2) comparing alternative models provides a means to objectively test hypotheses regarding the impact of events on the history of viral dispersal (*i.e.*, by assessing the relative fit of data to competing models that include/exclude the impact of a putative event on the average and/or relative viral dispersal rates). We can assess the *relative* fit of two or more candidate phylodynamic models to a given dataset using Bayes factors; this requires that we first estimate the marginal likelihood for each model (which represents the average fit of a model to a dataset), and then compute the Bayes factor as twice the difference in the log marginal likelihoods of the competing models (Kass and Raftery 1995).

However, even the best candidate model may fail to provide an adequate description of the process that gave rise to our study data. We can leverage our ability to simulate histories under interval-specific phylodynamic models to develop new methods to assess the *absolute*

fit of a candidate phylodynamic model using posterior-predictive assessment (Gelman et al. 1996). This Bayesian approach for assessing model adequacy is based on the following premise: if our inference model provides an adequate description of the process that gave rise to our observed data, then we should be able to use that model to simulate datasets that resemble our original data. The resemblance between the observed and simulated datasets is quantified using a summary statistic. Accordingly, posterior-predictive simulation requires: (1) the ability to simulate geographic datasets under interval-specific phylodynamic models for a given set of parameter values, and; (2) summary statistics that allow us to compare the resulting simulated datasets to the observed dataset. We describe each of these components below.

Simulating under interval-specific phylodynamic models.

We draw m random samples from the joint posterior distribution of the model; each sample i consists of a fully specified phylodynamic model, $\theta_i = \{\Psi_i, \mathbf{Q}_i, \mu_i\}$. For each sample, we simulate a new geographic dataset on the sampled tree, Ψ_i, given the sampled parameters of the geographic model, $\{\mathbf{Q}_i, \mu_i\}$; we label the newly simulated dataset G_i^{sim}. Under a constant phylodynamic model, we simulate full dispersal histories forward in time over a tree using the `sim.history()` function in the `R` package `phytools` (Revell 2012). We implemented an extension of the `sim.history()` function to simulate dispersal histories under interval-specific phylodynamic models. These functions allow us to perform posterior-predictive simulation to assess the adequacy of both the constant and the interval-specific phylodynamic models.

Summary statistics.

We define a summary statistic, which we generically denote $T(G \mid \theta_i)$, where G is either the simulated or observed dataset. For each simulated dataset, we compute a discrepancy statistic,

$$D_i = T(G_i^{\text{sim}} \mid \theta_i) - T(G^{\text{obs}} \mid \theta_i),$$

where G^{obs} is the observed geographic dataset and G^{sim} is a simulated dataset. We developed two summary statistics to assess the adequacy of interval-specific phylodynamic models: (1) the *parsimony statistic*, and; (2) the *tipwise-multinomial statistic*. The parsimony statistic is calculated as the difference in the parsimony score for the observed areas and the simulated areas across the tips of the tree (where the parsimony score is the minimum number of dispersal events required to explain the distribution of areas across the tips of a tree). We compute

parsimony scores using the `parsimony()` function in the `R` package, `phangorn` (Schliep 2010). The tipwise-multinomial statistic is inspired by the multinomial statistic that was proposed by Goldman (1993) and later used by Bollback (2002) to assess the adequacy (absolute fit) of substitution models to sequence alignments. Our tipwise statistic treats the set of states (areas) across the tips of the tree as an outcome of a multinomial trial. Specifically, we calculate the tipwise-multinomial statistic as the difference in the multinomial probabilities for the observed set of areas versus the simulated set of areas across the tips of the tree. We calculate each multinomial probability as:

$$T(G \mid \theta_i) = \sum_{i=1}^{k} n_i \ln(n_i/n),$$

where n is the number of tips in the tree, and n_i is the number of tips that occur area i.

Time-slice summary statistics.

To assess the ability of phylodynamic models to describe the temporal distribution of dispersal events, we extend the parsimony and tipwise-multinomial summary statistics to assess time slices of the geographic history[1]. We calculate these summary statistics for k pre-specified time slices, resulting in k parsimony statistics and k tipwise-multinomial statistics for each simulated dataset. We compute the time-slice variant of the parsimony summary statistic as follows: (1) we first infer the most-parsimonious dispersal history (*i.e.*, the minimum number of dispersal events) for a given simulated dataset and the observed dataset using the `ancestral.pars()` function in the `R` package, `phangorn` (Schliep 2010); (2) we then assign each inferred dispersal event to one of the k time slices based on the time span of the branch along which the dispersal event was inferred (when a dispersal event is inferred to occur along a branch that spans two or more time slices, we locate the event uniformly along the branch, and then assign it to the corresponding slice), and finally; (3) we compute the difference in the number of dispersal events between the simulated and observed dataset for each time slice. We compute the time-slice variant of the tipwise-multinomial summary statistic in a similar manner; *i.e.*, we first find the set of tips in each time slice, and then compute the tipwise-multinomial statistic for that time slice (as described above) for the corresponding set of tips. Further details regarding the

[1]Note that the time slices that we define for summary statistics are distinct from the intervals specified in an interval-specific phylodynamic model. The time slices are motivated to better assess the adequacy of a phylodynamic model, whereas the intervals are motivated to accommodate variation in dispersal dynamics in the empirical data. Accordingly, we might use time-slice summary statistics to assess the adequacy of both constant or interval-specific phylodynamic models.

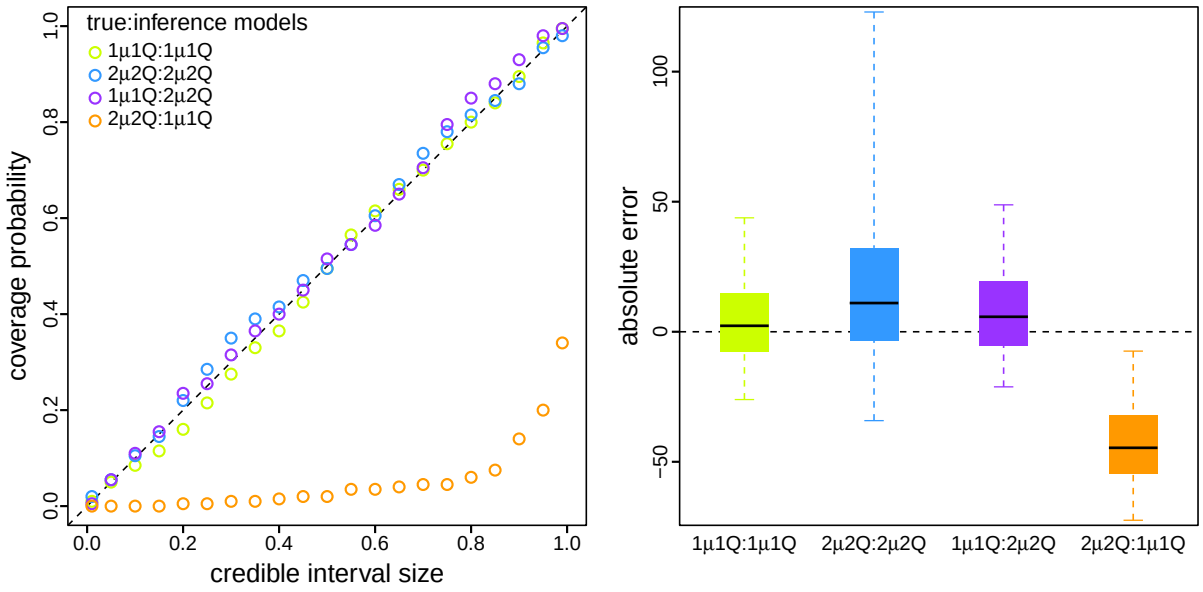

Figure 3.3: Simulation demonstrates that reliable inference of viral dispersal history requires a correctly specified phylodynamic model. We simulated 200 geographic datasets under each of two models: one that assumed a constant μ and $\mathbf{Q}$ ($1\mu1\mathbf{Q}$), and one that allowed μ and $\mathbf{Q}$ to vary over two intervals ($2\mu2\mathbf{Q}$). For each simulated dataset, we separately inferred the total number of dispersal events under each model, resulting in four true:inference model combinations ($1\mu1\mathbf{Q}$:$1\mu1\mathbf{Q}$, $2\mu2\mathbf{Q}$:$2\mu2\mathbf{Q}$, $1\mu1\mathbf{Q}$:$2\mu2\mathbf{Q}$, and $2\mu2\mathbf{Q}$:$1\mu1\mathbf{Q}$). Left) For each combination of true and inference model, we computed the coverage probability (the frequency with which the true number of dispersal events was contained in the corresponding X% credible interval; y-axis) as a function of the size of the credible interval (x-axis). When the model is true, we expect the coverage probability to be equal to the size of the credible interval (Cook et al. 2006). As expected, coverage probabilities fall along the one-to-one line when the model is correctly specified (green and blue). Moreover, coverage probabilities are also appropriate when the inference model is overspecified (*i.e.*, the inference model includes interval-specific parameters not included in the true model; purple). However, coverage probabilities are extremely unreliable when the inference model is underspecified (*i.e.*, the inference model excludes interval-specific parameters of the true model; orange). Right) For each true:inference model combination, we summarized the absolute error (estimated minus true number of dispersal events) as boxplots (median [horizontal bar], 50% probability interval [boxes], and 95% probability interval [whiskers]). Again, when the model is underspecified (orange) inferences are strongly biased compared to those under the correctly specified (green and blue) and overspecified (purple) models.

computation of these summary statistics are available in an `R` script provided in our GitHub and Dryad repositories.

Simulation Study

We performed a simulation study to explore the statistical behavior of the interval-specific phylodynamic models. Specifically, the goals of this simulation study are to assess: (1) our ability to perform reliable inference under interval-specific models; (2) the impact of model misspecification, and; (3) our ability to identify the correct model. To this end, we simulated 200 geographic datasets under each of two models: the first assumes a constant μ and $\mathbf{Q}$ ($1\mu1\mathbf{Q}$), and the second allows μ and $\mathbf{Q}$ to vary over two intervals ($2\mu2\mathbf{Q}$). For each simulated dataset, we separately

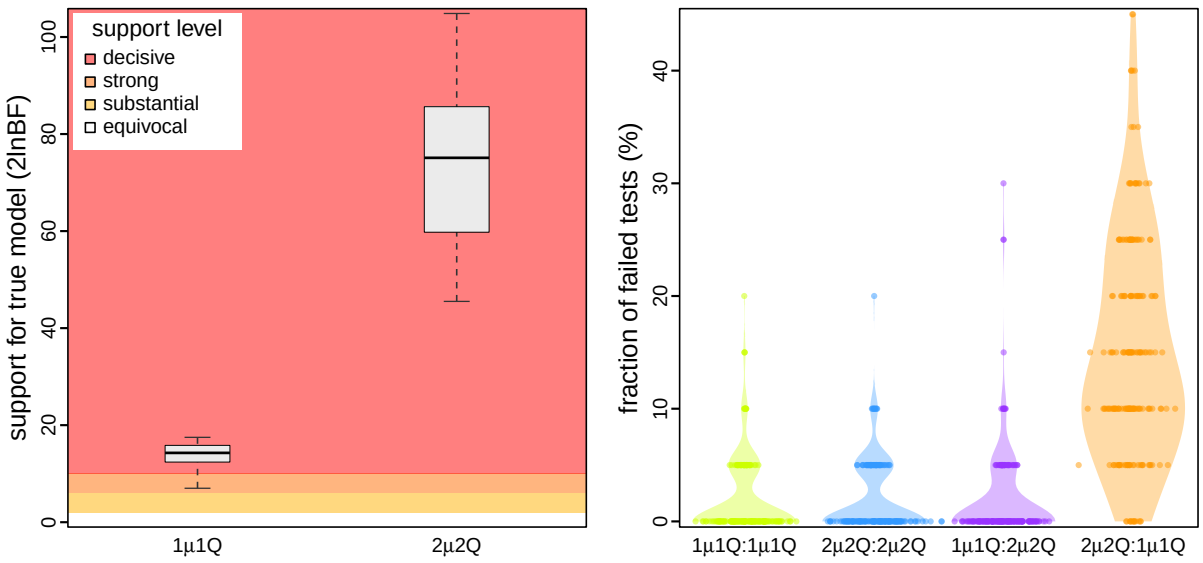

Figure 3.4: Simulation demonstrates our ability to accurately identify a correctly specified phylodynamic model. We assessed the relative and absolute fit of alternative models to the simulated datasets described in Fig. 3.3. Left) For each simulated dataset, we compared the relative fit of the true and alternative models using Bayes factors. The boxplots summarize Bayes factors for datasets simulated under the constant ($1\mu1\mathbf{Q}$, left) and interval-specific ($2\mu2\mathbf{Q}$, right) models, which demonstrate that we are able to decisively identify the true phylodynamic model. Right) For each combination of true:inference model, we assessed absolute model fit using posterior-predictive simulation with a set of 20 summary statistics. Each dot represents the fraction of those 20 summary statistics for which the corresponding inference model provides an inadequate fit to a single simulated dataset. The violin plots summarize the distribution of these values for all datasets under each true:inference model combination. As expected, the true model is overwhelmingly inferred to be adequate (green and blue). Encouragingly, model overspecification appears to have a negligible impact on model adequacy (purple). By contrast, an underspecified model severely impacts model adequacy (orange).

inferred the history of viral dispersal under each model, resulting in four true:inference model combinations: $1\mu1\mathbf{Q}$:$1\mu1\mathbf{Q}$, $2\mu2\mathbf{Q}$:$2\mu2\mathbf{Q}$, $1\mu1\mathbf{Q}$:$2\mu2\mathbf{Q}$, and $2\mu2\mathbf{Q}$:$1\mu1\mathbf{Q}$. We provide detailed descriptions of the simulation analyses and results in Section 2 of the Supplementary Material.

Ability to reliably estimate parameters of interval-specific phylodynamic models

Interval-specific phylodynamic models are inherently more complex than their constant counterparts, and therefore contain many more parameters that must be inferred from geographic datasets that contain minimal information; these datasets only include a single observation (*i.e.*, the area in which each virus was sampled). These considerations raise concerns about our ability to reliably estimate parameters of interval-specific phylodynamic models. Encouragingly, when the inference model is correctly specified (*i.e.*, where both the true and inference models include [or exclude] interval-specific parameters, $2\mu2\mathbf{Q}$:$2\mu2\mathbf{Q}$ and $1\mu1\mathbf{Q}$:$1\mu1\mathbf{Q}$), our simulation study demonstrates that estimates under interval-specific models are as reliable as those under constant models (Fig. 3.3, green, blue). Moreover, when the inference model is overspec-

ified (*i.e.*, it includes interval-specific parameters not included in the true model) inferences are comparable to those under correctly specified models (Fig. 3.3, purple). However, when the inference model is underspecified (*i.e.*, it excludes interval-specific parameters of the true model) inferences are severely biased estimates (Fig. 3.3, orange).

Ability to accurately identify an appropriately specified phylodynamic model

Our simulation study demonstrates the importance of identifying scenarios where an inference model is underspecified; failure to accommodate interval-specific variation in the study data can severely bias parameter estimates. Fortunately, our simulation study demonstrates that we can reliably identify when a given model is correctly specified, overspecified, or underspecified using a combination of Bayes factors (to assess the relative fit of competing models to the data; Fig. 3.4, left) and posterior-predictive simulation (to assess the absolute fit of each candidate model to the data; Fig. 3.4, right). Using a combination of Bayes factors and posterior-predictive simulation allows us to not only identify the best of the candidate models, but also to ensure that the best model provides an adequate description of the true process that gave rise to our study data.

Empirical Application

We illustrate our new phylodynamic methods with analyses of all publicly available SARS-CoV-2 genomes sampled during the early phase of the COVID-19 pandemic (with 2598 viral genomes collected from 23 geographic areas between Dec. 24, 2019–Mar. 8, 2020 [downloaded from GISAID, Shu and McCauley 2017]). We used our study dataset to estimate the parameters of—and assess the relative and absolute fit to—nine candidate phylodynamic models. These models assign interval-specific parameters—for the average rate of viral dispersal, μ, and/or relative rates of viral dispersal, $\mathbf{Q}$—to one, two, four, or five pre-specified time intervals; *i.e.*, $1\mu 1\mathbf{Q}$, $2\mu 1\mathbf{Q}$, $1\mu 2\mathbf{Q}$, $2\mu 2\mathbf{Q}$, $4\mu 1\mathbf{Q}$, $1\mu 4\mathbf{Q}$, $4\mu 4\mathbf{Q}$, $5\mu 5\mathbf{Q}$, and $5\mu 5\mathbf{Q}^*$. We specified interval boundaries based on external information regarding events within the study period that might plausibly impact viral dispersal dynamics, including: (A) start of the Spring Festival travel season in China (the highest annual period of domestic travel, Jan. 12); (B) onset of mitigation measures in Hubei province, China (Jan. 26); (C) onset of international air-travel restrictions against China (Feb. 2), and; (D) relaxation of domestic travel restrictions in China (Feb. 16). Phylodynamic models with two intervals include event C, models with four intervals include events

A, C, and D, and the 5μ5**Q** model includes all four events. The final candidate model, 5μ5**Q***, includes five arbitrary and uniform (bi-weekly) intervals. We provide detailed descriptions of our empirical data collection, analyses, and results in Section 3 of the SI Appendix.

An interval-specific model best describes viral dispersal in the early phase of the pandemic

Our phylodynamic analyses of the SARS-CoV-2 dataset reveal that the early phase of the COVID-19 pandemic exhibits significant variation in both the average and relative rates of viral dispersal over four time intervals. Bayes factor comparisons (Fig. 3.5, left) demonstrate that the 4μ4**Q** interval-specific model is decisively preferred both over all less complex candidate models—including models that allow *either* the average dispersal rate *or* relative dispersal rates to vary over the same four intervals (4μ1**Q** and 1μ4**Q**, respectively)—and also over more complex candidate models (5μ5**Q**, and 5μ5**Q***). Posterior-predictive analyses (Fig. 3.5, right) demonstrate that the preferred model, 4μ4**Q**, also provides an adequate description of the process that gave rise to our SARS-CoV-2 dataset. Below, we will use the preferred (4μ4**Q**) interval-

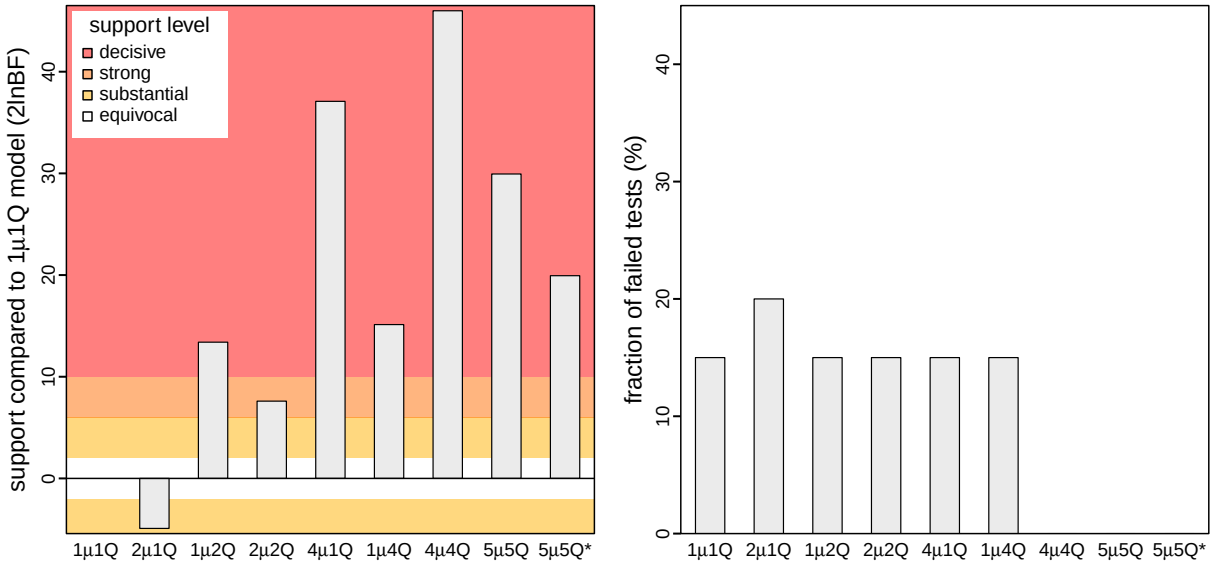

Figure 3.5: An interval-specific model provides the best relative and absolute fit to our SARS-CoV-2 dataset. We assessed the relative and absolute fit of nine candidate phylodynamic models to our study dataset (comprised of all publicly available SARS-CoV-2 genomes from the early phase of the COVID-19 pandemic). Left) We compared the relative fit of each candidate model to the constant (1μ1**Q**) phylodynamic model using Bayes factors, which indicate that the 4μ4**Q** interval-specific model outcompetes both less complex and more complex models. Right) We performed posterior-predictive simulation for each candidate model using 20 summary statistics, plotting the fraction of those summary statistics indicating that a given candidate model was inadequate. Our results indicate that three candidate models (4μ4**Q**, 5μ5**Q**, and 5μ5**Q***) provide an adequate fit to our SARS-CoV-2 dataset. The simplest of these adequate models (4μ4**Q**) also provides the best relative fit. Collectively, these results identify the 4μ4**Q** model as the clear choice for phylodynamic analyses of our study dataset.

specific phylodynamic model to explore various aspects of viral dispersal during the early phase of the COVID-19 pandemic and—for the purposes of comparison—we also present corresponding results inferred using the (underspecified) constant ($1\mu 1\mathbf{Q}$) phylodynamic model.

Variation in global viral dispersal rates

Between late 2019 and early March, 2020, COVID-19 emerged (in Wuhan, China) and established a global distribution—with reported cases in 83% of the study areas by this date (WHO 2020)—despite the implementation of numerous intervention efforts to slow the spread of the causative SARS-CoV-2 virus (Hsiang et al. 2020). This crucial early phase of the pandemic provides a unique opportunity to explore the dispersal dynamics that led to the worldwide establishment of the virus and to assess the efficacy of key public-health measures to mitigate the spread of COVID-19. The constant ($1\mu 1\mathbf{Q}$) model infers a static rate of global viral dispersal throughout the study period (Fig. 3.6, orange). By contrast, inferences under the preferred ($4\mu 4\mathbf{Q}$) model reveal significant variation in global viral dispersal rates over four intervals, exhibiting both increases and decreases over the early phase of the pandemic (Fig. 3.6, dark blue). The significant decrease in the global viral dispersal rate between the second and third interval

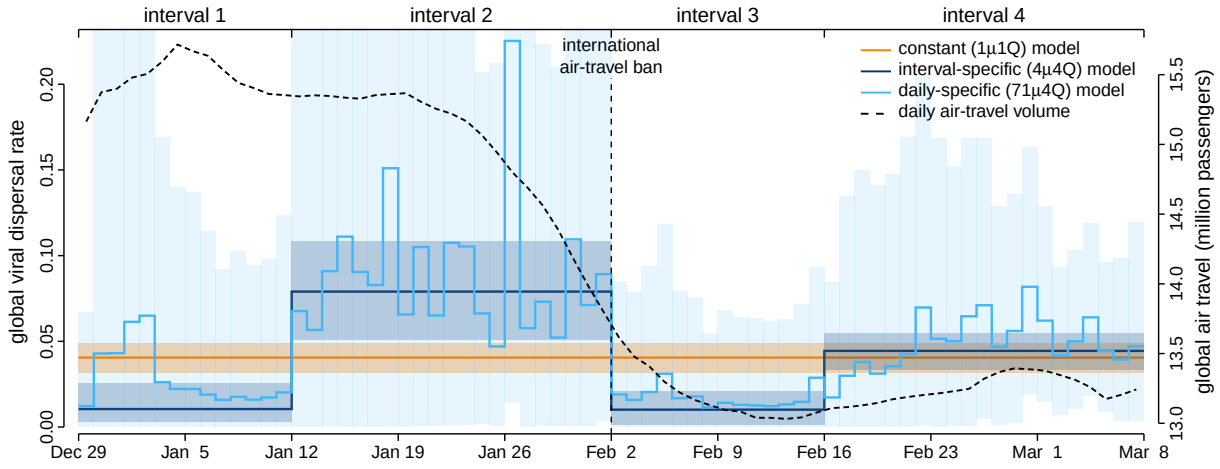

Figure 3.6: Patterns and correlates of variation in global viral dispersal rate early in the COVID-19 pandemic. The COVID-19 pandemic emerged in Wuhan, China, in late 2019, and established a global distribution by Mar. 8, 2020. Our phylodynamic analyses of this critical early phase of the pandemic provide estimates of the average rate of viral dispersal across all 23 study areas, μ (posterior mean [solid lines], 95% credible interval [shaded areas]). By assumption, the constant ($1\mu 1\mathbf{Q}$) model infers a static rate of global viral dispersal (orange). By contrast, the preferred interval-specific ($4\mu 4\mathbf{Q}$) model reveals significant variation in the global viral dispersal rate (dark blue). Notably, the global viral dispersal rate decreases sharply on Feb. 2, which coincides with the onset of international air-travel bans with China. The efficacy of these air-travel restrictions is further corroborated by estimates of daily global viral dispersal rates (light blue)—inferred under a more granular, interval-specific ($71\mu 4\mathbf{Q}$) model—that are significantly correlated with independent information on daily global air-travel volume (dashed line, obtained from FlightAware).

(with a boundary at Feb. 2) coincides with the initiation of international air-travel bans with China (imposed by 34 countries and nation states by this date). To further explore the possible impact of the air-travel ban on the global spread of COVID-19, we inferred daily rates of global viral dispersal under a more granular interval-specific model (71μ4**Q**; Fig. 3.6, light blue). Our estimates of daily rates of global viral dispersal are significantly correlated with independent information on daily global air-travel volume (Fig. 3.6, dashed) over the interval from Jan. 31 (when the virus first achieved a cosmopolitan distribution; WHO 2020) to the end of our study period (see the supplementary material for detailed descriptions of the correlation test and results).

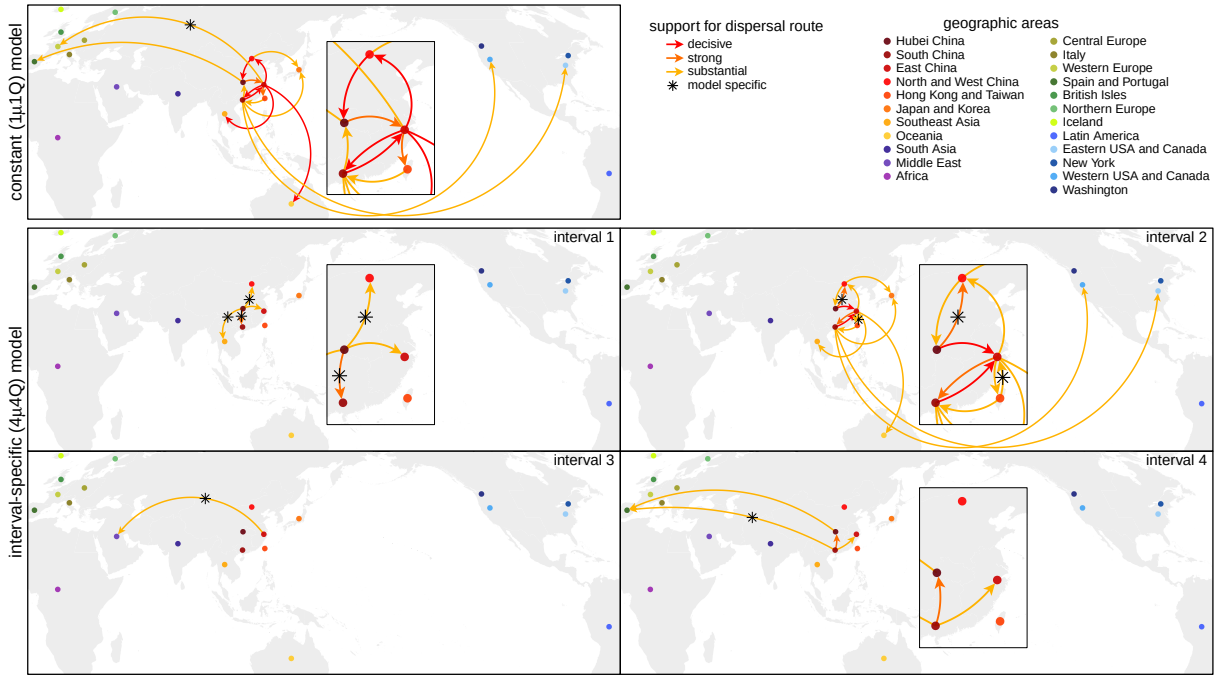

Figure 3.7: Variation in viral dispersal routes involving China during the early phase of the pandemic. Arrows indicate routes inferred to play a significant role in viral dispersal to/from China during the early phase of the COVID-19 pandemic; colors indicate the level of evidential support for each dispersal route (as 2 ln Bayes factors). We focus on dispersal routes involving China both because it was the point of origin, and because it was the area against which travel bans were imposed. The number, duration, and significance of dispersal routes inferred under the constant (1μ1**Q**) model differ strongly from those inferred under the preferred (4μ4**Q**) interval-specific model. By assumption, the constant (1μ1**Q**) model implies an invariant set of dispersal routes. By contrast, the preferred (4μ4**Q**) interval-specific model reveals that the number and intensity of dispersal routes varied over the four intervals. The first interval (Nov. 17–Jan. 12) is dominated by dispersal from Hubei to other areas in China, and the second interval (Jan. 12–Feb. 2) exhibits more widespread international dispersal originating from China. The third interval (Feb. 2–Feb. 16)—immediately following the onset of international air-travel bans with China—exhibits a sustained reduction in the number of dispersal routes. Note that the constant model infers a spurious dispersal route from East China to West Europe. Conversely, the preferred interval-specific model reveals six significant dispersal routes (not detected under the constant model) that imply a more significant role for Hubei as a source of viral spread in the first and second intervals, and also reveals additional dispersal routes emanating from China (to the Middle East in the third interval and to Spain/Portugal in the fourth interval).

Variation in viral dispersal routes

In addition to revealing differences in the global viral dispersal rate, our interval-specific phylodynamic models allow us to explore how relative dispersal rates vary through time. Specifically, our analyses allow us to identify the dispersal routes by which the SARS-CoV-2 virus achieved a global distribution during the early phase of the COVID-19 pandemic. We focus on dispersal routes involving China both because it was the point of origin, and because it was the area against which travel bans were imposed. Inferences under the constant ($1\mu1\mathbf{Q}$) and preferred ($4\mu4\mathbf{Q}$) phylodynamic models imply strongly contrasting viral dispersal dynamics (Fig. 3.7). In contrast to the invariant set of dispersal routes identified by the constant model, the preferred interval-specific model reveals that the number and intensity of dispersal routes varied significantly over the four intervals, with a sharp decrease in the number of dispersal routes following the onset of air-travel bans on Feb. 2. Moreover, the constant model infers one spurious dispersal route, while failing to identify six significant dispersal routes; the preferred model implies a more significant role for Hubei as a source of viral spread in the first and second intervals and reveals additional viral dispersal routes originating from China in the third and fourth intervals. The patterns of variation in dispersal routes among all 23 study areas are similar to—but more pronounced than—those involving China; *e.g.*, where the constant model infers a total of nine spurious dispersal routes, and the interval-specific model reveals a total of ten significant dispersal routes that were not detected by the constant model (Figs. S.3.16 and S.3.17).

Variation in the number of viral dispersal events

Our phylodynamic analyses also allow us to infer the number of SARS-CoV-2 dispersal events between areas during the early phase of the COVID-19 pandemic. Here, we focus on the number of viral dispersal events originating from China because it was the point of origin and primary source of SARS-CoV-2 spread early in the pandemic. The constant ($1\mu1\mathbf{Q}$) and preferred ($4\mu4\mathbf{Q}$) phylodynamic models infer distinct trends in—and support different conclusions regarding the impact of mitigation measures on—the number of viral dispersal events out of China. The constant model infers a gradual decrease in the number of dispersal events from late Jan. through mid-Feb. (Fig. 3.8, orange). By contrast, the preferred interval-specific model reveals a sharp decrease in the number of dispersal events on Feb. 2, which coincides with the

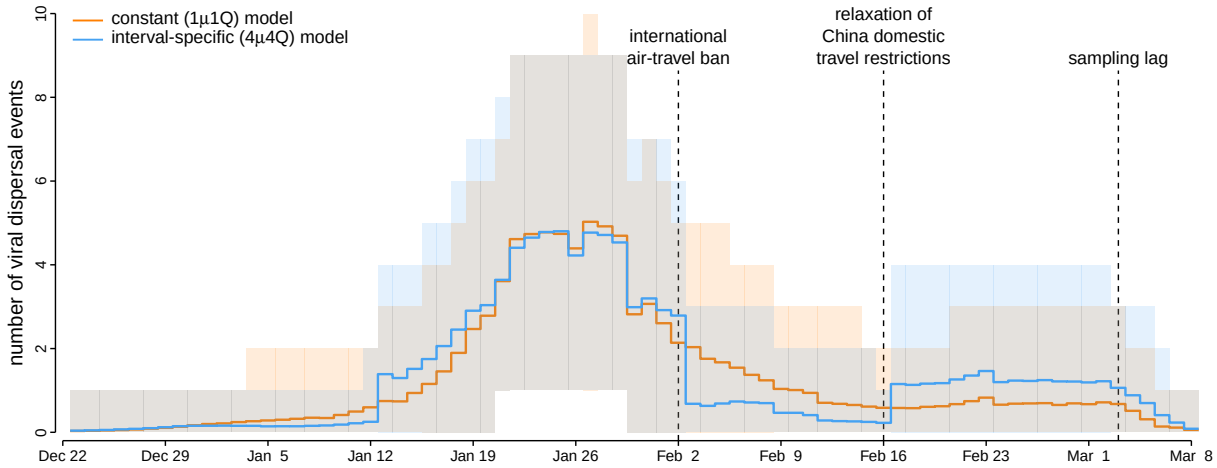

Figure 3.8: Variation in the number of viral dispersal events out of China early in the COVID-19 pandemic. Our phylogenetic analyses of SARS-CoV-2 genomes sampled during the early phase of the COVID-19 pandemic allow us to estimate the number of viral dispersal events from China to all other study areas (posterior mean [solid lines], 95% credible interval [shaded areas]). The constant ($1\mu\mathbf{1Q}$) model implies that the number of viral dispersal events emanating from China remained relatively high following the onset of international air-travel bans on Feb. 2 (orange). By contrast, the preferred interval-specific ($4\mu\mathbf{4Q}$) model reveals that the number of viral dispersal events emanating from China decreased sharply on Feb. 2 (blue), which supports the efficacy of these international air-travel restrictions. The preferred model also infers an uptick in the number of viral dispersal events on Feb. 17 (not detected by the constant model), which coincides with the relaxation of domestic travel restrictions in China. Note that sampling lag causes the number of dispersal events near the end of the sampling period to be underestimated.

onset of air-travel bans imposed against China (Fig. 3.8, blue). Moreover, the preferred phylodynamic model infers an uptick in the number of viral dispersal events on Feb. 17 (not detected by the constant model), which coincides with the lifting of domestic travel restrictions within China (except for Hubei, where the travel restrictions were enforced through late Mar.).

DISCUSSION

Phylodynamic methods increasingly inform our understanding of the spatial and temporal dynamics of viral spread. The vast majority of discrete-geographic phylodynamic studies assume—despite direct (and compelling) evidence to the contrary—that disease outbreaks are intrinsically constant: ≈98% of all such studies are based on constant phylodynamic models. These considerations have motivated previous extensions of phylodynamic models that allow *either* the average (Membrebe et al. 2019) *or* relative (Bielejec et al. 2014) dispersal rates to vary, and our development of more complex phylodynamic models that allow *both* the average and relative dispersal rates to vary independently over two or more pre-specified intervals. By accommodating ubiquitous temporal variation in the dynamics of disease outbreaks—and by allowing us to incorporate independent information regarding events that may impact viral

dispersal—our new interval-specific phylodynamic models are more realistic (providing a better description of the processes that gives rise to empirical datasets), thereby enhancing the accuracy of our epidemiological inferences based on these models.

Our simulation study demonstrates that (in principle): (1) we are able to accurately identify when phylodynamic models are correctly specified, overspecified, or underspecified (Fig. 3.4); (2) when the phylodynamic model is correctly specified, we are able to reliably estimate parameters of these more complex interval-specific phylodynamic models (Fig. 3.3), and; (3) when the phylodynamic model is underspecified, failure to accommodate interval-specific variation in the study data can bias parameter estimates and mislead inferences about viral dispersal history based on those biased estimates (Fig. 3.3).

Our empirical study of SARS-CoV-2 data from the early phase of the COVID-19 pandemic demonstrates that (in practice): (1) our interval-specific phylodynamic model (where *both* the global rate of viral dispersal *and* the relative rates of viral dispersal vary over four distinct intervals) significantly improves the relative and absolute fit to our study dataset compared to constant phylodynamic models (Lemey et al. 2009; Edwards et al. 2011) and to phylodynamic models that allow *either* the average dispersal rate (Membrebe et al. 2019) *or* the relative dispersal rates (Bielejec et al. 2014) to vary over the same four intervals; (2) the preferred interval-specific phylodynamic model provides qualitatively different insights on key aspects of viral dynamics during the early phase of the pandemic—on global rates of viral dispersal (Fig. 3.6), viral dispersal routes (Fig. 3.7), and the number of viral dispersal events (Fig. 3.8)—compared to conventional estimates based on constant (and underspecified) phylodynamic models, and; (3) inferences under the preferred interval-specific phylodynamic model support qualitatively different conclusions regarding the impact of mitigation measures to limit the spread of the COVID-19 pandemic; *e.g.*, the variation in global viral dispersal rate, viral dispersal routes, and number of viral dispersal events revealed by the interval-specific model (but masked by the constant model) collectively support the efficacy of the international air-travel bans in slowing the progression of the COVID-19 pandemic.

Our interval-specific models promise to enhance the accuracy of phylodynamic inferences not only by virtue of their increased realism, but also by allowing us to incorporate additional information (related to events in the history of disease outbreaks) in our phylodynamic inferences. The ability to incorporate independent/external information is particularly valuable

for phylodynamic inference—where many parameters must be estimated from datasets with limited information—which has also motivated the development of other innovative phylodynamic approaches for incorporating external information (Lemey et al. 2014; Bielejec et al. 2016). The potential benefit of harnessing external information is evident in our empirical study: our inference model—$4\mu 4\mathbf{Q}$, with four intervals that we specified based on external evidence regarding events that might plausibly impact viral dispersal dynamics—is decisively preferred ($2\ln \mathrm{BF} = 27.3$) over a substantially more complex model, $5\mu 5\mathbf{Q}^*$, with five *arbitrarily* specified (14-day) intervals.

Importantly, comparison of alternative interval-specific phylodynamic models provides a powerful framework for testing hypotheses about the impact of various events (*i.e.*, assessing the efficacy of mitigation measures) on viral dispersal dynamics. Our empirical study allows us, for example, to assess the impact of domestic mitigation measures imposed in the Hubei province of China. This simply involves comparing the relative fit of our data to two candidate phylodynamic models; $4\mu 4\mathbf{Q}$ and $5\mu 5\mathbf{Q}$. The $5\mu 5\mathbf{Q}$ model adds an interval (corresponding to the onset of the Hubei lockdown on Jan. 26) to the otherwise identical $4\mu 4\mathbf{Q}$ model. In contrast to the international air-travel ban, this domestic mitigation measure does not appear to have significantly impacted global SARS-CoV-2 dispersal dynamics: the $5\mu 5\mathbf{Q}$ model is decisively rejected when compared to the $4\mu 4\mathbf{Q}$ model ($2\ln \mathrm{BF} = -15.9$).

We have focused on interval-specific models where each interval involves a change in both the average and relative dispersal rates. For example, the scenario depicted in Fig. 3.1 involves two events that define three intervals, where both $\mathbf{Q}$ and μ are impacted by each event, such that the interval-specific parameters are $(\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_3)$ and (μ_1, μ_2, μ_3). However, our interval-specific models also allow the average and relative dispersal rates to vary *independently* across intervals. For example, under an alternative scenario for Fig. 3.1, the first event may have impacted both the relative and average dispersal rates, $\mathbf{Q}$ and μ, whereas the second event may have only changed the relative dispersal rates, $\mathbf{Q}$; in this case, the interval-specific parameters would be $(\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_3)$ and (μ_1, μ_2, μ_2). Allowing dispersal rates to vary independently enables these models to accommodate more complex patterns of variation in empirical datasets (and thereby improve estimates from these more realistic models), and also provides tremendous flexibility for testing hypotheses about the impact of various mitigation measures on *either* the relative and/or average rates of viral dispersal.

Nevertheless, this flexibility comes at a cost: interval-specific models are inherently more complex than their constant counterparts, with many parameters that must be estimated from minimal data (*i.e.*, the geographic location of each virus). Accordingly, careful model selection and validation is necessary to avoid specification of an over-parameterized model. Moreover, the space of phylodynamic models expands rapidly as we increase the number of intervals. For a model with three intervals, for example, we can specify five allocations for the average dispersal rate parameter, μ—(μ_1, μ_1, μ_1), (μ_1, μ_1, μ_2), (μ_1, μ_2, μ_1), (μ_1, μ_2, μ_2), and (μ_1, μ_2, μ_3)—and, similarly, five allocations for the relative dispersal rate parameter, $\mathbf{Q}$: $(\mathbf{Q}_1, \mathbf{Q}_1, \mathbf{Q}_1)$, $(\mathbf{Q}_1, \mathbf{Q}_1, \mathbf{Q}_2)$, $(\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_1)$, $(\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_2)$, and $(\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_3)$. We can therefore specify 25 unique three-interval phylodynamic models (representing all combinations of the two parameter-allocation vectors), 225 unique four-interval models, 2704 unique five-interval models, 41209 unique six-interval models, etc. Accordingly, the effort required to identify the best interval-specific phylodynamic model quickly becomes prohibitive, particularly because this search requires that we estimate the marginal likelihood for each candidate model using computationally intensive methods (Xie et al. 2011; Baele et al. 2012). Nevertheless, our interval-specific models establish a foundation for developing more computationally efficient methods; *e.g.*, we could pursue a finite-mixture approach (Kazmi and Rodrigue 2019) that averages inferences of dispersal dynamics over the space of all possible interval-specific phylodynamic models with a given number of intervals.

We are optimistic that—by increasing (and providing a means to assess) model realism, incorporating additional information, and providing a powerful and flexible means to test alternative models/hypotheses—our phylodynamic methods will greatly enhance our ability to understand the dynamics of viral spread, and thereby inform policies to mitigate the impact of disease outbreaks.

Data and Code Availability

GISAID accession IDs of the SARS-CoV-2 sequences used in this study, as well as the flight-volume data (obtained from FlightAware, LLC) and intervention-measure data, are maintained in the GitHub repository (`https://github.com/jsigao/interval_specific_phylodynamic_models_supparchive`) and archived in the Dryad repository (`https://doi.org/10.25338/B89P9K`). Our repositories also contain `BEAST` XML scripts used to perform the phylodynamic

analyses, `R` scripts used to perform simulations and post processing, and a modified version of the `BEAST` program used for some of the analyses in this study.

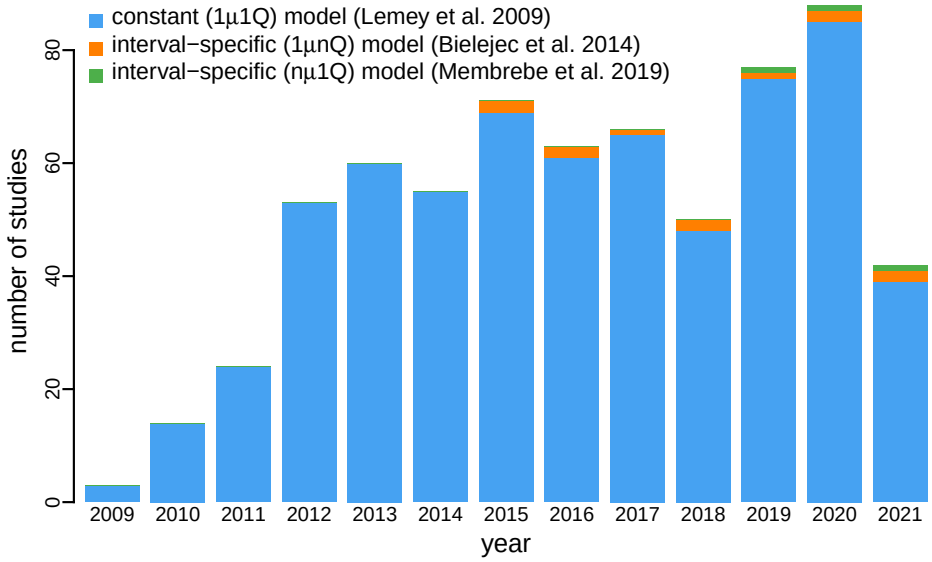

Figure S.3.1: Cited-reference search of empirical discrete-geographic phylodynamic studies. The bar plot summarizes the number of published discrete-geographic studies (obtained from Google Scholar on June 30, 2021) that have inferred biogeographic history using the constant phylodynamic model (blue; Lemey et al. 2009), an interval-specific model that allows only relative dispersal rates to vary (orange; Bielejec et al. 2014), or an interval-specific model that allows only the average dispersal rate to vary (green; Membrebe et al. 2019). The vast majority of these studies (651 of 666, 97.7%) are based on the constant model, whereas only ∼2% of the studies used either of the interval-specific models.

Extending Phylodynamic Methods

We presented the theoretical base of the interval-specific phylodynamic models in the main text; here we focus on the implementation extensions of `BEAST` that enable the inference and simulation under the model. Specifically, our implementation allows the transition-probability matrix, **P**, to be computed and the dispersal history to be simulated correctly along a branch when it spans multiple time intervals with different relative and/or absolute dispersal rates. We provide an executable `BEAST` program with our extensions in our GitHub and Dryad repositories.

Computing the transition-probability matrix when both average and relative dispersal rates vary across intervals

Prior to our extension, `BEAST` computed the **P** matrix correctly (except for a couple of programming issues that we will describe below) along a branch spanning multiple relative-rate intervals *or* multiple average-rate intervals, but not *both*. Let a given branch of length t time units

span m relative-rate intervals and n average-rate intervals. The mean of the average dispersal rate over this branch is thus computed as:

$$\bar{\mu} = \frac{\sum_{p=1}^{n} \mu_p t_p}{t}, \tag{S.3.1}$$

where μ_p and t_p are the average dispersal rate and the time that the branch spent in average-rate interval p, respectively. `BEAST` previously computed the transition-probability matrix for each relative-rate interval l, $\mathbf{P}_l$, as:

$$\mathbf{P}_l = \exp(\mathbf{Q}_l \bar{\mu} t_l), \tag{S.3.2}$$

where $\mathbf{Q}_l$ is the instantaneous-rate matrix in relative-rate interval l and t_l represents the time that the branch spent in relative-rate interval l. In other words, this way of computing P_l effectively assumed that the average dispersal rate in each relative-rate interval was identical across all relative-rate intervals, which is not correct when the branch spans *both* multiple relative-rate intervals *and* multiple average-rate intervals.

To correctly compute the $\mathbf{P}$ matrix, we modified `BEAST` source code to set t_p to the time that *relative-rate interval* l (instead of the *branch*) spent in average-rate interval p, rather than the time that the *branch* spent in average-rate interval p. Implementation details and source-code edits are available in this pull request to the source-code repository of `BEAST`. Putting all the steps together, the transition-probability matrix is now computed as:

$$\mathbf{P} = \prod_{l=1}^{m} \exp[\mathbf{Q}_l (\sum_{p=1}^{n_l} \mu_p t_p)], \tag{S.3.3}$$

where n_l is the number of average-rate intervals spanned by interval l, μ_p is the dispersal rate in average-rate interval p, and t_p is the time that relative-rate interval l spent in average-rate interval p.

Other relevant programming limitations that prohibited inferences under the interval-specific model

In addition, we identified and fixed two programming bugs that hinder correct inferences under interval-specific phylodynamic models: the first one matters when the relative dispersal rates vary across intervals, while the second one can be problematic even when the phylodynamic model is constant, but may be exacerbated when both average and relative dispersal rates vary across intervals.

Programming bug 1: **P**-matrix ordering along a branch

The transition-probability matrix for the entire branch is computed as the matrix product of interval-specific transition-probability matrices:

$$\mathbf{P} = \prod_{l=1}^{m} \mathbf{P}_l, \tag{S.3.4}$$

Since matrix multiplication is not commutative, these **P** matrices should be ordered from the parent node to the child node of the branch, *i.e.*, forward in time, as shown in Eq. 7 of Bielejec et al. (2014). However, the vector of **Q**-matrix indices along a branch returned by the `getBranchModelMapping` function of the interval-specific relative dispersal rates (`EpochalBranchModel`) class was ordered from the child node to the parent node; there was no reversal of the order prior to or during the computation of **P** matrices, resulting in an incorrect transition-probability matrix computed for the entire branch. We fixed this issue by reversing the order of **Q** matrices that will be returned by the `getBranchModelMapping` function. Implementation details and source-code edits of this fix are available in this pull request to the source-code repository of `BEAST`.

Programming bug 2: rescaling an asymmetric **Q** matrix

By convention, we rescale the **Q** matrix such that the average rate of dispersal between all areas is μ, which is computed as:

$$\mu = -\sum_{i=1}^{k} \pi_i q_{ii}, \tag{S.3.5}$$

where k is the number of discrete areas, π_i is the stationary frequency of area i, and q_{ii}—the diagonal element of row i of **Q**—is the negative of the total rate of leaving area i (*i.e.*, $q_{ii} = -\sum_{j \neq i} q_{ij}$). (Note that the implicit assumption here is that the **Q** matrix is irreducible, which guarantees the existence and uniqueness of the stationary distribution, π.) After rescaling, **Q** becomes an instantaneous-rate matrix of relative dispersal rates whose average rate of dispersal is one; thus μ represents the average dispersal rate (among areas) in units of expected number of dispersal events per unit time. As the **Q** matrix is now constrained, μ is a free parameter of the model.

Therefore, to rescale **Q**, we need to know π. π can be determined from **Q** by solving:

$$\pi \mathbf{Q} = 0. \tag{S.3.6}$$

This way of determining π is unnecessary when $\mathbf{Q}$ is symmetric (where the rate of dispersal from area i to area j is identical to the rate of dispersal from area j to area i) or time reversible (*e.g.*, under the GTR substitution model; Tavaré 1986). Depending on the specified model, `BEAST` thus treats π either as a constant uniform vector (when $\mathbf{Q}$ is symmetric) or a model parameter that will be used to construct $\mathbf{Q}$ and directly sampled in the MCMC (when $\mathbf{Q}$ is not symmetric but time reversible).

Conversely, when $\mathbf{Q}$ is asymmetric (Edwards et al. 2011)—where the rate of dispersal from area i to area j is different from the rate of dispersal from area j to area i (*i.e.*, q_{ij} is different from q_{ji})—π is necessarily nonuniform nor an explicit model parameter, so it needs to be computed using (S.3.6). However, previously `BEAST` did not allow this option; π had to be explicitly specified as either a constant vector or model parameter in the XML file, and this specified π would be used to rescale $\mathbf{Q}$. As a result, the average rate of disperal for the $\mathbf{Q}$-matrix would depart from 1, which risks conflating relative-rate matrix variation from overall dispersal-rate variation. We modified `BEAST` source code to allow π to be provided optionally, and to rescale an asymmetric $\mathbf{Q}$ using a stationary distribution computed from (S.3.6) through LU decomposition by adding a `computeStationaryDistribution` function to the asymmetric substitution model (`ComplexSubstitutionModel`) class. Implementation details and source-code edits of this fix are available in this pull request to the source-code repository of `BEAST`.

Stochastic mapping when both average and relative dispersal rates vary across intervals

Stochastic mapping, initially proposed by Nielsen (2002; see also Huelsenbeck et al. 2003; Bollback 2006), is commonly used to sample dispersal histories over branches of a phylogeny conditioned on the observed tip geographic areas. `BEAST` implements the endpoint-conditioned uniformization stochastic-mapping algorithm (Rodrigue et al. 2007; Fearnhead and Sherlock 2006; Hobolth and Stone 2009) to simulate full dispersal histories over the phylogeny, and a simulation-free algorithm (Minin and Suchard 2008a,b) to compute the expected number of dispersal events ('Markov jumps') and the expected time spent in each geographic area ('Markov rewards'). Here we focus on inferring the full dispersal history using the simulation-based stochastic-mapping algorithm.

A full dispersal history is sampled every certain number of generations (specified in the XML file) during an MCMC. At a sampling generation, the geographic state at each inter-

nal node of the phylogeny is first simulated using the ancestral-state-reconstruction algorithm (Yang et al. 1995; Huelsenbeck and Bollback 2001; Pagel et al. 2004) based on parameter values sampled at that generation. (Note that the implementation extensions and programming issues described above in this section that underlie the computation of transition-probability matrix under the interval-specific model also affect the ancestral-state reconstruction as the probability of transitioning from the start state to the end state of a branch is used to sample the end state conditioning on the start state, propagating the sampled root state to the tips of the phylogeny.) The stochastic-mapping algorithm is then responsible for simulating the dispersal history over each branch of the phylogeny, conditioning on the start and end states of each branch.

Let a given branch of length t time units start at time T_0 with state i and end at time T_m with state k. Further, let the dispersal process change (either by changing the average or relative dispersal rates) $m-1$ times on the branch at times $\{T_1, \ldots, T_{m-1}\}$, resulting in m intervals. For interval l, denote the average dispersal rate as μ_l, the instantaneous-rate matrix as $\mathbf{Q}_l$, and the duration as t_l. Prior to our extension, `BEAST` performed stochastic mapping along the branch using the same routine regardless whether the model was constant or interval-specific; *i.e.*, it used a single μ and a single $\mathbf{Q}$ to perform the simulation. Specifically, the single μ was assumed to be the average of the average dispersal rates spanned by the branch, computed as:

$$\bar{\mu} = \frac{\sum_{l=1}^{m} \mu_l t_l}{t}, \tag{S.3.7}$$

and the single $\mathbf{Q}$ was assumed to be $\mathbf{Q}_m$, the instantaneous-rate matrix of the last interval spanned by the branch.

We resolved this issue by adding a new routine to the `MarkovJumpsBeagleTreeLikelihood` class of `BEAST`, which simulates a dispersal history along the branch following a two-step procedure: (1) first, we sample the state (area) at each of the $m-1$ time points along the branch, and; (2) then we simulate the history between each time point, conditional on the states sampled in the first step; the second step is based on the fact that both the average and relative dispersal rates spanned by interval l are constant. To sample states at each time point, we first compute a transition-probability matrix for each interval:

$$\mathbf{P}_l = \exp\left(\mathbf{Q}_l \mu_l t_l\right). \tag{S.3.8}$$

We then calculate the probability of state j at the first time point, T_1, given that the branch

begins in state i and ends in state k, as:

$$\text{conditional probability of } j = \frac{\text{joint probability of } i, j, k}{\text{marginal probability of } i \text{ to } k \text{ transition}}$$

$$P(j \mid i, k) \propto \mathbf{P}_{ij,1} \times \left[\prod_{l=2}^{m} \mathbf{P}_l\right]_{jk}, \tag{S.3.9}$$

where the first term is the probability of transitioning from state i (the state at the beginning of the branch) to state j at the first time point, and the second term is the probability of transitioning from state j to state k (the state at the end of the branch) over the remaining time intervals. We compute this for each state j, and sample the state in proportion to these probabilities. We then repeat this process for each remaining time point, recursively conditioning on the state sampled at the previous time point and the state at the end of the branch.

Once the state at each time point is sampled, we invoke the existing endpoint-conditioned uniformization stochastic-mapping routine (Rodrigue et al. 2007; Fearnhead and Sherlock 2006; Hobolth and Stone 2009) to simulate the history in each interval conditional on its start and end states. The resulting simulated histories across intervals along the branch are then pasted together so that the history output format leaves unchanged.

(Note that in principle the simulation-free stochastic-mapping algorithm implemented in `BEAST` could be modified very similarly to work under the interval-specific model, but we did not make such changes as it was unclear to us in the first place what would be the issues with the current implementation of the algorithm when μ and/or $\mathbf{Q}$ vary across intervals.)

Implementation details and source-code edits of this fix are available in this pull request to the source-code repository of `BEAST`. We also implemented this stochastic-mapping function in `R`, which uses the model parameters and ancestral state of each internal node sampled during `BEAST` MCMC as the input. These two independent implementations provide a means to validate our methods (we include both implementations in our GitHub and Dryad repositories). These two independent implementations produce effectively identical estimates of the number of viral dispersal events (Fig. S.3.2).

Assessing adequacy of interval-specific phylodynamic models

We use posterior-predictive simulation (Gelman et al. 1996; Bollback 2002) to assess the adequacy of our interval-specific phylodynamic models. Posterior-predictive simulation requires: (1) the ability to simulate geographic datasets under interval-specific phylodynamic

models for a given set of parameter values, and; (2) summary statistics that allow us to compare the resulting simulated datasets to the observed dataset. We describe each of these components below.

Simulating under interval-specific phylodynamic models

We draw m random samples from the joint posterior distribution of the model; each sample i consists of a fully specified phylodynamic model, $\theta_i = \{\Psi_i, \mathbf{Q}_i, \mu_i\}$. For each sample, we simulate a new geographic dataset on the sampled tree, Ψ_i, given the sampled parameters of the geographic model, $\{\mathbf{Q}_i, \mu_i\}$; we label the newly simulated dataset G_i^{sim}.

Under a constant phylodynamic model, we simulate full dispersal histories forward in time over a tree using the `sim.history()` function in the `R` package `phytools` (Revell 2012). We implemented an extension of the `sim.history()` function to simulate dispersal histories under interval-specific phylodynamic models. These functions allow us to perform posterior-predictive simulation to assess the adequacy of both the constant and the interval-specific phylodynamic models. We provide these `R` scripts in our GitHub and Dryad repositories.

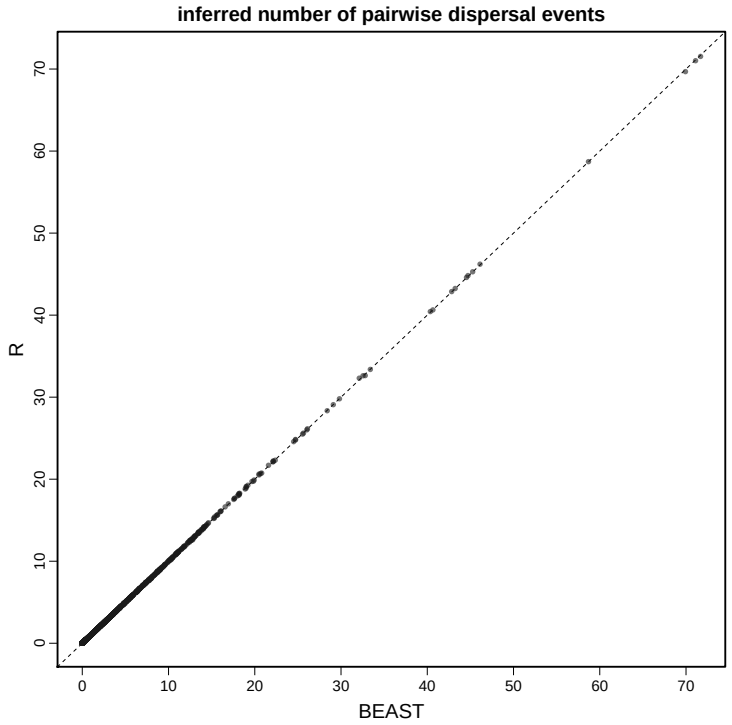

Figure S.3.2: Comparison of the estimated number of pairwise dispersal events in two independent implementations. Each dot represents the mean estimate of the number of dispersal events in a given time interval between a given pair of areas. Independent implementations in `BEAST` and `R` produce effectively identical estimates.

Summary statistics

We define a summary statistic, which we generically denote $T(G \mid \theta_i)$, where G is either the simulated or observed dataset. (We note that the dependence of the statistic on θ_i—while often suppressed or ignored in phylogenetic applications of posterior-predictive simulation—is consistent with posterior-predictive discrepancy analysis, as described by Gelman et al. 1996.) For each simulated dataset, we compute a discrepancy statistic,

$$D_i = T(G_i^{\text{sim}} \mid \theta_i) - T(G^{\text{obs}} \mid \theta_i),$$

where G^{obs} is the observed geographic dataset and G^{sim} is a simulated dataset.

If the inference model provides an adequate description of the true data-generating process, the posterior-predictive distribution of D should contain zero with high probability. Accordingly, for the m predictive datasets for a given model and dataset combination, we calculate the posterior-predictive p value as:

$$P = \frac{1}{m} \sum_{i=1}^{m} D_i \geq 0.$$

Values between 0.025 and 0.975 indicate that the model is adequate and cannot be rejected (*i.e.*, zero falls within the 95% posterior-predictive interval).

We developed two summary statistics to assess the adequacy of interval-specific phylodynamic models: (1) the *parsimony statistic*, and; (2) the *tipwise-multinomial statistic*. The parsimony statistic is calculated as the difference in the parsimony score for the observed areas and the simulated areas across the tips of the tree (where the parsimony score is the minimum number of dispersal events required to explain the distribution of areas across the tips of a tree). We compute parsimony scores using the `parsimony()` function in the `R` package, `phangorn` (Schliep 2010). The tipwise-multinomial statistic is inspired by the multinomial statistic that was first proposed by Goldman (1993) and later used by Bollback (2002) to assess the adequacy (absolute fit) of substitution models to sequence alignments. Our tipwise statistic treats the set of states (areas) across the tips of the tree as an outcome of a multinomial trial. Specifically, we calculate the tipwise-multinomial statistic as the difference in the multinomial probabilities for the observed the set of areas and the simulated the set of areas across the tips of the tree. We calculate each multinomial probability as:

$$T(G \mid \theta_i) = \sum_{i=1}^{k} n_i \ln(n_i/n),$$

where n is the number of tips in the tree, and n_i is the number of tips that occur area i. (Note that this statistic is also similar to the entropy statistic used to assess genetic variability along sequences; Shannon 1948; Schneider et al. 1986.)

Time-slice summary statistics

We could use our summary statistics to assess the ability of a phylodynamic model to describe the entire history of dispersal; *i.e.*, to assess whether the model adequately describes the process that generated the data over the entire phylogeny. Of potential concern, however, is the sufficiency of the summary statistics to detect variation in the data-generating process over time. For example, it is conceivable that a phylodynamic model may be capable of simulating datasets with parsimony scores that are very similar to those for the observed dataset—implying that the model provides an adequate description of the process that generated the entire geographic history—but the underlying dispersal events simulated under this model may nevertheless occur at inappropriate times.

To assess the ability of phylodynamic models to describe the temporal distribution of dispersal events, we extend the parsimony and tipwise-multinomial summary statistics to assess time slices of the geographic history[2]. We calculate these summary statistics for k prespecified time slices, resulting in k parsimony statistics and k tipwise-multinomial statistics for each simulated dataset. We compute the time-slice variant of the parsimony statistic as follows: (1) we first infer the most-parsimonious dispersal history (*i.e.*, the minimum number of dispersal events) for a given simulated dataset and the observed dataset using the `ancestral.pars()` function in the `R` package, `phangorn` (Schliep 2010); (2) we then assign each inferred dispersal event to one of the k time slices based on the time span of the branch along which the dispersal event was inferred (when a dispersal event is inferred to occur along a branch that spans two or more time slices, we locate the event uniformly along the branch, and then assign it to the corresponding slice), and finally; (3) we compute the difference in the number of dispersal events between the simulated and observed dataset for each time slice. We compute the time-slice variant of the tipwise-multinomial statistic in a similar manner; *i.e.*, we first find the set of tips in each time slice, and then compute the

[2]Note that the time slices that we define for summary statistics are distinct from the intervals specified in an interval-specific phylodynamic model. The former are motivated to better assess the adequacy of a phylodynamic model, the latter are motivated to accommodate variation in dispersal dynamics in the empirical data. Accordingly, we might use time-slice summary statistics to assess the adequacy of both constant or interval-specific phylodynamic models.

tipwise-multinomial statistic for that time slice (as described above) for the corresponding set of tips. Further details regarding the computation of these summary statistics are available in an `R` script, `posterior_predictive_teststatistics_functions.R`, included in our GitHub and Dryad repositories.

Running `BEAST` analyses under interval-specific phylodynamic models

To aid the application of our newly developed interval-specific phylodynamic models with `BEAST`, we provide a hands-on tutorial—that describes how to specify the model in an XML file—in our GitHub and Dryad repositories.

Simulation Study

Simulation design

We performed a simulation study to explore the statistical behavior of our interval-specific phylodynamic models. Specifically, we sought to assess: (1) our ability to perform reliable inference under interval-specific models; (2) the impact of model misspecification, and; (3) our ability to identify the correct model. To provide a meaningful evaluation of the statistical behavior of a method, it is critical for a simulation study to explore realistic parameter space (*i.e.*, to subject the method to simulated datasets that are similar to those it will actually encounter in empirical analyses). Accordingly, we focused our simulation study on simulated datasets that resemble our empirical SARS-CoV-2 reduced dataset. That is, our simulation study explored a region of parameter space that is centered on the joint posterior probability distribution of phylodynamic model parameters estimated from our empirical analyses. To that end, we first analyzed our empirical dataset under each of two models, $1\mu 1\mathbf{Q}$ and $2\mu 2\mathbf{Q}$, and then centered the parameter values of our simulation on the resulting posterior median estimates of the corresponding parameters (μ and $\mathbf{Q}$). Specifically, we used these empirically based parameter values to simulate 200 geographic datasets under each of two models, $1\mu 1\mathbf{Q}$ and $2\mu 2\mathbf{Q}$. Finally, we performed separate analyses of each simulated dataset under each of the two models, resulting in four true:inference model combinations ($1\mu 1\mathbf{Q}$:$1\mu 1\mathbf{Q}$, $2\mu 2\mathbf{Q}$:$2\mu 2\mathbf{Q}$, $1\mu 1\mathbf{Q}$:$2\mu 2\mathbf{Q}$, and $2\mu 2\mathbf{Q}$:$1\mu 1\mathbf{Q}$). Below, we provide additional details of how we generated simulated datasets, analyzed these simulated datasets, and summarized results from our analyses of the simulated datasets.

Generating simulated datasets

Model specification for empirical analyses

We performed analyses of our reduced SARS-CoV-2 dataset under a constant ($1\mu 1\mathbf{Q}$) model and an interval-specific phylodynamic ($2\mu 2\mathbf{Q}$) model. Our reduced SARS-CoV-2 dataset has 1271 sequences. To reduce the computational burden of our simulation study we: (1) aggregated our 23 study areas into three more coarsely defined areas: China, North America, and the rest of the world, and; (2) conditioned our analyses on the MCC summary phylogeny inferred from our reduced SARS-CoV-2 dataset (described in this section). We specified the single boundary for the $2\mu 2\mathbf{Q}$ model at February 2 (corresponding to the onset of international air-

travel bans against China). For both the constant ($1\mu 1\mathbf{Q}$) and interval-specific ($2\mu 2\mathbf{Q}$) models, we used an asymmetric $\mathbf{Q}$ matrix (Edwards et al. 2011) that allows the relative rate of dispersal from area i to area j to be different from the relative rate of dispersal from area j to area i. We specified diffuse priors on the parameters (for the average dispersal rate, μ, and relative dispersal rates, $\mathbf{Q}$) for each phylodynamic model (Table S.3.1).

We also specified a root-frequency vector, ω, which represents the prior probability that the tree begins in each of the geographic areas. For models with a single $\mathbf{Q}$, it is possible to use the stationary frequency implied by the $\mathbf{Q}$ as this root-frequency vector. However, interval-specific models do not have a global stationary frequency (*i.e.*, each interval has a separate stationary frequency); in this case, it is conventional to treat ω as a free parameter and estimate it from the data. To be as consistent as possible between the constant and interval-specific models, we specified ω as a free parameter for both models.

Table S.3.1: Priors used in analyses of the reduced SARS-CoV-2 dataset.

Parameter	Description	Prior
μ_l	Average dispersal rate in interval l	$\text{Exp}(1/\lambda)$; $\lambda \sim \Gamma(0.5, 0.5)$
$r_{ij,l}$	Relative dispersal rate from i to j in interval l	$\Gamma(1,1)$
$\boldsymbol{\omega}$	Root frequencies	$\text{Dir}(1,1,1)$

Parameter estimation

For each of model ($1\mu 1\mathbf{Q}$ and $2\mu 2\mathbf{Q}$), we inferred the joint posterior distribution of parameters from our empirical dataset by running four independent MCMC simulations using our modified version of `BEAST` (see this section) with the `BEAGLE` library (compiled from the 'hmc-clock' branch, commit 'dd36bf5'; Ayres et al. 2019). We ran each replicate MCMC simulation for 700000–800000 generations, sampling every 500–1000 generations. We discarded the initial 100000–250000 generations (as burn-in) from each replicate MCMC simulation, and then combined the remaining posterior samples from all replicate simulations using `LogCombiner` version 1.10.5. (The number of generations, sampling frequency, and the length of the burn-in are presented as ranges here and below because we deliberately ran the analyses under more complex models longer and sampled less frequently.) We then assessed MCMC performance for the resulting composite posterior sample by inspecting the log files using `Tracer` (Rambaut et al. 2018) version 1.7.1, and using the `coda` package (Plummer et al. 2006) in `R` (R Core Team 2020). Specifically, we ensured that the computed ESS values for all continuous parameters

were $\gg$2000. Details of these analyses are available in the XML scripts included in our GitHub and Dryad repositories.

Simulating geographic datasets using the inferred parameter values

We simulated a total of 200 geographic datasets under each of two models: the constant ($1\mu 1\mathbf{Q}$) and interval-specific ($2\mu 2\mathbf{Q}$) phylodynamic models, using our forward-in-time simulator (the simulator `R` script is included in our GitHub and Dryad repositories). We simulated these geographic datasets over the MCC summary tree inferred from our reduced SARS-CoV-2 dataset (see this section), with parameter values (for μ and $\mathbf{Q}$) of the simulating models ($1\mu 1\mathbf{Q}$ and $2\mu 2\mathbf{Q}$) set to the corresponding posterior median estimates from our empirical analyses (described in this section). Specifically, we used the following parameter values for constant ($1\mu 1\mathbf{Q}$) model:

$$\mu = 0.0320,\ \mathbf{Q} = \begin{pmatrix} - & 1.4018 & 0.1740 \\ 0.0190 & - & 0.7366 \\ 0.1104 & 1.2983 & - \end{pmatrix},$$

and the following parameter values for the interval-specific ($2\mu 2\mathbf{Q}$) model:

$$\mu_1 = 0.0242,\ \mathbf{Q}_1 = \begin{pmatrix} - & 1.4214 & 1.2043 \\ 0.0107 & - & 0.7309 \\ 0.0654 & 1.3779 & - \end{pmatrix};\quad \mu_2 = 0.0602,\ \mathbf{Q}_2 = \begin{pmatrix} - & 0.6974 & 0.0814 \\ 1.0008 & - & 0.2428 \\ 0.8077 & 0.5744 & - \end{pmatrix}.$$

Values of the diagonal elements are specified in the usual manner (*i.e.*, set equal to the negative sum of the off-diagonal elements in the corresponding row).

Analyzing simulated datasets

Model specification

For each simulated dataset, we inferred the joint posterior distribution under each of the two models, $1\mu 1\mathbf{Q}$ and $2\mu 2\mathbf{Q}$, using the same priors (listed in Table S.3.1) specified in the empirical analyses that generated parameter values used to simulate the datasets.

Parameter estimation

For each inference model, $1\mu 1\mathbf{Q}$ and $2\mu 2\mathbf{Q}$, we estimated the joint posterior distribution for each simulated dataset by running two to four independent MCMC simulations using our

modified version BEAST (see this section) with the BEAGLE library (compiled from the 'hmc-clock' branch, commit 'dd36bf5'; Ayres et al. 2019). We ran each replicate MCMC simulation for 300000–800000 generations, sampling every 500–1000 generations. When a sample was drawn, we performed stochastic mapping using the endpoint-conditioned uniformization algorithm (Rodrigue et al. 2007; Fearnhead and Sherlock 2006; Hobolth and Stone 2009) and our modified algorithm for interval-specific phylodynamic models (see this section) implemented in BEAST to simulate a dispersal history over the MCC phylogeny. We discarded the initial 50000–100000 generations (as burn-in) from each replicate MCMC simulation, and then combined the remaining posterior samples from all replicate simulations using LogCombiner version 1.10.5. We then assessed MCMC performance for the resulting composite posterior sample by inspecting the log files using Tracer (Rambaut et al. 2018) version 1.7.1, and using the coda package (Plummer et al. 2006) in R (R Core Team 2020). Specifically, we ensured that the computed ESS values for all continuous parameters were $\gg$500. Details of these analyses are available in the XML scripts included in our GitHub and Dryad repositories.

Summarizing results of the simulation analyses

For each of the four true:inference model combinations, we first computed the coverage probability (the frequency with which the true value was contained in the X% posterior credible interval) as a function of the size (X) of the credible interval for all model parameters (μ and $\mathbf{Q}$) and for the pairwise and total number of dispersal events. For each true:inference model combination, we also summarized the absolute error (estimated minus true values) for the model parameters (μ and $\mathbf{Q}$) and the pairwise and total number of dispersal events.

When the inference model is correctly specified (*i.e.*, scenarios $1\mu1\mathbf{Q}$:$1\mu1\mathbf{Q}$, and $2\mu2\mathbf{Q}$:$2\mu2\mathbf{Q}$), there is a one-to-one correspondence between the true:inference model parameters, which allows us to simply compare the true:estimated values for each parameter. By contrast, when the inference model is misspecified (*i.e.*, scenarios $1\mu1\mathbf{Q}$:$2\mu2\mathbf{Q}$, and $2\mu2\mathbf{Q}$:$1\mu1\mathbf{Q}$), there is a lack of direct correspondence between the true:inference model parameters. When the inference model is *overspecified* (*i.e.*, where $2\mu2\mathbf{Q}$:$1\mu1\mathbf{Q}$), we compared the inferred parameter value for each interval to the true, time-constant value (*e.g.*, we compared interval-specific estimates of μ_1 and μ_2 to the time-constant true value, μ). Conversely, when the inference model is *underspecified* (*i.e.*, where $2\mu2\mathbf{Q}$:$1\mu1\mathbf{Q}$), we compared estimates of the time-constant param-

eter values to each true, interval-specific parameter values (*e.g.*, we compared time-constant estimates of μ to both interval-specific true values, μ_1 and μ_2).

Results

When the inference model is correctly specified (*i.e.*, where both the true and inference models include [or exclude] interval-specific parameters, $2\mu2\mathbf{Q}$:$2\mu2\mathbf{Q}$ and $1\mu1\mathbf{Q}$:$1\mu1\mathbf{Q}$), our simulation study demonstrates that estimates under interval-specific models are as reliable as those under constant models (Figs. 3.3–S.3.7, green and blue). Moreover, when the inference model is over-specified (*i.e.*, it includes interval-specific parameters not included in the true model) inferences are comparable to those under correctly specified models (Figs. 3.3–S.3.7, purple). However, when the inference model is underspecified (*i.e.*, it excludes interval-specific parameters of the true model) inferences are severely biased (Figs. 3.3–S.3.7, orange).

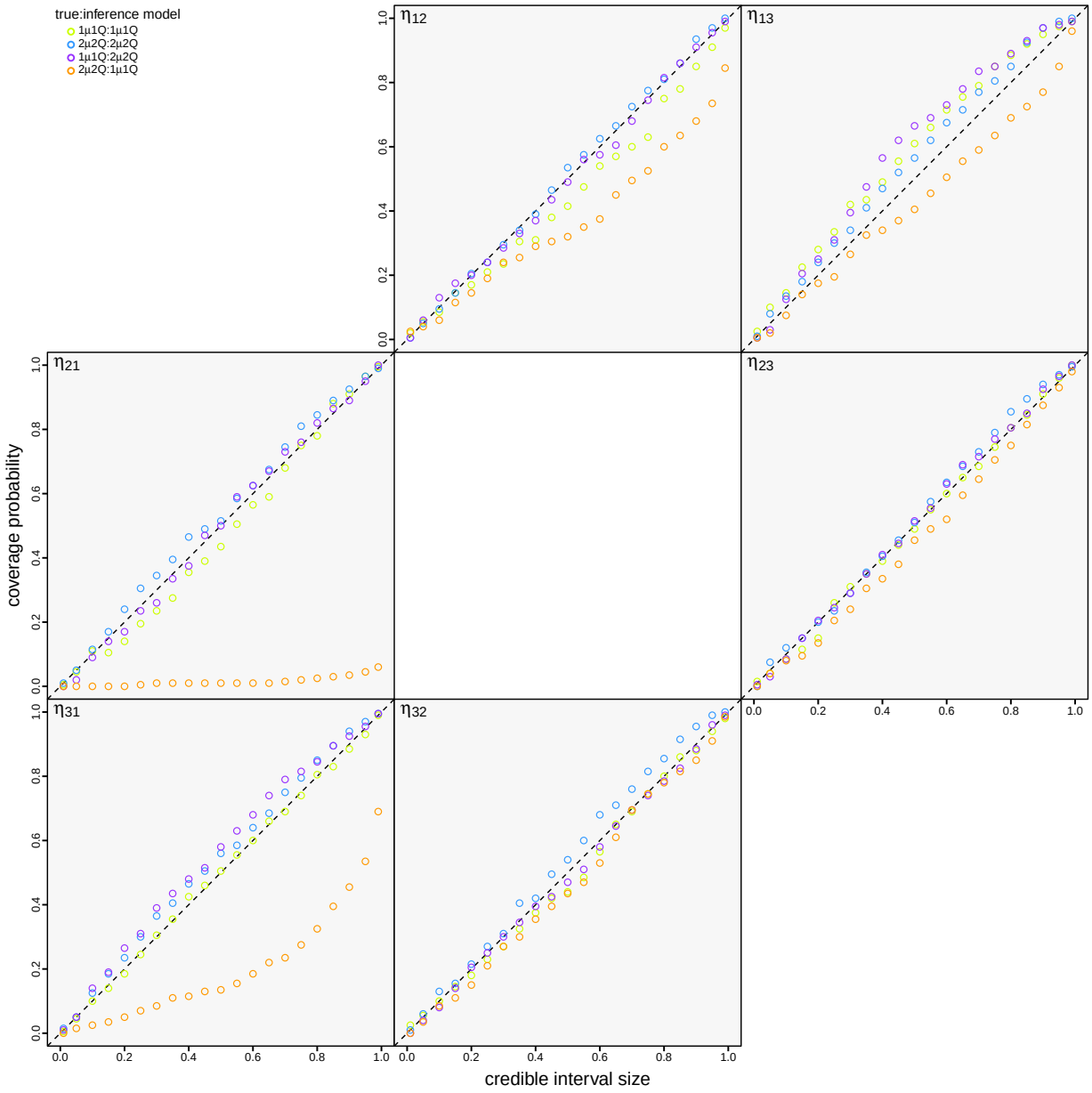

Figure S.3.3: Coverage probabilities for the number of dispersal events between each pair of areas. For each true:inference model combination, we computed the coverage probability (the frequency with which the true number of dispersal events was contained in the X% credible interval; y-axis) as a function of the size (X) of the credible interval (x-axis). Each panel summarizes the estimated coverage probabilities for the number of dispersal events, η, between areas i and j, η_{ij}. The panels are arranged to mirror the six pairwise, off-diagonal dispersal routes of the **Q** matrix; *e.g.*, the cell in the first row and second column depicts the estimated coverage probability for the number of dispersal events from area 1 to area 2, etc. When the model is true, we expect the coverage probability to be equal to the size of the credible interval (Cook et al. 2006). As expected, coverage probabilities fall along the one-to-one line when the model is correctly specified (green and blue). Moreover, coverage probabilities are also appropriate when the inference model is overspecified (*i.e.*, the inference model includes interval-specific parameters not included in the true model; purple). However, coverage probabilities are extremely unreliable when the inference model is underspecified (*i.e.*, the inference model excludes interval-specific parameters of the true model; orange).

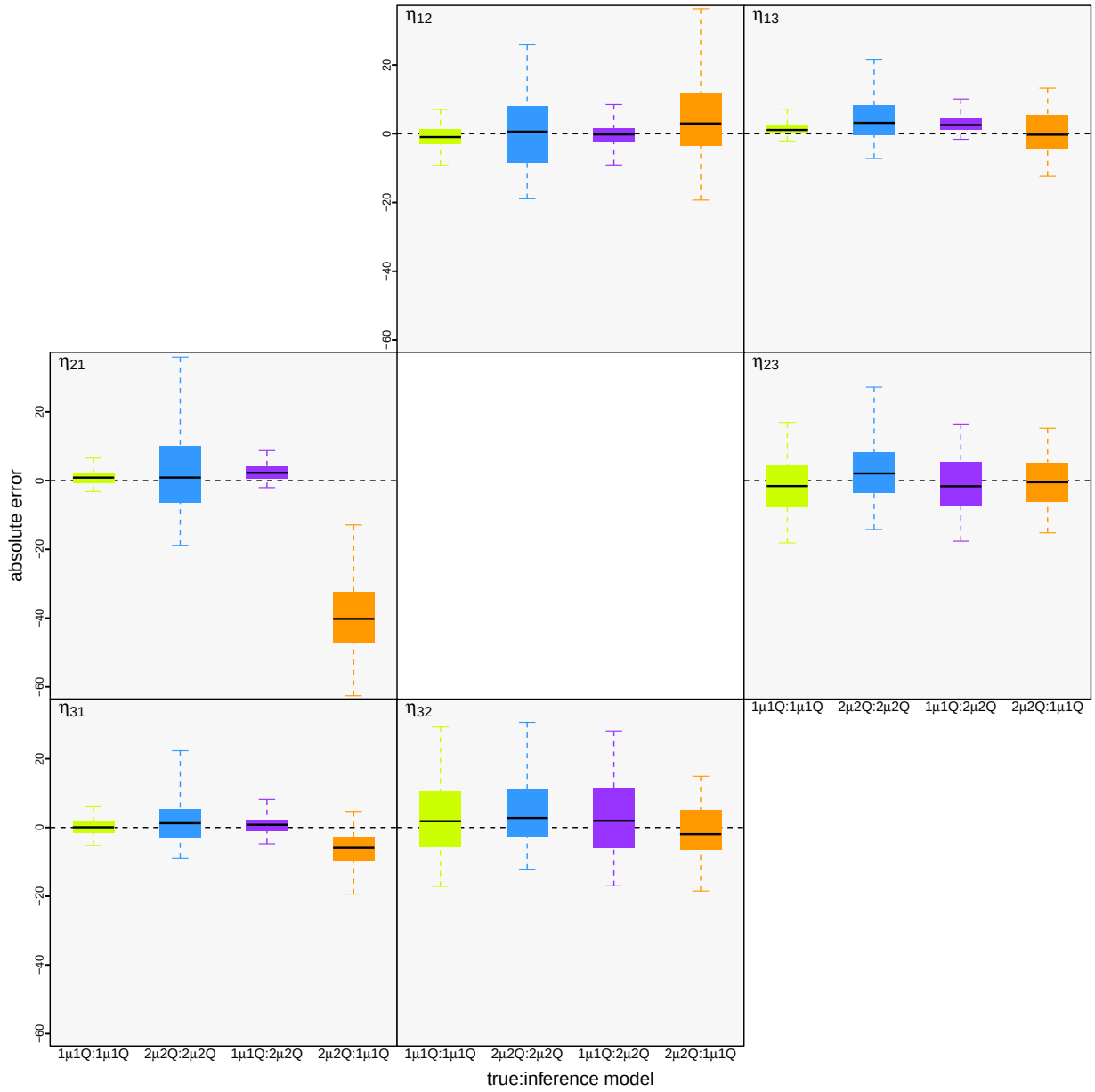

Figure S.3.4: Absolute error (estimated−true values) for the number of dispersal events between pairs of areas. For each true:inference model combination, we summarized the absolute error for the number of dispersal events as boxplots (median [horizontal bar], 50% probability interval [boxes], and 95% probability interval [whiskers]). Each panel summarizes the absolute error for the number of dispersal events, η, between areas i and j, η_{ij}. The six panels are arranged to mirror the corresponding six off-diagonal elements of the **Q** matrix (*c.f.*, Figure S.3.3). Again, when the model is underspecified (orange) inferences are strongly biased compared to those under the correctly specified (green and blue) and overspecified (purple) models.

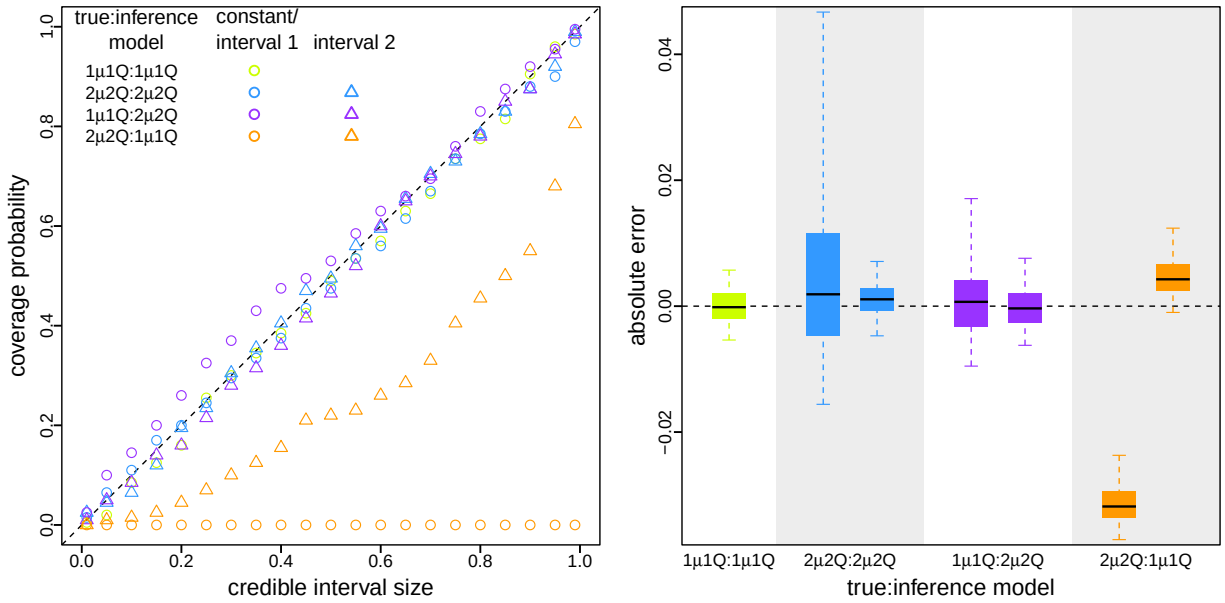

Figure S.3.5: Reliable inference of average dispersal rate requires a correctly specified phylodynamic model. We simulated 200 geographic datasets under each of two models, 1μ1**Q**) 2μ2**Q**. For each simulated dataset, we separately inferred the average dispersal rate under each model, resulting in four true:inference model combinations (1μ1**Q**:1μ1**Q**, 2μ2**Q**:2μ2**Q**, 1μ1**Q**:2μ2**Q**, and 2μ2**Q**:1μ1**Q**). For 1μ1**Q**:2μ2**Q**, we compared interval-specific parameter estimates to the true, time-constant parameter value (*i.e.*, we compared estimates of μ_1 and μ_2 to the true, time-constant value, μ). Conversely, for 2μ2**Q**:1μ1**Q**, we compared the time-constant parameter estimates to each of the true, interval-specific values (*i.e.*, we compared estimates of μ to each of the true values, μ_1 and μ_2). Left) For each true:inference model combination, we plotted the coverage probability (y-axis) as a function of the size of the credible interval (x-axis). When the true or inference model is interval-specific, we plot separate true:inference comparisons for the first (circles) and second (triangles) time intervals. As expected (Cook et al. 2006), coverage probabilities fall along the one-to-one line when the model is correctly specified (green and blue); additionally, coverage probabilities are also appropriate when the inference model is overspecified (purple). However, coverage probabilities are extremely unreliable when the inference model is underspecified (orange). Right) For each true:inference model combination, we summarized the absolute error (estimated−true values) for the average dispersal rate as boxplots (median [horizontal bar], 50% probability interval [boxes], and 95% probability interval [whiskers]). When the true or inference model is interval-specific, we separately plot absolute error for the first (left) and second (right) intervals. Again, when the model is underspecified (orange) inferences are strongly biased compared to those under the correctly specified (green and blue) and overspecified (purple) models.

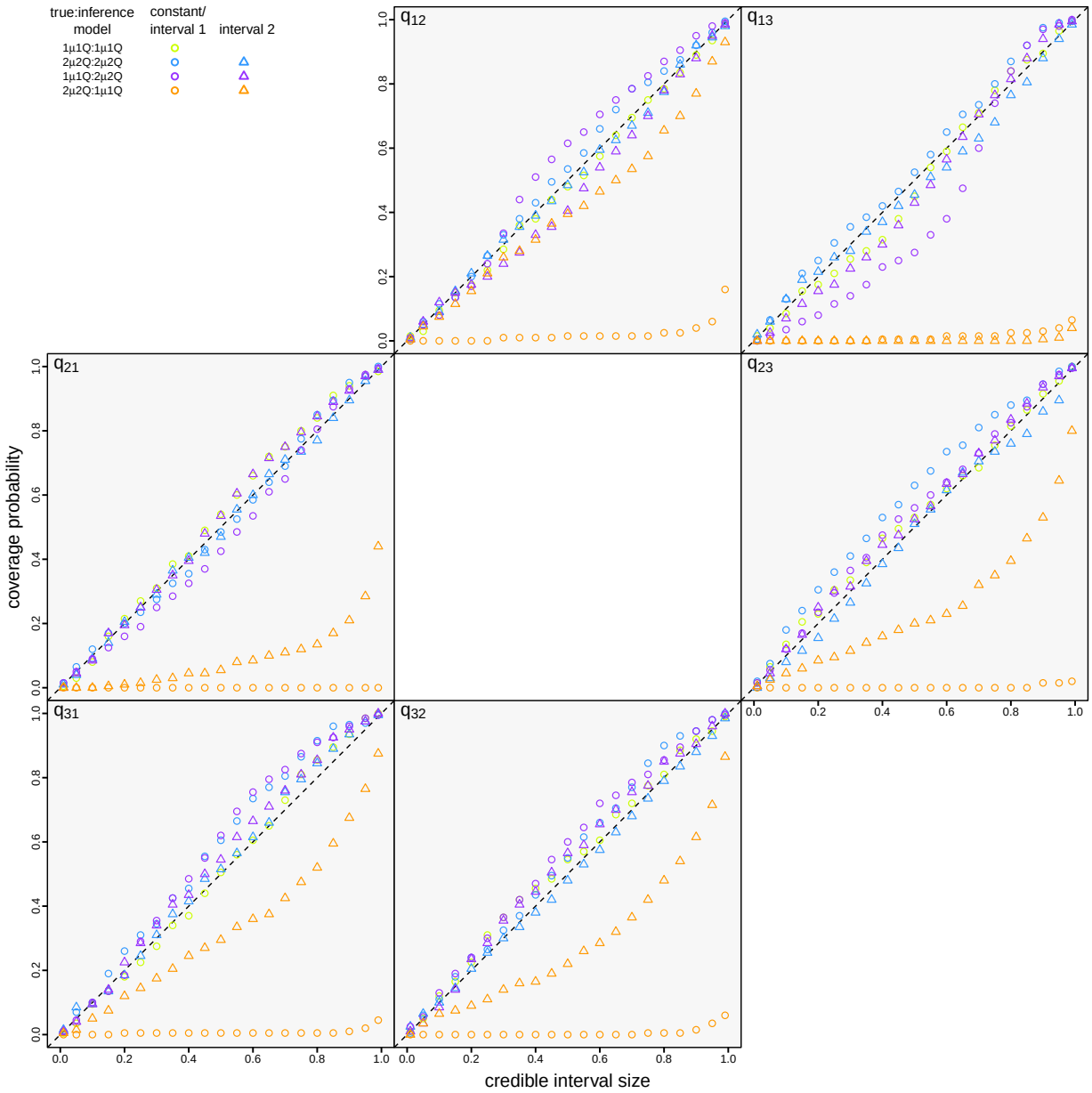

Figure S.3.6: Coverage probabilities for the relative dispersal rates between each pair of areas. For each of the four true:inference model combinations, we plotted estimates of the coverage probability (y-axis) as a function of the size of the credible interval (x-axis). Each panel summarizes the estimated coverage probabilities for the relative dispersal rate between areas i and j, q_{ij}. The panels are arranged to mirror the corresponding off-diagonal elements of the **Q** matrix (*c.f.*, Figure S.3.3). For $1\mu1\mathbf{Q}$:$2\mu2\mathbf{Q}$, we compared interval-specific parameter estimates to the true, time-constant parameter value (*i.e.*, we compared estimates of $q_{ij,1}$ and $q_{ij,2}$ to the true, time-constant value, q_{ij}). Conversely, for $2\mu2\mathbf{Q}$:$1\mu1\mathbf{Q}$, we compared the time-constant parameter estimates to each of the true, interval-specific values (*i.e.*, we compared estimates of q_{ij} to each of the true values, $q_{ij,1}$ and $q_{ij,2}$). When the true or inference model is interval-specific, we separately plot true:inference comparisons for the first (circles) and second (triangles) intervals. As expected (Cook et al. 2006), coverage probabilities fall along the one-to-one line when the model is correctly specified (green and blue); additionally, coverage probabilities are also appropriate when the inference model is overspecified (purple). However, coverage probabilities are extremely unreliable when the inference model is underspecified (orange).

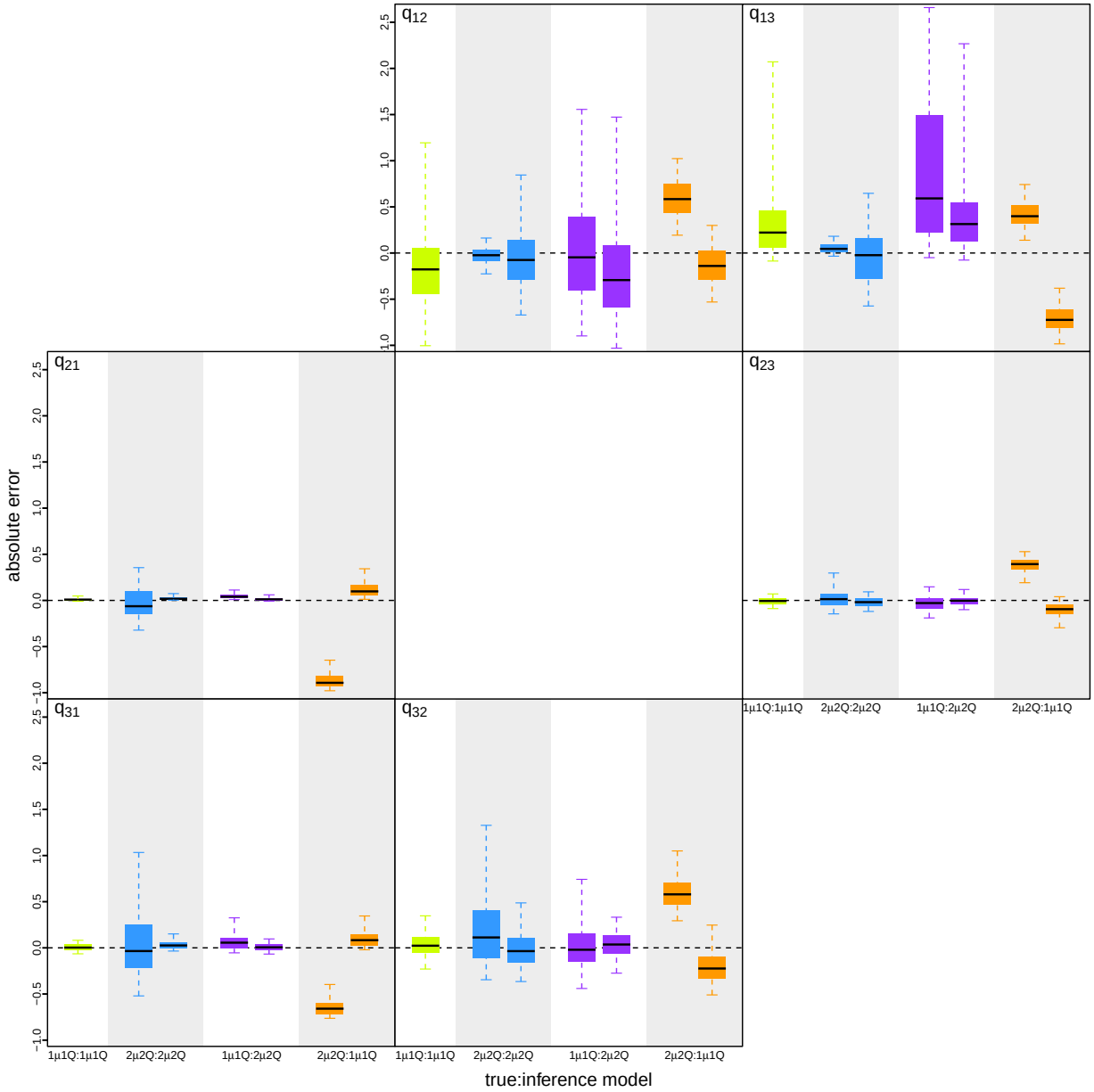

Figure S.3.7: Absolute error (estimated minus true values) for the relative dispersal rates between pairs of areas. For each combination of true:inference models, we summarized the absolute error for the relative dispersal rates as boxplots (median [horizontal bar], 50% probability interval [boxes], and 95% probability interval [whiskers]). Each panel summarizes the absolute error for the relative dispersal rates between areas i and j, q_{ij}. The panels are arranged to mirror the six off-diagonal relative dispersal rates of the **Q** matrix (*c.f.*, Figure S.3.3). For $1\mu 1\mathbf{Q}$:$2\mu 2\mathbf{Q}$, we compared interval-specific parameter estimates to the true, time-constant parameter value (*i.e.*, we compared estimates of $q_{ij,1}$ and $q_{ij,2}$ to the true, time-constant value, q_{ij}). Conversely, for $2\mu 2\mathbf{Q}$:$1\mu 1\mathbf{Q}$, we compared the time-constant parameter estimates to each of the true, interval-specific values (*i.e.*, we compared estimates of q_{ij} to each of the true, interval-specific values, $q_{ij,1}$ and $q_{ij,2}$). When the true or inference model is interval-specific, we separately plot absolute error for the first (left) and second (right) intervals. Again, when the model is underspecified (orange) inferences are strongly biased compared to those under the correctly specified (green and blue) and overspecified (purple) models.

Assessing model fit to simulated datasets

Assessing relative fit of the true and alternative models using Bayes factors

We used Bayes factors to assess the relative fit of the true and alternative models to each simulated dataset. Specifically, for each dataset, we first estimated the marginal likelihood under each of two models ($1\mu 1\mathbf{Q}$ and $2\mu 2\mathbf{Q}$), and then computed the Bayes factor as twice the difference in the resulting log marginal likelihoods (Kass and Raftery 1995). We estimated marginal likelihoods for each inference model using both thermodynamic-integration (Lartillot and Philippe 2006) and stepping-stone (Xie et al. 2011; Baele et al. 2012) estimators. Our analyses to estimate marginal likelihoods conditioned on the MCC summary phylogeny (see this section).

For each simulated dataset, we ran two replicate power-posterior MCMC simulations for both models ($1\mu 1\mathbf{Q}$ and $2\mu 2\mathbf{Q}$) using our modified version of `BEAST` (see this section) with the `BEAGLE` library (compiled from the 'hmc-clock' branch, commit 'dd36bf5'; Ayres et al. 2019). For each replicate power-posterior MCMC simulation, we used 24 powers placed at evenly-spaced quantiles of a Beta$(0.3, 1.0)$ distribution. For each power, we discarded the initial 70000–80000 generations as burn-in and then sampled every 100 generations during the remaining 160000–180000 generations. (The number of generations and the length of the burn-in of each power are presented as ranges here and below because we deliberately ran the MCMC longer at each power for the power-posterior analyses under more complex models, and also set up replicate MCMCs with increasing length as one way of assessing the reliability of our marginal-likelihood estimates.) We assessed the reliability of our marginal-likelihood estimates by comparing values from all the replicate power-posterior MCMCs. Details of these analyses (*e.g.*, proposal weights) are available in the XML scripts included in our GitHub and Dryad repositories.

Assessing absolute fit of each model using posterior-predictive simulation

For each true:inference model combination, we assessed absolute model fit to each simulated dataset using posterior-predictive simulation (Gelman et al. 1996) with a set of 20 time-slice summary statistics. For each *simulated* dataset, we simulated $m = 800-1000$ *predictive* datasets using the parameter values that were randomly sampled from the inferred joint posterior distribution of the corresponding *simulated* dataset under each of the two models. We then gener-

ated posterior-predictive distributions from each set of m predictive datasets under 20 separate time-slice summary statistics, *i.e.*, for all combinations of the two types of summary statistics (parsimony and tipwise-multinomial statistics) and 10 time slices spanning the entire dispersal history of the early phase of COVID-19 (including week 0 that covers the duration from the origin of SARS-CoV-2 to January 5, 2020, and weeks 1–9 corresponding to the nine weeks between January 6 and March 8, 2020). For each posterior-predictive distribution, we computed the posterior-predictive p value (see this section) to assess the adequacy of (*i.e.*, absolute fit) the corresponding inference model.

Results

Our simulation study demonstrates the importance of identifying scenarios where an inference model is underspecified; failure to accommodate interval-specific variation in the study data will severely bias parameter estimates. Fortunately, our simulation study demonstrates that we can reliably identify when a given model is correctly specified, overspecified, or underspecified using a combination of Bayes factors (to assess the relative fit of competing models to the data; Fig. 3.4, left) and posterior-predictive simulation (to assess the absolute fit of each candidate model to the data; Fig. 3.4, right, Fig. S.3.8 and table S.3.2). Using a combination of Bayes factors and posterior-predictive simulation allows us to not only identify the best of the candidate models, but also to ensure that the best model provides an adequate description of the true process that gave rise to our study data.

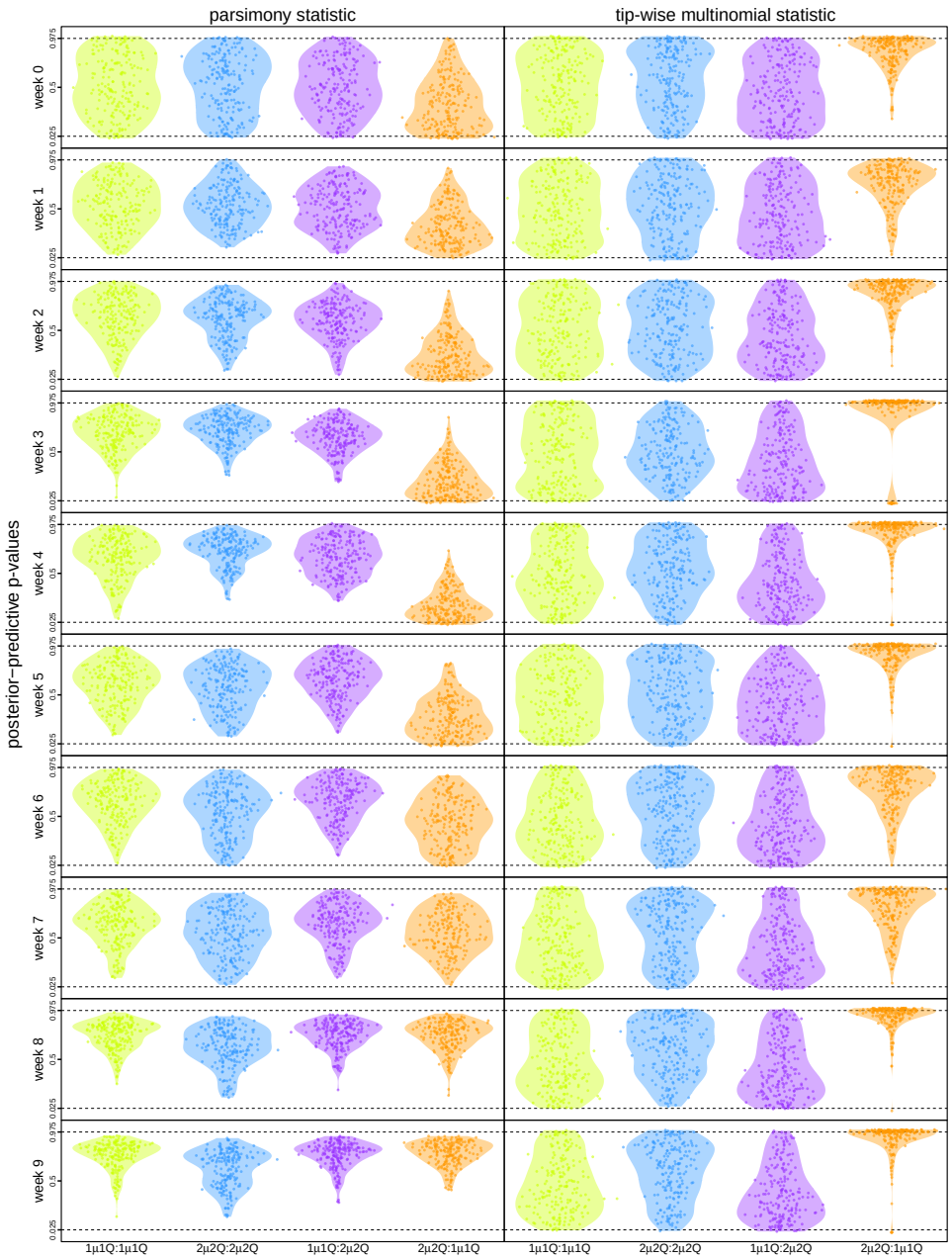

Figure S.3.8: Assessing the absolute fit of true and alternative phylodynamic models to simulated datasets. We assessed the absolute fit of alternative models to the simulated datasets. For each combination of true:inference model, we assessed absolute model fit (*i.e.*, model adequacy) using posterior-predictive simulation with a set of 20 time-slice summary statistics. We group results based on the parsimony (left column) and tipwise-multinomial (right column) summary statistics; the rows in each column corresponds to one of the 10 (weekly) time slices, and each cell plots the posterior-predictive distributions of the corresponding statistic for each of the four true:inference model combinations. Each dot represents the posterior-predictive p value for a single dataset, and the violin plots summarize the distribution of these p values for all datasets under the corresponding true:inference model combination. Dashed lines indicate critical posterior-predictive p values (of 0.025 and 0.975); a dot above the top dashed line or below the bottom dashed line indicates that the corresponding inference model provides an inadequate description of the true process that gave rise to that dataset. As expected, the true model is overwhelmingly inferred to be adequate (green and blue). Encouragingly, model overspecification appears to have a negligible impact on model adequacy (purple). By contrast, an underspecified model severely impacts model adequacy (orange).

Table S.3.2: Percent of simulated datasets that were inadequately modeled. The organization of the table mirrors that of Fig. S.3.8. Each cell of the table indicates the percent of simulated datasets for which the inferene model was inferred to provide an inadequate description of the true process that generated the simulated datasets. That is, each cell indicates the percent of the posterior-predictive p values that fall outside the critical (0.025 and 0.975) thresholds (*i.e.*, the corresponding percent of dots above the top dashed line or below the bottom dashed line in Fig. S.3.8). Values indicating significant model inadequacy (*i.e.*, $\geq$5%) are indicated in red text.

	parsimony statistic				tipwise-multinomial statistic			
	1μ1**Q**:1μ1**Q**	2μ2**Q**:2μ2**Q**	1μ1**Q**:2μ2**Q**	2μ2**Q**:1μ1**Q**	1μ1**Q**:1μ1**Q**	2μ2**Q**:2μ2**Q**	1μ1**Q**:2μ2**Q**	2μ2**Q**:1μ1**Q**
week 0	4.5	3.5	3.0	**6.0**	4.5	4.0	5.0	**11.5**
week 1	0.0	0.5	0.0	0.5	3.0	**6.5**	4.0	3.0
week 2	0.5	0.0	0.0	3.5	4.0	3.5	2.5	**16.0**
week 3	0.0	0.0	0.0	**6.0**	4.0	1.0	4.0	**74.5**
week 4	0.0	0.0	0.5	**7.5**	2.5	4.0	3.0	**37.5**
week 5	0.0	0.0	0.5	4.0	3.0	3.5	3.0	**31.5**
week 6	0.0	0.5	0.0	1.0	3.5	5.0	2.5	**7.5**
week 7	0.0	0.0	0.0	0.0	5.0	2.5	5.0	**6.5**
week 8	0.0	0.0	0.0	0.0	1.0	1.0	1.5	**36.0**
week 9	0.0	0.0	0.0	0.0	2.0	2.5	2.0	**45.5**
average	0.5	0.45	0.4	2.85	3.25	3.35	3.25	**26.95**

Empirical Application

Overview

The results of our empirical study are based on a complex and comprehensive series of computationally intensive analyses. In this section, we provide a high-level overview of our data collection and data analyses to clarify the rationale of our empirical study, while directing readers to the corresponding subsections below that provide additional details on the various analyses that we performed.

Data Acquisition and Curation

Epidemiological data

We used two types of epidemiological information in this study: (1) the number of confirmed COVID-19 cases, and; (2) the intervention measures involving China that were enacted during the early phase of the pandemic. We compiled a dataset of the number of confirmed COVID-19 cases recorded on each day for each country/province/state based on various sources (WHO 2020; DXY 2020; NHCPRC 2020; ECDC 2020; USCDC 2020) via two intermediate portals (Wu et al. 2020b; Dong et al. 2020; see this section). We used these case-number data to assess the fraction of total cases represented by our genomic sequences, and to estimate the approximate date by which SARS-CoV-2 had spread to most geographic areas. We collected information on international travel bans with China and domestic mitigation measures within China from multiple sources (Wikipedia 2020a,b; Kraemer et al. 2020; Tian et al. 2020; Hsiang et al. 2020; Lai et al. 2020; see this section).

Travel data

We used the daily number of commercial passenger flights obtained from FlightAware as a proxy for the global air-travel volume (see this section).

Delineation of time intervals and geographic areas

To explore the dynamics of viral geographic dispersal in the early phase of the COVID-19 pandemic, we partitioned the study period into five time intervals: (1) interval 1 from late 2019 (origin time) to Jan. 12, 2020; (2) interval 2 from Jan. 1 3, 2020 to Jan. 25, 2020; (3) interval 3 from Jan. 26, 2020 to Feb. 2, 2020; (4) interval 4 from Feb. 3, 2020 to Feb. 16, 2020, and; (5) interval 5 from Feb. 16, 2020 to Mar. 8, 2020. Boundaries between these intervals coincide with the initiation of containment measures (*e.g.*, international travel bans with China) or other events

associated with changes in the level of population movement (*e.g.*, start of the Spring Festival travel season); see this section. For our phylodynamic analyses, we discretized the globe into geographic areas to study the early spread of SARS-CoV-2. We grouped geographically adjacent countries/territories (for non-focal regions) or states/provinces (for focal regions) to specify a total of 23 geographic areas (Fig. S.3.10).

SARS-CoV-2 genomic sequence data

We curated two genomic sequence datasets for this study, one with 1271 sequences (the "reduced dataset") and the other with 2598 sequences (the "entire dataset"). The reduced dataset was produced on Apr. 19, 2020, based on all available SARS-CoV-2 genomic sequences from the Global Initiative on Sharing All Influenza Data (GISAID, Shu and McCauley 2017) as of that date. The entire dataset was produced by adding sequences that were available on GISAID as of Sept. 22, 2020. See this section for details on the sequence curation and alignment, and differences between the reduced and entire datasets.

Phylodynamic Analyses

Our objective is to infer the joint posterior probability distribution of the viral phylogeny, divergence times, and biogeographic history under a composite phylodynamic model that is appropriate for the entire SARS-CoV-2 dataset. The composite phylodynamic model is comprised of four main components: (1) a substitution model that describes the evolution of nucleotide

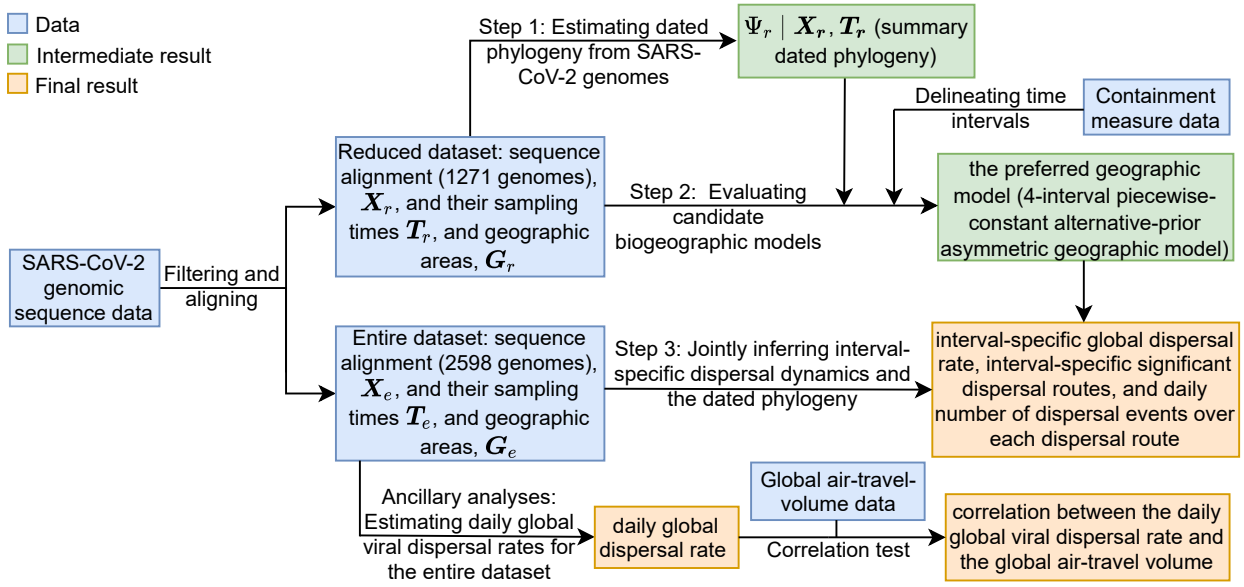

Figure S.3.9: Workflow of the empirical analyses in our study.

sequences over the tree; (2) a branch-rate prior model that characterizes how rates of substitution vary across branches of the tree; (3) a branching-process model that specifies the prior distribution of the tree topologies and divergence times, and; (4) a biogeographic model that describes how viruses disperse between geographic areas.

For each of these components, there are numerous candidate models; the vast space of composite phylodynamic models makes it computationally prohibitive to evaluate the fit of each candidate model to the entire dataset. Accordingly, we adopt a three-step model-selection procedure: (1) we first estimate the dated phylogeny for the reduced dataset under a relaxed-clock model with biologically motivated specification of the substitution model, branch-rate prior model, and branching-process prior model; (2) we then condition on the resulting dated phylogeny to select among candidate biogeographic models using the reduced dataset, and; (3) finally, we perform joint inference of the dated phylogeny and biogeographic history for the entire dataset using the preferred composite phylodynamic model.

Step 1: Estimating the dated phylogeny of the reduced dataset

We inferred a dated phylogeny by performing Bayesian analyses of the reduced SARS-CoV-2 sequence dataset under a relaxed-clock model, which includes the first three of the four model components of the composite phylodynamic model: (1) a substitution model; (2) a branch-rate prior model, and; (3) a branching-process prior model.

Specifically, we specified a partitioned substitution model to accommodate possible variation in the evolutionary process across genomic regions. The SARS-CoV-2 genome is comprised of 11 gene regions (we list these gene regions and their corresponding coordinates in the reference genome in Table S.3.3). We partitioned the SARS-CoV-2 genomes into six data subsets, with three subsets for the ORF1ab gene region (one for each codon position), and three subsets for the remaining ten combined gene regions (one for each codon position). For each of these data subsets, we specified an independent TN93 substitution model (Tamura and Nei 1993). We used partition-specific rate multipliers to capture differences in the substitution rate across the six data subsets. We specified a discrete-gamma model to accommodate substitution-rate variation across sites within each data subset. Our preliminary analyses specified an independent discrete-gamma model for each of the six data subsets, which revealed a similar degree of among-site rate variation within each data subset (*i.e.*, with similar posterior estimates of the six α-shape parameters). Accordingly, to decrease model complexity, we specified a shared,

discrete-gamma model (Yang 1994) to accommodate substitution-rate variation across sites of the entire alignment. We specified an uncorrelated lognormal (UCLN) branch-rate prior model (Drummond et al. 2006; Li and Drummond 2012; Rannala and Yang 2007) by drawing i.i.d. rate multipliers for each branch from a shared underlying lognormal distribution, where the parameters of this distribution (mean and standard deviation) are estimated from the data. For the branching-process prior model, we used a coalescent model with exponential population growth.

We performed MCMC simulations to approximate the joint posterior distribution of the relaxed-clock model parameters and the dated phylogeny using `BEAST` (Suchard et al. 2018). We then used `TreeAnnotator` to generate a summary phylogeny from the combined posterior sample of dated phylogenies as a maximum clade credibility (MCC) tree. We provide a more detailed description of these analyses in this section. The phylogeny inferred from these analyses was used both for our simulation study and also in the next step to evaluate candidate biogeographic models.

Step 2: Evaluating candidate biogeographic models using the reduced dataset

We explored a pool of nine candidate biogeographic models. These models assign interval-specific parameters—for the average rate of viral dispersal, μ, and/or relative rates of viral dispersal, $\mathbf{Q}$—to one, two, four, or five pre-specified time intervals; *i.e.*, $1\mu 1\mathbf{Q}$, $1\mu 2\mathbf{Q}$, $2\mu 1\mathbf{Q}$, $2\mu 2\mathbf{Q}$, $1\mu 4\mathbf{Q}$, $4\mu 1\mathbf{Q}$, $4\mu 4\mathbf{Q}$, $5\mu 5\mathbf{Q}$, and $5\mu 5\mathbf{Q}^*$. We specified interval boundaries based on external information regarding events within the study period that might plausibly impact viral dispersal dynamics, including: (A) start of the Spring Festival travel season in China (the highest annual period of domestic travel, January 12); (B) onset of mitigation measures in Hubei province, China (January 26); (C) onset of international air-travel restrictions against China (February 2), and; (D) relaxation of domestic travel restrictions in China (February 16). Phylodynamic models with two intervals include event C, models with four intervals include events A, C, and D, and the $5\mu 5\mathbf{Q}$ model includes all four events. The final candidate model, $5\mu 5\mathbf{Q}^*$, includes five arbitrary and uniform (bi-weekly) intervals.

We assessed both the *relative* and *absolute* fit of these candidate biogeographic models to our reduced SARS-CoV-2 dataset. We assessed the *relative fit* of competing biogeographic models by computing Bayes factors based on their marginal-likelihood estimates. We performed power-posterior MCMC simulations using `BEAST` (Suchard et al. 2018) to estimate marginal

likelihoods using both thermodynamic-integration (Lartillot and Philippe 2006) and stepping-stone (Xie et al. 2011; Baele et al. 2012) estimators. We also assessed the *absolute fit* of each model using posterior-predictive simulation (Gelman et al. 1996). For each model, we first inferred the joint posterior distribution from the observed biogeographic data (*i.e.*, the geographic location of each sampled sequence) by performing MCMC simulations using `BEAST` (Suchard et al. 2018). We then simulated predictive datasets by repeatedly sampling at random from the corresponding joint posterior probability distribution for a given the model. Finally we generated posterior-predictive distributions from each predictive dataset under various summary statistics (as described in this section), which measure the discrepancy between the observed dataset and the simulated dataset. We provide a more detailed description of these analyses in this section. The preferred biogeographic model identified by these analyses was then used in our subsequent joint phylodynamic analyses, described below.

Step 3: Joint phylodynamic inference of the entire dataset

We performed joint inference of the phylogeny, divergence times, and biogeographic history using the entire SARS-CoV-2 dataset based on a phylodynamic model that includes (1) a relaxed-clock model, and (2) a biogeographic model. The relaxed-clock model specified in these joint analyses is identical to that specified in Step 1 (with minor changes in the prior specification to reflect differences in viral sampling). The biogeographic model specified in these joint analyses is identical to the biogeographic model selected in Step 2: the 4-interval ($4\mu4\mathbf{Q}$) model.

We performed MCMC simulations to approximate the joint posterior distribution of the phylodynamic-model parameters using `BEAST` (Suchard et al. 2018). We also performed posterior-predictive simulation to confirm that the preferred biogeographic model provides an adequate fit to the entire SARS-CoV-2 dataset under the joint inference. We provide a more detailed description of these analyses in this section.

Ancillary analyses: Estimating daily global viral dispersal rates for the entire dataset

We performed additional analyses to explore the correlation between daily global air-travel volume and daily average global SARS-CoV-2 dispersal rate during the early phase of the COVID-19 pandemic. We first estimated the daily global dispersal rate using the entire SARS-CoV-2 dataset under a more granular interval-specific phylodynamic model that allows daily

variation in the average viral dispersal rate. We then computed the correlation between these estimates and independent information on the daily volume of global air travel during this period. The phylodynamic model we specified for these analyses was identical to that used in Step 3, except that we further discretized the time intervals for the average global dispersal rate to vary daily. In these analyses, we accommodated phylogenetic uncertainty by averaging over the marginal posterior probability distribution of dated phylogenies inferred in Step 3. We performed MCMC simulations to approximate the joint posterior distribution of the daily-rate model parameters using `BEAST` (Suchard et al. 2018). We then performed a standard correlation test between the daily global air-travel volume and the estimated mean daily global SARS-CoV-2 dispersal rates by computing Pearson's r and the corresponding p value, focussing on the period spanning from Jan. 31 (by which date the virus achieved a global distribution) to Mar. 8, 2020. We provide a more detailed description of these analyses in this section.

Data and Code Availability

GISAID accession IDs of the SARS-CoV-2 sequences used in this study, as well as the flight-volume data (obtained from FlightAware, LLC) and intervention-measure data, are maintained in the GitHub repository (`https://github.com/jsigao/interval_specific_phylodynamic_models_supparchive`) and archived in the Dryad repository (`https://datadryad.org/stash/share/vTbeDwLq2uSL9rL4NCe_Cocp2bY7BgWTI2tUgoNrLDA`). Our repositories also contain `BEAST` XML scripts used to perform the phylodynamic analyses, `R` scripts used to perform simulations and post processing, and a modified version of the `BEAST` program used for some of the analyses in this study.

Detailed Description of Data Acquisition and Curation

Epidemiological Data

COVID-19 case numbers

We obtained the number of confirmed COVID-19 cases from five major sources: (1) the WHO COVID-19 situation reports (WHO 2020), (2) the COVID-19 dashboard published on a Chinese medical website, Ding Xiang Yuan (DXY), that integrates data from local governmental reports (DXY 2020), (3) the National Health Commission of the People's Republic of China (NHCPRC) COVID-19 situation reports (NHCPRC 2020), (4) the European Centre for Disease Prevention and Control (ECDC) COVID-19 situation update (ECDC 2020), and (5) the US Centers for Disease Control and Prevention (USCDC) COVID-19 data tracker (USCDC 2020). Rather than directly collecting data from these sources, we accessed them via two intermediate portals: the `R` (R Core Team 2020) package `nCov2019` (Wu et al. 2020b), and the COVID-19 Data Repository by the Center for Systems Science and Engineering at Johns Hopkins University (Dong et al. 2020).

Intervention measures

We focused on two types of intervention measures enacted during the early phase of the COVID-19 pandemic that involved China: targeted-containment measures involving China (*i.e.*, international air-travel bans), and domestic-mitigation measures within China. We compiled information on these containment measures from various news reports, Wikipedia pages (Wikipedia 2020a,b), and peer-reviewed publications (Kraemer et al. 2020; Tian et al. 2020; Hsiang et al. 2020; Lai et al. 2020). For domestic measures within China, we focused on measures that were likely to interrupt travel among regions, including lockdowns at city or province levels, inter-city travel restrictions, and home or neighborhood isolation. See `international_airtravelban_withchina.csv` for a collection of countries or territories that enacted international travel bans with China (and the associated initiation date), and `china_domestic.csv` for a collection of provinces or cities that enacted mitigation measures in China (including the associated implementation period and type of measure); these spreadsheets are included in our GitHub and Dryad repositories.

Travel Data

We acquired global air-travel-volume data from FlightAware, detailing the number of all commercial passenger flights (subdivided by each aircraft type) per day between Dec. 30, 2019 and Mar. 8, 2020. We transformed these daily aircraft-volume data to provide an estimate of the daily air-travel passenger volume (Fig. 5, dashed line) by multiplying the number of flights for each type of aircraft by the capacity (seat number) for the corresponding type of aircraft. The original air-travel-volume data are contained in `nflights_daily_byaircraft.csv`, and the number of seats for each aircraft type is provided in `aircraft_nseats.csv` (included in our GitHub and Dryad repositories).

Definition of Time Intervals

We partitioned the early phase of the COVID-19 pandemic into five time intervals: (1) interval 1 from late 2019 (origin time) to Jan. 12, 2020; (2) interval 2 from Jan. 13 to Jan. 25; (3) interval 3 from Jan. 26 to Feb. 2; (4) interval 4 from Feb. 3 to Feb. 16, and; (5) interval 5 from Feb. 17 to Mar. 8, 2020.

The Jan. 12 boundary coincides with the start of the Spring Festival travel season in China (the highest annual period of domestic travel). The Jan. 26 boundary coincides with onset of widespread mitigation measures in China to restrict domestic travel: these measures began with the city-wide lockdown of Wuhan on Jan. 23 (that were extended to the entire Hubei province in the following days), followed by the declaration of level-1 emergency in all mainland provinces between Jan. 24–29, the extension of the Spring Festival national holiday (effectively school and workplace closure) announced on Jan. 27, and the enactment of stringent home- or neighborhood-isolation orders in various cities outside Hubei beginning Feb. 2. The Feb. 2 boundary coincides with the initiation of international air-travel bans with China (imposed by 34 countries by this date) and the cancellation (or significant reduction) of international air services involving China (by over 130 airlines; International Civil Aviation Organization 2020; Wikipedia 2020b), following the declaration of Public Health Emergency of International Concern (PHEIC) by the World Health Organization (WHO) on Jan. 31. The Feb. 17 boundary coincides with the lifting of travel restrictions in China (except in Hubei, where the travel restrictions were not lifted until late Mar.).

SARS-CoV-2 Genomic Sequence Data

We curated two SARS-CoV-2 genomic sequence datasets for our study, one with 1271 sequences (the "reduced dataset") and the other with 2598 sequences (the "entire dataset").

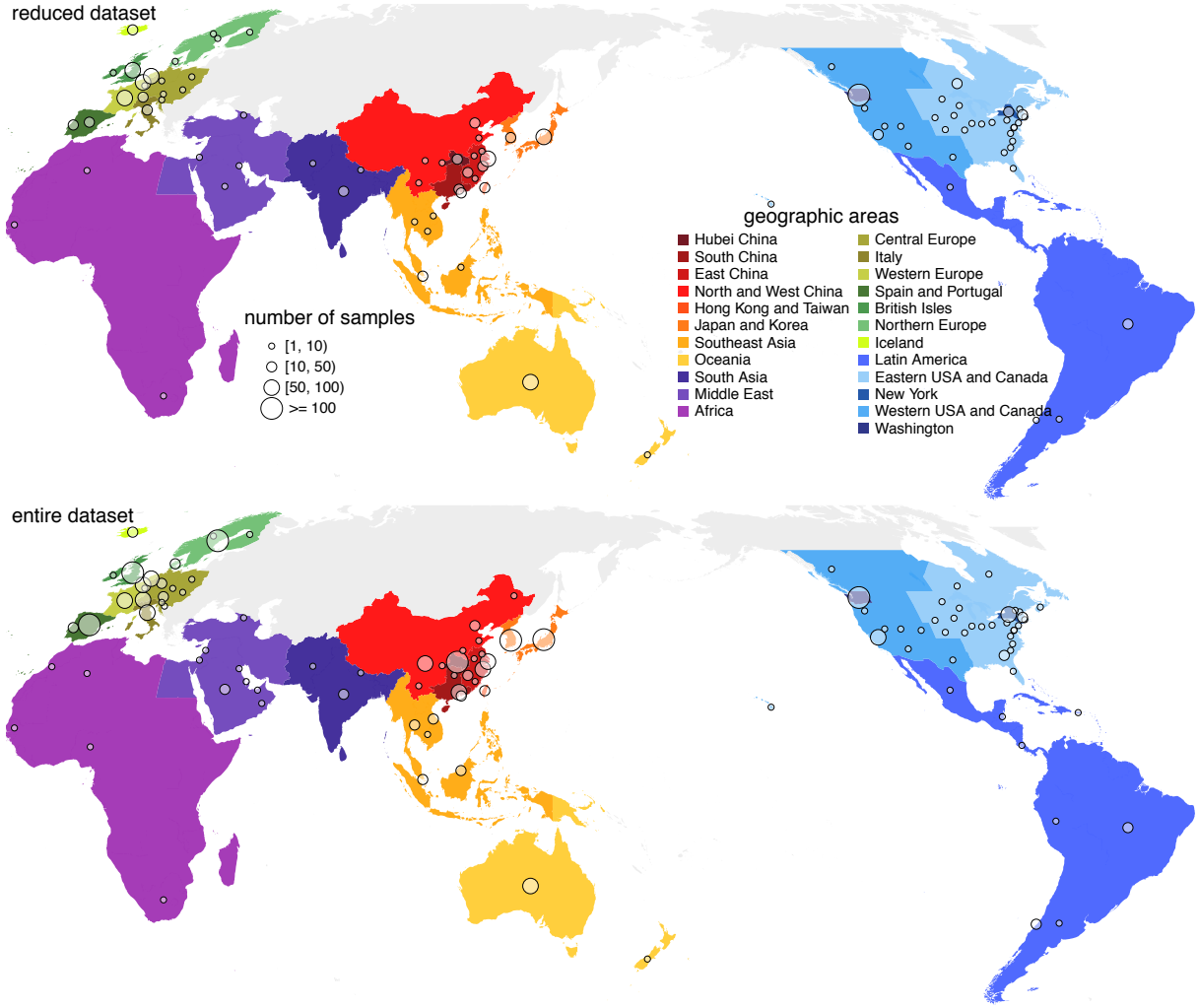

Figure S.3.10: SARS-CoV-2 genomes sampled from each discretized geographic area. Our study includes two SARS-CoV-2 datasets—the reduced (1271 sequences, top) and entire (2598 sequences, bottom)—that comprise viral genomes collected between Dec. 24, 2019–Mar. 8, 2020 from 23 discrete geographic areas (colored regions); circles indicate the number and location of samples in our study.

Assembling the reduced dataset

The reduced dataset consists of all available SARS-CoV-2 genomic sequences available as of Apr. 19, 2020 from GISAID (https://www.gisaid.org/; Shu and McCauley 2017). As our focus is on the crucial early phase of the COVID-19 pandemic, we excluded sequences that were collected after Mar. 8, leaving 2003 sequences in the dataset. We first filtered the dataset

by excluding sequences that fit any of the following conditions: (1) fewer than 29000 sites (not counting missing or gap sites); (2) lacking associated metadata (*e.g.*, sampling time or location); (3) lacking the precise sampling date or geographic location (state/province for sequences from China, Canada, or U.S.A. and country for the others); (4) sampled from a non-human host; (5) multiple sequences from the same individual (in which case we randomly selected one of sequence and discarded the others), or; (6) duplicates of other sequences in the dataset [for this purpose, we used the "exclude list" used by `Nextstrain` (https://github.com/nextstrain/ncov/blob/master/defaults/exclude.txt) as a reference]. Application of these filters resulted in a genomic dataset consisting of 1620 sequences.

We then inferred an alignment of these nucleotide sequences using `MUSCLE` version 3.8 (Edgar 2004). We performed a second round of filtration of the resulting alignment. First, we excluded sequences that appeared to be anomalously divergent; this was achieved by comparing each sequence to the reference genome while assuming that the rate of mutation accu-

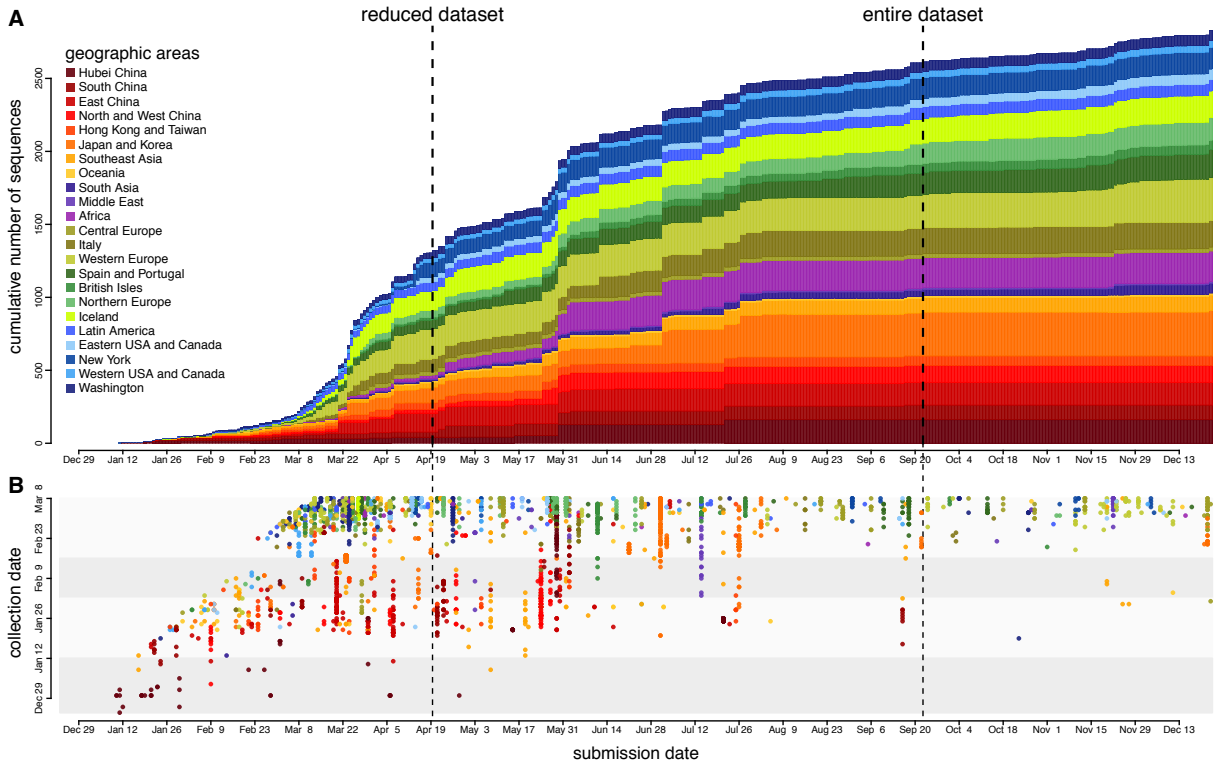

Figure S.3.11: Submission and collection dates for SARS-CoV-2 genomic sequences. (A) Cumulative number of SARS-CoV-2 sequences submitted to GISAID that were collected during the early phase of the COVID-19 pandemic. The color of each segment in the stacked bar plot indicates the number of sequences submitted from the corresponding geographic area on that day. **(B)** Submission and collection dates for each SARS-CoV-2 sequence included in our study. The deposition rate of sequences collected prior to Mar. 8 drastically decreased in Sept. 2020.

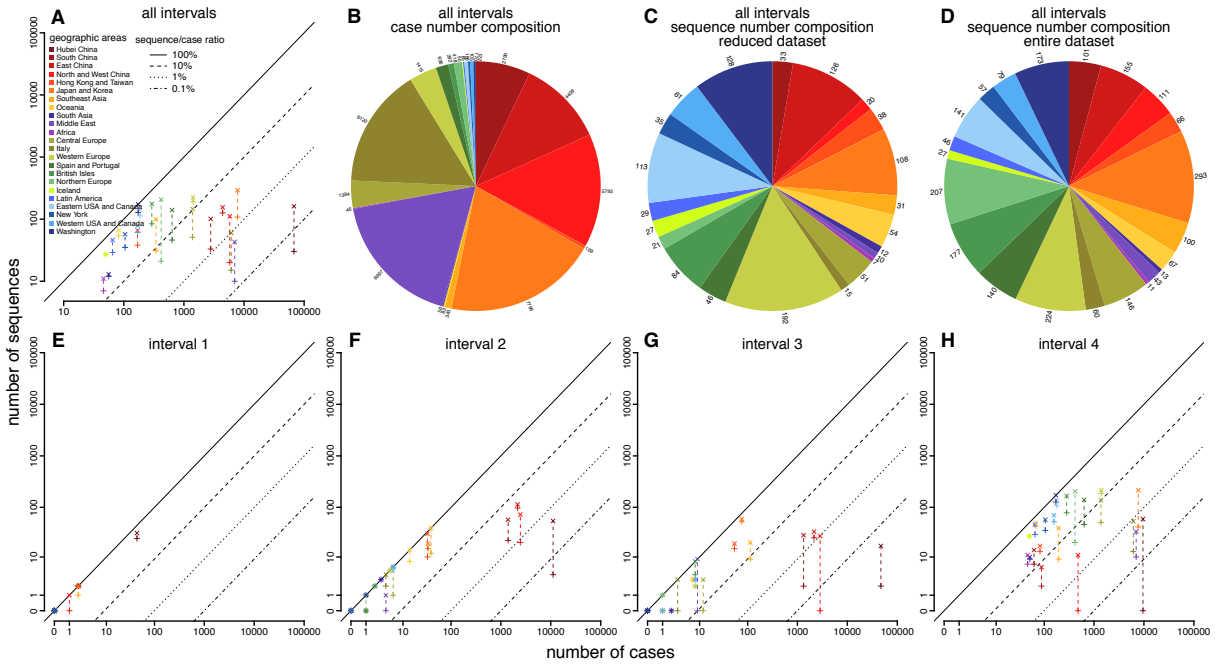

Figure S.3.12: Number of SARS-CoV-2 genome sequences versus number of confirmed COVID-19 cases. (A) Number of sequences versus number of confirmed cases across the early phase of the pandemic. + and x indicate the reduced and entire datasets, respectively; they are connected by dashed line for each geographic area to show the increase of the number of sequences (and thus the sequence/case ratio) in the entire dataset. **(B)** Geographic distribution of confirmed case numbers (excluding Hubei). **(C)** Geographic distribution of sequence number (excluding Hubei; reduced dataset). **(D)** Geographic distribution of sequence number (excluding Hubei; entire dataset). **(E–H)** Number of sequences versus number of confirmed cases for each interval of the early phase, respectively.

mulation should not exceed 10 mutations per genome per month. We also excluded sequences with many ambiguous sites (*i.e.*, sites for which the nucleotide could not be unambiguously identified); specifically, we discarded sequences with more than 15 ambiguous sites, and sequences with at least 10 ambiguous sites and fewer than 10 sites differing from the reference genome. Next, we excluded sequences with nonsense mutations; to this end, we translated the nucleotide alignment into an amino-acid alignment using the `seqinr` package (Charif and Lobry 2007) in `R` (R Core Team 2020) to identify sequences with premature stop codons. We assumed that the rate of amino-acid substitution accumulation should not exceed 4 substitutions per genome per month; we therefore discarded sequences with more than 6 ambiguous amino-acid sites, and sequences with at least 3 ambiguous amino-acid sites but fewer than 3 sites differing from the reference genome. After filtering, our reduced dataset included 1271 sequences.

Assembling the entire dataset

We also compiled a more comprehensive dataset by curating all sequences available from GISAID as of Sept. 22, 2020. Specifically, we downloaded an alignment from GISAID, which was inferred using `MAFFT` (Katoh and Standley 2013). After excluding sequences that were collected after Mar. 8, 2020, the alignment included 4012 sequences. We then performed the same two-step filtration procedure that we applied to the reduced dataset, culminating in an alignment ("entire dataset") with 2598 sequences.

The entire dataset is more comprehensive than the reduced dataset: it contains more than twice the number of sequences (Fig. S.3.11), and is also more evenly sampled, as the sequence-to-case ratios of many undersampled geographic areas are significantly higher, especially for the third and fourth intervals of our study (Fig. S.3.12). Moreover, the entire dataset contains SARS-CoV-2 genomic sequences that are likely to represent the vast majority of such data that will ever be available; the deposition rate of sequences collected from the early phase of the pandemic drastically decreased in Sept., 2020 (Fig. S.3.11).

Trimming and partitioning the curated alignment

For each curated dataset, we trimmed the 5'UTR and 3'UTR as well as the other non-coding regions, retaining only coding regions in the alignment. Table S.3.3 lists the coding regions and

Table S.3.3: Genomic coordinates of the SARS-CoV-2 coding regions.

Region	Starting coordinate	Ending coordinate
ORF1ab*	266	21555
S	21563	25384
ORF3a	25393	26220
E	26245	26472
M	26523	27191
ORF6	27202	27387
ORF7a	27394	27759
ORF7b	27756	27887
ORF8	27894	28259
N	28274	29533
ORF10	29558	29674

*During translation, ORF1ab experiences a -1 ribosomal frameshift at site 13468, so the range is (266–13468, 13468–21555).

their corresponding coordinates in the reference genome (Wuhan-Hu-1, Wu et al. 2020a). After removing the stop codon for each coding region, both the reduced and entire alignments for the complete coding region included 29,232 nucleotide sites.

Detailed Description of Phylodynamic Analyses

Estimating a Dated Phylogeny for the Reduced SARS-CoV-2 Dataset

Overview

In this section, we describe the analyses that we performed to infer a dated phylogeny for the reduced sample of COVID-19 viruses. We use the phylogeny inferred from these analyses both in our simulation study (see this section) and also in our subsequent analyses to evaluate candidate biogeographic models (see *Evaluating Candidate Biogeographic Models*).

Model specification

We inferred a dated phylogeny by performing Bayesian analyses of the reduced SARS-CoV-2 sequence dataset under a relaxed-clock model, which includes three main components: (1) a substitution model; (2) a branch-rate prior model; and (3) a branching-process prior model. Below, we describe each of these model components and the corresponding priors for the parameters of those models (note that we used an empirical Bayesian approach to specify non-default priors for several parameters; *i.e.*, where the results of preliminary analyses and/or published results were used to specify the parameters of priors. Details of the priors are described in Table S.3.4.

Substitution model.—The substitution model collectively describes the process of molecular evolution of the SARS-CoV-2 genomes over the branches of the phylogeny. The process of molec-

Table S.3.4: Priors used to estimate a dated phylogeny of the sampled SARS-CoV-2 sequences.

Parameter	Description	Prior
κ_1	Ratio of the A $\rightarrow$ G rate to the transversion rate	Lognormal$(\mu = 1.0, \sigma = 0.8)^*$
κ_2	Ratio of the C $\rightarrow$ T rate to the transversion rate	Lognormal$(\mu = 1.0, \sigma = 0.8)$
$\boldsymbol{\pi}$	Nucleotide stationary frequencies	Dir$(1, 1, 1, 1)$
$\boldsymbol{m}$	Partition-specific rate multipliers	Dir$(1, 1, 1, 1, 1, 1)$
α	Shape and scale parameter of the Γ_4 distribution	Lognormal$(\mu = -2.1, \sigma = 0.5874)$
$\mathbb{E}[\boldsymbol{r}]$	Mean of the UCLN	Lognormal$(\mu = -12.7, \sigma = 0.5874)$
$SD(\boldsymbol{r})$	Standard deviation of the UCLN	Exp$(\lambda = 1/(2.0\text{e}{-6}))$
N_T	Effective number of infected individuals at sampling time, T	Lognormal$(\mu = 7.5, \sigma = 1.0)$
r	Exponential growth rate of the coalescent model	Laplace$(0.07, 0.01)$

*μ and σ in this table are the mean and standard deviation of the normal distribution.

ular evolution is apt to vary among regions of these viral genomes. For example, the ORF1ab gene of SARS-CoV-2 encodes nonstructural proteins and is therefore likely to have been subjected to strong purifying selection (Li et al. 2020b), whereas other genes, such as the spike (S) gene, encode structural proteins that determine antigenicity and other immune properties of SARS-CoV-2, and are therefore likely to have been subjected to strong positive selection (Korber et al. 2020; Plante et al. 2020; Hou et al. 2020; Volz et al. 2021). Accordingly, we specified a partitioned substitution model to accommodate possible variation in the evolutionary process across genomic regions. Specifically, we partitioned the SARS-CoV-2 genomes into six data subsets, with three subsets for the ORF1ab gene region (one for each codon position), and three subsets for the remaining ten combined gene regions (one for each codon position). (For a complete list of the gene regions and their corresponding coordinates in the reference genome, see Table S.3.3).

For each of these data subsets, we specified an independent TN93 substitution model (Tamura and Nei 1993), with transition-transversion rate-ratio parameters κ_1 and κ_2 (the instantaneous rates of A to G and C to T substitutions, respectively, relative to the transversion rate) and π (the stationary frequency of each nucleotide). For each transition-transversion rate-ratio parameter, we specified lognormal priors with a prior mean of 3.74 and a 95% prior interval of $[0.56, 13.04]$.

To accommodate possible variation in the overall rate of substitution *between* gene regions, we specified independent rate multipliers for each of the six data subsets. To accommodate variation in substitution rates across sites *within* each data subset, we specified a discrete-gamma model (Yang 1994). Our preliminary analyses specified an independent discrete-gamma model for each of the six data subsets, which revealed a similar degree of among-site rate variation within each data subset (*i.e.*, with similar posterior estimates for the six α-shape parameters). Accordingly, to decrease model complexity, we specified a shared, discrete-gamma model for the entire alignment. We specified a lognormal hyperprior on the α-shape parameter, with a prior mean of 0.15 and a 95% prior interval spanning one order of magnitude around the mean, $[0.039, 0.39]$. This prior reflects our expectation of a high degree of substitution-rate variation across sites, motivated by our observation that most sites in our SARS-CoV-2 alignment are invariant, while a small number of sites appear to be highly variable.

Branch-rate model.—The branch-rate model describes how the overall substitution rate varies across branches of the tree. Our composite relaxed-clock model specifies the uncorrelated log-normal (UCLN) branch-rate prior model (Drummond et al. 2006; Li and Drummond 2012; Rannala and Yang 2007), which accommodates variation in the overall substitution rate across branches by drawing i.i.d. rate multipliers for each branch from a shared underlying log-normal distribution, where the parameters of this distribution (mean and standard deviation) are estimated from the data. For the mean of the UCLN, we specified a lognormal hyperprior with an expectation of 3.63e−6 substitutions/site/day and 95% prior interval of [0.96e−6, 9.64e−6], motivated by published substitution-rate estimates of approximately 30 substitutions/genome/year (*cf.* Duchene et al. 2020). We specified an exponential hyperprior on the standard deviation of the UCLN such that the branch-specific substitution rates are expected to vary over approximately one order of magnitude.

Branching-process model.—The branching-process model describes the prior distribution of tree topologies and divergence times. We used a coalescent model with exponential population growth as our branching-process model. This model assumes that the viral population size grows as a deterministic exponential function (Beaumont 1999; Drummond et al. 2002), which is motivated by the fact that our SARS-CoV-2 dataset was sampled from the early, explosive stage of the COVID-19 pandemic. This model is completely described by two free parameters: N_T, the effective number of infected individuals in the population at the sampling time, T (*i.e.*, the last sampling date in our dataset, Mar. 8, 2020), and r, the exponential growth rate. We specified empirically informed and biologically realistic priors on these parameters.

Prior on the exponential growth rate, r.—We specified a prior on r using external information about the R_0, the basic reproductive number, and τ, the duration of the infectious period; these quantities are related through the equation $r = (R_0 - 1)/\tau$. We specified a Laplace prior on r, with location and scale parameter values specified according to published estimates of R_0 and τ for COVID-19 (Chinazzi et al. 2020; Li et al. 2020a; Hao et al. 2020; Vaughan et al. 2020; Nadeau et al. 2021; Wölfel et al. 2020; van Kampen et al. 2021; Byrne et al. 2020). This prior on r directly translates to an expected population doubling time, $t = \ln(2)/r$, of 10.6 days (with 95% prior interval ranging from 6.9 to 17.2 days).

Prior on the effective population size, N_T.—Given a population doubling time of t, the expected number of individuals at time T is $N_T = N_0 2^{(T-T_0)/t}$, where N_0 is the number of individuals at the beginning of the process, and T_0 is the origin time of the process. (This equation follows from the fact that there are $(T - T_0)/t$ doubling cycles in a period of duration $T - T_0$.) We therefore specified a prior on N_T informed by our previously determined prior on t, as well as several realistic values for N_0 (1 or 2) and T_0 (some time in late Nov., 2019). Based on these values, we chose a lognormal prior on N_T such that the mean was 2981 and 95% prior interval spanned $[254, 12840]$. Given that there were at least 3940 reported cases on Mar. 8, 2020 alone, it may seem unreasonable to specify a prior such that the expected number of individuals is as low as 2981. However, we note that N_T represents the *effective* number of infected individuals in the population, which is typically substantially smaller than the *total* number of infected individuals in the population. Additionally, we note that: (1) the posterior-mean estimate of N_T under this prior is $\approx$801, indicating that, if anything, this prior mean is too high, and; (2) sensitivity analyses suggested that posterior estimates of N_T were not very sensitive to this prior (results not shown).

Parameter estimation

We performed six independent MCMC simulations to approximate the joint posterior distribution of the relaxed-clock model parameters using `BEAST` version 1.10.5 (Suchard et al. 2018) with the `BEAGLE` library (compiled from the 'hmc-clock' branch, commit 'dd36bf5'; Ayres et al. 2019) to accelerate computation. We ran each replicate MCMC simulation for 50 million generations, sampling continuous parameters every 1000 generations and trees every 10,000 generations. Details of these analyses (*e.g.*, proposal weights) are available in the XML scripts included in our GitHub and Dryad repositories. After discarding the first 20% as the burn-in from each replicate simulation, we combined the remaining posterior samples of trees from all the replicates and then down-sampled every 50,000 generations using `LogCombiner` version 1.10.5. Following initial inspection of the log files using `Tracer` (Rambaut et al. 2018) version 1.7.1, we further evaluated MCMC performance using the `coda` package (Plummer et al. 2006) in `R` (R Core Team 2020). We assessed convergence of replicate MCMC simulations by calculating the ESS for each continuous parameter for the combined posterior samples; ensuring that values for the substitution-model parameters were all $\gg$10000 and those for the branch-rate and branching-process models were all $\gg$200. We then used `TreeAnnotator` version 1.10.5 to

generate a summary phylogeny from the combined posterior sample of trees—as a maximum clade credibility (MCC) tree—where the age of each internal node is computed by marginalizing over the age of that node across all samples. Note that as the age of each node is summarized independently across the posterior distribution of trees, it is possible for the MCC summary tree to have negative branch lengths (*i.e.*, where an ancestral node is younger than its descendant node). To avoid potential issues caused by this phenomenon in downstream analyses, we assigned a small positive value (0.001 days) as the duration of these "time-traveling" branches.

Evaluating Candidate Biogeographic Models

Overview

In this section, we describe our analyses to explore candidate biogeographic models that describe the geographic progression of the SARS-CoV-2 virus during the early phase of the COVID-19 pandemic. We begin by defining the space of candidate models that we will evaluate, and then describe the analyses that we performed to assess both the *relative fit* (by computing Bayes factors to compare competing models) and the *absolute fit* (using posterior-predictive simulation) of these candidate biogeographic models to our reduced SARS-CoV-2 dataset. In evaluating candidate biogeographic models, we condition on the MCC summary phylogeny inferred using the reduced dataset described above (see this section: *Estimating a Dated Phylogeny for the Reduced SARS-CoV-2 Dataset*).

Candidate biogeographic models

Specifying priors for biogeographic models.—For a biogeographic history with k discrete areas, the stochastic process of geographic dispersal over the branches of the tree is fully specified by a $k \times k$ instantaneous-rate matrix, $\mathbf{Q}$, where an element of the matrix, q_{ij}, is the instantaneous rate of change between state i and state j (*i.e.*, the instantaneous rate of dispersal from area i to area j). Each element, q_{ij}, of the instantaneous-rate matrix, $\mathbf{Q}$, is specified as:

$$q_{ij} = r_{ij}\delta_{ij},$$

where r_{ij} is the rate of dispersal between areas i and j, and δ_{ij} is an indicator variable that takes one of two states (1 or 0); when $\delta_{ij} = 1$, a dispersal route from area i to area j exists, when $\delta_{ij} = 0$ it does not. The total number of dispersal routes, $\sum \delta_{ij}$, for a given biogeographic model is denoted Δ. We used an asymmetric $\mathbf{Q}$ matrix (Edwards et al. 2011) that allows the rate of dispersal from area i to area j to be different from the rate of dispersal from area j to area i (*i.e.*, r_{ij} can be different from r_{ji}, and δ_{ij} can also be different from δ_{ji}). By convention, we rescale the $\mathbf{Q}$ matrix such that the expected number of dispersal events in one time unit is equal to the parameter μ (Yang 2014). We specified the root frequency ω—the prior probability of the geographic area at the root—as a stochastic random variable to be estimated from the data.

Prior on the number of dispersal routes..—We specified a Poisson prior on the number of dispersal routes, Δ, with the rate parameter of the Poisson distribution, $\lambda = \binom{k}{2}$, representing a prior belief that half of all possible dispersal routes are included in the biogeographic model; this

results in a relatively flat prior probability that any given dispersal route exists for all values of k.

Prior on the average dispersal rate..—Recall that the rate matrix, $\mathbf{Q}$, is rescaled so that the average rate of dispersal between all areas is μ. For a tree of length T (*i.e.*, the sum of the durations of all branches in the tree), the expected number of dispersal events is $\mu \times T$. Therefore, the prior on μ represents our prior belief about the number of dispersal events over the tree. We specified an exponential prior on μ with rate parameter θ, and a mean of $1/\theta$. Rather than assuming a fixed value for the mean of the exponential prior, we treat it as a random variable to be estimated from the data. Specifically, we specified a gamma hyperprior on $1/\theta$; this gamma hyperprior has shape parameter $\alpha = 0.5$ and rate parameter $\beta = 0.5$ (enforcing the shape and rate parameters to be equal ensures that the resulting prior on μ is proper). The resulting prior—known as the *K*-distribution (Jakeman and Pusey 1978)—is a rather diffuse prior on μ, as is the resulting prior distribution on the number of dispersal events.

Space of candidate biogeographic models.—We explored a pool of nine candidate biogeographic models. These models assign interval-specific parameters—for the average rate of viral dispersal, μ, and/or relative rates of viral dispersal, $\mathbf{Q}$—to one, two, four, or five pre-specified time intervals; *i.e.*, $1\mu1\mathbf{Q}$, $2\mu1\mathbf{Q}$, $1\mu2\mathbf{Q}$, $2\mu2\mathbf{Q}$, $4\mu1\mathbf{Q}$, $1\mu4\mathbf{Q}$, $4\mu4\mathbf{Q}$, $5\mu5\mathbf{Q}$, and $5\mu5\mathbf{Q}^*$. For example, $4\mu1\mathbf{Q}$ is an interval-specific biogeographic model that allows the average dispersal rate to vary among the four time intervals (but assumes that the relative dispersal rates are constant among intervals). Conversely, $1\mu4\mathbf{Q}$ assumes interval-specific relative dispersal rates (but assumes a constant average dispersal rate across the four time intervals). The $4\mu4\mathbf{Q}$ model may be viewed as a composite of former two models, as it allows *both* the average *and* relative dispersal rates to vary independently among the four intervals.

We specified interval boundaries based on external information regarding events within the study period that might plausibly impact viral dispersal dynamics, including: (A) start of the

Table S.3.5: Priors used in evaluating candidate biogeographic models.

Parameter	Description	Prior
Δ_l	Number of dispersal routes in interval l	$\text{Pois}(253)$
μ_l	Average dispersal rate in interval l	$\text{Exp}(1/\lambda)$; $\lambda \sim \Gamma(0.5, 0.5)$
$r_{ij,l}$	Relative dispersal rate from i to j in interval l	$\Gamma(1,1)$
$\boldsymbol{\omega}$	Root frequencies	$\text{Dir}(1,1,\ldots,1)$

Spring Festival travel season in China (the highest annual period of domestic travel, Jan. 12); (B) onset of mitigation measures in Hubei province, China (Jan. 26); (C) onset of international air-travel bans against China (Feb. 2), and; (D) relaxation of domestic travel restrictions in China (Feb. 16). Biogeographic models with two intervals include event C, models with four intervals include events A, C, and D, and the $5\mu5\mathbf{Q}$ model includes all four events. The final candidate model, $5\mu5\mathbf{Q}^*$, includes five arbitrary and uniform (bi-weekly) intervals. (See this section for additional details on these time intervals.)

Evaluating the models

Assessing relative fit of candidate biogeographic models using Bayes factors.—We evaluated the *relative fit* of each candidate biogeographic model to our SARS-CoV-2 dataset using Bayes factors. This Bayesian model-comparison approach requires that we first estimate the marginal likelihood for each candidate biogeographic model, and then compute the Bayes factor for each pair of competing models as twice the difference in their log marginal likelihoods (Kass and Raftery 1995). We estimated marginal likelihoods for each candidate biogeographic model using both thermodynamic-integration (Lartillot and Philippe 2006) and stepping-stone (Xie et al. 2011; Baele et al. 2012) estimators. These marginal-likelihood estimators tend to be unstable when inferring the phylogeny and biogeographic history jointly, owing to the diffuse (hyper)priors on node-age and branch-rate model parameters, as well as the vast tree space (see Baele et al. 2015). Accordingly, we estimated marginal likelihoods for our candidate biogeographic models by conditioning on the summary phylogeny (the MCC tree) that we inferred using sequence data alone (see this section).

For each candidate biogeographic model, we first ran eight replicate power-posterior MCMC simulations in `BEAST` (Suchard et al. 2018) with the `BEAGLE` library (version 3.2.0; Ayres et al. 2019). For constant phylodynamic models, we used `BEAST` version 1.10.5.; for interval-specific phylodynamic models, we used our extended version of `BEAST` (see this section). The accuracy of marginal-likelihood estimates using power posteriors depends on the number of powers as well as the number of generations per power (Xie et al. 2011). Therefore, to assess the reliability of our marginal-likelihood estimates, we used an increasing number of powers and an increasing number of generation per power (so the specific values are represented as ranges below) across replicates and checked the variation of estimates among replicates. Specifically, for each replicate power-posterior MCMC simulation, we used 36–64 powers placed at

evenly-spaced quantiles of a Beta$(0.3, 1.0)$ distribution. For each power, we discarded the initial 65000–160000 generations as burn-in and then sampled every 100 generations over the remaining 210000–480000 generations.

To assess the stability of the marginal-likelihood estimates, we also set up a "golden run" for the power-posterior analysis under each model with a large number of powers (128, placed at evenly-spaced quantiles of a Beta$(0.3, 1.0)$ distribution) and a large number of generations (three million) per power. In the interest of time, here we ran the `BEAST` analyses under each power in parallel by specifying a single XML script per power and running them independently; for each analysis we ran four replicate MCMCs, each of length one million generations with the first 25% discarded as burnin. We combined the output of each independent run to produce the output of the golden run. We then subsampled the golden-run output under each model either by the number of powers or by the number of generations per power to produce a sequence of shorter runs with either fewer powers $(8, 16, 32, 64)$ while holding the number of generations per power same as the golden run, or fewer generations per power $(10000, 25000, \ldots, 2000000)$ while holding the number of powers same as the golden run. We examined the sequence of the estimated marginal likelihood under each candidate biogeographic model as a function of the number of powers and as a function of the number of generations per power to check the convergence behavior.

After confirming the marginal-likelihood estimate under each model converged, we combined the output from all the power-posterior analyses (including the initial eight replicates as well as the golden run) to compute a single marginal likelihood for each model. Details of these analyses (*e.g.*, proposal weights) are available in the XML scripts included in our GitHub and Dryad repositories.

Assessing absolute fit of candidate biogeographic models using posterior-predictive simulation.—We assessed the *absolute fit* of each candidate biogeographic model to our reduced SARS-CoV-2 dataset using posterior-predictive simulation (Gelman et al. 1996). We first estimated the joint posterior probability distribution of parameters for the candidate model from the observed biogeographic dataset, and then we performed simulations using the parameter estimates randomly drawn from the inferred joint posterior distribution. We used the time-slice parsimony and tipwise-multinomial statistics (as described in this section) to assess the adequacy (*i.e.*, absolute fit) of each candidate model.

Estimating the joint posterior probability distribution for each candidate biogeographic model.—For each of the candidate biogeographic models, we first inferred the joint posterior distribution from the observed biogeographic data (*i.e.*, the geographic location of each of the sequences in our reduced SARS-CoV-2 dataset) by performing four independent MCMC simulations using `BEAST` (Suchard et al. 2018) with the `BEAGLE` library (version 3.2.0; Ayres et al. 2019). Specifically, the analyses under the constant ($1\mu 1\mathbf{Q}$) biogeographic models were performed using `BEAST` version 1.10.5, whereas those under the interval-specific biogeographic models were performed using our modified version of `BEAST` (see this section). For each replicate MCMC simulation, we ran 10 million generations, sampling every 2000 generations. We discarded the initial 10% of samples (as burn-in) from each replicate MCMC, and then combined the remaining posterior samples from all the replicates using `LogCombiner` version 1.10.5. We then assessed MCMC performance for the resulting composite posterior sample by inspecting the log files using `Tracer` (Rambaut et al. 2018) version 1.7.1 and the `coda` package (Plummer et al. 2006) in `R` (R Core Team 2020). We ensured that the computed ESS values for all continuous parameters were $\gg$100. Details of these analyses are available in the XML scripts included in our GitHub and Dryad repositories.

Posterior-predictive simulations.—For each candidate biogeographic model, we simulated $m = 2500$ predictive datasets by repeatedly sampling at random from the corresponding joint posterior probability distribution. We then generated posterior-predictive distributions from each set of m predictive datasets for 20 separate summary statistics. These 20 statistics include time-slice variants (with 10 time slices) of the two (parsimony and tip-wise multinomial) summary statistics. We specified 10 (weekly) time slices spanning the early phase of COVID-19 (where the first slice covers the period from the origin of SARS-CoV-2 to Jan. 5, 2020, and the remaining nine weekly slices spanning the period between Jan. 6 and Mar. 8). For each posterior-predictive distribution, we computed the posterior-predictive p value (see this section) to assess the adequacy (*i.e.*, absolute fit) of the corresponding biogeographic model.

Results

Our golden-run experiments demonstrate that our marginal-likelihood estimates converged to stable values (Fig. S.3.13). Bayes-factor comparisons of all candidate models decisively support (*i.e.*, $2\ln \mathrm{BF} \gg 10$; Table S.3.6) the 4-interval ($4\mu 4\mathbf{Q}$) biogeographic model. The preference for this model is corroborated by the results of our posterior-predictive simulations: $4\mu 4\mathbf{Q}$ was

inferred to provide an adequate absolute fit to our reduced SARS-CoV-2 dataset for every summary statistic, whereas all less complex models (*i.e.*, with fewer interval-specific parameters for the average dispersal rates, μ, and/or the relative dispersal rates $\mathbf{Q}$ (*i.e.*, were all inferred to be inadequate by at least two of the 20 time-slice summary statistics (Fig. S.3.14). Accordingly, we use the $4\mu4\mathbf{Q}$ model for our joint phylodynamic analyses of the entire SARS-CoV-2 dataset described below (see this section).

The relative fit of competing biogeographic models to the reduced SARS-CoV-2 dataset

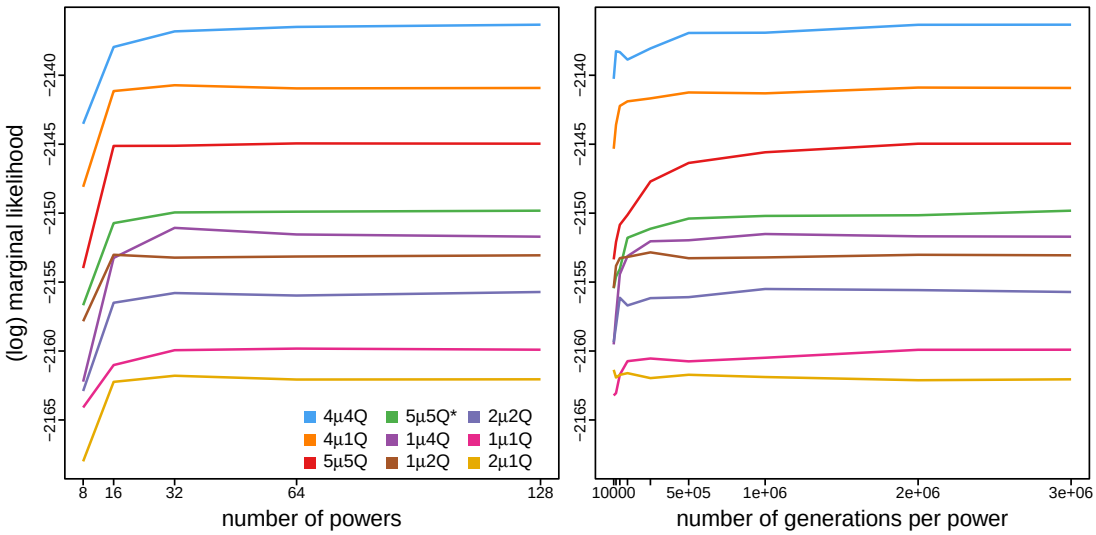

Figure S.3.13: Convergence of marginal-likelihood estimates of the candidate biogeographic models. Convergence of the marginal-likelihood estimate under each candidate biogeographic model as a function of the number of powers (left panel) and the number of generations per power (right panel). To assess the convergence of the marginal-likelihood estimate, we set up a 'golden run' under each model where a large number of powers (128, placed at evenly-spaced quantiles of a Beta$(0.3, 1.0)$ distribution) and a large number of generations (three million) per power are used. We then subsampled each golden run either by the number of powers or by the number of generations per power to produce a sequence of shorter runs with either fewer powers $(8, 16, 32, 64$; while holding the number of generations per power same as the golden run) or fewer generations $(10000, 25000, 50000, \ldots, 2000000$; while holding the number of powers same as the golden run) per power. Each colored line shows the sequence of the estimated log marginal likelihood under a given model plateauing as the number of powers increases (left) or the number of generations per power increases (right). The settings (including number of powers and the number of generations per power) of the golden run appear to be sufficient to obtain stable marginal-likelihood estimates. The 4-interval ($4\mu4\mathbf{Q}$) model appears to be consistently preferred over all the other models across all settings.

Table S.3.6: Marginal-likelihood estimates of (and Bayes factor comparisons among) the candidate biogeographic models. Column 1 lists the candidate biogeographic models. Column 2 lists the composite marginal-likelihood estimates (computed by combining the samples from replicate power-posterior MCMC simulations). The last column lists the inferred support ($2\ln \mathrm{BF}$) of the alternative interval-specific models compared to the constant model ($1\mu 1\mathbf{Q}$). The preferred biogeographic model ($4\mu 4\mathbf{Q}$) is indicated in bold text.

model	ln marginal likelihood	$2\ln \mathrm{BF}$ compared to $1\mu 1\mathbf{Q}$
$1\mu 1\mathbf{Q}$	-2159.57	—
$2\mu 1\mathbf{Q}$	-2162.03	-4.92
$1\mu 2\mathbf{Q}$	-2152.87	13.40
$2\mu 2\mathbf{Q}$	-2155.77	7.60
$4\mu 1\mathbf{Q}$	-2141.03	37.08
$1\mu 4\mathbf{Q}$	-2152.01	15.13
$4\mu 4\mathbf{Q}$	**-2136.58**	**45.99**
$5\mu 5\mathbf{Q}$	-2144.60	29.94
$5\mu 5\mathbf{Q}^*$	-2149.60	19.94

Absolute fit of competing biogeographic models to the reduced SARS-CoV-2 dataset

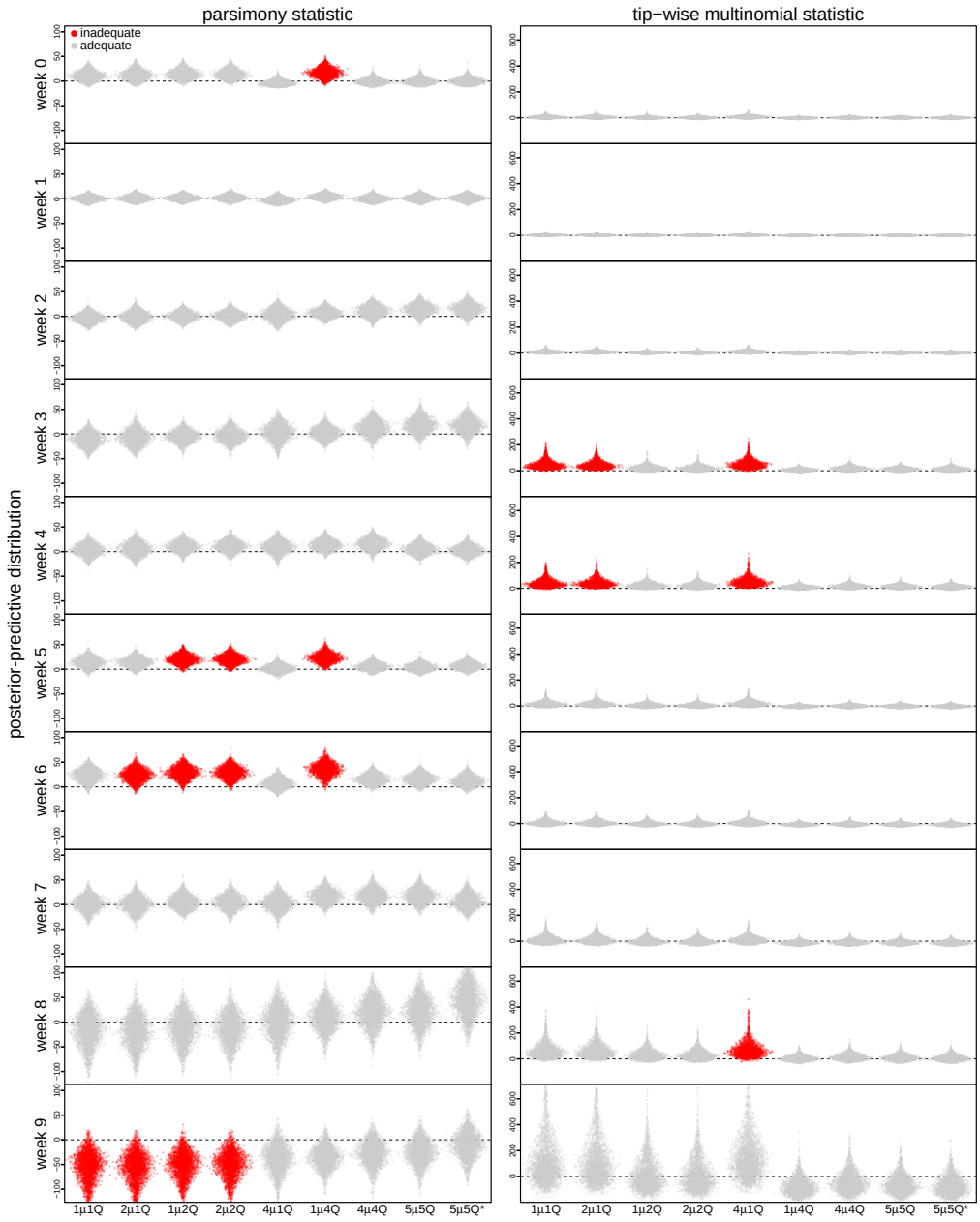

Figure S.3.14: Posterior-predictive distributions under the candidate biogeographic models. Each column of panels corresponds to one of the two types of summary statistics (parsimony and tip-wise multinomial); each row of panels corresponds to one of the 10 (weekly) time slices. Each panel includes a set of nine violin plots (one for each of the candidate biogeographic models listed in Table S.3.6). Each violin plot depicts the posterior-predictive distribution of the 2500 replicate simulations for the corresponding summary statistic under the corresponding candidate model. Each dot represents the value of the corresponding summary statistic for a single replicate posterior-predictive simulations, where the value is the discrepancy between the summary statistic for the observed dataset and the single simulated dataset. The horizontal dashed line indicates the value of the summary statistic under identical fit of the simulated and observed datasets. The violin plots in red indicate that the corresponding model provides an inadequate fit to the SARS-CoV-2 dataset (*i.e.*, it is incapable of generating geographic datasets that are similar to the observed data) under the corresponding time-slice summary statistic, as its 95% posterior-predictive interval does not overlap with the dashed line.

Joint Analyses of the Entire SARS-CoV-2 Dataset

Overview

In this section, we describe the analyses we performed to infer the joint posterior probability distribution of the phylodynamic model—comprising all parameters of the component relaxed-clock and biogeographic models—from the entire SARS-CoV-2 dataset (which includes the viral genome sequences, and the geographic areas and dates of viral sampling). For these analyses, we specified a relaxed-clock model that was similar to that used to estimate the dated phylogeny for the reduced SARS-CoV-2 dataset (see this section), and specified the biogeographic model that was selected based on analyses of the reduced dataset (this section). Below, we provide details on: (1) the specified phylodynamic model; (2) the MCMC simulations we performed to estimate the joint posterior under this model, and; (3) the posterior-predictive simulations we performed to assess the absolute fit of the biogeographic model to the entire SARS-CoV-2 dataset.

Model specification

Our joint analyses of the entire SARS-CoV-2 dataset are based on a phylodynamic model that includes (1) a relaxed-clock model, and (2) a biogeographic model. The relaxed-clock model that we specified for our joint analyses of the entire SARS-CoV-2 sequence dataset is identical to that specified previously in our analyses of the reduced SARS-CoV-2 sequence dataset (see this section) with minor changes to the priors to accommodate differences in viral sampling (see Table S.3.7). The biogeographic model that we specified for our joint analyses of the entire SARS-CoV-2 geographic dataset is identical to the biogeographic model that we selected previously based on analyses of the reduced SARS-CoV-2 dataset (see this section); specifically, the 4-interval ($4\mu 4\mathbf{Q}$) model.

Data analysis

Estimating the joint posterior of phylodynamic model parameters using MCMC simulation.—We performed 20–30 independent MCMC simulations to approximate the joint posterior distribution of the phylodynamic-model parameters—including the phylogeny, divergence times, and biogeographic history—from the entire SARS-CoV-2 dataset using our modified version of `BEAST` (see this section) with the `BEAGLE` library (compiled from the 'hmc-clock' branch, commit 'dd36bf5'; Ayres et al. 2019) to accelerate computation. We ran each replicate MCMC simulation

Table S.3.7: Priors used to jointly infer SARS-CoV-2 phylogeny and biogeographic history for the entire dataset.

Parameter	Description	Prior
κ_1	Ratio of the A $\rightarrow$ G rate to the transversion rate	Lognormal$(\mu = 1.0, \sigma = 0.8)^*$
κ_2	Ratio of the C $\rightarrow$ T rate to the transversion rate	Lognormal$(\mu = 1.0, \sigma = 0.8)$
$\boldsymbol{\pi}$	Nucleotide stationary frequencies	Dir$(1, 1, 1, 1)$
$\boldsymbol{m}$	Partition-specific rate multipliers	Dir$(1, 1, 1, 1, 1, 1)$
α	Shape and scale parameter of the Γ_4 distribution	Lognormal$(\mu = -2.1, \sigma = 0.5874)$
$\mathbb{E}[\boldsymbol{r}]$	Mean of the UCLN	Lognormal$(\mu = -12.5, \sigma = 0.5)$
$SD(\boldsymbol{r})$	Standard deviation of the UCLN	Exp$(\lambda = 1/(4.0\text{e}{-6}))$
N_T	Effective number of infected individuals at sampling time, T	Lognormal$(\mu = 7.0, \sigma = 1.0)$
r	Exponential growth rate of the coalescent model	Laplace$(0.07, 0.01)$
Δ_l	Number of dispersal routes in interval l	Pois(253)
μ_l	Average dispersal rate in interval l	Exp$(1/\lambda)$; $\lambda \sim \Gamma(0.5, 0.5)$
$r_{ij,l}$	Relative dispersal rate from i to j in interval l	$\Gamma(1, 1)$
$\boldsymbol{\omega}$	Root frequencies	Dir$(1, 1, \ldots, 1)$

$^*\mu$ and σ in this table are the mean and standard deviation of the normal distribution.

for 10–20 million generations, sampling continuous parameters every 1000 generations and the dated phylogeny every 10,000 generations. When a phylogeny was sampled, we performed stochastic mapping using the endpoint-conditioned uniformization algorithm (Rodrigue et al. 2007; Fearnhead and Sherlock 2006; Hobolth and Stone 2009) and our modified algorithm to perform stochastic mapping under interval-specific models (see this section) to simulate dispersal histories over the sampled tree. Details of these analyses are available in the XML scripts included in our GitHub and Dryad repositories.

After discarding the first 10–75% of samples from each replicate MCMC simulation (as burn-in), we combined the remaining posterior samples of trees from all replicates and then down-sampled every 50,000 generations using `LogCombiner` version 1.10.5. Following initial inspection of the log files using `Tracer` (Rambaut et al. 2018) version 1.7.1, we further evaluated MCMC performance using the `coda` package (Plummer et al. 2006) in `R` (R Core Team 2020). We assessed convergence of replicate MCMC simulations by calculating the ESS for each continuous parameter for the combined posterior samples; ensuring that values for the substitution-model parameters were all $\gg$4000, those for the geographic model parameters were all $\gg$200, and that the ESS values for all parameters of the branch-rate and branching-

process models were all $\gg$100.

(Re)assessing adequacy of the biogeographic model using posterior-predictive simulation.—We previously established that the preferred biogeographic model provides an adequate description of the process of geographic dispersal during the early phase of the COVID-19 pandemic (see this section). However, those analyses were based on the reduced (rather than entire) SARS-CoV-2 dataset, and also conditioned on a single dated phylogeny—the MCC tree inferred in this section—rather than integrating over the posterior probability distribution of dated phylogenies. Accordingly, we performed additional posterior-predictive simulation to confirm that the preferred biogeographic model provides an adequate fit to the entire SARS-CoV-2 dataset under an inference scenario where geographic history is jointly integrated over the posterior distribution of dated phylogenies.

We performed a series of posterior-predictive simulations to assess the adequacy of the preferred biogeographic model ($4\mu4\mathbf{Q}$). As a point of reference, we also assessed the absolute fit of the constant ($1\mu1\mathbf{Q}$) biogeographic model (*c.f.*, Table S.3.6). For both biogeographic models, we simulated $m = 2500$ posterior-predictive datasets by repeatedly sampling at random from the corresponding joint posterior distribution of phylodynamic model parameters inferred from the entire SARS-CoV-2 dataset. We then generated posterior-predictive distributions from each set of m predictive datasets under 20 separate statistics include the two (parsimony and tip-wise multinomial) summary statistics, each computed over 10 time slices. We specified 10 (weekly) time slices spanning the early phase of COVID-19 (where the time first slice covers the period from the origin of SARS-CoV-2 to Jan. 5, 2020, and the remaining nine weekly time slices spanning the period between Jan. 6 and Mar. 8). For each posterior-predictive distribution, we computed the posterior-predictive p value (see this section) to assess the adequacy (*i.e.*, absolute fit) of the corresponding biogeographic model.

Quantifying differences between prior and posterior distributions.—The discrete-geographic phylodynamic model has many parameters, which raises questions about our ability to infer the parameters from a single set of biogeographic observations. If there is insufficient information in the biogeographic data to estimate the parameters, we expect the posterior distribution of each model parameter to resemble its prior distribution. To quantify the degree to which the posterior distribution of the interval-specific model is updated by the data, we computed Kullback–Leibler (KL) divergence between the marginal posterior and the prior distributions

of each pairwise relative dispersal rate under each of the constant and preferred models. We represent the KL divergence as $D_{KL}(P \,||\, Q)$, where P indicates the posterior distribution and Q indicates the prior distribution.

We also used a symmetric version of the KL divergence—$D_{KL}(P \,||\, Q) + D_{KL}(Q \,||\, P)$—to quantify the difference in the inferred posterior distributions of the pairwise relative dispersal rates between candidate biogeographic models. In this case, P represents the posterior distribution of one model, and Q represents the posterior distribution of the other model. We focused on three pairs of models ($4\mu4\mathbf{Q}$ versus $1\mu1\mathbf{Q}$, $4\mu4\mathbf{Q}$ versus $4\mu1\mathbf{Q}$, and $1\mu4\mathbf{Q}$ versus $1\mu1\mathbf{Q}$) with different relative-rate intervals to assess the impact of allowing relative rates of dispersal to vary across intervals; we also examined such difference using two pairs of models ($4\mu1\mathbf{Q}$ versus $1\mu1\mathbf{Q}$, and $4\mu4\mathbf{Q}$ versus $1\mu4\mathbf{Q}$) with identical relative-rate intervals to assess the impact of allowing average dispersal rate to vary across intervals on the relative-rate estimates.

As we used BSSVS in our inferences, each pairwise relative dispersal rate, q_{ij}, is drawn from a mixture of discrete (when $\delta_{ij} = 0$) and continuous (when $\delta_{ij} = 1$) distributions. The probability density function is:

$$P(q_{ij}) = P(\delta_{ij} = 0) + P(\delta_{ij} = 1)P(r_{ij}). \tag{S.3.10}$$

The KL divergence of distribution P from distribution Q for parameter q_{ij} is then computed as:

$$\begin{aligned} D_{KL}^{q_{ij}}(P \,||\, Q) &= \int P(q_{ij}) \log \frac{P(q_{ij})}{Q(q_{ij})} dq_{ij} \\ &= P(\delta_{ij} = 0) \log \frac{P(\delta_{ij} = 0)}{Q(\delta_{ij} = 0)} + \int P(\delta_{ij} = 1)P(r_{ij}) \log \frac{P(\delta_{ij} = 1)P(r_{ij})}{Q(\delta_{ij} = 1)Q(r_{ij})} dr_{ij} \\ &= P(\delta_{ij} = 0) \log \frac{P(\delta_{ij} = 0)}{Q(\delta_{ij} = 0)} + P(\delta_{ij} = 1) \log \frac{P(\delta_{ij} = 1)}{Q(\delta_{ij} = 1)} \\ &\quad + P(\delta_{ij} = 1) \int P(r_{ij}) \log \frac{P(r_{ij})}{Q(r_{ij})} dr_{ij}. \end{aligned} \tag{S.3.11}$$

We computed each component of (S.3.11) from the sampled distribution. The last component is the KL divergence between two continuous distributions; we computed it using a conventional approach based on the empirical cumulative distribution function (Pérez-Cruz 2008).

Parameter summary.—We summarized the number of viral dispersal events between a given pair of geographic areas by counting the number of dispersal events from the source region (*e.g.*, China) to the destination region (*e.g.*, North America) that occurred on that day for a given

simulated history, and then looped over all of the histories to obtain the posterior distribution of the number of pairwise dispersal events. Mean and 95% credible intervals for the daily number of viral dispersal events were then computed from the corresponding posterior distribution.

Results

Posterior-predictive simulations confirm that the preferred interval-specific biogeographic model ($4\mu 4\mathbf{Q}$) provides an adequate fit to the entire SARS-CoV-2 dataset, whereas the constant biogeographic model is inferred to be inadequate (Fig. S.3.15). The results of these joint analyses of the entire SARS-CoV-2 dataset are presented in the main text (Figs. 3.6–3.8) and Figs. S.3.16–S.3.18.

The computed KL divergence between the posterior and prior distributions shows that, under the interval-specific ($4\mu 4\mathbf{Q}$) model, the most recent interval—the interval with much longer total branch length and more dispersal events than the previous three intervals—appears to contain the most information in inferring the relative dispersal rates, and be comparable to the counterpart under the constant model. The average amount of information gain in the first three intervals appear to be much more limited than the last interval, with noticeable exceptions (*e.g.*, Hubei to East China in interval 2, Japan and Korea to West USA and Canada in interval 3) which also show less information gain under the constant model (Fig. S.3.19).

The inferred posterior distributions of the relative rates of dispersal appear to be much more similar between the pair of comparing models who share the relative-rate intervals (Fig. S.3.23) than between the pair of models with different relative-rate intervals (Figs. S.3.20–S.3.22), indicating that the observed differences in the relative-rate estimates between preferred interval-specific ($4\mu 4\mathbf{Q}$) and the constant ($1\mu 1\mathbf{Q}$) models result from allowing the relative dispersal rates, instead of the average dispersal rate, to vary across intervals.

Absolute fit of the constant (1μ1**Q**) and preferred (4μ4**Q**) models to the SARS-CoV-2 dataset

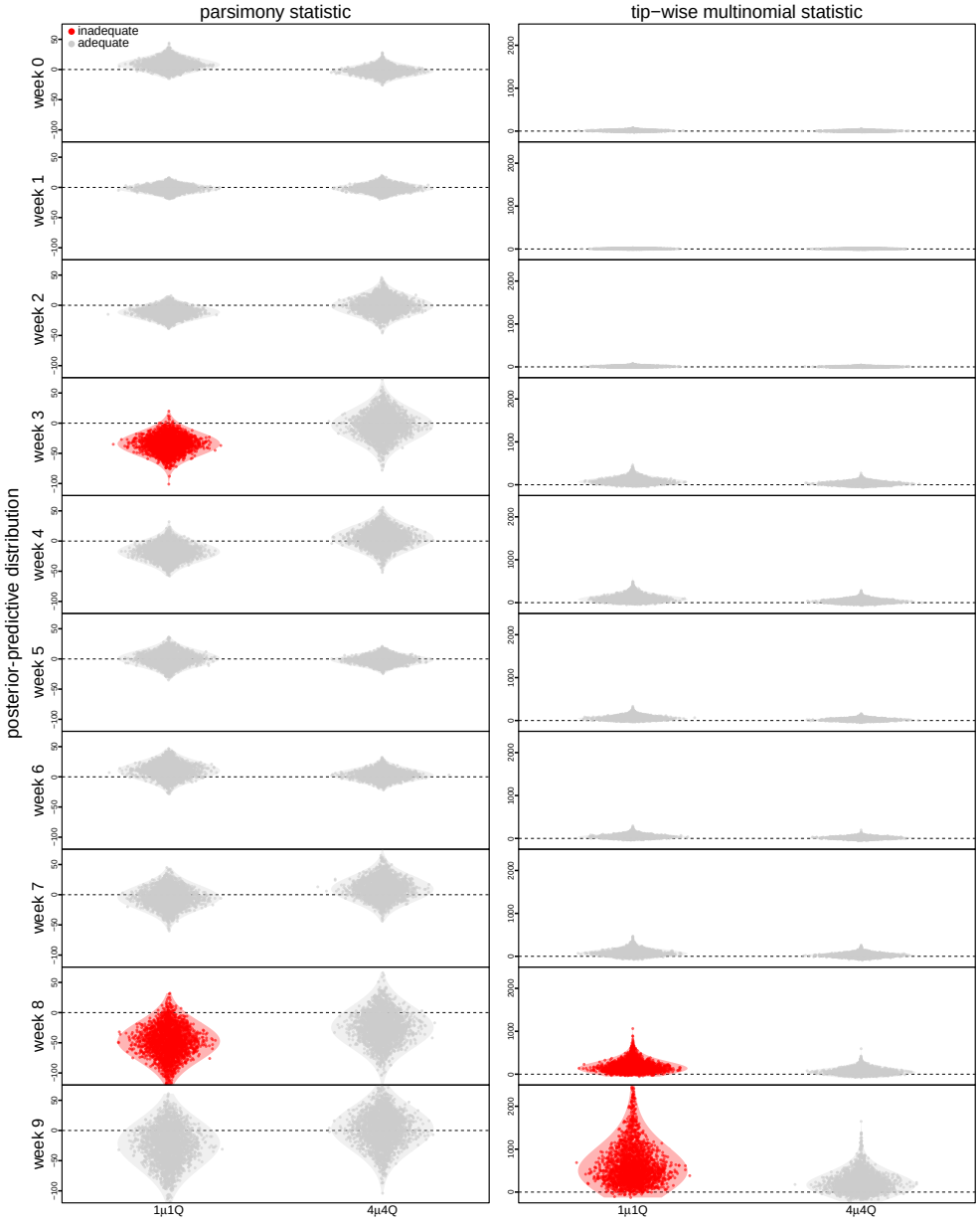

Figure S.3.15: Posterior-predictive distributions of biogeographic models under joint inference of the entire SARS-CoV-2 dataset. Each column of panels corresponds to one of the two types of summary statistics (parsimony and tip-wise multinomial); each row of panels corresponds to one of the 10 (weekly) time slices. Each panel includes a two violin plots for the preferred (4μ4**Q**, right) and constant (1μ1**Q**, left) biogeographic models. Each violin plot depicts the posterior-predictive distribution of the 2500 replicate simulations for the corresponding summary statistic under the corresponding candidate model. Each dot represents the value of the corresponding summary statistic for a single replicate posterior-predictive simulations, where the value is the discrepancy between the summary statistic for the observed dataset and the single simulated dataset. The horizontal dashed line indicates the value of the summary statistic under identical fit of the simulated and observed datasets. The violin plots in red indicate that the corresponding model provides an inadequate fit to the SARS-CoV-2 dataset (*i.e.*, it is incapable of generating geographic datasets that are similar to the observed data) under the corresponding time-slice summary statistic, as its 95% posterior-predictive interval does not overlap with the dashed line.

Inferred support for dispersal routes under the constant (1μ1**Q**) and preferred (4μ4**Q**) models

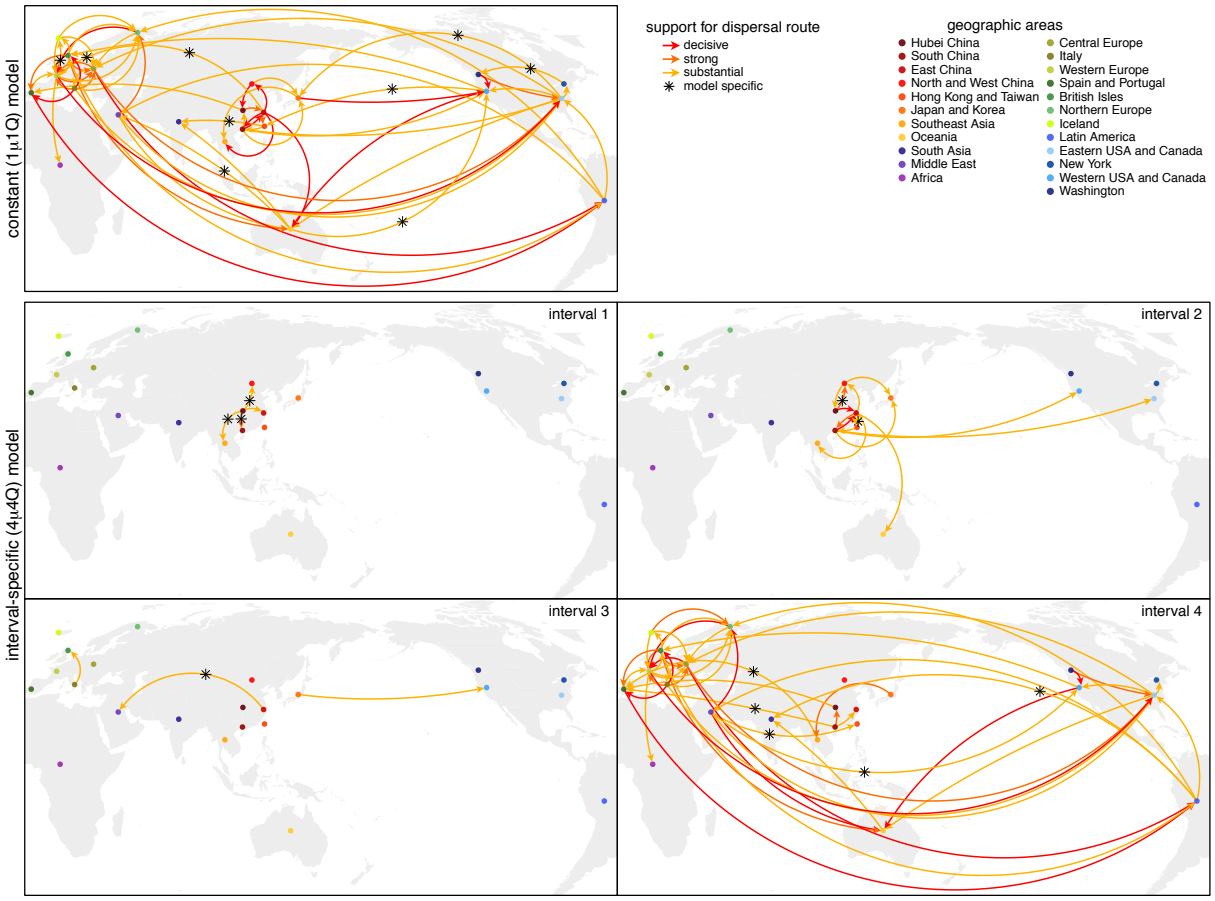

Figure S.3.16: Variation in viral dispersal routes during the early phase of the COVID-19 pandemic. Arrows indicate routes inferred to play a significant role in viral dispersal during the early phase of the COVID-19 pandemic; colors indicate the level of evidential support for each dispersal route (as 2ln Bayes factors). The number, duration, and significance of dispersal routes inferred under the constant (1μ1**Q**) model differ strongly from those inferred under the preferred (4μ4**Q**) interval-specific model. By assumption, the constant (1μ1**Q**) model implies an invariant set of dispersal routes. By contrast, the preferred (4μ4**Q**) interval-specific model reveals that the number and intensity of dispersal routes varied over the four intervals. The first interval (Nov. 17–Jan. 12) is dominated by dispersal from Hubei to other areas in China, and the second interval (Jan. 12–Feb. 2) exhibits more widespread international dispersal originating from China. The third interval (Feb. 2–Feb. 16)—immediately following the onset of international air-travel bans with China—exhibits a sustained reduction in the number of dispersal routes. Note that the constant model infers nine spurious dispersal routes (not detected under the interval-specific model). Conversely, the preferred interval-specific model reveals ten significant dispersal routes (not detected under the constant model) that imply a more significant role for Hubei as a source of viral spread in the first and second intervals, and also reveals additional dispersal routes emanating from China (to the Middle East in the third interval and to Spain/Portugal in the fourth interval).

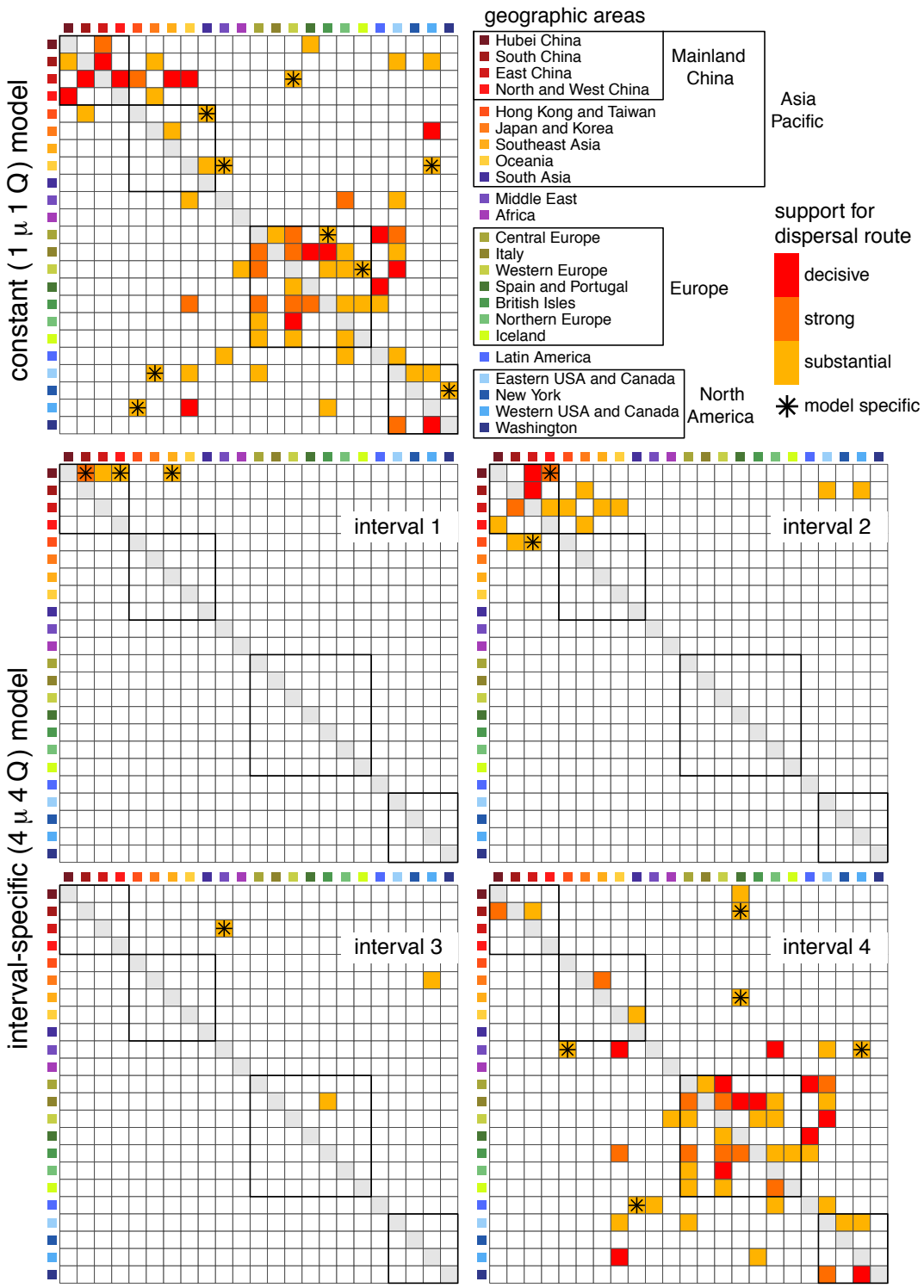

Figure S.3.17: Variation in viral dispersal routes during the early phase of the COVID-19 pandemic. This is simply a heatmap representation of Fig. S.3.16 that may improve the clarity of the evidential support for dispersal routes among all 23 study areas during the rarely phase of the COVID-19 pandemic. Each panel is a 23-by-23 matrix, where row i indicates the 'source' area and each j column indicates the 'destination' area, such that each δ_{ij} element of the matrix indicates the evidential support (as 2ln BF, see inset) for the dispersal route from area i to area j. The boxes within each matrix indicate groups of areas within a region (*e.g.*, the four geographic regions of mainland China).

Inferred pairwise dispersal parameters under the preferred (4μ4**Q**) interval-specific model

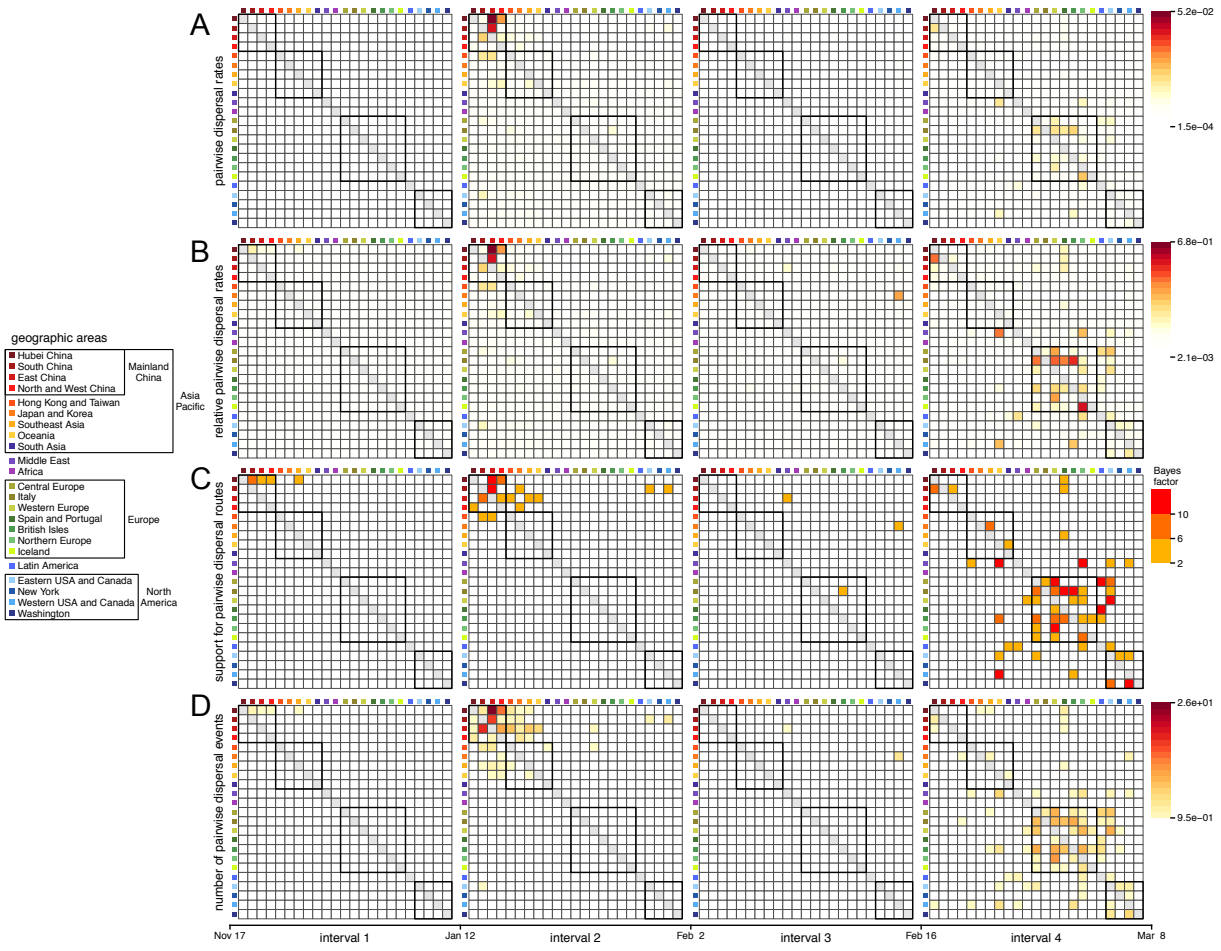

Figure S.3.18: Summary of dispersal parameters inferred under the preferred interval-specific (4μ4Q) model. The four time intervals exhibit distinct dispersal dynamics. **(A)** Absolute viral dispersal rate between each pair of discrete geographic areas. **(B)** Relative viral dispersal rates (*i.e.*, the absolute rates in panel A divided by the inferred global dispersal rate for the corresponding interval) between each pair of discrete geographic areas. **(C)** The evidential support (Bayes factors, inset legend, panel C, right) that a given dispersal route played a significant role in the spread of the virus. **(D)** Number of viral dispersal events between each pair of discrete geographic areas. Boxes in each panel indicate groups of areas (inset legend, left). The first interval is dominated by dispersal from Hubei to other areas in China, the second interval by more widespread dispersal within Asia and by dispersal from China to North America, culminating in cosmopolitan dispersal in the fourth interval. Note that interval three—immediately following the onset of international air-travel bans with China—exhibits a large reduction in the number of viral dispersal routes, including disruption of the dispersal routes from China to North America.

Information gain under the constant ($1\mu 1\mathbf{Q}$) and preferred ($4\mu 4\mathbf{Q}$) models

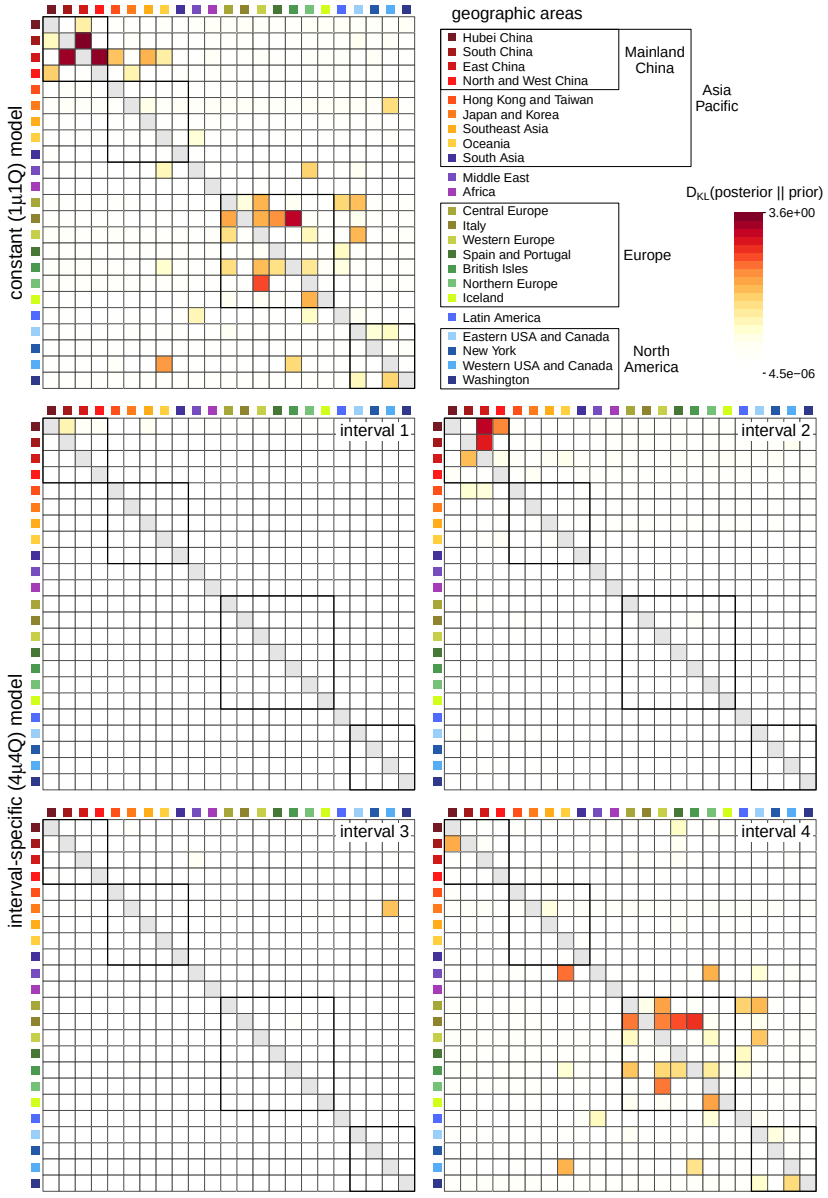

Figure S.3.19: The information gain on the pairwise relative dispersal rates under the constant ($1\mu 1\mathbf{Q}$) and preferred ($4\mu 4\mathbf{Q}$) models. We use Kullback–Leibler (KL) divergence to measure the information gain in moving from the prior to posterior distribution of the pairwise relative dispersal rates. Each panel is a 23-by-23 matrix, where row i indicates the 'source' area and column j indicates the 'destination' area, such that each element of the matrix indicates the KL divergence (colored according to the inset legend bar) between the inferred posterior distribution and the specified prior distribution (which is the same for all pairs) for the relative rate of dispersal from area i to area j. The top row shows the information gain under the constant model, while the remaining two rows show such measure under the preferred interval-specific ($4\mu 4\mathbf{Q}$) model. Under the interval-specific ($4\mu 4\mathbf{Q}$) model, the last interval appears to contain the most information in inferring the relative dispersal rates, and be comparable to the counterpart under the constant model. The average amount of information gain in the first three intervals appear to be much more limited than the last interval, with noticeable exceptions (*e.g.*, Hubei to East China in interval 2, Japan and Korea to West USA and Canada in interval 3) which also show less information gain under the constant model.

Difference between inferred pairwise dispersal rates under biogeographic models

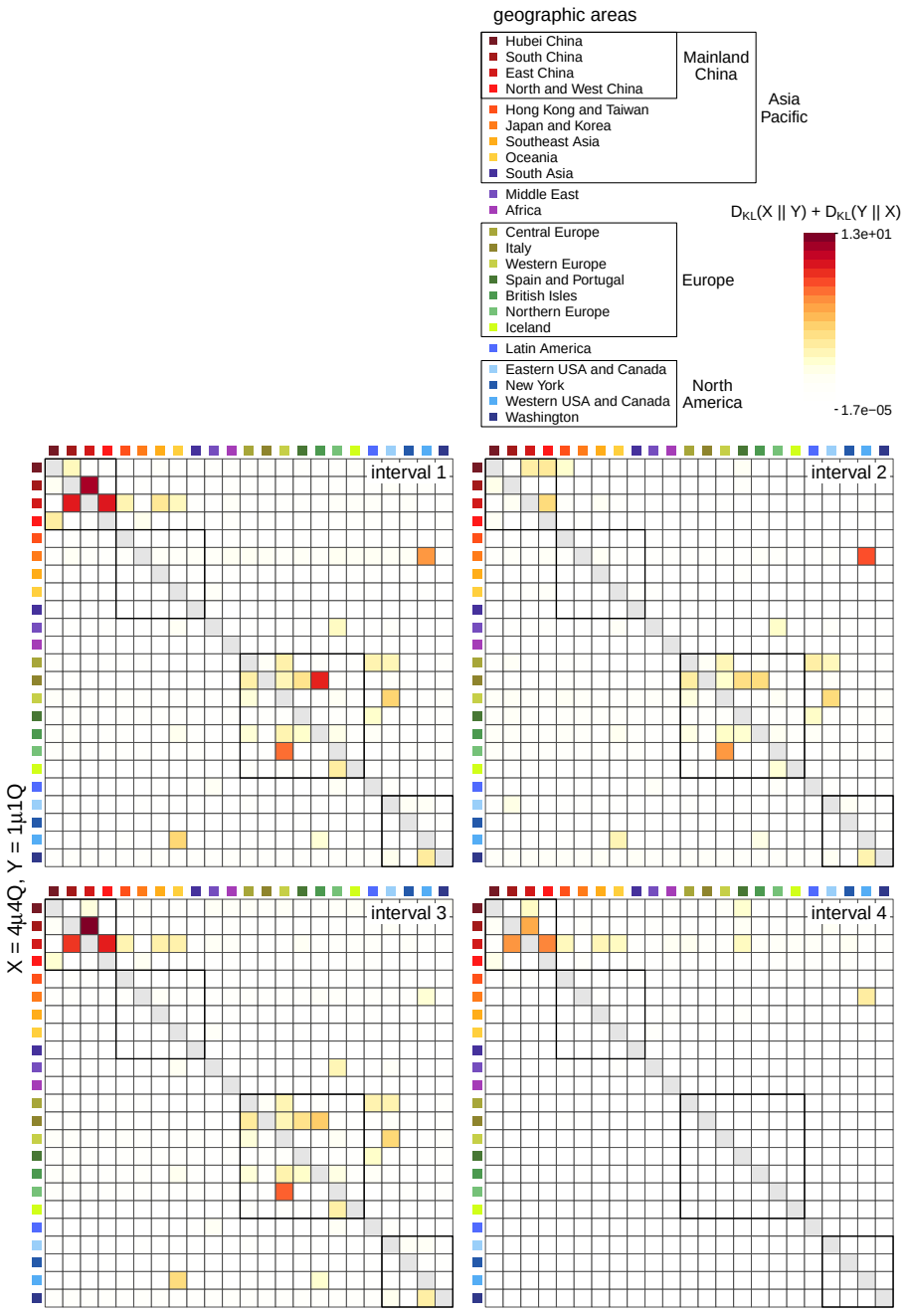

Figure S.3.20: Difference between inferred pairwise relative dispersal rates under models with different relative-rate intervals. We use symmetric KL divergence to measure the difference in the inferred posterior distribution of the pairwise relative dispersal rates between the preferred interval-specific ($4\mu 4\mathbf{Q}$) and constant ($1\mu 1\mathbf{Q}$) models. Each panel is a 23-by-23 matrix, where row i indicates the 'source' area and column j indicates the 'destination' area, such that each element of the matrix indicates the KL divergence (colored according to the inset legend bar; note that the heatmap color scale is shared among Figs. S.3.20–S.3.23) between the inferred posterior distribution of the relative rate of dispersal from area i to area j in the corresponding interval under the interval-specific model and the counterpart under the constant model.

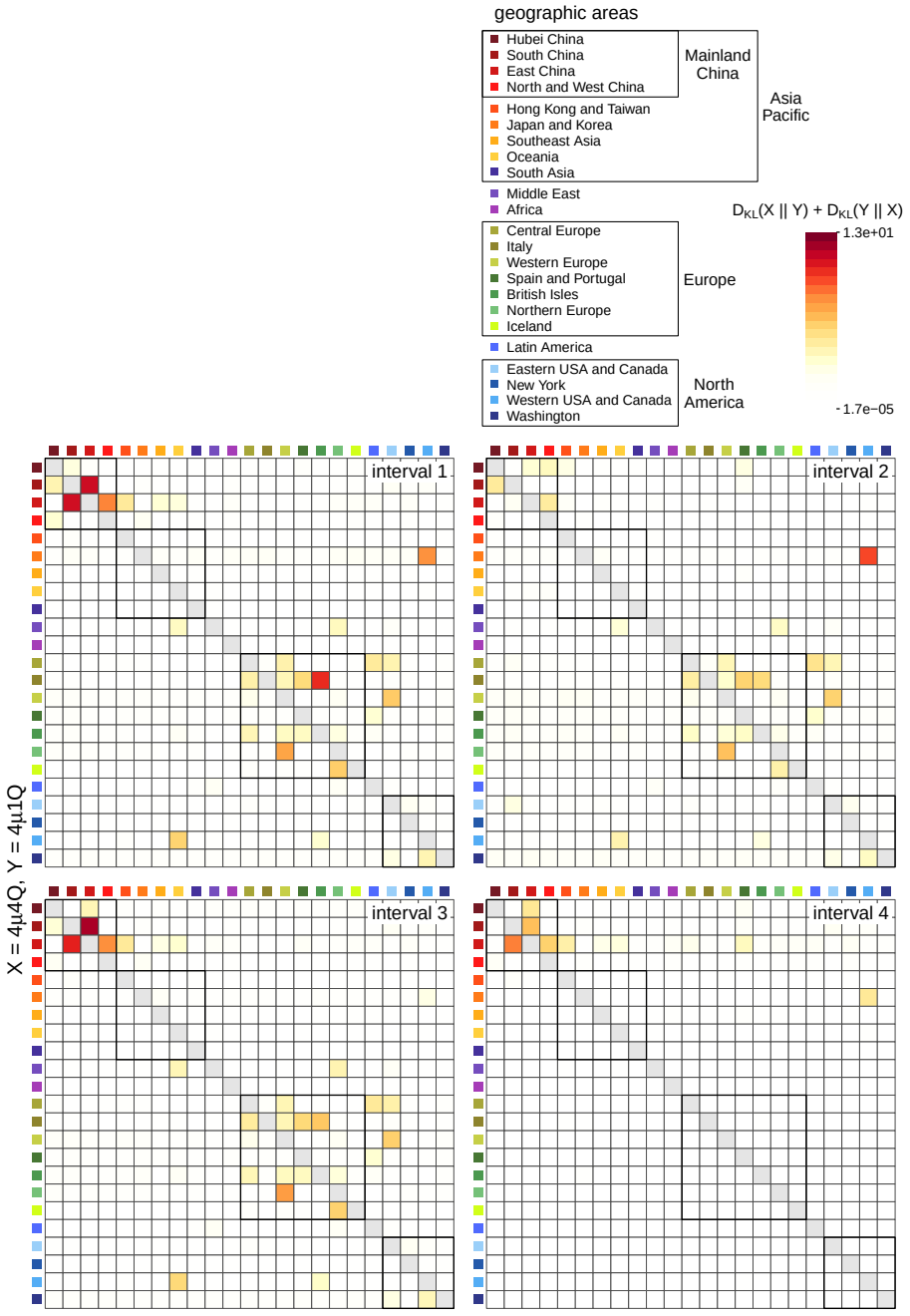

Figure S.3.21: Difference between inferred pairwise relative dispersal rates under models with different relative-rate intervals. We use symmetric KL divergence to measure the difference in the inferred posterior distribution of the pairwise relative dispersal rates between two interval-specific ($4\mu 4\mathbf{Q}$ and $4\mu 1\mathbf{Q}$) models. Each panel is a 23-by-23 matrix, where row i indicates the 'source' area and column j indicates the 'destination' area, such that each element of the matrix indicates the KL divergence (colored according to the inset legend bar; note that the heatmap color scale is shared among Figs. S.3.20–S.3.23) between the inferred posterior distribution of the relative rate of dispersal from area i to area j in the corresponding interval under the $4\mu 4\mathbf{Q}$ model and the counterpart under the $4\mu 1\mathbf{Q}$ model.

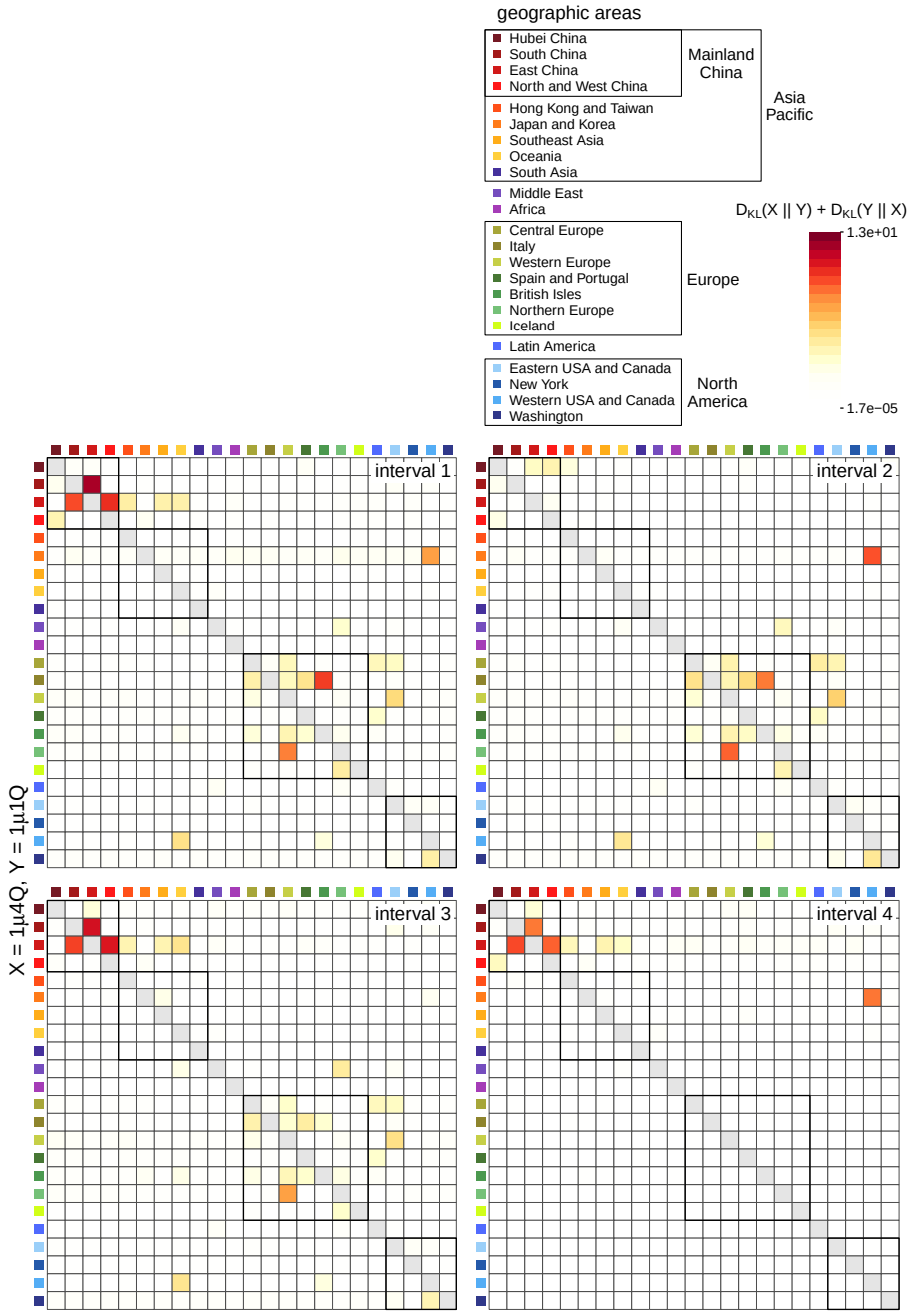

Figure S.3.22: Difference between inferred pairwise relative dispersal rates under models with different relative-rate intervals. We use symmetric KL divergence to measure the difference in the inferred posterior distribution of the pairwise relative dispersal rates between an interval-specific ($1\mu 4\mathbf{Q}$) and the constant ($1\mu 1\mathbf{Q}$) models. Each panel is a 23-by-23 matrix, where row i indicates the 'source' area and column j indicates the 'destination' area, such that each element of the matrix indicates the KL divergence (colored according to the inset legend bar; note that the heatmap color scale is shared among Figs. S.3.20–S.3.23) between the inferred posterior distribution of the relative rate of dispersal from area i to area j in the corresponding interval under the interval-specific model and the counterpart under the constant model.

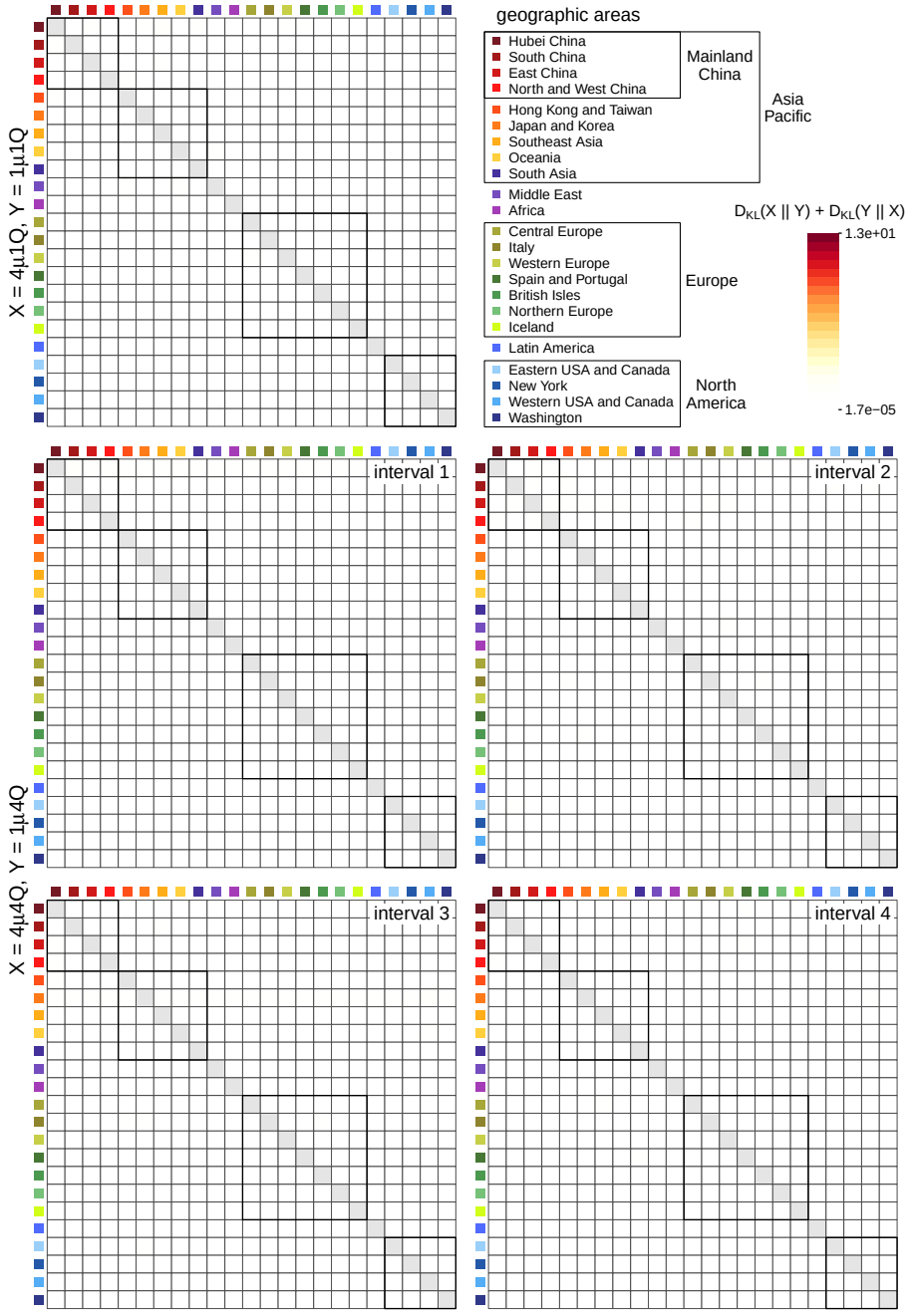

Figure S.3.23: Difference between inferred pairwise relative dispersal rates under models with the same relative-rate intervals. We use symmetric KL divergence to measure the difference in the inferred posterior distribution of the pairwise relative dispersal rates between the models with constant relative dispersal rates ($4\mu 1\mathbf{Q}$ and $1\mu 1\mathbf{Q}$; top row), and between the models with interval-specific relative dispersal rates ($4\mu 4\mathbf{Q}$ and $1\mu 4\mathbf{Q}$; bottom two rows). Each panel is a 23-by-23 matrix, where row i indicates the 'source' area and column j indicates the 'destination' area, such that each element of the matrix indicates the KL divergence (colored according to the inset legend bar; note that the heatmap color scale is shared among Figs. S.3.20–S.3.23) between the inferred posterior distribution of the relative rate of dispersal from area i to area j under the comparing models. The posterior distributions appear to be much more similar when the pair of comparing models share the relative-rate intervals than the pairs with different relative-rate intervals (see Figs. S.3.20–S.3.22).

Estimating Daily Global Viral Dispersal Rates

Overview

In this section, we describe our analyses to explore the correlation between daily global air-travel volume and global SARS-CoV-2 dispersal rates during the early phase of the COVID-19 pandemic. Because we were able to obtain data on the *daily* volume of global air travel during this period, we performed an analysis of the entire SARS-CoV-2 dataset under a more granular phylodynamic model that allows the average dispersal rate to vary from day to day. We then performed standard statistical tests to assess the degree of correlation between the inferred daily global viral dispersal rates and daily global air-travel volume.

Model specification

Our estimates of daily variation in global viral dispersal rates are based on the phylodynamic model ($4\mu4\mathbf{Q}$) that we previously used to infer the joint posterior of SARS-CoV-2 phylogeny and biogeographic history (see this section), except that we further discretized the number of time intervals in which the global average dispersal rate, μ, was free to vary. Specifically, rather than allowing global average dispersal rate to vary between four time intervals, we specified an independent μ for each of the 70 days between Dec. 30, 2019 and Mar. 8, 2020 (with an additional independent μ for the time spanning the origin of SARS-CoV-2 to Dec. 29). The prior on each μ is specified according to the posterior estimates inferred in this section. We computed the posterior mean of the global viral dispersal rate across the entire history inferred from the joint analyses and used it as the prior mean, and we specified standard deviation of the prior distribution so that the 95% prior interval spans three orders of magnitude around the mean. Details of the priors are described in Table S.3.8. Our inferences of geographic history under this model were averaged over the marginal posterior probability distribution of dated phylogenies inferred in this section. Details of these analyses are available in the XML scripts

Table S.3.8: Priors used to infer the daily global viral dispersal rates for the entire dataset.

Parameter	Description	Prior
Δ_l	Number of dispersal routes in interval l	$\text{Pois}(253)$
μ_p	Global dispersal rate in day p	$\text{Lognormal}(\mu = -4.76, \sigma = 1.7622)^*$
$r_{ij,l}$	Relative dispersal rate from i to j in interval l	$\Gamma(1,1)$
$\boldsymbol{\omega}$	Root frequencies	$\text{Dir}(1,1,\ldots,1)$

*μ and σ in this table are the mean and standard deviation of the normal distribution.

included in our GitHub and Dryad repositories.

Data analysis

Parameter estimation.—We performed 15 independent MCMC simulations to approximate the joint posterior distribution of the biogeographic-model parameters from the entire SARS-CoV-2 dataset using our modified version of `BEAST` (see this section) and `BEAGLE` version 3.2.0 (Ayres et al. 2019). We ran each replicate MCMC simulation for 5 million generations, sampling continuous parameters every 2000 generations and the dated phylogeny every 10000 generations. When a phylogeny was sampled, we performed stochastic mapping using the endpoint-conditioned uniformization algorithm (Rodrigue et al. 2007; Fearnhead and Sherlock 2006; Hobolth and Stone 2009) and our modified algorithm to perform stochastic mapping under interval-specific models (see this section) to simulate dispersal histories over the sampled tree. Details of these analyses are available in the XML scripts included in our GitHub and Dryad repositories.

For each replicate MCMC simulation, we discarded the first one million generations (as burn-in), and then combined the remaining posterior samples from all replicates using `LogCombiner` version 1.10.5. Following initial inspection of the log files using `Tracer` (Rambaut et al. 2018) version 1.7.1, we further evaluated MCMC performance using the `coda` package (Plummer et al. 2006) in `R` (R Core Team 2020). We assessed convergence of replicate MCMC simulations by calculating the ESS for each continuous parameter for the combined posterior samples, ensuring that values for all parameters were $\gg$700.

Correlation test.—We tested for correlation between the volume of daily global air travel, $\boldsymbol{V} = \{v_i\}$ (where $i = \{1, 2, \ldots, m\}$ and m represents the total number of days included in our dataset), and the mean estimate of daily global SARS-CoV-2 dispersal rate, $\boldsymbol{\mu} = \{\mu_i\}$. To remove potential trend or seasonality in the time series of $\boldsymbol{V}$ and $\boldsymbol{\mu}$, we first transformed each of the two time series by taking the difference between each value of the time series and the value a week prior to it. Specifically, we computed $\boldsymbol{v'}$ as $\{v_j - v_{j-7}\}$ (where $j = \{8, \ldots, m\}$) and $\boldsymbol{\mu'}$ as $\{\mu_j - \mu_{j-7}\}$. We then generated various truncated dataset by including values from each of the two differenced time series ($\boldsymbol{v'}$ and $\boldsymbol{\mu'}$) with different start dates (ranging from Jan. 6 to Feb. 17, 2020; *i.e.*, j ranges from 8 to 49) to the same end date (the end of our study period, Mar. 8, 2020; $j = 70$). Finally, we assessed the correlation for each truncated dataset by computing Pearson's r and the corresponding p-value to determine the time that the correlation first established.

Results

The daily global dispersal rate estimates are presented in the main text (Fig. 3.6, light blue). Pearson's r and the corresponding p-value between the volume of daily global air travel and the estimated mean rate of daily global SARS-CoV-2 dispersal are presented in Fig. S.3.24. The correlation appears to increase when we focus on the time series of February (*i.e.*, discarding the January values), possibly reflecting that the geographic distribution of SARS-CoV-2 was still confined to specific regions (*e.g.*, China and some other Asian countries) prior to this point. The p-value increases quickly when we fewer than 25 time points are included in the correlation test, presumably reflecting a decrease in power as the number of data points decreases. Therefore, we report Pearson's r and the corresponding p value between the two time series over the interval from Jan. 31 (when the virus first achieved a cosmopolitan distribution; WHO 2020) to the end of our study period (Mar. 8, 2020) in the main text.

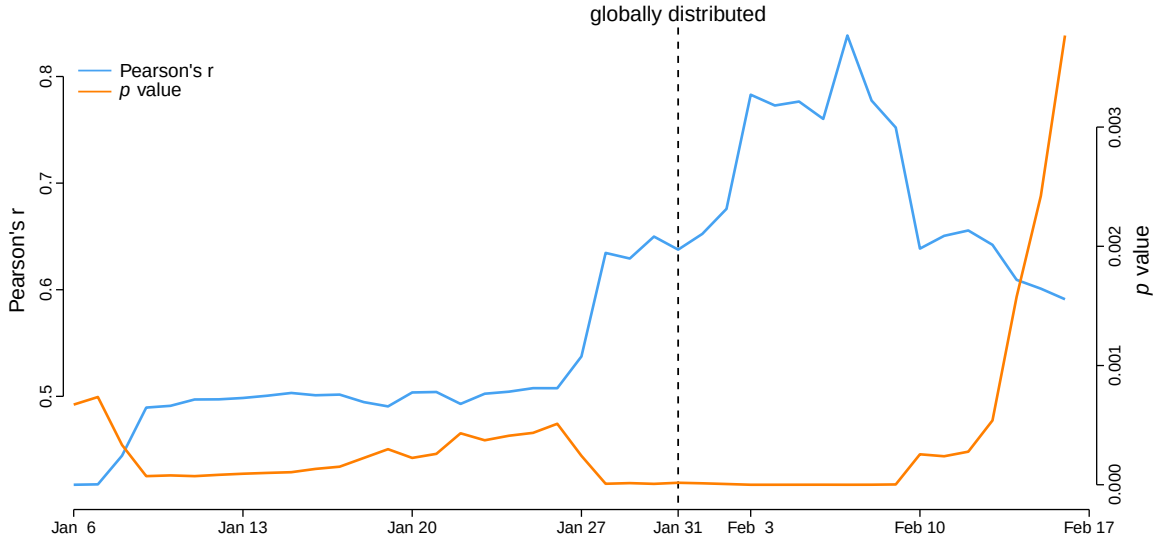

Figure S.3.24: Correlation between daily global air travel volume and the estimated mean daily global SARS-CoV-2 dispersal rate. We generated time-series datasets by including values with different start dates (along the x axis) to the same end date (the end of our study period, Mar. 8, 2020). We then computed Pearson's r and the corresponding p value for each dataset, and plotted them as a function of the corresponding start date. The virus first achieved a cosmopolitan distribution on Jan. 31 (dashed line; WHO 2020).

Alpert, T., Brito, A. F., Lasek-Nesselquist, E., Rothman, J., Valesano, A. L., MacKay, M. J., Petrone, M. E., Breban, M. I., Watkins, A. E., Vogels, C. B., et al. (2021). Early introductions and transmission of SARS-CoV-2 variant B. 1.1. 7 in the United States. *Cell*, 184(10):2595–2604.

Ayres, D. L., Cummings, M. P., Baele, G., Darling, A. E., Lewis, P. O., Swofford, D. L., Huelsenbeck, J. P., Lemey, P., Rambaut, A., and Suchard, M. A. (2019). Beagle 3: Improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. *Systematic biology*, 68(6):1052–1061.

Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M. A., and Alekseyenko, A. V. (2012). Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular Biology and Evolution*, 29(9):2157–2167.

Baele, G., Lemey, P., and Suchard, M. A. (2015). Genealogical working distributions for bayesian model testing with phylogenetic uncertainty. *Systematic biology*, 65(2):250–264.

Baele, G., Suchard, M. A., Rambaut, A., and Lemey, P. (2017). Emerging concepts of data integration in pathogen phylodynamics. *Systematic Biology*, 66(1):e47–e65.

Bayes, T. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philosophical Transactions of the Royal Society of London*, (53):370–418.

Beaumont, M. A. (1999). Detecting population expansion and decline using microsatellites. *Genetics*, 153(4):2013–2029.

Bedford, T., Greninger, A. L., Roychoudhury, P., Starita, L. M., Famulare, M., Huang, M.-L., Nalla, A., Pepper, G., Reinhardt, A., Xie, H., Shrestha, L., Nguyen, T. N., Adler, A., Brandstetter, E., Cho, S., Giroux, D., Han, P. D., Fay, K., Frazar, C. D., Ilcisin, M., Lacombe, K., Lee, J., Kiavand, A., Richardson, M., Sibley, T. R., Truong, M., Wolf, C. R., Nickerson, D. A., Rieder, M. J., Englund, J. A., , Hadfield, J., Hodcroft, E. B., Huddleston, J., Moncla, L. H., Müller, N. F., Neher, R. A., Deng, X., Gu, W., Federman, S., Chiu, C., Duchin, J. S., Gautom, R., Melly, G., Hiatt, B., Dykema, P., Lindquist, S., Queen, K., Tao, Y., Uehara, A., Tong, S., MacCannell, D., Armstrong, G. L., Baird, G. S., Chu, H. Y., Shendure, J., and Jerome, K. R. (2020). Cryptic transmission of SARS-CoV-2 in Washington state. *Science*, 370(6516):571–575.

Bedford, T., Riley, S., Barr, I. G., Broor, S., Chadha, M., Cox, N. J., Daniels, R. S., Gunasekaran, C. P., Hurt, A. C., Kelso, A., et al. (2015). Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*, 523(7559):217.

Bielejec, F., Baele, G., Rodrigo, A. G., Suchard, M. A., and Lemey, P. (2016). Identifying predictors of time-inhomogeneous viral evolutionary processes. *Virus Evolution*, 2(2):vew023.

Bielejec, F., Lemey, P., Baele, G., Rambaut, A., and Suchard, M. A. (2014). Inferring heterogeneous evolutionary processes through time: from sequence substitution to phylogeography. *Systematic Biology*, 63(4):493–504.

Bollback, J. P. (2002). Bayesian model adequacy and choice in phylogenetics. *Molecular Biology and Evolution*, 19(7):1171–1180.

Bollback, J. P. (2006). Simmap: stochastic character mapping of discrete traits on phylogenies. *BMC bioinformatics*, 7(1):88.

Byrne, A. W., McEvoy, D., Collins, A. B., Hunt, K., Casey, M., Barber, A., Butler, F., Griffin, J., Lane, E. A., McAloon, C., et al. (2020). Inferred duration of infectious period of SARS-CoV-2: rapid scoping review and analysis of available evidence for asymptomatic and symptomatic COVID-19 cases. *BMJ open*, 10(8):e039856.

Candido, D. S., Claro, I. M., de Jesus, J. G., Souza, W. M., Moreira, F. R. R., Dellicour, S., Mellan,

T. A., du Plessis, L., Pereira, R. H. M., Sales, F. C. S., Manuli, E. R., Thézé, J., Almeida, L., Menezes, M. T., Voloch, C. M., Fumagalli, M. J., Coletti, T. M., da Silva, C. A. M., Ramundo, M. S., Amorim, M. R., Hoeltgebaum, H. H., Mishra, S., Gill, M. S., Carvalho, L. M., Buss, L. F., Prete, C. A., Ashworth, J., Nakaya, H. I., Peixoto, P. S., Brady, O. J., Nicholls, S. M., Tanuri, A., Rossi, Á. D., Braga, C. K. V., Gerber, A. L., de C. Guimarães, A. P., Gaburo, N., Alencar, C. S., Ferreira, A. C. S., Lima, C. X., Levi, J. E., Granato, C., Ferreira, G. M., Francisco, R. S., Granja, F., Garcia, M. T., Moretti, M. L., Perroud, M. W., Castiñeiras, T. M. P. P., Lazari, C. S., Hill, S. C., de Souza Santos, A. A., Simeoni, C. L., Forato, J., Sposito, A. C., Schreiber, A. Z., Santos, M. N. N., de Sá, C. Z., Souza, R. P., Resende-Moreira, L. C., Teixeira, M. M., Hubner, J., Leme, P. A. F., Moreira, R. G., Nogueira, M. L., , Ferguson, N. M., Costa, S. F., Proenca-Modena, J. L., Vasconcelos, A. T. R., Bhatt, S., Lemey, P., Wu, C.-H., Rambaut, A., Loman, N. J., Aguiar, R. S., Pybus, O. G., Sabino, E. C., and Faria, N. R. (2020). Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science*, 369(6508):1255–1260.

Charif, D. and Lobry, J. (2007). SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In Bastolla, U., Porto, M., Roman, H., and Vendruscolo, M., editors, *Structural approaches to sequence evolution: Molecules, networks, populations*, Biological and Medical Physics, Biomedical Engineering, pages 207–232. Springer Verlag, New York. ISBN : 978-3-540-35305-8.

Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., y Piontti, A. P., Mu, K., Rossi, L., Sun, K., et al. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science*, 368(6489):395–400.

Cook, S. R., Gelman, A., and Rubin, D. B. (2006). Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15(3):675–692.

Dash, P. K., Sharma, S., Soni, M., Agarwal, A., Sahni, A. K., and Parida, M. (2015). Complete genome sequencing and evolutionary phylogeography analysis of Indian isolates of Dengue virus type 1. *Virus Research*, 195:124–134.

Davies, N. G., Abbott, S., Barnard, R. C., Jarvis, C. I., Kucharski, A. J., Munday, J. D., Pearson, C. A. B., Russell, T. W., Tully, D. C., Washburne, A. D., Wenseleers, T., Gimma, A., Waites, W., Wong, K. L. M., van Zandvoort, K., Silverman, J. D., null null, null null, Diaz-Ordaz, K., Keogh, R., Eggo, R. M., Funk, S., Jit, M., Atkins, K. E., and Edmunds, W. J. (2021). Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science*, 372(6538).

De Maio, N., Wu, C.-H., O'Reilly, K. M., and Wilson, D. (2015). New routes to phylogeography: a bayesian structured coalescent approximation. *PLoS genetics*, 11(8):e1005421.

Dellicour, S., Durkin, K., Hong, S. L., Vanmechelen, B., Martí-Carreras, J., Gill, M. S., Meex, C., Bontems, S., André, E., Gilbert, M., Walker, C., Maio, N. D., Faria, N. R., Hadfield, J., Hayette, M.-P., Bours, V., Wawina-Bokalanga, T., Artesi, M., Baele, G., and Maes, P. (2021). A phylodynamic workflow to rapidly gain insights into the dispersal history and dynamics of SARS-CoV-2 lineages. *Molecular Biology and Evolution*, 38(4):1608–1613.

Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*, 20(5):533–534.

Douglas, J., Mendes, F. K., Bouckaert, R., Xie, D., Jiménez-Silva, C. L., Swanepoel, C., de Ligt, J., Ren, X., Storey, M., Hadfield, J., Simpson, C. R., Geoghegan, J. L., Drummond, A. J., and Welch, D. (2021). Phylodynamics reveals the role of human travel and contact tracing in controlling the first wave of COVID-19 in four island nations. *Virus Evolution*, 7(2):veab052.

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

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., and Solomon, W. (2002). Estimating muta-

tion parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, 161(3):1307–1320.

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular biology and evolution*, 22(5):1185–1192.

Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29(8):1969–1973.

du Plessis, L., McCrone, J. T., Zarebski, A. E., Hill, V., Ruis, C., Gutierrez, B., Raghwani, J., Ashworth, J., Colquhoun, R., Connor, T. R., et al. (2021). Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science*, 371(6530):708–712.

Duchene, S., Featherstone, L., Haritopoulou-Sinanidou, M., Rambaut, A., Lemey, P., and Baele, G. (2020). Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evolution*, 6(2):veaa061.

DXY (2020). Global COVID-19 realtime map. https://ncov.dxy.cn/ncovh5/view/pneumonia. (Accessed on 12/30/2020).

ECDC (2020). COVID-19 situation update worldwide. https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases. (Accessed on 12/30/2020).

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797.

Edwards, C. J., Suchard, M. A., Lemey, P., Welch, J. J., Barnes, I., Fulton, T. L., Barnett, R., O'Connell, T. C., Coxon, P., Monaghan, N., et al. (2011). Ancient hybridization and an Irish origin for the modern polar bear matriline. *Current Biology*, 21(15):1251–1258.

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., Tatem, A. J., Sousa, J. D., Arinaminpathy, N., Pépin, J., et al. (2014). The early spread and epidemic ignition of HIV-1 in human populations. *Science*, 346(6205):56–61.

Fauver, J. R., Petrone, M. E., Hodcroft, E. B., Shioda, K., Ehrlich, H. Y., Watts, A. G., Vogels, C. B., Brito, A. F., Alpert, T., Muyombwe, A., Razeq, J., Downing, R., Cheemarla, N. R., Wyllie, A. L., Kalinich, C. C., Ott, I. M., Quick, J., Loman, N. J., Neugebauer, K. M., Greninger, A. L., Jerome, K. R., Roychoudhury, P., Xie, H., Shrestha, L., Huang, M.-L., Pitzer, V. E., Iwasaki, A., Omer, S. B., Khan, K., Bogoch, I. I., Martinello, R. A., Foxman, E. F., Landry, M. L., Neher, R. A., Ko, A. I., and Grubaugh, N. D. (2020). Coast-to-coast spread of SARS-CoV-2 during the early epidemic in the United States. *Cell*, 181(5):990–996.

Fearnhead, P. and Sherlock, C. (2006). An exact gibbs sampler for the markov-modulated poisson process. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(5):767–784.

Gao, J., May, M. R., Rannala, B., and Moore, B. R. (2022). New phylogenetic models incorporating interval-specific dispersal dynamics improve inference of disease spread. *Molecular Biology and Evolution*.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, pages 733–760.

Gelman, A. and Rubin, D. B. (1992). Inferences from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511.

Gill, M. S., Lemey, P., Bennett, S. N., Biek, R., and Suchard, M. A. (2016). Understanding past population dynamics: Bayesian coalescent-based modeling with covariates. *Systematic biology*, 65(6):1041–1056.

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B., and Suchard, M. A. (2013). Improving bayesian population dynamics inference: a coalescent-based model for multiple loci. *Molecular biology and evolution*, 30(3):713–724.

Gill, M. S., Tung Ho, L. S., Baele, G., Lemey, P., and Suchard, M. A. (2017). A relaxed directional random walk model for phylogenetic trait evolution. *Systematic biology*, 66(3):299–319.

Goldman, N. (1993). Statistical tests of models of DNA substitution. *Journal of Molecular Evolution*, 36(2):182–198.

Gu, X., Fu, Y.-X., and Li, W.-H. (1995). Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Molecular Biology and Evolution*, 12(4):546–557.

Hao, X., Cheng, S., Wu, D., Wu, T., Lin, X., and Wang, C. (2020). Reconstruction of the full transmission dynamics of COVID-19 in Wuhan. *Nature*, 584(7821):420–424.

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

Hasegawa, M., Yano, T.-a., and Kishino, H. (1984). A new molecular clock of mitochondrial DNA and the evolution of Hominoids. *Proceedings of the Japan Academy, series B*, 60(4):95–98.

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

Hobolth, A. and Stone, E. A. (2009). Simulation from endpoint-conditioned, continuous-time markov chains on a finite state space, with applications to molecular evolution. *The annals of applied statistics*, 3(3):1204.

Hou, Y. J., Chiba, S., Halfmann, P., Ehre, C., Kuroda, M., Dinnon, K. H., Leist, S. R., Schäfer, A., Nakajima, N., Takahashi, K., Lee, R. E., Mascenik, T. M., Graham, R., Edwards, C. E., Tse, L. V., Okuda, K., Markmann, A. J., Bartelt, L., de Silva, A., Margolis, D. M., Boucher, R. C., Randell, S. H., Suzuki, T., Gralinski, L. E., Kawaoka, Y., and Baric, R. S. (2020). SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science*, 370(6523):1464–1468.

Hsiang, S., Allen, D., Annan-Phan, S., Bell, K., Bolliger, I., Chong, T., Druckenmiller, H., Huang, L. Y., Hultgren, A., Krasovich, E., Lau, P., Lee, J., Rolf, E., Tseng, J., and Wu, T. (2020). The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature*, 584(7820):262–267.

Huelsenbeck, J. P. and Bollback, J. P. (2001). Empirical and hierarchical bayesian estimation of ancestral states. *Systematic biology*, 50(3):351–366.

Huelsenbeck, J. P., Nielsen, R., and Bollback, J. P. (2003). Stochastic mapping of morphological characters. *Systematic Biology*, 52(2):131–158.

International Civil Aviation Organization (2020). Effects of novel coronavirus (COVID-19) on civil aviation:economic impact analysis. https://www.icao.int/sustainability/Documents/ICAO_Coronavirus_Econ_Impact.pdf. (Accessed on 01/04/2021).

Jakeman, E. and Pusey, P. (1978). Significance of K distributions in scattering experiments. *Physical Review Letters*, 40(9):546.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780.

Kazmi, S. O. and Rodrigue, N. (2019). Detecting amino acid preference shifts with codon-level mutation-selection mixture models. *BMC Evolutionary Biology*, 19(62).

Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E. E., Bhattacharya, T., Foley, B., Hastie, K. M., Parker, M. D., Partridge, D. G., Evans,

C. M., Freeman, T. M., de Silva, T. I., Angyal, A., Brown, R. L., Carrilero, L., Green, L. R., Groves, D. C., Johnson, K. J., Keeley, A. J., Lindsey, B. B., Parsons, P. J., Raza, M., Rowland-Jones, S., Smith, N., Tucker, R. M., Wang, D., Wyles, M. D., McDanal, C., Perez, L. G., Tang, H., Moon-Walker, A., Whelan, S. P., LaBranche, C. C., Saphire, E. O., and Montefiori, D. C. (2020). Tracking changes in SARS-CoV-2 spike: evidence that d614g increases infectivity of the COVID-19 virus. *Cell*, 182(4):812–827.

Kraemer, M. U., Yang, C.-H., Gutierrez, B., Wu, C.-H., Klein, B., Pigott, D. M., Du Plessis, L., Faria, N. R., Li, R., Hanage, W. P., et al. (2020). The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science*, 368(6490):493–497.

Kraemer, M. U. G., Hill, V., Ruis, C., Dellicour, S., Bajaj, S., McCrone, J. T., Baele, G., Parag, K. V., Battle, A. L., Gutierrez, B., Jackson, B., Colquhoun, R., Áine O'Toole, Klein, B., Vespignani, A., null null, Volz, E., Faria, N. R., Aanensen, D. M., Loman, N. J., du Plessis, L., Cauchemez, S., Rambaut, A., Scarpino, S. V., and Pybus, O. G. (2021). Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B. 1.1. 7 emergence. *Science*, 373(6557):889–895.

Kühnert, D., Stadler, T., Vaughan, T. G., and Drummond, A. J. (2016). Phylodynamics with migration: a computational framework to quantify population structure from genomic data. *Molecular biology and evolution*, 33(8):2102–2116.

Lai, S., Ruktanonchai, N. W., Zhou, L., Prosper, O., Luo, W., Floyd, J. R., Wesolowski, A., Santillana, M., Zhang, C., Du, X., Yu, H., and Tatem, A. J. (2020). Effect of non-pharmaceutical interventions to contain COVID-19 in China. *Nature*, 585(7825):410–413.

Lartillot, N. and Philippe, H. (2006). Computing Bayes factors using theromodynamic integration. *Systematic Biology*, 55:195–207.

Lele, S. R., Dennis, B., and Lutscher, F. (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology Letters*, 10(7):551–563.

Lemey, P., Rambaut, A., Bedford, T., Faria, N., Bielejec, F., Baele, G., Russell, C. A., Smith, D. J., Pybus, O. G., Brockmann, D., and Suchard, M. A. (2014). Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS Pathogens*, 10(2):e1003932.

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

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

Lemey, P., Ruktanonchai, N., Hong, S. L., Colizza, V., Poletto, C., Van den Broeck, F., Gill, M. S., Ji, X., Levasseur, A., Oude Munnink, B. B., Koopmans, M., Sadilek, A., Lai, S., Tatem, A. J., Baele, G., Suchard, M. A., and Dellicour, S. (2021). Untangling introductions and persistence in COVID-19 resurgence in Europe. *Nature*, 595(7869):713–717.

Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., and Shaman, J. (2020a). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science*, 368(6490):489–493.

Li, W. L. S. and Drummond, A. J. (2012). Model averaging and bayes factor calculation of relaxed molecular clocks in bayesian phylogenetics. *Molecular biology and evolution*, 29(2):751–761.

Li, X., Giorgi, E. E., Marichannegowda, M. H., Foley, B., Xiao, C., Kong, X.-P., Chen, Y., Gnanakaran, S., Korber, B., and Gao, F. (2020b). Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Science Advances*, 6(27):eabb9153.

Membrebe, J. V., Suchard, M. A., Rambaut, A., Baele, G., and Lemey, P. (2019). Bayesian in-

ference of evolutionary histories under time-dependent substitution rates. *Molecular Biology and Evolution*, 36(8):1793–1803.

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

Minin, V. N., Bloomquist, E. W., and Suchard, M. A. (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular Biology and Evolution*, 25(7):1459–1471.

Minin, V. N. and Suchard, M. A. (2008a). Counting labeled transitions in continuous-time Markov models of evolution. *Journal of Mathematical Biology*, 56(3):391–412.

Minin, V. N. and Suchard, M. A. (2008b). Fast, accurate and simulation-free stochastic mapping. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1512):3985–3995.

Müller, N. F., Dudas, G., and Stadler, T. (2019). Inferring time-dependent migration and coalescence patterns from genetic sequence and predictor data in structured populations. *Virus evolution*, 5(2):vez030.

Müller, N. F., Rasmussen, D. A., and Stadler, T. (2017). The structured coalescent and its approximations. *Molecular biology and evolution*, 34(11):2970–2981.

Müller, N. F., Wagner, C., Frazar, C. D., Roychoudhury, P., Lee, J., Moncla, L. H., Pelle, B., Richardson, M., Ryke, E., Xie, H., Shrestha, L., Addetia, A., Rachleff, V. M., Lieberman, N. A. P., Huang, M.-L., Gautom, R., Melly, G., Hiatt, B., Dykema, P., Adler, A., Brandstetter, E., Han, P. D., Fay, K., Ilcisin, M., Lacombe, K., Sibley, T. R., Truong, M., Wolf, C. R., Boeckh, M., Englund, J. A., Famulare, M., Lutz, B. R., Rieder, M. J., Thompson, M., Duchin, J. S., Starita, L. M., Chu, H. Y., Shendure, J., Jerome, K. R., Lindquist, S., Greninger, A. L., Nickerson, D. A., and Bedford, T. (2021). Viral genomes reveal patterns of the SARS-CoV-2 outbreak in Washington state. *Science Translational Medicine*, 13(595).

Nadeau, S. A., Vaughan, T. G., Scire, J., Huisman, J. S., and Stadler, T. (2021). The origin and early spread of SARS-CoV-2 in Europe. *Proceedings of the National Academy of Sciences*, 118(9).

NHCPRC (2020). COVID-19 reports. http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml. (Accessed on 12/30/2020).

Nielsen, R. (2002). Mapping mutations on phylogenies. *Systematic Biology*, 51(5):729–739.

O'Brien, J. D., Minin, V. N., and Suchard, M. A. (2009). Learning to count: robust estimates for labeled distances between molecular sequences. *Molecular Biology and Evolution*, 26(4):801–814.

Pagel, M., Meade, A., and Barker, D. (2004). Bayesian estimation of ancestral character states on phylogenies. *Systematic biology*, 53(5):673–684.

Pérez-Cruz, F. (2008). Kullback-leibler divergence estimation of continuous distributions. In *2008 IEEE international symposium on information theory*, pages 1666–1670. IEEE.

Plante, J. A., Liu, Y., Liu, J., Xia, H., Johnson, B. A., Lokugamage, K. G., Zhang, X., Muruato, A. E., Zou, J., Fontes-Garfias, C. R., Mirchandani, D., Scharton, D., Bilello, J. P., Ku, Z., An, Z., Kalveram, B., Freiberg, A. N., Menachery, V. D., Xie, X., Plante, K. S., Weaver, S. C., and Shi, P.-Y. (2020). Spike mutation D614G alters SARS-CoV-2 fitness. *Nature*, pages 1–6.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: convergence diagnosis and output analysis for mcmc. *R news*, 6(1):7–11.

Ponciano, J. M., Burleigh, J. G., Braun, E. L., and Taper, M. L. (2012). Assessing parameter identifiability in phylogenetic models using data cloning. *Systematic Biology*, 61(6):955–972.

Ponciano, J. M., Taper, M. L., Dennis, B., and Lele, S. R. (2009). Hierarchical models in ecology: confidence intervals, hypothesis testing, and model selection using data cloning. *Ecology*,

90(2):356–362.

Pybus, O. G., Suchard, M. A., Lemey, P., Bernardin, F. J., Rambaut, A., Crawford, F. W., Gray, R. R., Arinaminpathy, N., Stramer, S. L., Busch, M. P., et al. (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the national academy of sciences*, 109(37):15066–15071.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarization in bayesian phylogenetics using tracer 1.7. *Systematic biology*, 67(5):901.

Rannala, B. and Yang, Z. (2007). Inferring speciation times under an episodic molecular clock. *Systematic Biology*, 56(3):453–466.

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2):217–223.

Robert, C. P. (1993). Prior feedback: A bayesian approach to maximum likelihood estimation. *Computational Statistics*, 8:279–294.

Rodrigue, N., Philippe, H., and Lartillot, N. (2007). Uniformization for sampling realizations of markov processes: applications to bayesian implementations of codon substitution models. *Bioinformatics*, 24(1):56–62.

Schliep, K. P. (2010). phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593.

Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *Journal of molecular biology*, 188(3):415–431.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Shu, Y. and McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, 22(13):30494.

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using beast 1.10. *Virus evolution*, 4(1):vey016.

Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10(3):512–526.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17(2):57–86.

Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., San, E. J., Msomi, N., Mlisana, K., von Gottberg, A., Walaza, S., Allam, M., Ismail, A., Mohale, T., Glass, A. J., Engelbrecht, S., Van Zyl, G., Preiser, W., Petruccione, F., Sigal, A., Hardie, D., Marais, G., Hsiao, N.-y., Korsman, S., Davies, M.-A., Tyers, L., Mudau, I., York, D., Maslo, C., Goedhals, D., Abrahams, S., Laguda-Akingba, O., Alisoltani-Dehkordi, A., Godzik, A., Wibmer, C. K., Sewell, B. T., Lourenço, J., Alcantara, L. C. J., Kosakovsky Pond, S. L., Weaver, S., Martin, D., Lessells, R. J., Bhiman, J. N., Williamson, C., and de Oliveira, T. (2021). Emergence of a SARS-CoV-2 variant of concern with mutations in spike glycoprotein. *Nature*, pages 1–8.

Tian, H., Liu, Y., Li, Y., Wu, C.-H., Chen, B., Kraemer, M. U., Li, B., Cai, J., Xu, B., Yang, Q., et al. (2020). An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science*, 368(6491):638–642.

USCDC (2020). United States COVID-19 cases and deaths by state over time. https://data.cdc.gov/Case-Surveillance/

United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36. (Accessed on 12/30/2020).

van Kampen, J. J., van de Vijver, D. A., Fraaij, P. L., Haagmans, B. L., Lamers, M. M., Okba, N., van den Akker, J. P., Endeman, H., Gommers, D. A., Cornelissen, J. J., et al. (2021). Duration and key determinants of infectious virus shedding in hospitalized patients with coronavirus disease-2019 (COVID-19). *Nature Communications*, 12(1):1–6.

Vaughan, T. G., Sciré, J., Nadeau, S. A., and Stadler, T. (2020). Estimates of outbreak-specific SARS-CoV-2 epidemiological parameters from genomic data. *medRxiv*.

Volz, E., Hill, V., McCrone, J. T., Price, A., Jorgensen, D., Áine O'Toole, Southgate, J., Johnson, R., Jackson, B., Nascimento, F. F., Rey, S. M., Nicholls, S. M., Colquhoun, R. M., da Silva Filipe, A., Shepherd, J., Pascall, D. J., Shah, R., Jesudason, N., Li, K., Jarrett, R., Pacchiarini, N., Bull, M., Geidelberg, L., Siveroni, I., COG-UK Consortium, Goodfellow, I., Loman, N. J., Pybus, O. G., Robertson, D. L., Thomson, E. C., Rambaut, A., and Connor, T. R. (2021). Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. *Cell*, 184(1):64–75.

Washington, N. L., Gangavarapu, K., Zeller, M., Bolze, A., Cirulli, E. T., Schiabor Barrett, K. M., Larsen, B. B., Anderson, C., White, S., Cassens, T., Jacobs, S., Levan, G., Nguyen, J., Ramirez, J. M., Rivera-Garcia, C., Sandoval, E., Wang, X., Wong, D., Spencer, E., Robles-Sikisaka, R., Kurzban, E., Hughes, L. D., Deng, X., Wang, C., Servellita, V., Valentine, H., De Hoff, P., Seaver, P., Sathe, S., Gietzen, K., Sickler, B., Antico, J., Hoon, K., Liu, J., Harding, A., Bakhtar, O., Basler, T., Austin, B., MacCannell, D., Isaksson, M., Febbo, P. G., Becker, D., Laurent, M., McDonald, E., Yeo, G. W., Knight, R., Laurent, L. C., de Feo, E., Worobey, M., Chiu, C. Y., Suchard, M. A., Lu, J. T., Lee, W., and Andersen, K. G. (2021). Emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States. *Cell*, 184(10):2587–2594.

WHO (2020). Coronavirus disease (COVID-19) situation reports. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports. (Accessed on 12/19/2020).

Wikipedia (2020a). COVID-19 pandemic in mainland China. https://en.wikipedia.org/wiki/COVID-19_pandemic_in_mainland_China. (Accessed on 12/19/2020).

Wikipedia (2020b). COVID-19 pandemic lockdown in Hubei. https://en.wikipedia.org/wiki/COVID-19_pandemic_lockdown_in_Hubei. (Accessed on 12/19/2020).

Wilfert, L., Long, G., Leggett, H., Schmid-Hempel, P., Butlin, R., Martin, S., and Boots, M. (2016). Deformed wing virus is a recent global epidemic in honeybees driven by Varroa mites. *Science*, 351(6273):594–597.

Wilkinson, E., Giovanetti, M., Tegally, H., San, J. E., Lessells, R., Cuadros, D., Martin, D. P., Rasmussen, D. A., Zekri, A.-R. N., Sangare, A. K., Ouedraogo, A.-S., Sesay, A. K., Priscilla, A., Kemi, A.-S., Olubusuyi, A. M., Oluwapelumi, A. O. O., Hammami, A., Amuri, A. A., Sayed, A., Ouma, A. E. O., Elargoubi, A., Ajayi, N. A., Victoria, A. F., Kazeem, A., George, A., Trotter, A. J., Yahaya, A. A., Keita, A. K., Diallo, A., Kone, A., Souissi, A., Chtourou, A., Gutierrez, A. V., Page, A. J., Vinze, A., Iranzadeh, A., Lambisia, A., Ismail, A., Rosemary, A., Sylverken, A., Femi, A., Ibrahimi, A., Marycelin, B., Oderinde, B. S., Bolajoko, B., Dhaala, B., Herring, B. L., Njanpop-Lafourcade, B.-M., Kleinhans, B., McInnis, B., Tegomoh, B., Brook, C., Pratt, C. B., Scheepers, C., Akoua-Koffi, C. G., Agoti, C. N., Peyrefitte, C., Daubenberger, C., Morang'a, C. M., Nokes, D. J., Amoako, D. G., Bugembe, D. L., Park, D., Baker, D., Doolabh, D., Ssemwanga, D., Tshiabuila, D., Bassirou, D., Amuzu, D. S. Y., Goedhals, D., Omuoyo, D. O., Maruapula, D., Foster-Nyarko, E., Lusamaki, E. K., Simulundu, E., Ong'era, E. M., Ngabana, E. N., Shumba, E., Fahime, E. E., Lokilo, E., Mukantwari, E., Philomena,

E., Belarbi, E., Simon-Loriere, E., Anoh, E. A., Leendertz, F., Ajili, F., Enoch, F. O., Wasfi, F., Abdelmoula, F., Mosha, F. S., Takawira, F. T., Derrar, F., Bouzid, F., Onikepe, F., Adeola, F., Muyembe, F. M., Tanser, F., Dratibi, F. A., Mbunsu, G. K., Thilliez, G., Kay, G. L., Githinji, G., van Zyl, G., Awandare, G. A., Schubert, G., Maphalala, G. P., Ranaivoson, H. C., Lemriss, H., Anise, H., Abe, H., Karray, H. H., Nansumba, H., Elgahzaly, H. A., Gumbo, H., Smeti, I., Ayed, I. B., Odia, I., Boubaker, I. B. B., Gaaloul, I., Gazy, I., Mudau, I., Ssewanyana, I., Konstantinus, I., Lekana-Douk, J. B., Makangara, J.-C. C., Tamfum, J.-J. M., Heraud, J.-M., Shaffer, J. G., Giandhari, J., Li, J., Yasuda, J., Mends, J. Q., Kiconco, J., Morobe, J. M., Gyapong, J. O., Okolie, J. C., Kayiwa, J. T., Edwards, J. A., Gyamfi, J., Farah, J., Nakaseegu, J., Ngoi, J. M., Namulondo, J., Andeko, J. C., Lutwama, J. J., O'Grady, J., Siddle, K., Adeyemi, K. T., Tumedi, K. A., Said, K. M., Hae-Young, K., Duedu, K. O., Belyamani, L., Fki-Berrajah, L., Singh, L., de O. Martins, L., Tyers, L., Ramuth, M., Mastouri, M., Aouni, M., el Hefnawi, M., Matsheka, M. I., Kebabonye, M., Diop, M., Turki, M., Paye, M., Nyaga, M. M., Mareka, M., Damaris, M.-M., Mburu, M. W., Mpina, M., Nwando, M., Owusu, M., Wiley, M. R., Youtchou, M. T., Ayekaba, M. O., Abouelhoda, M., Seadawy, M. G., Khalifa, M. K., Sekhele, M., Ouadghiri, M., Diagne, M. M., Mwenda, M., Allam, M., Phan, M. V. T., Abid, N., Touil, N., Rujeni, N., Kharrat, N., Ismael, N., Dia, N., Mabunda, N., yuan Hsiao, N., Silochi, N. B., Nsenga, N., Gumede, N., Mulder, N., Ndodo, N., Razanajatovo, N. H., Iguosadolo, N., Judith, O., Kingsley, O. C., Sylvanus, O., Peter, O., Femi, O., Idowu, O., Testimony, O., Chukwuma, O. E., Ogah, O. E., Onwuamah, C. K., Cyril, O., Faye, O., Tomori, O., Ondoa, P., Combe, P., Semanda, P., Oluniyi, P. E., Arnaldo, P., Quashie, P. K., Dussart, P., Bester, P. A., Mbala, P. K., Ayivor-Djanie, R., Njouom, R., Phillips, R. O., Gorman, R., Kingsley, R. A., Carr, R. A. A., Kabbaj, S. E., Gargouri, S., Masmoudi, S., Sankhe, S., Lawal, S. B., Kassim, S., Trabelsi, S., Metha, S., Kammoun, S., Lemriss, S., Agwa, S. H. A., Calvignac-Spencer, S., Schaffner, S. F., Doumbia, S., Mandanda, S. M., Aryeetey, S., Ahmed, S. S., Elhamoumi, S., Andriamandimby, S., Tope, S., Lekana-Douki, S., Prosolek, S., Ouangraoua, S., Mundeke, S. A., Rudder, S., Panji, S., Pillay, S., Engelbrecht, S., Nabadda, S., Behillil, S., Budiaki, S. L., van der Werf, S., Mashe, T., Aanniz, T., Mohale, T., Le-Viet, T., Schindler, T., Anyaneji, U. J., Chinedu, U., Ramphal, U., Jessica, U., George, U., Fonseca, V., Enouf, V., Gorova, V., Roshdy, W. H., Ampofo, W. K., Preiser, W., Choga, W. T., Bediako, Y., Naidoo, Y., Butera, Y., de Laurent, Z. R., Sall, A. A., Rebai, A., von Gottberg, A., Kouriba, B., Williamson, C., Bridges, D. J., Chikwe, I., Bhiman, J. N., Mine, M., Cotten, M., Moyo, S., Gaseitsiwe, S., Saasa, N., Sabeti, P. C., Kaleebu, P., Tebeje, Y. K., Tessema, S. K., Happi, C., Nkengasong, J., and de Oliveira, T. (2021). A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa. *Science*, 374(6566):423–431.

Wölfel, R., Corman, V. M., Guggemos, W., Seilmaier, M., Zange, S., Müller, M. A., Niemeyer, D., Jones, T. C., Vollmar, P., Rothe, C., et al. (2020). Virological assessment of hospitalized patients with COVID-2019. *Nature*, 581(7809):465–469.

Worobey, M., Pekar, J., Larsen, B. B., Nelson, M. I., Hill, V., Joy, J. B., Rambaut, A., Suchard, M. A., Wertheim, J. O., and Lemey, P. (2020). The emergence of SARS-CoV-2 in Europe and North America. *Science*, 370(6516):564–570.

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., Yuan, M.-L., Zhang, Y.-L., Dai, F.-H., Liu, Y., Wang, Q.-M., Zheng, J.-J., Xu, L., Holmes, E. C., and Zhang, Y.-Z. (2020a). A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798):265–269.

Wu, T., Ge, X., Yu, G., and Hu, E. (2020b). Open-source analytics tools for studying the COVID-19 coronavirus outbreak. *medRxiv*.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., and Chen, M.-H. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology*, 60:150—160.

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

Yang, Z. (2014). *Molecular Evolution: a Statistical Approach*. Oxford University Press.

Yang, Z., Kumar, S., and Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, 141(4):1641–1650.

Yao, H.-W., Yang, Y., Liu, K., Li, X.-L., Zuo, S.-Q., Sun, R.-X., Fang, L.-Q., and Cao, W.-C. (2015). The spatiotemporal expansion of human rabies and its probable explanation in mainland China, 2004-2013. *PLoS Neglected Tropical Diseases*, 9(2):e0003502.