PNAS RESEARCH ARTICLE | EVOLUTION APPLIED MATHEMATICS

Modeling the velocity of evolving lineages and predicting dispersal patterns

Paul Bastide[a,b], Pauline Rocu[c], Johannes Wirtz[d], Gabriel W. Hassler[e], François Chevenet[f], Denis Fargette[g], Marc A. Suchard[h,i,j], Simon Dellicour[k,l], Philippe Lemey[l], and Stéphane Guindon[c,1]

Affiliations are included on p. 8.

Edited by Eugene Koonin, NIH, Bethesda, MD; received June 10, 2024; accepted October 14, 2024

Accurate estimation of the dispersal velocity or speed of evolving organisms is no mean feat. In fact, existing probabilistic models in phylogeography or spatial population genetics generally do not provide an adequate framework to define velocity in a relevant manner. For instance, the very concept of instantaneous speed simply does not exist under one of the most popular approaches that models the evolution of spatial coordinates as Brownian trajectories running along a phylogeny. Here, we introduce a family of models—the so-called Phylogenetic Integrated Velocity (PIV) models—that use Gaussian processes to explicitly model the velocity of evolving lineages instead of focusing on the fluctuation of spatial coordinates over time. We describe the properties of these models and show an increased accuracy of velocity estimates compared to previous approaches. Analyses of West Nile virus data in the United States indicate that PIV models provide sensible predictions of the dispersal of evolving pathogens at a one-year time horizon. These results demonstrate the feasibility and relevance of predictive phylogeography in monitoring epidemics in time and space.

phylogeography | Bayesian inference | West Nile virus | integrated velocity models

Significance

Measuring the speed at which pathogens disperse during an epidemic is challenging. Classical epidemiological approaches rely on incidence and prevalence counts. Yet, these data provide limited information about the dispersal process. The comparison of pathogen genomes combined with spatial coordinates can alleviate this issue. This study introduces a model that is specially designed to infer velocities from the analysis of geo-referenced genetic sequences. The reconstruction of ancestral and present-day velocities provides a detailed picture of the dispersal patterns underlying an epidemic. Importantly, present-day velocity estimates serve as a basis for predicting future dispersal events. Application of this technique to the West Nile virus in the United States demonstrates the potential of this approach for monitoring the spread of an epidemic.

Evaluating the pace at which organisms move in space during the course of evolution is an important endeavor in biology. When considering deep evolutionary time scales, understanding past dispersal events is key to explaining the spatial diversity of contemporaneous species. Over shorter time frames, making sense of the migration patterns of closely related organisms is crucial in building a detailed picture of a population's demographic past, present, and future dynamics. Tracking the spatial dynamics of pathogens during a pandemic, in particular, is of utmost interest as it conveys useful information about the means and the rapidity at which a disease is spreading in a population. Epidemiological data generally consist in records of incidence of the disease at various points in time and space. Yet, estimating the speed at which an organism spreads at the onset of an epidemic from count data is challenging (1, 2). Similarly, characterizing the migration process from occurrence data in cases where the organism under scrutiny is already well-established in a region is not feasible. These difficulties mainly stem from the fact that count or occurrence data do not convey information about the nonindependence between observations due to their shared evolutionary paths.

Genomes carry useful information about the relationships between pathogens. Observed differences between homologous genetic sequences are at the core of phylogenetic and population genetics approaches which provide a sound framework to account for the nonindependence between data points in downstream analyses. This framework also accommodates for situations where nucleotide (or protein) sequences are sampled at various points in time (3). Heterochronous samples combined with the molecular clock hypothesis (4) may then serve as a basis to infer the rate at which substitutions accumulate and to reconstruct the time scale of past demographic trajectories of the population under scrutiny (see, e.g., ref. 5 for a review).

Designing models for the joint analysis of genetic sequences and their locations of collection was initiated in the middle of the last century by Wright and Malécot who brought forward the isolation by distance model (6, 7). The rise of statistical phylogeography over the last decade proposed alternatives that are less mechanistic but still aim at capturing the main features of the spatial diffusion process. These approaches are also well suited to deal with heterochronous data and handle cases where the population of interest is scattered along a spatial continuum rather than structured into discrete demes. Lemey et al. (8), in particular, described a hierarchical model whereby spatial coordinates evolve along a phylogenetic tree according to a Brownian diffusion

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: stephane.guindon@lirmm.fr.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2411582121/-/DCSupplemental.

Published November 15, 2024.

process with branch-specific diffusion rates. The so-called Relaxed Random Walk (RRW) model has since then been used to characterize the spatial dynamics of several pathogens of high public health, societal, and agricultural impacts, including Ebola (9) and the rice yellow mottle (10) viruses for instance.

One of the key objectives of the RRW model is to infer the rate at which organisms disperse. Pybus et al. (11) suggested using a diffusion coefficient which derives from the ratio of the estimated squared displacement between the start and the end of a branch and the corresponding elapsed time. The branch-level ratios are then averaged over the edges in the phylogeny. Pybus et al. (11) and then Dellicour et al. (12) later introduced wavefront-through-time plots, deriving from the displacement between the estimated root location and the most distant tip locations at various points in time. Trovão et al. (13) considered instead dispersal rates which are defined as ratios of estimated displacements (using great-circle distances) by the elapsed time.

These statistics generally provide a rough characterization of the dispersal process. The limitations of the dispersal statistics mainly stem from the very nature of the RRW model: Because Brownian trajectories are nowhere differentiable, the concept of instantaneous speed is simply not defined under that family of models. Also, the sum of displacements deriving from the observation of a Brownian particle at various points in time grows with the square root of the number of (equally spaced in time) observations, making the estimation of an average speed sampling inconsistent. Finally, the analysis of spatial data simulated under the Brownian motion model along birth–death trees shows that the standard dispersal statistics often fail to provide accurate estimates of speed (14).

The present study tackles the issue of dispersal velocity and speed estimation by introducing an approach that models the instantaneous velocity of lineages explicitly. Under these models, the spatial coordinates of lineages derive from integrating their velocities so that we refer to Phylogenetic Integrated Velocity (PIV) models throughout this article. This study uses the integrated velocity models in a phylogenetic context. Integrated processes are common however in a variety of applications, ranging from population biology (15) to financial economics (16). In virology, longitudinal studies measuring CD4 T cell numbers in cohorts of patients with AIDS have used them to test the hypothesis of "derivative tracking," in which an individual's measurements over time tend to maintain the same trajectory (17). Closer to phylogeography, integrated processes are instrumental in the field of animal movement ecology (18, 19). Unlike simple random walks, these processes are not Markovian as the entire track provides information about the next step through the integration. They are thus relevant for accounting for directional persistence. Furthermore, the integrated processes are related to physical models of particles moving on a potential surface (20, 21), therefore permitting fine-grained modeling of animal movement telemetry data. One of the goals of the present work is to explore the potential of such approaches in the context of phylogeography, starting with the two simplest and most common models, namely the integrated Brownian and Ornstein–Uhlenbeck processes.

Although velocity is not directly observable from heterochronous and geo-referenced genetic sequences, our results indicate that this quantity can be estimated reliably. Using simulations under realistic spatial population genetics models, we show that the velocity inferred with PIV models are more accurate than those deriving from the RRW approach. Velocities estimated from the analysis of multiple West Nile virus datasets were also used to predict the spatial distribution of the pathogen over a one-year time horizon in the United States. Comparison of these predictions to incidence data at the county level suggests that important features of the spatial dynamics are indeed amenable to reasonably accurate predictions.

Our ability to efficiently monitor and anticipate the spread of emerging epidemics depends on the accuracy with which the pace of dispersal can be quantified. The family of models introduced in this study provides a relevant tool to achieve this objective. While important aspects of viral evolution may escape prediction indefinitely (22) and predicting the time and/or location of the next virus outbreak remains out of reach (23, 24), the present study shows how predictive phylogeography may complement classical approaches in epidemiology.

Results

PIV Models: Rationale. The main attributes of models that belong to the PIV family are presented first. We focus on the process of interest along a given time interval $[0, t]$, corresponding to the length (in calendar time units) of a given branch in the phylogeny of a sample of the organism of interest. Let $X(s)$ be the random variable representing the location (i.e., the coordinates) of a lineage at time $0 \leq s \leq t$. $Y(s)$ is its velocity, i.e., the vector that is made of the instantaneous rate at which a lineage changes its position along each dimension of the habitat at time s. In all the following, we reserve the term velocity for the vector, and speed for its scalar norm. Both X and Y are typically vectors of length two, corresponding to latitude and longitude. The location $X(t)$ at the end of the branch may then be expressed as follows:

$$X(t) = x(0) + \int_0^t Y(s)\mathrm{d}s, \qquad [1]$$

where $x(0)$, the location at the time of origin, is fixed. The Brownian Motion (BM) and the RRW models focus on $\{X(s), 0 \leq s \leq t\}$, i.e., the process describing the evolution of the location during a time interval. While, in one dimension, BM models have a single dispersal parameter that applies to all edges in the phylogeny, the RRW model has branch-specific dispersal parameters, in a manner similar to the relaxed clock model (25) used in molecular dating.

Instead of modeling the fluctuation of coordinates, PIV models deal with $\{Y(s), 0 \leq s \leq t\}$, i.e., the process describing the variation of velocity in that interval. The dynamics of spatial coordinates then derive from the integration over the velocity as stated in Eq. **1** above, hence the name "phylogenetic integrated velocity." In the following, we introduce two stochastic processes for $\{Y(s), 0 \leq s \leq t\}$ and characterize the corresponding distributions of $X(t)$. In order to simplify the presentation, we provide formulas for univariate processes only in the main text. Formulas for bivariate (and, more generally, multivariate) processes are given in (*SI Appendix*, sections C and D).

Behavior of PIV Models.

Velocities. The Integrated Brownian Motion (IBM) model relies on a Wiener process with shift and scale parameters $y(0)$ and σ, respectively, to model $\{Y(t); t > 0\}$. That process is Gaussian and we have (26):

$$\mathrm{E}(Y(t) \mid y(0)) = y(0), \qquad [2]$$

$$\mathrm{Cov}(Y(u), Y(v)) = \sigma^2 u, \text{ with } 0 < u \leq v. \qquad [3]$$

The Integrated Ornstein–Uhlenbeck (IOU) model uses instead an Ornstein–Uhlenbeck (OU) process to describe the

evolution of velocity. The mean and variance of velocity at time t are given below (26):

$$\mathrm{E}(Y(t) \mid y(0)) = y(0)e^{-\theta t} + \mu(1 - e^{-\theta t}), \quad [4]$$

$$\mathrm{Cov}(Y(u), Y(v)) = \frac{\sigma^2 e^{-\theta v}}{\theta} \sinh(\theta u), \text{ with } 0 < u \leq v. \quad [5]$$

The parameter θ in the OU model governs the strength with which $Y(t)$ is pulled toward the trend μ.

Spatial coordinates. We now examine the evolution of spatial coordinates under the PIV models. Characterizing the process governing the evolution of spatial coordinates will shed light on the biological relevance of the proposed approach and exhibit the main difference in behavior in comparison with the BM and, by extension, the RRW models. The stochastic processes modeling the fluctuation of velocity being Gaussian, the coordinates also follow a Gaussian process (15). We give below the mean and variance of the distribution of $X(t)$ given $x(0)$ and $y(0)$, the coordinates and velocity at time 0.

When velocity follows a Brownian process (IBM process), we have

$$\mathrm{E}(X(t) \mid x(0), y(0)) = x(0) + y(0)t, \quad [6]$$

$$\mathrm{V}(X(t) \mid x(0), y(0)) = \frac{\sigma^2 t^3}{3}. \quad [7]$$

A linear increase of the spatial coordinates is thus expected with a direction that is determined by the initial velocity (Eq. **6**). Because of the inertia deriving from their velocity, spatial coordinates of lineages evolving under IBM thus tend to resist changes in their direction of motion, i.e., they exhibit directional persistence (18). This mean drift is similar to the directional random walk, used e.g., in ref. 27 to model the spatial spread of HIV-1. The BM model has a distinct behavior as it authorizes sudden changes of direction. The RRW can even lead to large discontinuous "jumps" from one place to another (28). In contrast, the IBM is smoother (differentiable) by design, and well suited to model autocorrelated movements. Moreover, as suggested by Eq. **7** above, the variance of coordinates grows cubically in time, thereby allowing the IBM model to accommodate for dispersal events over long distances in short periods of time. This process is thus able to handle fast spatial range expansion, yet with continuous and differentiable trajectories.

The corresponding expectation and variance for the IOU model are given in (*SI Appendix*, section A). Here again, the average coordinates at the end of the branch of focus are determined by the coordinates at the start of that branch ($x(0)$) plus the expected displacement ($y(0)t$) along that same edge. In this simple IOU model, the velocity of the process converges to the central value μ, leading to trajectories with a clear directional trend that are well suited for dispersal along an established spatial gradient. While for small values of θ the IOU has a behavior that is similar to the IBM, for larger values of that parameter, its variance grows linearly in time and the process behaves like a directional BM (27). The autocorrelation (or strength) parameter θ is thus interpreted as the amount of directional persistence present in the data (18), with small values indicating more dependence to the trajectory of the elapsed path for future moves.

Fig. 1 illustrates the behavior of the classical random walk and integrated models along a 5-tip tree. Trajectories of coordinates generated with the BM and OU versions of the random walk model are intricate, showing abrupt changes of directions in the movements (Fig. 1 *B* and *C*). The same behavior is displayed by the velocity trajectories under the IBM and IOU models (Fig. 1 *D* and *F*) as the models are here identical to that used for the BM and OU models indeed. Yet, integrating over these rugged paths gives smooth (differentiable) trajectories of coordinates under the corresponding models (Fig. 1 *E* and *G*), with particles moving swiftly away from their initial points, illustrating the cubic variance pointed above. The IOU model presented here converges to a (1,1) velocity so that the coordinates of the five lineages show a clear directionality, stronger than that obtained with the OU model (Fig. 1 *G* vs. *C*).

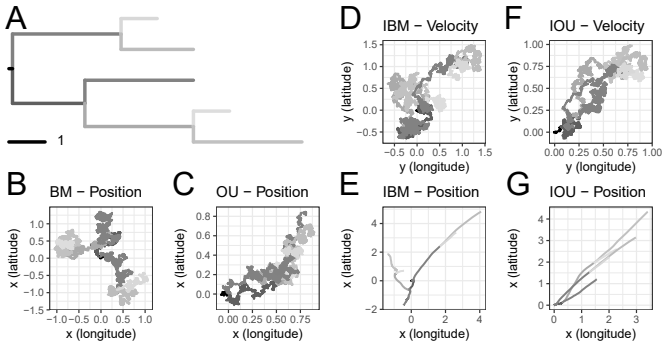

Fig. 1. Simulated trajectories of classical random walk and PIV models on a simple tree. Each process was simulated on the tree (*A*), with branches of matching colors. Movements along the latitude and longitude axes were simulated independently. (*B*) Random walk using Brownian motion (BM) with variance $\sigma^2 = 0.1$, and starting point $x(0) = (0, 0)$. (*C*) Ornstein–Uhlenbeck process (OU) with stationary variance of $\sigma^2/(2\theta) = 0.1$, strength $\theta = 0.17$, starting at $x(0) = (0, 0)$, and converging to its central value $\mu = (1, 1)$. (*D* and *E*) Velocity y and position x of an Integrated Brownian Motion (IBM) with variance $\sigma^2 = 0.1$, starting point $x(0) = (0, 0)$, and starting velocity $y(0) = (0, 0)$. (*F* and *G*) Velocity y and position x of an Integrated Ornstein–Uhlenbeck (IOU) with stationary variance $\sigma^2/(2\theta) = 0.1$, strength $\theta = 0.17$, central trend of $\mu = (1, 1)$, starting point $x(0) = (0, 0)$, and starting velocity $y(0) = (0, 0)$.

Accuracy of Speed Estimation. Datasets were simulated under the spatial Lambda-Fleming-Viot (SLFV) model (29, 30) and an agent-based spatially explicit transmission chain simulator which aimed at mimicking outbreaks of the Ebola virus in West Africa (31). 100 datasets were analyzed for each of these two simulation settings. As traditional speed statistics are typically computed over the whole tree (32), we assessed the ability of PIV models to estimate tree-level speed by averaging node-level velocities across the tree. The classical weighted lineage dispersal velocity (WLDV) (33) was used instead for all analyses performed under the RRW model. As shown recently (14), we expect the WLDV statistic on RRW models to perform poorly and would like to assess the ability of PIV models to provide more accurate speed estimates.

Examination of the estimated vs. true speed relationship (Fig. 2) indicates that the RRW model systematically underestimates speed and the bias worsens with increasing speed. This bias is strong with data simulated under the SLFV (Fig. 2*A*) and milder with the Ebola datasets (Fig. 2*B*), which is expected since the transmission trees generated in the latter case are sampled in time and not ultrametric, making the temporal signal to estimate speed stronger. Nonetheless, true speed values are, on average, 1.6 times larger than those estimated with the RRW for the Ebola data and 22 times larger for the SLFV data (the ratios for IBM are 1.2 for both simulation settings). The SLFV model assumes a finite-size habitat (a square here) and boundary effects, which occur for large and small dispersal values, are expected to impact the estimation of speed under models that ignore

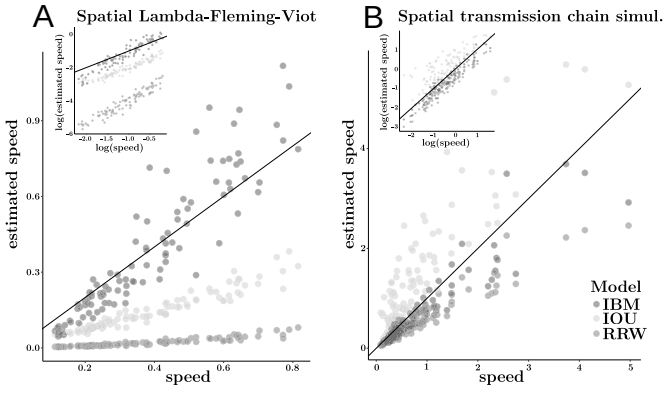

Fig. 2. Accuracy of speed estimation under the RRW and PIV models. True (*x*-axis) vs. estimated (*y*-axis) speed. Estimates were obtained under the IBM, IOU, and RRW models. (*A*) 100 datasets were simulated using the spatial Lambda-Fleming-Viot process on a 10-by-10 square. (*B*) 100 datasets were simulated under a random walk model inspired by the Ebola epidemic in West Africa (see the main text). The *Insets* give the log–log scatterplots of the estimated vs. true speeds. The $y = x$ line is shown in black on each plot.

this constraint. Yet, the IBM model is largely immune to this issue. While the IOU model underestimates speed for the SLFV datasets, its estimates are less biased than those deriving from the RRW model. The IOU model also tends to overestimate speed on the Ebola datasets. Further examination of these results shows a clear influence of the prior distribution on the strength parameter in the IOU model, a phenomenon already observed in ref. 34.

Dispersal Dynamics of the West Nile Virus (WNV) in the United States. The phylogeography of the WNV in the United States has been studied extensively (see, e.g., ref. 32). The origin of this epidemic took place in New York City during the summer 1999 (35, 36). By 2004, human infections, veterinary disease cases, or infections in mosquitoes, birds, or sentinel animals had been reported to the Centers for Disease Control and Prevention (CDC) in most counties.

We fitted the PIV and RRW models to several subsets of the 801 geo-referenced sequence dataset analyzed in ref. 32. PIV models are less flexible than the RRW approach as they do not authorize sudden changes of direction, as noted earlier (and see *SI Appendix*, section B). Therefore, ensuring that both approaches nonetheless provide comparable fit to the data is a prerequisite to further analyses. We then used the IBM model to predict the dispersal patterns and evaluate these predictions through the comparison with incidence data for the 2000 to 2007 time period.

Model comparison. We compared the fit of the RRW and PIV models to the WNV data using cross-validation of location information. Cross-validation is a powerful model comparison technique in the context of phylogenetic factor analysis (37). Using a subset of 150 data points chosen uniformly at random among the 801 available observations, a leave-one-out procedure was applied to the sample coordinates. Each tip location was first hidden and its posterior density was estimated using Markov chain Monte Carlo (MCMC) from the remaining 149 locations and all 150 sequences (*SI Appendix*, section G).

Fig. 3 shows the distributions of the great circle distances between the observed and reconstructed tip locations as inferred under the PIV and the RRW models, along with that of uniform at random predictions. The three phylogeographic models have similar behavior overall with a majority of distances between true and reconstructed tip locations ranging between 238 km (25% quantile of distribution from MCMC output pooled across models) and 950 km (75% quantile) with a median of 450 km. In contrast, if inferred locations are uniform at random within the United States (excluding Alaska and Hawaii), the median distance is 1,564 km, i.e., more than three times that estimated with the phylogenetic models. This result demonstrates the ability of these models to extract meaningful signal from the data, even though these approaches do not account for habitat borders (while the uniform predictor does so). Examination of the posterior distribution deriving from each model taken separately indicates that the median distances obtained under the IBM, IOU, and RRW models are 474, 496, and 416 km, respectively. While the fit of the RRW model is superior to that of the PIV models, the performance of the three models are nonetheless qualitatively similar.

Predicting dispersal using PIV models. PIV models enable the estimation of dispersal velocity of each sampled lineage. These velocities may then serve as a basis to predict the spatial distribution of the underlying population in the near future. Here, we tested the ability of the IBM model to anticipate the dynamics of dispersal of the WNV in the early and later stages of the epidemic.

Sequences collected earlier than December of year Y were randomly subsampled from the complete dataset with exponentially increasing weights given to recent samples. Datasets with 150 sequences were obtained except for years 2000 to 2002 where smaller sample sizes were considered due to a lack of observations in this time period. Estimated posterior distributions of velocities at the tips of the obtained phylogeny under the IBM model were then used as predictors of the spatial distribution of the virus in year $Y+1$ (see *Materials and Methods*). The predicted occurrences were compared to yearly incidence data collected at the county level.

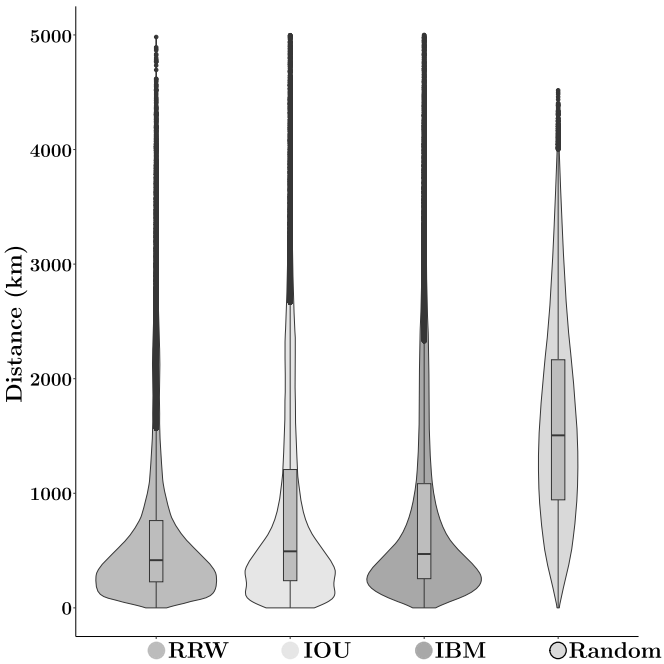

Fig. 3. Distribution of the distance between true and estimated tip coordinates under the PIV, the RRW models and uniform at random predictions (WNV data). Cross-validation was used to predict the locations of held-one-out tip lineages under the RRW and PIV models. "Random" gives the distance between two locations selected uniformly at random within the United States. The *y* axis gives the great circle distance between coordinates (in km).

Fig. 4 shows the incidence and the predicted occurrence of the WNV in the early stages in the epidemic. Samples for years 2000 to 2002 included only 7, 19, and 68 geo-referenced sequences, thereby making any prediction inherently challenging. For instance, predictions for year 2000 are overly dispersed and sensitive to priors (*SI Appendix*, section H). Also, while the virus had reached Florida by 2001, our model failed to predict its presence south of North Carolina. Predictions for subsequent years rely on larger numbers of observations and demonstrate the relevance of our approach. Indeed, the PIV model successfully predicted the arrival of the pathogen along the west coast of the United States by the end of 2002. It also correctly predicted that the north west corner of the country would remain largely virus-free until the end of 2003. Predictions deriving from the RRW show qualitatively distinct patterns with a widespread presence of the virus for years 2003 and 2004 that contrasts with incidence data (*SI Appendix*, section H). Overall, the RRW shows a higher sensitivity (average of 0.89 over all years for the RRW, vs. 0.72 for the IBM), but a lower specificity (average of 0.36 for the RRW, vs. 0.56 for the IBM), consistent with wider and rather vague predicted regions.

By 2004 the virus reached an endemic state and the spatial dynamics of the epidemic diverged from that of the early stages. Fig. 5 shows the results for the 2004 to 2007 time period. Prediction at local spatial scales has limited accuracy. For instance, a high probability of occurrence was systematically estimated for the states in the North East corner of the country and the south of Texas while incidence was generally mild in these areas. Note that the difference between predicted and observed occurrence could reflect a relatively lower ecological suitability of these regions to host local WNV circulation, thereby serving a useful purpose. Moreover, the IBM model correctly predicts the expansion of the epidemic north of California and Nevada between 2005 and 2006. Also, according to our predictions, the pathogen covered limited distances in the 2004 to 2007 period compared to the early stages of the epidemic. This quasi-stasis is confirmed by the largely similar distributions of yearly incidences. Hence, here again, our approach manages to capture changes in the spatial dynamics of the pandemic that are central in the context of pathogen surveillance.

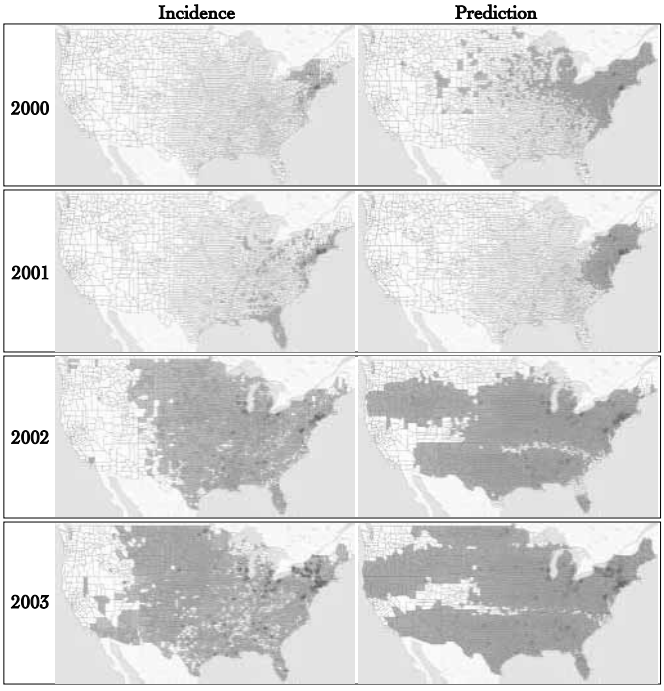

Fig. 4. Incidence and predicted occurrence of WNV in the early phase of the epidemic (model for prediction: IBM). Purple dots correspond to sampled locations. Incidence data (*Left*) for each year and each county was obtained from the CDC. For year *Y*, predicted occurrence of the WNV (*Right*) was inferred using data collected earlier than the end of December of year $Y-1$. The maps were generated with EvoLaps2 (38).

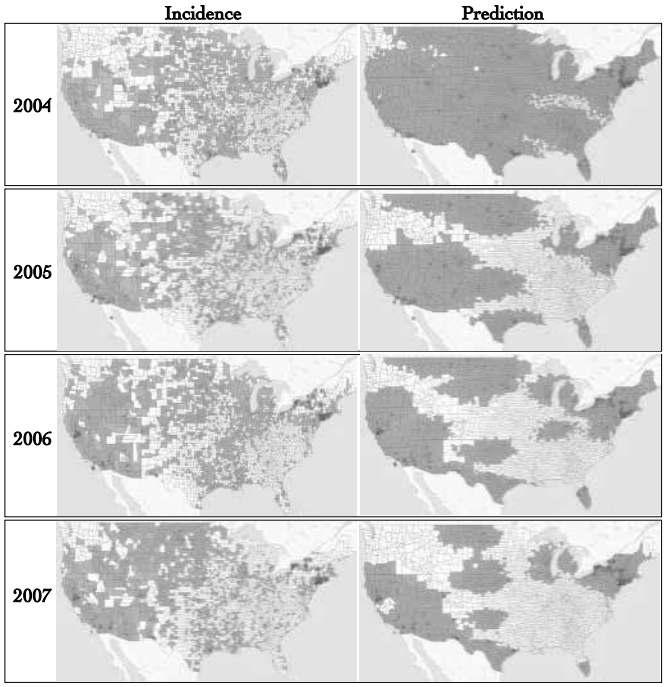

Fig. 5. Incidence and predicted occurrence of WNV in an endemic regime (model for prediction: IBM). See the caption of Fig. 4.

Discussion

The present study addresses shortcomings in the estimation of the velocity of lineages using popular models in phylogeography. These approaches rest on the probabilistic modeling of the coordinates of lineages along their phylogeny. Yet, the central concept of instantaneous speed does not exist under the most popular RRW model. As a consequence, measuring speed as a ratio between a displacement and the corresponding elapsed time leads to difficulties. In order to circumvent these limitations, we introduce PIV models. The originality of this family of models lies in their modeling of the velocity of evolving lineages instead of their coordinates. This approach enables a proper definition of instantaneous speed, which can be inferred anywhere along the tree, including at its tips.

Datasets were simulated under two models of spatial evolution that are distinct from that underlying the PIV and RRW approaches. Results show that speed estimates obtained with PIV models are generally more accurate than those deriving from the RRW approach, especially in cases where the pace of dispersal is high. Also, unlike RRW, PIV models produce velocity vector estimates at each node of the tree. We assessed the accuracy of these estimates at tip nodes in the IBM case and found that the velocity vectors were well estimated, with highest posterior density intervals having good coverage (*SI Appendix*, section I). Yet, PIV models are less flexible than RRW in their description of the movement of lineages during the course of evolution.

In particular, sudden changes in the direction of dispersal are not well accounted for by PIV models. These changes would indeed require "breaks" in the trajectory of velocities, which the underlying Gaussian processes do not allow. However, our analysis of West Nile virus data in the United States indicates that the movements of lineages display here enough inertia so that rapid changes in the spatial trajectories are seldom observed. Cross-validation suggests in fact that the two PIV models tested here provide a fit to the data similar to that obtained with the RRW. Moreover, the analytical expressions of the variance of coordinates under the IBM model grow with time in a superlinear manner, thereby allowing large displacements in short amounts of time.

Estimates of tip velocities can serve as a basis to model future dispersal events. Here, we evaluate the accuracy of predicted movements through the analysis of subsets of a large dataset of West Nile virus geo-referenced sequences and county-level yearly incidence data in the United States. Our predictions focus on deciding whether the pathogen will occupy (or be absent from) a given county at a given time interval in the future, i.e., a modest, yet challenging and critical endeavor compared to predicting future incidence. The proposed approach accurately predicted the arrival of the virus along the west coast of the United States in 2002 from the analysis of data collected before the end of December 2001. Furthermore, the predictions clearly point to a change of dispersal dynamics around 2004 to 2005 with a transition from an expansion phase to an endemic regime whereby rapid east-to-west dispersal events are replaced with short-distance migrations. While the proposed predictions have limited accuracy in the early stages of the pandemic where data is scarce and sampling likely to be biased, the PIV models successfully anticipate dispersal events in many instances. Altogether, our results indicate that the predictive phylogeography approach put forward in the present study could indeed serve a useful purpose in real-time forecasting of the spread of an epidemic. Future work could aim at incorporating data on the ecological suitability of the investigated areas in order to improve predictions, in a manner similar to that used in "landscape phylogeography" (39).

In addition to prediction, the PIV models are also expected to prove useful in many cases where the RRW model has been applied to quantify and compare dispersal velocity. These applications range from animal and human viruses to plant viruses. For instance, lower rates of dengue virus dispersal in urban as opposed to rural settings have implicated a major role for mosquito-mediated dispersal (40). Also, dispersal velocity has often been estimated for rabies lineages with dogs as the main host species, resulting in hypotheses of their spread being impacted by human activities (41). More recently, a slow dispersal has been estimated for Lassa virus in its rodent reservoir, which could in part explain the restricted distributions of the virus (42). Finally, increasing dispersal rates of the rice yellow mottle virus in Africa has led to the suggestion that intensification of rice cultivation could have enhanced the spread of that virus (43). Applications of the PIV models could increase the credibility of these and many more hypotheses of viral spread.

The proposed models and predictions have limitations, however. In a manner similar to that of the classical RRW framework, PIV models assume that i) the geographical position does not impact the fitness or the molecular evolution of the pathogen, ii) all the lineages are independent from one another, excluding any competition effect and iii) the geographical spread of the pathogen is independent from its current position. While limiting, these assumptions permit efficient computations and provided a sound methodological framework for important phylodynamics studies (see, e.g., ref. 44 for a review). In some specific contexts such as discrete phylogeography, some of these assumptions were relaxed, see, e.g., refs. 45 and 46 and references therein for i), (47–49) for ii), and ref. 50 for iii). Similar extensions to the PIV framework proposed here should be considered. In particular, models of animal movement also rely on integrated processes, with an additional potential function that links the dynamics of velocity evolution of an individual to its position at each point in time (20, 21). Such a potential function could be extended to include prior knowledge on the environmental layers impacting the spread of pathogens, including natural barriers such as coastline, or could be used to test the impact of specific environmental variables on the dispersion (51). However, the pruning algorithm used here (*Materials and Methods*) would not apply to these kinds of models, which are thus likely to be highly computationally intensive.

Furthermore, sampling is likely to impact the results in case it is driven by practical aspects (e.g., the distribution of genomic surveillance facilities is not uniform throughout the habitat) and does not reflect the underlying spatial distribution of the population under scrutiny (52). Recent work (53) shows how different sampling strategies can be incorporated in the RRW model. A similar framework could apply to PIV models and mitigate the impact of sampling. Additionally, when available, incidence data convey information about the demographic dynamics of an epidemic. Hence, increased accuracy of the predictions may be achievable through the incorporation of past incidence data in the models presented in this work.

Materials and Methods

Likelihood Calculation and Bayesian Inference. Let $\mathbf{X}^*$ and $\mathbf{X}$ correspond to random variables denoting the vectors of positions at the tips and the internal nodes, respectively. $\mathbf{x}^* = \{x_1, \ldots, x_n\}$ and $\mathbf{x} = \{x_{n+1}, \ldots, x_{2n-1}\}$ are realizations of the corresponding random variables, where n is the number of tips and $2n-1$ is the index of the root node. $\mathbf{Y}^*$ and $\mathbf{Y}$ are the vectors of velocities at tip and ancestral nodes, respectively. Here, we describe two different approaches for the Bayesian inference of PIV model parameters.

Data augmentation: Sampling velocities. The first method, implemented in PhyREX (53), relies on data augmentation. It starts with the computation of $p(\mathbf{x}^*, \mathbf{x}, \mathbf{y}^*, \mathbf{y})$, i.e., the joint density of all (i.e., ancestral and tip) locations and velocities. This density is also conditioned on the phylogeny, i.e., a rooted tree topology with node heights, which is not included in the formula below for the sake of conciseness. Given the locations and velocities at all nodes in the tree, the evolutionary process taking place along every branch is independent from that happening along the other edges. The likelihood is then evaluated as follows:

$$\begin{aligned} &p(\mathbf{x}^*, \mathbf{x}, \mathbf{y}^*, \mathbf{y}) \\ &\quad = \pi(y_\rho, x_\rho) \prod_{i=1}^{2n-2} p(x_i \mid x_{\mathrm{pa}(i)}, y_i, y_{\mathrm{pa}(i)}) p(y_i \mid y_{\mathrm{pa}(i)}, x_{\mathrm{pa}(i)}) \\ &\quad = \pi(y_\rho, x_\rho) \prod_{i=1}^{2n-2} p(x_i \mid x_{\mathrm{pa}(i)}, y_i, y_{\mathrm{pa}(i)}) \prod_{i=1}^{2n-2} p(y_i \mid y_{\mathrm{pa}(i)}), \end{aligned} \quad [8]$$

where the subscript $\mathrm{pa}(i)$ corresponds to the direct parent of node i. Also, $\pi(y_\rho, x_\rho)$ is the velocity and location density at the root node. In the present work, we use a normal density for the corresponding distribution. Since $(X_i \mid x_{\mathrm{pa}(i)}, y_i, y_{\mathrm{pa}(i)})$ is normally distributed, we can use the pruning algorithm as described in ref. 11 to integrate over $\mathbf{X}$, giving the following likelihood:

$$p(\mathbf{x}^*, \mathbf{y}^*, \mathbf{y}) = \pi(y_\rho, x_\rho) \phi(\mathbf{x}^*; \mathbf{y}^*, \mathbf{y}) \prod_{i=1}^{2n-2} p(y_i \mid y_{\mathrm{pa}(i)}), \quad [9]$$

where $\phi(\mathbf{x}^*; \mathbf{y}^*, \mathbf{y})$ is obtained through a postorder tree traversal, assuming the movements along both spatial axes are independent from one another and using the means and variances for either the IBM or the IOU model (*SI Appendix*, section B). The calculation just described relies on augmented data since velocities at all nodes in the tree are considered as known. Uncertainty around these latent variables is non-negligible. Samples from the joint posterior distribution of all model parameters, including ancestral and contemporaneous velocities, were obtained through MCMC integration.

Direct likelihood computation with the pruning algorithm. The second method for evaluating the likelihood of PIV models is implemented in the BEAST phylogenetic software package (54). It relies on the direct computation of $p(\mathbf{x}^*)$, the likelihood of the observed positions at the tips conditionally on the tree. It uses the fact that the stochastic process $\mathbf{Z}(t) = (\mathbf{Y}(t), \mathbf{X}(t))$ that describes the joint evolution of both the velocity and position is a multivariate Markov process, that can be framed as linear Gaussian as in refs. 55 and 56. Indeed, as shown in (*SI Appendix*, section C), for any node i with parent $\mathrm{pa}(i)$, the joint velocity-position vector $\mathbf{Z}_i$ can be written conditionally on $\mathbf{Z}_{\mathrm{pa}(i)}$ the vector at the parent node $\mathrm{pa}(i)$, as $\mathbf{Z}_i = \mathbf{q}_i \mathbf{Z}_{\mathrm{pa}(i)} + \mathbf{r}_i + \boldsymbol{\epsilon}_i$, with $\boldsymbol{\epsilon}_i$ a Gaussian random variable with variance $\boldsymbol{\Sigma}_i$ that is independent from $\mathbf{Z}_{\mathrm{pa}(i)}$, and $\sum_i$, $\mathbf{q}_i$, and $\mathbf{r}_i$, two matrices and a vector of dimension 4 that only depend on the tree and the parameters of the PIV process considered. In this approach, all the velocities at the tips are considered as missing: We only observe the last two entries of vector $\mathbf{Z}$ corresponding to the position, but the velocities are unknown. In ref. 56, a general pruning algorithm is described to deal with this kind of process (with missing values), that provides not only the likelihood (one postorder traversal) but also the conditional distribution of nonobserved traits conditioned on observed traits at the tips (one additional preorder traversal). This algorithm hence readily gives the posterior distribution of velocities without the need to sample from them using MCMC. Moreover, it does not need to assume that movements along the spatial axes are independent from one another.

Phylogeographic Bayesian inference. In both approaches, standard operators were used to update the topology of the phylogenetic tree, the node ages along with the parameters of a Hasegawa, Kishino, Yano (HKY) (57) nucleotide substitution model. The diffusion parameters of the Brownian process were also updated using standard Metropolis–Hastings steps. Most results in this study were derived with PhyREX even though BEAST outperformed PhyREX in terms of speed of parameter inference (*SI Appendix*, section E). The two independent implementations of Bayesian samplers under the same models provide a robust validation of most results presented in this study.

Simulations.

Spatial Lambda-Fleming-Viot model. Genealogies and the accompanying spatial coordinates were first generated according to the "individual-based" SLFV model (29, 30). In this model, individuals give birth to descendants which locations are normally distributed. Death events are also governed by the same kernel so that the spatial density of the population is constant, on average, during the course of evolution. The normal kernel is truncated, allowing the SLFV model to accommodate habitats of finite size, as opposed to most continuous phylogeographic models. We selected the SLFV model as it describes the evolution of a population of related individuals along a spatial continuum as opposed to discrete demes. It is not subject to the shortcomings that hinder other popular spatial population genetics models such as sampling inconsistency (58) or Felsenstein's infamous "pain in the torus" (59). Finally and most importantly, because lineages' coordinates evolve here according to a jump process, the exact spatial coordinates of each lineage at each point in time can be monitored. This information may then serve as a basis to evaluate the total distance covered by all lineages in the genealogy. The ratio of this distance by the corresponding elapsed time gives an (average) speed that genuinely reflects the dispersal ability of the organisms under scrutiny.

50 individuals were sampled on a 10-by-10 square defining the habitat of the corresponding population. The rate of events where lineages die and/or give birth to descendants (the so-called REX events in ref. 60) was set to 10^3 events per unit of time per unit area and the variance of the normal density that defines the radius parameter in the SLFV model was chosen uniformly at random in $[0.1, 0.3]$. These parameter values are such that lineage jumps are short and frequent, thereby mimicking the behavior of a Brownian process (61).

Ebola-like simulations. Here, we used the agent-based spatially explicit simulator implemented in the R package nosoi (31). Parameters were chosen so as to mimic the Ebola epidemic in West Africa over a time period of 365 d, starting from a single infected host in Guéckédou (Guinea). nosoi is a discrete time, continuous space simulator that explicitly models within-host dynamics and between-host transmissions. It can exploit a geographic raster to simulate a full transmission tree where the geographic position of each infected host is tracked at all time. We simulated datasets using the same parameters as in ref. 31 which are informed by the literature describing human infections by Ebola. Spatial demographic data from WorldPop (www.worldpop.org) was also taken into account for these simulations.

Each host had a probability of 20% to move every day. These migrations were governed by a bivariate Gaussian distribution centered at the location of the lineage under scrutiny, with diagonal covariance matrix and equal SDs for longitude and latitude. The SD was set constant for each simulation, and drawn from a log-normal distribution with mean and SD equal to approximately 15 km in each direction. We used a raster of the entire West Africa, ensuring that no epidemic reached the border of the map within the time frame of the simulation.

As previously, we sampled 50 infected individuals randomly from the transmission tree, and extracted the sampled genealogy as well as the realized speed, that exploits the simulated position at each time of the chain. Note that the genealogies produced by these simulations are sampled through time and not ultrametric, making the estimation of speed easier.

Sequence simulation. In both simulation settings, edges in the obtained genealogy were rescaled so that the average length of an edge after scaling was 0.05 nucleotide substitutions per site. Nucleotide sequences were then generated under a strict clock model according to the HKY model of evolution (57) with transition/transversion ratio set to 4.0. 100 genealogies along with the corresponding spatial coordinates and homologous nucleotide sequences were generated this way for the SLFV and Ebola simulations.

Statistical inference. Each simulated dataset was processed using the RRW, IBM, and IOU models with independent coordinates. When considering their spatial components only, these models have 3, 2, and 6 parameters, respectively. The RRW model used a log-normal distribution of branch-specific dispersal rates, which is the standard parametrization for that model. The nucleotide substitution rate was set to its simulated value by taking the ratio of the tree length as expressed in molecular and calendar units. The tree-generating process was assumed to be Kingman's coalescent (62) with constant effective population size and a flat (improper) prior distribution on that parameter. Although sequences evolved according to a strict clock model, we used an uncorrelated relaxed clock model (25) with a log-normal distribution of edge-specific substitution rate multipliers. An exponential prior with rate set to 100 was used for the variance of this log-normal density.

For each dataset, the true average speed was taken as the actual Euclidean (SLFV) and great-circle (Ebola) distance covered by every lineage divided by the tree length in calendar time unit. For RRW, distances between the (observed or estimated) coordinates at each end of every branch in the tree were used to derive the dispersal rate through the "weighted lineage dispersal velocity" statistic (33). The posterior median of that statistic was used as our speed estimate. For PIV models, speed at the tree level was obtained by averaging the speed estimated at each node, the latter deriving from the corresponding velocities. Here again, we obtained the posterior distribution of the tree-level speed and use the median as our estimate. Note that none of the processes used for inference is the "true" process used for simulation, but simplified versions of it.

Predictive Phylogeography. The PIV models provide an adequate framework to estimate velocities at the tips of the inferred phylogenies. It thus makes sense to apply them to predicting dispersal patterns. Here, we designed a prediction technique which goal is to assess whether the organism under scrutiny may be found in a given region at a given point in time after the most recent sample was collected. Our approach utilizes the posterior distribution of the velocities estimates at each tip of the phylogeny to build a predictor. The latter is obtained by linear extrapolation of the estimated velocity at each tip in the tree that

assumes a constant speed of lineages after their sampling. Survival of these linear trajectories is taken into account so that older samples are less likely than recent ones to survive to a given time point in the future. This approach therefore puts more weight on recent samples to predict dispersal patterns (*SI Appendix*, section F). Incidence data used for comparison were extracted from https://www.cdc.gov/west-nile-virus/data-maps/historic-data.html.

Data, Materials, and Software Availability. The data and code to reproduce all analyses and figures displayed in this study are available at https://github.com/pbastide/integrated_phylogenetic_models. The PhyREX and BEAST programs are open source and freely available from https://github.com/stephaneguindon/phyml and https://github.com/beast-dev/beast-mcmc, respectively. All other data are included in the manuscript and/or *SI Appendix*. Previously published data were used for this work (32).

ACKNOWLEDGMENTS. S.G. thanks the Institut Français de Bioinformatique for computational resources. The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 725422–ReservoirDOCS) and the US NIH (R01 AI153044 and F31 AI154824). P.R.'s internship at the University of Montpellier was founded by the project "Montpellier Université d'Excellence" (I-SITE MUSE) through the Key Initiative "Data and Life Sciences." P.B. thanks Pierre Gloaguen for useful discussions on the integrated models. S.D. acknowledges support from the Fonds National de la Recherche Scientifique (Belgium; grant no. F.4515.22) and the Research Foundation–Flanders (Fonds voor Wetenschappelijk Onderzoek–Vlaanderen, Belgium; grant no. G098321N). P.L. acknowledges support by the Research Foundation–Flanders (Fonds voor Wetenschappelijk Onderzoek–Vlaanderen, G051322N and G005323N).

Author affiliations: [a]Institut Montpelliérain Alexander Grothendieck, Université de Montpellier, CNRS, Montpellier 34090, France; [b]Université Paris Cité, CNRS, Mathématiques appliquées 'a Paris 5, Paris F-75006, France; [c]Équipe Méthodes et Algorithmes pour la Bioinformatique, Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, CNRS—UMR 5506, Montpellier 34095, France; [d]Centre d'Ecologie Fonctionnelle et Evolutive, Université de Montpellier, CNRS, Ecole Pratique des Hautes Etudes, Institut de Recherche pour le Développement, Montpellier 34293, France; [e]Department of Economics, Sociology, and Statistics, RAND, Santa Monica, CA 90407-2138; [f]Maladies Infectieuses et Vecteurs : Ecologie, Génétique, Evolution et Contrôle, IRD, CNRS, Université de Montpellier, Montpellier 34394, France; [g]Plant Health Institute of Montpellier, IRD, Institut national de recherche pour l'agriculture, l'alimentation et l'environnement, Centre de coopération Internationale en Recherche Agronomique pour le Développement, Université de Montpellier, Montpellier 34394, France; [h]Department of Biostatistics, Jonathan and Karin Fielding School of Public Health, University of California, Los Angeles, CA 90095-1772; [i]Department of Human Genetics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, CA 90095; [j]Department of Computational Medicine, David Geffen School of Medicine at UCLA, University of California, Los Angeles, CA 90095-1766; [k]Spatial Epidemiology Lab, Université Libre de Bruxelles, Brussels B-1050, Belgium; and [l]Department of Microbiology, Immunology and Transplantation, Rega Institute, Laboratory for Clinical and Epidemiological Virology, Katholieke Universiteit Leuven, Leuven B-3000, Belgium

Author contributions: P.B. and S.G. designed research; P.B. and S.G. performed research; P.B., P.R., J.W., G.W.H., F.C., M.A.S., and S.G. contributed new reagents/analytic tools; P.B. and S.G. analyzed data; and P.B., D.F., S.D., P.L., and S.G. wrote the paper.

1. F. van den Bosch, R. Hengeveld, J. Metz, Analysing the velocity of animal range expansion. *J. Biogeogr.* **19**, 135–150 (1992).
2. C. Tisseuil *et al*., Evaluating methods to quantify spatial variation in the velocity of biological invasions. *Ecography* **39**, 409–418 (2016).
3. A. Drummond, O. Pybus, A. Rambaut, R. Forsberg, A. Rodrigo, Measurably evolving populations. *Trend. Ecol. Evol.* **18**, 481–488 (2003).
4. E. Zuckerkandl, L. Pauling, Molecules as documents of evolutionary history. *J. Theor. Biol.* **8**, 357–366 (1965).
5. S. Ho, B. Shapiro, Skyline-plot methods for estimating demographic history from nucleotide sequences. *Mol. Ecol. Res.* **11**, 423–434 (2011).
6. S. Wright, Isolation by distance. *Genetics* **28**, 114–138 (1943).
7. G. Malécot, *Mathematics of Heredity* (Masson et Cie., Paris, 1948).
8. P. Lemey, A. Rambaut, J. Welch, M. Suchard, Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* **27**, 1877–1885 (2010).
9. S. Dellicour *et al*., Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak. *Nat. Commun.* **9**, 1–9 (2018).
10. S. Issaka *et al*., Rivers and landscape ecology of a plant virus, Rice yellow mottle virus along the Niger Valley. *Virus Evol.* **7**, veab072 (2021).
11. O. Pybus *et al*., Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 15066–15071 (2012).
12. S. Dellicour *et al*., Using viral gene sequences to compare and explain the heterogeneous spatial dynamics of virus epidemics. *Mol. Biol. Evol.* **34**, 2563–2571 (2017).
13. N. Trovão *et al*., Host ecology determines the dispersal patterns of a plant virus. *Virus Evol.* **1**, vev016 (2015).
14. S. Dellicour *et al*., How fast are viruses spreading in the wild? bioRxiv [Preprint] (2024). https://doi.org/10.1101/2024.04.10.588821 (Accessed 24 May 2024).
15. W. Cumberland, C. Rohde, A multivariate model for growth of populations. *Theor. Popul. Biol.* **11**, 127–139 (1977).
16. O. Barndorff-Nielsen, N. Shephard, Integrated OU processes and non-Gaussian OU-based stochastic volatility models. *Scand. J. Stat.* **30**, 277–295 (2003).
17. J. Taylor, W. Cumberland, J. Sy, A stochastic model for analysis of longitudinal AIDS data. *J. Am. Stat. Assoc.* **89**, 727–736 (1994).
18. D. Johnson, J. London, M. A. Lea, J. Durban, Continuous-time correlated random walk model for animal telemetry data. *Ecology* **89**, 1208–1215 (2008).
19. M. Hooten, D. Johnson, Basis function models for animal movement. *J. Am. Stat. Assoc.* **112**, 578–589 (2017).
20. H. Preisler, A. Ager, M. Wisdom, Analyzing animal movement patterns using potential functions. *Ecosphere* **4**, 1–13 (2013).
21. J. Russell, E. Hanks, M. Haran, D. Hughes, A spatially varying stochastic differential equation model for animal movement. *Ann. Appl. Stat.* **12**, 1312–1331 (2018).
22. E. Holmes, What can we predict about viral evolution and emergence? *Curr. Opin. Virol.* **3**, 180–184 (2013).
23. M. Wille, J. Geoghegan, E. Holmes, How accurately can we assess zoonotic risk? *PLoS Biol.* **19**, e3001135 (2021).
24. E. C. Holmes, A. Rambaut, K. G. Andersen, Pandemics: Spend on surveillance, not prediction. *Nature* **558**, 180–182 (2018).
25. A. Drummond, S. Ho, M. Phillips, A. Rambaut, Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).
26. C. Gardiner, *Stochastic Methods* (Springer, Berlin, Heidelberg, ed. 4, 2009).
27. M. Gill, L. S. Tung Ho, G. Baele, P. Lemey, M. Suchard, A relaxed directional random walk model for phylogenetic trait evolution. *Syst. Biol.* **66**, 299–319 (2016).
28. P. Bastide, G. Didier, The Cauchy process on phylogenies: A tractable model for pulsed evolution. *Syst. Biol.* **72**, 1296–1315 (2023).
29. A. Etheridge, Drift, draft and structure: Some mathematical models of evolution. *Banach center Publ.* **80**, 121–144 (2008).
30. N. Barton, A. Etheridge, A. Véber, A new model for evolution in a spatial continuum. *Electron. J. Probab.* **15**, 162–216 (2010).
31. S. Lequime, P. Bastide, S. Dellicour, P. Lemey, G. Baele, nosoi: A stochastic agent-based transmission chain simulation framework in R. *Methods Ecol. Evol.* **11**, 1002–1007 (2020).
32. S. Dellicour *et al*., Epidemiological hypothesis testing using a phylogeographic and phylodynamic framework. *Nat. Commun.* **11**, 5620 (2020).
33. S. Dellicour, R. Rose, N. R. Faria, P. Lemey, O. G. Pybus, SERAPHIM: Studying environmental rasters and phylogenetically informed movements. *Bioinformatics* **32**, 3204–3206 (2016).
34. J. Cornuault, Bayesian analyses of comparative data with the Ornstein-Uhlenbeck model: Potential pitfalls. *Syst. Biol.* **71**, 1524–1540 (2022).
35. R. Lanciotti *et al*., Origin of the West Nile virus responsible for an outbreak of encephalitis in the northeastern United States. *Science* **286**, 2333–2337 (1999).
36. G. Campbell, A. Marfin, R. Lanciotti, D. Gubler, West Nile virus. *Lancet Infect. Dis.* **2**, 519–529 (2002).
37. G. Hassler *et al*., Principled, practical, flexible, fast: A new approach to phylogenetic factor analysis. *Methods Ecol. Evol.* **13**, 2181–2197 (2022).
38. F. Chevenet, D. Fargette, P. Bastide, T. Vitré, S. Guindon, EvoLaps 2: Advanced phylogeographic visualization. *Virus Evol.* **10**, vead078 (2024).
39. S. Dellicour, B. Vrancken, N. Trovão, D. Fargette, P. Lemey, On the importance of negative controls in viral landscape phylogeography. *Virus Evol.* **4**, vey023 (2018).
40. J. Raghwani *et al*., Endemic dengue associated with the co-circulation of multiple viral lineages and localized density-dependent transmission. *PLoS Pathog.* **7**, e1002064 (2011).
41. C. Talbi *et al*., Phylodynamics and human-mediated dispersal of a zoonotic virus. *PLoS Pathog.* **6**, e1001166 (2010).
42. R. Klitting *et al*., Predicting the evolution of the Lassa virus endemic area and population at risk over the next decades. *Nat. Commun.* **13**, 5596 (2022).
43. M. Rakotomalala *et al*., Comparing patterns and scales of plant virus phylogeography: Rice yellow mottle virus in Madagascar and in continental Africa. *Virus Evol.* **5**, vez023 (2019).
44. G. Baele, S. Dellicour, M. Suchard, P. Lemey, B. Vrancken, Recent advances in computational phylodynamics. *Curr. Opin. Virol.* **31**, 24–32 (2018).
45. R. FitzJohn, Quantitative traits and diversification. *Syst. Biol.* **59**, 619–633 (2010).
46. N. Müller, D. Rasmussen, T. Stadler, The structured coalescent and its approximations. *Mol. Biol. Evol.* **34**, 2970–2981 (2017).
47. J. Drury, J. Clavel, M. Manceau, H. Morlon, Estimating the effect of competition on trait evolution using maximum likelihood inference. *Syst. Biol.* **65**, 700–710 (2016).
48. M. Manceau, A. Lambert, H. Morlon, A unifying comparative phylogenetic framework including traits coevolving across interacting lineages. *Syst. Biol.* **66**, 551–568 (2017).
49. K. Bartoszek, S. Glémin, I. Kaj, M. Lascoux, Using the Ornstein-Uhlenbeck process to model the evolution of interacting populations. *J. Theor. Biol.* **429**, 35–45 (2017).
50. P. Lemey *et al*., Unifying viral genetics and human transportation data to predict the global transmission dynamics of Human Influenza H3N2. *PLoS Pathog.* **10**, e1003932 (2014).
51. S. Dellicour *et al*., Incorporating heterogeneous sampling probabilities in continuous phylogeographic inference-application to H5N1 spread in the Mekong region. *Bioinformatics* **36**, 2098–2104 (2020).
52. A. Kalkauskas *et al*., Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk. *PLoS Comput. Biol.* **17**, e1008561 (2021).

53. S. Guindon, N. De Maio, Accounting for spatial sampling patterns in Bayesian phylogeography. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2105273118 (2021).
54. M. Suchard *et al*., Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).
55. V. Mitov, K. Bartoszek, G. Asimomitis, T. Stadler, Fast likelihood calculation for multivariate Gaussian phylogenetic models with shifts. *Theor. Popul. Biol.* **131**, 66–78 (2020).
56. P. Bastide, L. S. T. Ho, G. Baele, P. Lemey, M. Suchard, Efficient Bayesian inference of general Gaussian models on large phylogenetic trees. *Ann. Appl. Stat.* **15**, 971–997 (2021).
57. M. Hasegawa, H. Kishino, T. Yano, Dating of the Human-Ape splitting by a molecular clock of mitochondrial-DNA. *J. Mol. Evol.* **22**, 160–174 (1985).
58. N. Barton, A. Etheridge, A. Véber, Modelling evolution in a spatial continuum. *J. Stat. Mech. Theory Exp.* **2013**, P01002 (2013).
59. J. Felsenstein, A pain in the torus: Some difficulties with models of isolation by distance. *Am. Nat.* **109**, 359–368 (1975).
60. S. Guindon, H. Guo, D. Welch, Demographic inference under the coalescent in a spatial continuum. *Theor. Popul. Biol.* **111**, 43–50 (2016).
61. J. Wirtz, S. Guindon, On the connections between the spatial Lambda-Fleming-Viot model and other processes for analysing geo-referenced genetic data. *Theor. Popul. Biol.* **158**, 139–149 (2023).
62. J. Kingman, The coalescent. *Stoch. Process. Their Appl.* **13**, 235–248 (1982).