Phylogeographic and phylodynamic approaches to epidemiological hypothesis testing

Simon Dellicour[1,2,*], Sebastian Lequime[2], Bram Vrancken[2], Mandev S. Gill[2], Paul Bastide[2], Karthik Gangavarapu[3], Nate Matteson[3], Yi Tan[4,5], Louis du Plessis[6], Alexander A. Fisher[7], Martha I. Nelson[8], Marius Gilbert[1], Marc A. Suchard[7,9,10], Nathan D. Grubaugh[11], Kristian G. Andersen[3,12], Oliver G. Pybus[6], and Philippe Lemey[2]

[1] Spatial Epidemiology Lab (SpELL), Université Libre de Bruxelles, CP160/12 50, av. FD Roosevelt, 1050 Bruxelles, Belgium.

[2] Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium.

[3] Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA.

[4] Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA.

[5] Infectious Diseases Group, J. Craig Venter Institute, Rockville, Maryland, USA.

[6] Department of Zoology, University of Oxford, Oxford, UK.

[7] Department of Biomathematics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA.

[8] Fogarty International Center, National Institutes of Health, Bethesda, MD 20894, USA.

[9] Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, CA, USA.

[10] Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA.

[11] Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT 06510, USA.

[12] Scripps Research Translational Institute, La Jolla, CA 92037, USA.

∗ Corresponding author (simon.dellicour@ulb.ac.be)

Computational analyses of pathogen genomes are increasingly being used to unravel the dispersal history and transmission dynamics of epidemics. Here, we show how to go beyond historical reconstructions and use spatially-explicit phylogeographic and phylodynamic approaches to formally test epidemiological hypotheses. We focus on the spread and invasion of West Nile virus spread in North America that has been responsible for substantial impacts on public, veterinary and wildlife health. WNV isolates have been sampled at various times and locations across North America since its introduction to New York twenty years ago. We exploit this genetic data repository to demonstrate that factors hypothesised to affect viral dispersal and demography can be statistically tested. We find that WNV lineages tend to disperse faster in areas with higher temperatures and we identify temporal variation in temperature as a main predictor of viral genetic diversity through time. Finally, we compare inferred and simulated dispersal histories of lineages in order to assess the impact of migratory bird flyways on the rapid east-to-west continental spread of WNV. We find no evidence that viral lineages preferentially circulate within the same migratory flyway, suggesting a substantial role for non-migratory birds or mosquito dispersal along the longitudinal gradient. Our study demonstrates that the development and application of statistical approaches, coupled with comprehensive pathogen genomic data, can address epidemiological questions that might otherwise be difficult or unacceptably costly to answer.

Keywords: molecular epidemiology, landscape phylogeography, phylodynamic, environmental factors, West Nile virus.

The evolutionary analysis of rapidly evolving pathogens, particularly RNA viruses, allows us to establish the epidemiological relatedness of cases through time and space. Such transmission information can be difficult to detect using classical epidemiological approaches. The development of spatially-explicit phylogeographic models[1,2], which place time-referenced phylogenies in a geographical context, can provide a detailed spatiotemporal picture of the dispersal history of virus lineages[3]. Recent advances in methodology have moved beyond simple reconstructions of epidemic history and instead attempt to analyse the impact of underlying factors on the dispersal dynamics of virus lineages[4–6], giving rise to the concept of landscape phylogeography[7].

Similar improvements have been made to phylodynamic analyses that use flexible coalescent models to reconstruct virus demographic history[8,9]; these methods can now provide insights into epidemiological or environmental variables that might be associated with population size change[10]. In this study we aim to go beyond historical reconstructions and formally test epidemiological hypotheses by exploiting phylodynamic and spatially-explicit phylogeographic models. We illustrate our approach by examining the spread of West Nile virus (WNV) across North America, an emergent virus lineage that is responsible for substantial impacts on public, veterinary, and wildlife health[11].

WNV is the most widely-distributed encephalitic flavivirus transmitted by the bite of infected mosquitoes[12,13]. WNV is a single-stranded RNA virus that is maintained by an enzootic transmission cycle involving birds and mosquitoes that mainly belong to the *Culex* genus[14–17]. Humans are incidental terminal hosts, because viremia does not reach a sufficient level for subsequent transmission to mosquitoes[16,18]. WNV human infections are mostly subclinical although symptoms may range from fever to meningoencephalitis and can occasionally lead to death[16,19]. The WNV epidemic in North America likely resulted from a single introduction to the continent twenty years ago[20]. WNV persistence in North America is not the result of successive reintroductions to the territory, but rather of local overwintering and maintenance of long-term avian and/or mosquito transmission cycles[11]. Overwintering could also be facilitated by vertical transmission of WNV from infected female mosquitos to their offspring[21,22]. WNV represents one of the most important vector-borne diseases in North America[15]; there were an estimated 50,720 human WNV cases between 1999 to 2018, leading to 2,300 deaths (www.cdc.gov/westnile). In addition, WNV has had a notable impact on North American bird populations[23], with several species[24] such as the American crow (*Corvus brachyrhynchos*) and the blue jay (*Cyanocitta cristata*) being particularly severely affected.

Since the beginning of the epidemic in North America in 1999[20], WNV has received considerable attention from local and national health institutions and the scientific community. This had led to the sequencing of more than 2,000 complete viral genomes collected at various times and locations across the continent. The resulting availability of virus genetic data represents a unique opportunity to better understand the evolutionary history of WNV invasion into an

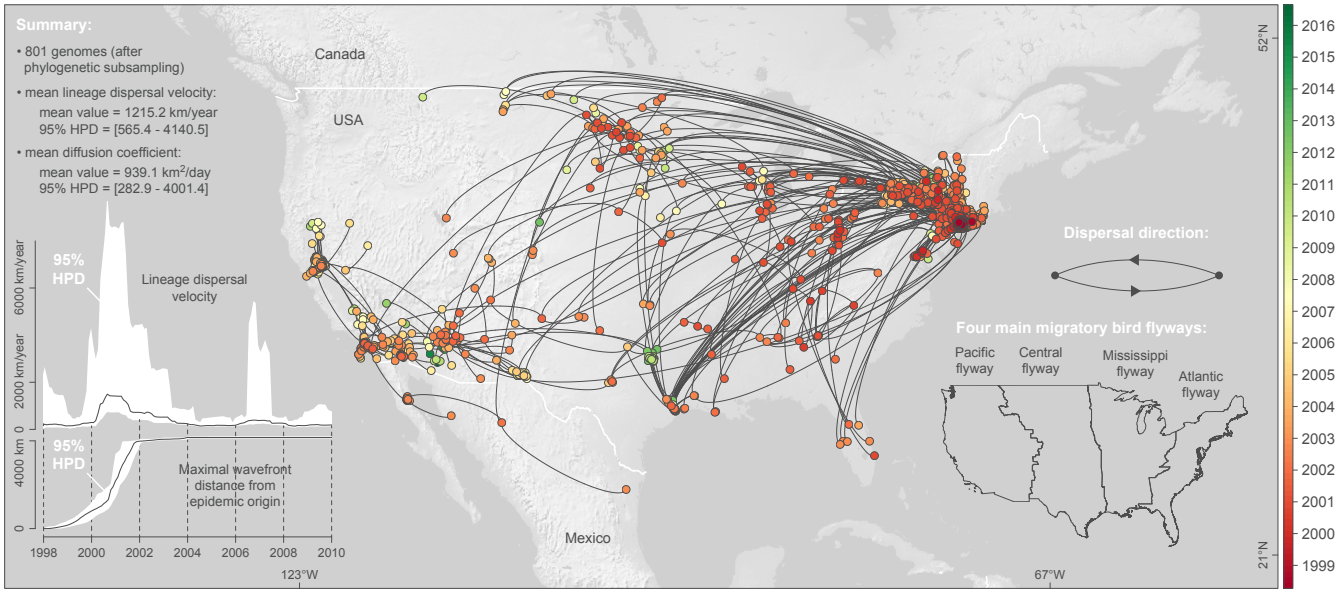

Figure 1. Spatiotemporal diffusion of WNV lineages in North America. Maximum clade credibility (MCC) tree obtained by continuous phylogeographic inference based on 100 posterior trees (see the text for further details). Nodes of the tree are coloured from red (the time to the most recent common ancestor, TMRCA) to green (most recent sampling time). Older nodes are plotted on top of younger nodes, but we provide also an alternative year-by-year representation in Figure S1. In addition, this figure reports global dispersal statistics (mean lineage dispersal velocity and mean diffusion coefficient) averaged over the entire virus spread, the evolution of the mean lineage dispersal velocity through time, the evolution maximal wavefront distance from the origin of the epidemic, as well as the delimitations of the North American Migratory Flyways (NAMF) considered in the USA.

originally non-endemic area. Here, we analyse a comprehensive data set of WNV genomes with the objective of unveiling the dispersal and demographic dynamics of the virus in North America. Specifically, we aim to (i) reconstruct the dispersal history of WNV on the continent, (ii) test the impact of environmental factors on the dispersal velocity of lineages, (iii) test the impact of migratory bird flyways on the dispersal history, and (iv) test the association between external covariates and WNV genetic diversity through time.

RESULTS

Reconstruction of WNV dispersal history

To infer the dispersal history of WNV lineages in North America, we performed a spatially-explicit phylogeographic analysis[1] of 801 viral genomes (Fig. 1), which is almost an order of magnitude larger than the early US-wide study by Pybus *et al.*[2] (104 WNV genomes). Year-by-year visualisation of the reconstructed invasion history highlighted frequent long-distance dispersal events across the continent (Fig. S1). To quantify the spatial dissemination of virus lineages, we extracted the spatio-temporal information embedded in molecular clock phylogenies sampled by Bayesian phylogeographic analysis. From the resulting collection of lineage movement vectors, we estimated several key statistics of spatial dynamics (Fig. 1). We estimated a mean lineage dispersal velocity of ~1,200 km/year and a mean diffusion coefficient of ~940 km^2/day, consistent with previous estimates[2]. We further inferred how the mean lineage dispersal velocity changed through time, and found that dispersal velocity was notably higher in the earlier years of the epidemic (Fig. 1). The early peak of lineage dispersal velocity around 2001 corresponds to the expansion phase of the epidemic. This is corroborated by our estimate of the maximal wavefront distance from the epidemic origin through time (Fig. 1). This expansion phase lasted until the end of 2001, when WNV lineages first reached the west coast (Figs. 1-S1).

Testing the impact of environmental factors on dispersal velocity

We employed two different statistical tests based on landscape phylogeography to investigate the impact of environmental factors on the dispersal dynamics of WNV. First, we analysed whether the heterogeneity observed in lineage dispersal velocity could be explained by specific environmental factors. For this purpose, we used a computational method that assesses the correlation between lineage dispersal durations and environmentally-scaled distances[4,25]. These distances were computed on rasters (geo-referenced grids) that summarise the different environmental factors to be tested (Fig. 2): elevation, land cover in the study area (forests, shrublands, savannas, grasslands, croplands, water areas), annual mean temperature, and annual precipitation. This analysis aimed to quantify the impact of each factor of virus movement by calculating a statistic, Q, that measures the correlation between lineage durations and environmentally-scaled distances. Specifically, the Q statistic describes difference in strength of the correlation when distances are scaled using the environmental raster versus when they are computed using a "null" raster (i.e. a uniform raster with a value of "1" assigned to all cells). As detailed in the Methods section, two alternative path models were used to compute these environmentally-scaled distances: the least-cost path model[26] and a model based on circuit theory[27]. The Q statistic was estimated for each posterior tree sampled during the phylogeographic analysis, yielding a posterior distribution of this metric. Finally, statistical support for Q was obtained by comparing inferred and simulated distributions of Q; the latter was obtained by estimating Q on the same set of tree topologies, along which a new stochastic diffusion history was simulated. This simulation procedure thereby generated a null model of dispersal, and the comparison between the inferred and simulated Q distributions enabled us to approximate a Bayes factor support (see Methods for further details).

As summarised in Table S1, we found strong support for one variable: annual temperature raster treated as a conductance factor. Using this factor, the association between lineage duration and environmentally-scaled distances was significant using the path model based on circuit theory[27]. As detailed in Figure 3, this environmental variable better explained the heterogeneity in lineage dispersal velocity than geographic distance alone (i.e. its Q distribution was positive). Furthermore, this result received strong statistical support (Bayes factor $>$20), obtained by comparing the distribution of Q values with that obtained under a null model (Fig. 3).

Testing the impact of migratory bird flyways on dispersal history

The second landscape phylogeography test we performed focused on the impact of migratory bird flyways on the dispersal history of WNV. For this purpose, we first tested whether virus lineages tended to remain within the same North American Migratory Flyway (NAMF; Fig. 1). As with the testing approach used in the section above, we again compared inferred and simulated trees (i.e. simulation of a

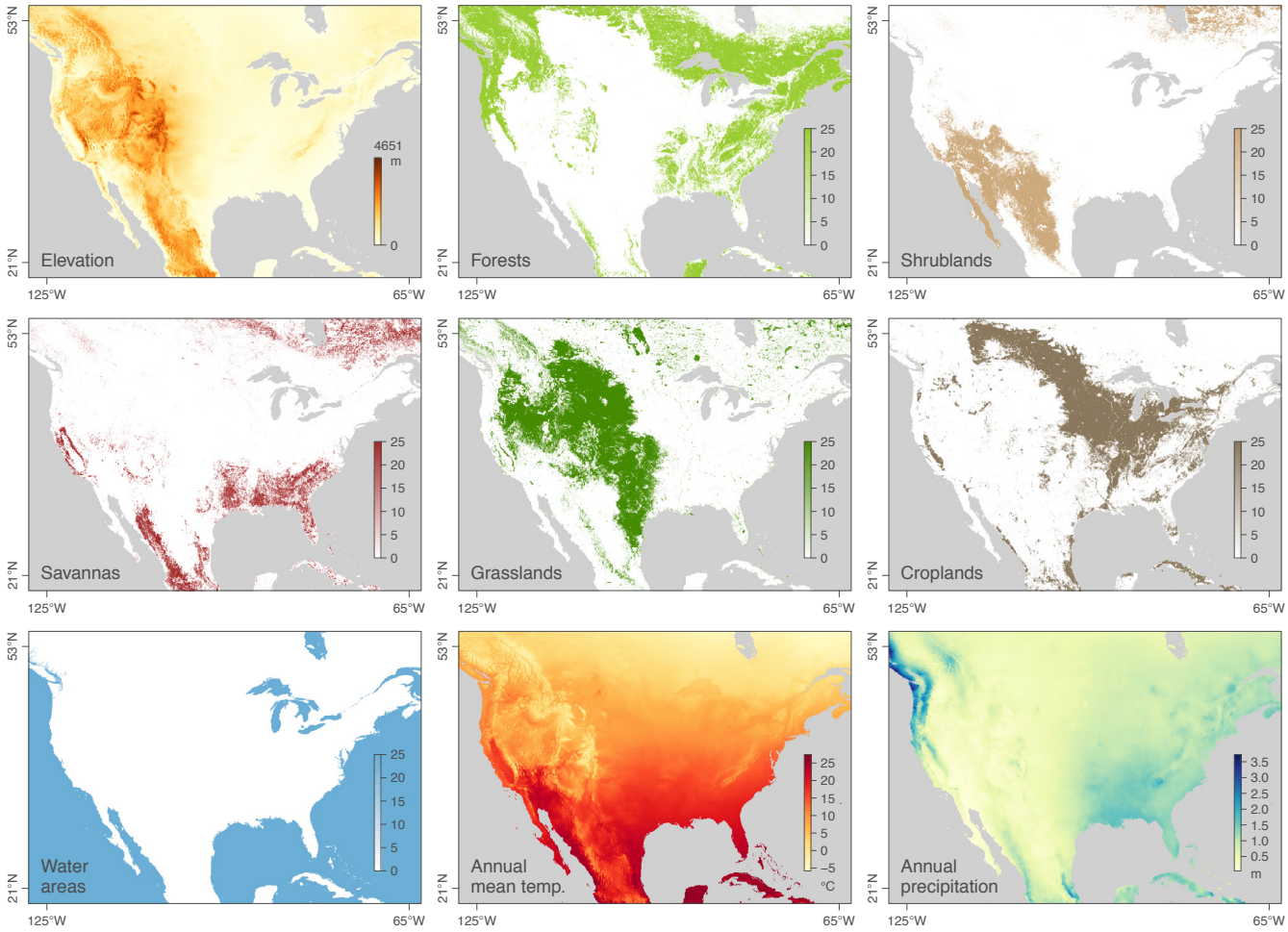

Figure 2. Environmental variables tested for their impact on the dispersal of West Nile virus lineages in North America.

new stochastic diffusion process along the estimated trees). Under the null hypothesis (i.e. NAMFs have no impact on WNV dispersal history), virus lineages should not transition between flyways less often than under the null dispersal model. Our test did not reject this null hypothesis (BF <1). Because the NAMF borders are based on administrative areas (US counties), we also performed a similar test using the alternative delimitation of migratory bird flyways estimated for terrestrial bird species by La Sorte *et al.*[28] (Fig. S2). Again, the null hypothesis was not rejected, confirming that inferred virus lineages did not tend to remain within specific flyways more often than expected by chance.

Testing the impact of environmental factors on the viral diversity through time

We next employed a phylodynamic approach to investigate predictors of the dynamics of viral genetic diversity through time. In particular, we used the generalised linear model (GLM) extension[10] of the skygrid coalescent model[9], hereafter referred to as the "skygrid-GLM" approach, to statistically test for associations between estimated dynamics of virus effective population size and several covariates. Coalescent models that estimate effective population size (Ne) typically assume a single panmictic population that encompasses all individuals. Because this assumption is frequently violated in practice, the estimated effective population size is sometimes interpreted as representing an estimate of the genetic diversity of the whole virus population[29]. The skygrid-GLM approach accounts for uncertainty in effective population size estimates when testing for associations with covariates; neglecting this uncertainty can lead to spurious conclusions[10].

We first performed univariate skygrid-GLM analyses of five distinct time-varying covariates: human WNV case counts (log-transformed), temperature, precipitation, a greenness index, and a bird observation index. For the human case count covariate, we modelled a possible lag period of one or two months, because a rise or reduction in cases may be observed potentially some time after the corresponding change in virus infections of mosquito vectors and bird hosts (e.g. due to incubation periods in hosts, vectors, and humans, and/or to the time needed to detect and report human cases). We obtained evidence for a significant lagged association between the time series of human cases and virus effective population size and the covariate (the former lagged the latter by two months). Therefore, in subsequent analyses we considered only the number of human cases advanced by two months (see below).

In addition, univariate analyses of temperature and precipitation time series were also associated with the virus genetic diversity dynamics (i.e. the posterior GLM coefficients for these covariates had 95% credible intervals that did not include zero; Fig. 4). To further assess the relative importance of each covariate, we performed multivariate skygrid-GLM analyses to rank covariates based on their inclusion probabilities[30]. The first multivariate analysis involved all covariates and suggested that the lagged human case counts best explain viral population size dynamics, with an inclusion probability close to 1. However, because human case counts are known to be a consequence rather than a potential causal driver of the WNV epidemic, we performed a second multivariate analysis after having excluded this covariate. This time, the temperature time series emerged as a covariate with the highest inclusion probability.

DISCUSSION

The spread of WNV in North America can be divided into an initial "invasion phase" and a subsequent "maintenance phase" (see Carrington *et al.*[31] for similar terminology used in the context of spatial invasion of dengue viruses). The invasion phase is characterised by an increase in virus effective population size until the west coast was reached, followed by a maintenance phase associated with a more

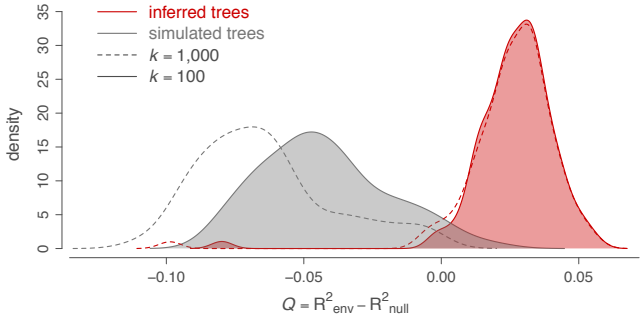

Figure 3. Impact of annual mean temperature acting as a conductance factor on lineage dispersal velocity. The graph displays the distribution of the correlation metric Q computed on 100 spatially-annotated trees obtained by continuous phylogeographic inference (red distributions). The metric Q measures to what extent considering a heterogeneous environmental raster, increases the correlation between lineage durations and environmentally-scaled distances compared to a homogeneous raster. If Q is positive and supported, it indicates that the heterogeneity in lineage dispersal velocity can be at least partially explained by the environmental factor under investigation. The graph also displays the distribution of Q values computed on the same 100 posterior trees along which we simulated a new forward-in-time diffusion process (grey distributions). These simulations are used as a null dispersal model to estimate the support associated with the inferred distribution of Q values. For both inferred and simulated trees, we report the Q distributions obtained while transforming the original environmental raster according to two different scaling parameter k values (100 and 1,000; respectively full and dashed line, see the text for further details on this transformation). The annual mean temperature raster, transformed in conductance values using these two k values, is the only environmental factor for which we detect a positive distribution of Q that is also associated with a strong statistical support (Bayes factor >20).

stable cyclic variation of effective population size (Fig. 4). In only 2-3 years, WNV rapidly spread from the east to the west coast of North America, despite the fact that the migratory flyways of its avian hosts are primarily north-south directed. This suggests potentially-important roles for non-migratory bird movements, as well as natural or human-mediated mosquito dispersal, in spreading WNV along a longitudinal gradient[32,33]. Furthermore, we uncover a higher lineage dispersal velocity during the invasion phase, which could reflect a consequence of increased bird immunity through time slowing down spatial dispersal. It has indeed been demonstrated that avian immunity can impact WNV transmission dynamics[34].

Here we formally test the hypothesis that WNV lineages are contained or preferentially circulate within the same migratory flyway; we find no statistical support for this hypothesis. This result contrasts with previously-reported phylogenetic clustering by flyways[35]. However, the clustering analysis of Di Giallonardo *et al.*[35] was based on a discrete phylogeographic analysis and, as recognised by the authors, it is difficult to distinguish the effect of these flyways from those of geographic distance. Here, we circumvent this issue by performing a spatial analysis that explicitly represents dispersal as a function of geographic distance. Although we cannot entirely exclude the possibility that WNV moved via rapid and successive north-south bird migrations, our phylogeographic analysis highlights the occurrence of several fast and long-distance dispersal events along a longitudinal gradient. A potential anthropogenic contribution to such long-distance dispersal (e.g. through commercial transport) warrants further investigation.

The WNV epidemic in North America is a powerful illustration of viral invasion and emergence in a new environment[35]. Our analyses find evidence for the impact of only one environmental factor on virus lineage dispersal velocity, namely annual mean temperature. The relevance of temperature is further demonstrated by the association between the virus genetic dynamics and several time-dependent covariates. Indeed, among the four environmental time-series we tested, temporal variation in temperature is the most important predictor of cycles in viral genetic diversity. Temperature is known to have a dramatic impact on the biology of arboviruses and their arthropod hosts[36], including WNV. Higher temperatures have been shown to impact directly the mosquito life cycle, by accelerating larval development[11], decreasing the interval between blood meals, and prolonging mosquito breeding season[37]. Higher temperatures have been also associated with shorter extrinsic incubation periods, accelerating WNV transmission by the mosquito vector[38,39]. Interestingly, temperature has also been suggested as a variable that can increase the predictive power of WNV forecast models[40]. The impact of temperature that we reveal here on both dispersal velocity and viral genetic diversity is particularly important in the context of global warming. In addition to altering mosquito species distribution[41,42], an overall temperature increase in North America could imply increased enzootic transmission and hence increased spill-over risk in different regions.

In addition to temperature, we find evidence for an association between viral genetic diversity dynamics and the number of human cases, but only when a lag period of two months is added to the model. To our knowledge, this represents the first empirical evidence for a lag period between viral genetic diversity and human case counts for WNV. This result is in line with a previously reported lag for WNV between the basic reproduction number (R0) and incidence of human cases[43]. Such lags can, at least in part, be explained by the time needed for mosquitos to become infectious, bite humans and subsequently for symptoms to be detected and reported[43,44]. This result also implies that monitoring viral genetic diversity in mosquitoes/birds may have predictive power, but this would require sufficiently fast sampling and real-time sequencing[45,46]. Finally, our evidence for a lag period between viral genetic diversity and reported cases will hopefully motivate further developments of the skygrid-GLM approach that would enable co-estimating a lag period rather than testing a set of pre-specified values.

Our study shows the utility of landscape phylogeographic and phylodynamic hypothesis tests when applied to a comprehensive data set of viral genomes sampled during an epidemic. Such spatially-explicit investigations are possible when only viral genomes (whether recently collected or available on public databases such as GenBank) are associated with sufficiently precise metadata, in particular the collection date and the sampling location. The availability of precise collection dates - ideally known to the day - for isolates obtained over a sufficiently long time-span enables reliable timing of epidemic events due to the accurate calibration of molecular clock models. Further, spatially-explicit phylogeographic inference is possible only when viral genomes are associated with sampling coordinates. However, geographic coordinates are frequently unknown or unreported. In practice this may not represent a limitation if a sufficiently precise descriptive sampling location is specified (e.g. a district or administrative area), as this information can be converted into geographic coordinates. The full benefits of comprehensive phylogeographic analyses of viral epidemics will be realised only when precise location and time metadata are made systematically available.

Although we use a comprehensive collection of WNV genomes in this study, it would be useful to perform analyses based on even larger data sets that cover regions under-sampled in the current study; this work is the focus of an ongoing collaborative project (westnile4k.org). While the resolution of phylogeographic analyses will always depend on the spatial granularity of available samples, they can still be powerful in elucidating the dispersal history of sampled lineages. Furthermore, when testing the impact of environmental factors on virus dispersal history, heterogeneous sampling density will primarily affect statistical power in detecting the impact of relevant environmental factors in under- or unsampled areas[25]. In this study, we note that heterogeneous sampling density across counties can be at least partially mitigated by performing phylogenetic subsampling (detailed in the Methods section).

By placing virus lineages in a spatio-temporal context, phylogeographic inference provides information on the linkage of infections through space and time. Mapping lineage dispersal can provide a valuable source of information for epidemiological investigations and can shed light on the ecological and environmental processes that have

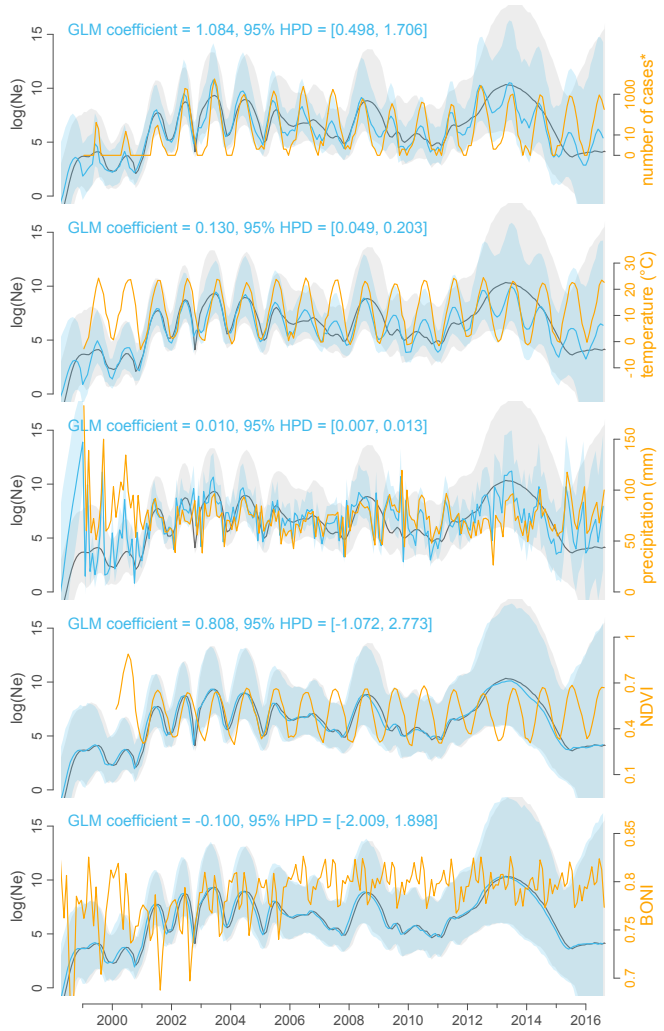

Figure 4. Associations between viral effective population size and potential covariates. These associations were tested with a generalised linear model (GLM) extension of the coalescent model used to infer the dynamics of the viral effective population size of the virus (Ne) through time. Specifically, we here tested the following time-series variables as potential covariates (orange curves): number of human cases (log-transformed and with a negative time period of two months), mean temperature, mean precipitation, Normalised Difference Vegetation Index (NDVI, a greenness index), and Birds Observation Normalised Index (BONI; see the text for further details). Posterior mean estimates of the viral effective population size based on both sequence data and covariate data are represented by blue curves, and the corresponding blue polygon reflects the 95% HPD region. Posterior mean estimates of the viral effective population size inferred strictly from sequence data are represented by grey curves and the corresponding grey polygon reflects the 95% HPD region. A significant association between the covariate and effective population size is inferred when the 95% HPD interval of the GLM coefficient excludes zero, which is the case for the case count, temperature, and precipitation covariates.

impacted the epidemic dispersal history and transmission dynamics. When complemented with phylodynamic testing approach, such as the skygrid-GLM approach used here, these methods offer new opportunities for epidemiological hypotheses testing. These tests can complement traditional epidemiological approaches that employ occurrence data. If coupled to real-time virus genome sequencing, landscape phylogeographic and phylodynamic testing approaches have the potential to inform epidemic control and surveillance decisions[47].

METHODS

Selection of viral sequences. We started by gathering all WNV sequences available on GenBank on the 20th November 2017. We then only selected sequences (i) of at least 10 kb, i.e. covering almost the entire viral genome (~11 kb), and (ii) associated with a sufficiently precise sampling location, i.e. at least an administrative area of level 2. Administrative areas of level 2 are hereafter abbreviated "admin-2" and corresponds to US counties. Finding the most precise sampling location (admin-2, city, village, or geographic coordinates), as well as the most precise sampling date available for each sequence, required a bibliographic screening because such metadata are often missing on GenBank. The resulting alignment of geo-referenced sequences of 993 genomic sequences of at least 10 kb was made using MAFFT[48] and manually edited in AliView[49]. Based on this alignment, we performed a first phylogenetic analysis using the maximum likelihood method implemented in the program FastTree[50] with 1,000 bootstrap replicates to assess branch supports. The aim of this preliminary phylogenetic inference was solely to identify monophyletic clades of sequences sampled from the same admin-2 area associated with a bootstrap support higher than 70%. Such clusters of sampled sequences largely represent lineage dispersal within a specific admin-2 area. Because we randomly draw geographic coordinates from an admin-2 polygon for sequences only associated with an admin-2 area of origin, keeping more than one sequence per cluster would not contribute any meaningful information in subsequent phylogeographic analyses[47]. Therefore, we subsampled the original alignment such that only one sequence is randomly selected per admin-2 location-specific cluster, leading to a final alignment of 801 genomic sequences.

Time-scaled phylogenetic analysis. Time-scaled phylogenetic trees were inferred using BEAST 1.10.4[51] and the BEAGLE 3 library[52] to improve computational performance. The substitution process was modelled according to a GTR+Γ parametrisation[53], branch-specific evolutionary rates were modelled according to a relaxed molecular clock with an underlying log-normal distribution[54], and the flexible skygrid model was specified as tree prior[9,10]. We ran and eventually combined ten independent analyses, sampling Markov chain Monte-Carlo (MCMC) chains every 2 x 10^8 generations. Combined, the different analyses were run for more than 10^{12} generations. For each distinct analysis, the number of sampled trees to discard as burn-in was identified using Tracer 1.7[55]. We used Tracer to inspect the convergence and mixing properties of the combined output, referred to as the "skygrid analysis" throughout the text, to ensure that estimated sampling size (ESS) values associated with estimated parameters were all >200.

Spatially-explicit phylogeographic analysis. The spatially-explicit phylogeographic analysis was performed using the relaxed random walk (RRW) diffusion model implemented in BEAST[1,2]. This model allows the inference of spatially- and temporally-referenced phylogenies while accommodating variation in dispersal velocity among branches[3]. Following Pybus *et al.*[2], we used a gamma distribution to model the among-branch heterogeneity in diffusion velocity. Even when launching multiple analyses and using GPU resources to speed-up the analyses, poor MCMC mixing did not permit reaching an adequate sample from the posterior in a reasonable amount of time. This represents a challenging problem that is currently under further investigation[56]. To circumvent this issue, we performed 100 independent phylogeographic analyses each based on a distinct fixed tree sampled from the posterior distribution of the skygrid analysis. We ran each analysis until ESS values associated with estimated parameters were all greater than 100. We then extracted the last spatially-annotated tree sampled in each of the 100 posterior distributions, which is the equivalent of randomly sampling a post-burn-in tree within each distribution. All the subsequent landscape phylogeographic testing approaches were based the resulting distribution of the 100 spatially-annotated trees. Given the computational limitations, we argue that the collection of 100 spatially-annotated trees, extracted from distinct posterior distributions each based on a different fixed tree topology, represents a reasonable approach to obtain a phylogeographic reconstruction that accounts for phylogenetic uncertainty. We note that this is similar to the approach of using a set of empirical trees that is frequently employed for discrete phylogeographic inference[57,58], but direct integration over such a set of trees is not appropriate for the RRW model because the proposal distribution for branch-specific scal-

ing factors does not hold in this case. We used TreeAnnotator 1.10.4[51] to obtain the maximum clade credibility (MCC) tree representation of the spatially-explicit phylogeographic reconstruction.

Phylogenetic branches, or "lineages", from spatially- and temporally-referenced trees can actually be treated as conditionally independent movement vectors[2]. We used the R package "seraphim"[4,59] to extract the spatio-temporal information embedded within each tree and to summarise lineages as movement vectors. We further used the package "seraphim" to estimate two dispersal statistics based on the collection of such vectors: the mean lineage dispersal velocity and the mean diffusion coefficient[2]. We also estimated the evolution of the maximal wavefront distance from the epidemic origin, as well as the evolution of the mean lineage dispersal velocity through time.

Generating a null dispersal model. To generate a null dispersal model we simulated a forward-in-time RRW diffusion process along each tree topology used for the phylogeographic analyses. These RRW simulations were performed with the "simulatorRRW1" function of the R package "seraphim" and based on the sampled precision matrix parameters estimated by the phylogeographic analyses[47]. For each tree, the RRW simulation started from the root node location inferred by the phylogeographic analysis. Furthermore, these simulations were constrained such that the simulated node locations remain within the study area, which is here defined by the minimum convex hull built around all node positions, minus non-accessible sea areas. As for the annotated trees obtained by phylogeographic inference, hereafter referred to as "inferred trees", we extracted the spatio-temporal information embedded within their simulated counterparts, hereafter referred as "simulated trees". Because RRW diffusion processes were simulated along fixed tree topologies, each simulated tree shares a common topology with an inferred tree. Such a pair of inferred and simulated trees thus only differs by the geographic coordinates associated with their nodes, except for the root node position that was fixed as starting points for the RRW simulation. The distribution of 100 simulated trees served as a null dispersal model for the landscape phylogeographic testing approaches described below.

Testing the impact of environmental factors on dispersal velocity. The first landscape phylogeographic testing approach aimed to test the association between environmental factors and the dispersal velocity of WNV lineages in North America. We tested several environmental rasters both as potential conductance factors (i.e. facilitating movement) or resistance factors (i.e. impeding movement). Each environmental factor was described by a raster that defines its spatial heterogeneity (Fig. 2). The original rasters presented a resolution of 0.5 arcmin, corresponding to cells ~1 km^2 (see Table S2 for the source of each original raster file). Starting from the original categorical land cover raster, we generated distinct land cover rasters by creating lower resolution rasters (10 arcmin) whose cell values equalled the number of occurrences of each land cover category within the 10 arcmin cells. The resolution of the other three other original rasters was also decreased to 10 arcmin for tractability. For each environmental factor, several distinct rasters were generated by transforming the original raster cell values with the following formula: $v_t = 1 + k(v_o/v_{max})$, where v_t and vo are the transformed and original cell values, and v_{max} the maximum cell value recorded in the raster. The rescaling parameter k here allows the definition and testing of different strengths of raster cell conductance or resistance, relative to the conductance/resistance of a cell with a minimum value set to "1". For each of the three environmental factors, we tested three different values for k (i.e. k = 10, 100 and 1000).

The following analytical framework can be summarised in three distinct steps[4]. First, we used each environmental raster to compute an environmentally-scaled distance for each branch in inferred and simulated trees. These distances were computed using two different path models: (i) the least-cost path model, which uses a least-cost algorithm to determine the route taken between the starting and ending points[26], and (ii) the Circuitscape path model, which uses circuit theory to accommodate uncertainty in the route taken[27]. Second, correlations between time elapsed on branches and environmentally-scaled distances are estimated with the statistic Q defined as the difference between two coefficients of determination: (i) the determination coefficient obtained when branch durations are regressed against environmentally-scaled distances computed on the environmental raster, and (ii) the determination coefficient obtained when branch durations are regressed against environmentally-scaled distances computed on a uniform null raster, i.e. an environmental raster with a value of "1" assigned to all the cells. A Q statistic was estimated for each tree and we subsequently obtained two distributions of Q values, one associated with inferred trees and one associated with simulated trees. An environmental factor was only considered as potentially explanatory if both its distribution of regression coefficients and its associated distribution of Q values were positive[5]. Finally, the statistical support associated with a positive Q distribution (i.e. with at least 90% of positive values) was evaluated by comparing it with its corresponding null of distribution of Q values based on simulated trees, and formalised by approximating a Bayes factor (BF) value[25]. For a particular environmental factor e, the Bayes factor BF_e associated with the statistic Q is approximated by the posterior odds that $Q_{estimated} > Q_{simulated}$ divided by the equivalent prior odds (the prior probability for $Q_{estimated} > Q_{simulated}$ is considered to be 0.5)[60]:

$$BF_e = \frac{p_e}{1-p_e} / \frac{0.5}{1-0.5}$$

where p_e is the posterior probability that $Q_{estimated} > Q_{simulated}$, i.e. the frequency at which $Q_{estimated} > Q_{simulated}$ in the samples from the posterior distribution. The prior odds is 1 because we can assume an equal prior expectation for $Q_{estimated}$ and $Q_{simulated}$[25].

Testing the impact of migratory bird flyways on the dispersal history. In the second landscape phylogeographic testing approach, we investigated the impact of migratory flyways on the dispersal frequency of viral lineages. We first performed a test based on the four North American Migratory Flyways (NAMF). Based on observed bird migration routes, these four administrative flyways (Fig. 1) were defined by the US Fish and Wildlife Service (USFWS; https://www.fws.gov/birds/ management/flyways.php) to facilitate management of migratory birds and their habitats. Although biologically questionable, we here used these administrative limits to discretise the study and investigate if viral lineages tended to remain within the same flyway. In practice, we analysed if viral lineages crossed NAMF borders less frequently than expected by chance, i.e. than expected in the null dispersal model in which simulated dispersal histories were not impacted by these borders. Following the procedure introduced by Dellicour *et al.*[47], we computed and compared the number N of changing flyway events for each pair of inferred and simulated tree. Each "inferred" N value ($N_{inferred}$) was thus compared to its corresponding "simulated" value ($N_{simulated}$) by approximating a BF value using the above formula, but this time defining p_e as the posterior probability that $N_{inferred} < N_{simulated}$, i.e. the frequency at which $N_{inferred} < N_{simulated}$ in the samples from the posterior distribution.

To complement the first test based on an administrative flyway delimitation, we performed a second test based on flyways estimated by La Sorte *et al.*[28] for terrestrial bird species: the Eastern, Central and Western flyways (Fig. S2). Contrary to the NAMF, these three flyways overlap with each other and are here defined by geo-referenced grids indicating the likelihood that studied species are in migration during spring or autumn (see La Sorte *et al.*[28] for further details). Because the spring and autumn grids are relatively similar, we built an averaged raster for each flyway. For our analysis, we then generated normalised rasters obtained by dividing each raster cell by the sum of the values assigned to the same cell in the three averaged rasters (Fig. 2).

Following a procedure similar to the first test based on NAMFs, we computed and compared the average difference D defined as follows:

$$D = \sum_{i=1}^{n} \frac{v_{i,end} - v_{i,start}}{n}$$

where n is the number of branches in the tree, $v_{i,start}$ the highest cell value among the three flyway normalised raster to be associated with the position of the starting (oldest) node of tree branch i, and $v_{i,end}$ the cell value extracted from the same normalised raster but associated with the position of the descendant (youngest) node of the tree branch i. D is thus a measure of the tendency of tree branches to remain within the same flyway. Each "inferred" D value ($D_{inferred}$) is then compared to its corresponding "simulated" value ($D_{simulated}$) by approximating a BF value using the above formula, but this time defining p_e as the posterior probability that $D_{simulated} < D_{inferred}$, i.e. the frequency at which $D_{simulated} < D_{inferred}$ in the samples from the posterior distribution.

Testing the impact of environmental factors on the viral diversity through time. We used the skygrid-GLM approach[9,10] implemented in BEAST 1.10.4 to measure the association between viral effective population size and five covariates: human case numbers, temperature, precipitation, a greenness index and a bird observation index. The monthly number of human cases were provided by the CDC and were considered with different lag times of one and two months (meaning that the viral effective population size was compared to case count data from one and two months later), as well as the absence of lag time. Preliminary skygrid-GLM analyses were used to determine the most relevant lag time to use in subsequent analyses. Data used to estimate the average temperature and precipitation time series were obtained from the database managed by the US National Oceanic and Atmospheric Administration (NOAA; https://data.nodc.noaa.gov). For each successive month, meteorological stations were selected based on their geographic location. To estimate the average temperature/precipitation value for a specific month, we only considered meteorological stations included in the corresponding monthly minimum convex polygon obtained from the continuous phylogeographic inference. For a given month, the corresponding minimum convex hull polygon was simply defined around all the tree node positions occurring before or during that month. In order to take the uncertainty related to the phylogeographic inference into account, the construction of these minimum convex hull polygons was based on the 100 posterior trees used in the phylogeographic inference (see above). The rationale behind this approach was to base the analysis on covariate values averaged only over measures originating from areas already reached by the epidemic. The greenness index values were based on monthly Normalised Difference Vegetation Index (NDVI) raster files obtained from the NASA Earth Observation database (NEO; https://neo.sci.gsfc.nasa.gov). Monthly NDVI values were here obtained by cropping the NDVI rasters with the series of minimum convex hull polygons introduced above, and then averaging the remaining raster cell values. Finally, the Bird Observation Normalised Index (BONI) was obtained by the ratio of the sum of WNV associated bird sightings over the sum of all bird sightings for each month of each year over the region of interest (either continental US or US counties covered by the minimum convex hull polygons). Raw data used to compute this index were obtained from the eBird database[61] (https://ebird.org), for which each individual bird sighting was given a weight of "1". The list of birds associated with WNV infection has been retrieved from a compiled list obtained from the CDC, excluding "exotic-captive" bird species. We analysed these covariates using both univariate and multivariate approaches. While univariate skygrid-GLM analyses only involved one covariate at a time, the multivariate analyses included all the five covariates and used inclusion probabilities to assess their relative importance[30]. To allow their inclusion within the same multivariate analysis, the covariates were all log-transformed and standardised.

REFERENCES

1. Lemey, P., Rambaut, A., Welch, J. J. & Suchard, M. A. Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology and Evolution* **27**, 1877–1885 (2010).
2. Pybus, O. G. *et al.* Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 15066–15071 (2012).
3. Baele, G., Dellicour, S., Suchard, M. A., Lemey, P. & Vrancken, B. Recent advances in computational phylodynamics. *Current Opinion in Virology* **31**, 24–32 (2018).
4. Dellicour, S., Rose, R. & Pybus, O. G. Explaining the geographic spread of emerging epidemics: A framework for comparing viral phylogenies and environmental landscape data. *BMC Bioinformatics* **17**, 1–12 (2016).
5. Jacquot, M., Nomikou, K., Palmarini, M., Mertens, P. & Biek, R. Bluetongue virus spread in Europe is a consequence of climatic, landscape and vertebrate host factors as revealed by phylogeographic inference. *Proceedings of the Royal Society B: Biological Sciences* **284**, 20170919 (2017).
6. Brunker, K. *et al.* Landscape attributes governing local transmission of an endemic zoonosis: Rabies virus in domestic dogs. *Molecular Ecology* **27**, 773–788 (2018).
7. Dellicour, S., Vrancken, B., TrovÃčo, N. S., Fargette, D. & Lemey, P. On the importance of negative controls in viral landscape phylogeography. *Virus Evolution* **4**, vey023 (2018).
8. Minin, V. N., Bloomquist, E. W. & Suchard, M. A. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular Biology and Evolution* **25**, 1459–1471 (2008).
9. Gill, M. S. *et al.* Improving Bayesian population dynamics inference: A coalescent-based model for multiple loci. *Molecular Biology and Evolution* **30**, 713–724 (2013).
10. Gill, M. S., Lemey, P., Bennett, S. N., Biek, R. & Suchard, M. A. Understanding past population dynamics: Bayesian coalescent-based modeling with covariates. *Systematic Biology* **65**, 1041–1056 (2016).
11. Reisen, W. K. Ecology of West Nile virus in North America. *Viruses* **5**, 2079–2105 (2013).
12. Hayes, E. B. *et al.* Epidemiology and transmission dynamics of West Nile virus disease. *Emerging Infectious Diseases* **11**, 1167–1173 (2005).
13. May, F. J., Davis, C. T., Tesh, R. B. & Barrett, A. D. T. Phylogeography of West Nile virus: from the cradle of evolution in Africa to Eurasia, Australia, and the Americas. *Journal of Virology* **85**, 2964–2974 (2011).
14. Kramer, L. D. & Bernard, K. A. West Nile virus in the western hemisphere. *Current Opinion in Infectious Diseases* **14**, 519–525 (2001).
15. Kilpatrick, A. M., Kramer, L. D., Jones, M. J., Marra, P. P. & Daszak, P. West Nile virus epidemics in North America are driven by shifts in mosquito feeding behavior. *PLoS Biology* **4**, 606–610 (2006).
16. Colpitts, T. M., Conway, M. J., Montgomery, R. R. & Fikrig, E. West Nile virus: Biology, transmission, and human infection. *Clinical Microbiology Reviews* **25**, 635–648 (2012).
17. Molaei, G., Andreadis, T. G., Armstrong, P. M., Anderson, J. F. & Vossbrinck, C. R. Host feeding patterns of *Culex* mosquitoes and West Nile virus transmission, northeastern United States. *Emerging Infectious Diseases* **12**, 468–474 (2006).
18. Bowen, R. A. & Nemeth, N. M. Experimental infections with West Nile virus. *Current Opinion in Infectious Diseases* **20**, 293–297 (2007).
19. Petersen, L. R. & Marfin, A. A. West Nile virus: A primer for the clinician. *Annals of Internal Medicine* **137**, 173–179 (2002).
20. Lanciotti, R. S. *et al.* Origin of the West Nile virus responsible for an outbreak of encephalitis in the northeastern United States. *Science* **286**, 2333–2337 (1999).
21. Goddard, L. B., Roth, A. E., Reisen, W. K. & Scott, T. W. Vertical transmission of West Nile virus by three California *Culex* (Diptera: Culicidae) species. *Journal of Medical Entomology* **40**, 743–746 (2003).
22. Lequime, S. & Lambrechts, L. Vertical transmission of arboviruses in mosquitoes: A historical perspective. *Infection, Genetics and Evolution* **28**, 681–690 (2014).
23. George, T. L. *et al.* Persistent impacts of West Nile virus on North American bird populations. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 14290–14294 (2015).
24. LaDeau, S. L., Kilpatrick, A. M. & Marra, P. P. West Nile virus emergence and large-scale declines of North American bird populations. *Nature* **447**, 710–713 (2007).
25. Dellicour, S. *et al.* Using viral gene sequences to compare and explain the heterogeneous spatial dynamics of virus epidemics. *Molecular Biology and Evolution* **34**, 2563–2571 (2017).
26. Dijkstra, E. W. A note on two problems in connexion with graphs. *Numerische Mathematik* **1**, 269–271 (1959).
27. McRae, B. H. Isolation by resistance. *Evolution* **60**, 1551–1561 (2006).
28. La Sorte, F. A. *et al.* The role of atmospheric conditions in the seasonal dynamics of North American migration flyways. *Journal of Biogeography* **41**, 1685–1696 (2014).
29. Holmes, E. C. & Grenfell, B. T. Discovering the phylodynamics of RNA viruses. *PLoS Computational Biology* **5**, e1000505 (2009).
30. Faria, N. R. *et al.* Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science* **361**, 894–899 (2018).

31. Carrington, C. V. F., Foster, J. E., Pybus, O. G., Bennett, S. N. & Holmes, E. C. Invasion and maintenance of dengue virus type 2 and type 4 in the Americas. *Journal of Virology* **79**, 14680–14687 (2005).

32. Rappole, J. H. *et al.* Modeling movement of West Nile virus in the western hemisphere. *Vector-Borne and Zoonotic Diseases* **6**, 128–139 (2006).

33. Goldberg, T. L., Anderson, T. K. & Hamer, G. L. West Nile virus may have hitched a ride across the Western United States on *Culex tarsalis* mosquitoes: News and views. *Molecular Ecology* **19**, 1518–1519 (2010).

34. Kwan, J. L., Kluh, S. & Reisen, W. K. Antecedent avian immunity limits tangential transmission of West Nile virus to humans. *PLoS One* **7**, e34127 (2012).

35. Di Giallonardo, F. *et al.* Fluid spatial dynamics of West Nile virus in the United States: Rapid spread in a permissive host environment. *Journal of Virology* **90**, 862–872 (2016).

36. Samuel, G. H., Adelman, Z. N. & Myles, K. M. Temperature-dependent effects on the replication and transmission of arthropod-borne viruses in their insect hosts. *Current Opinion in Insect Science* **16**, 108–113 (2016).

37. Paz, S. & Semenza, J. C. Environmental drivers of West Nile fever epidemiology in Europe and Western Asia - a review. *International Journal of Environmental Research and Public Health* **10**, 3543–3562 (2013).

38. Dohm, D. J., O'Guinn, M. L. & Turell, M. J. Effect of environmental temperature on the ability of *Culex pipiens* (Diptera: Culicidae) to transmit West Nile virus. *Journal of Medical Entomology* **39**, 221–225 (2002).

39. Kilpatrick, A. M., Meola, M. A., Moudy, R. M. & Kramer, L. D. Temperature, viral genetics, and the transmission of West Nile virus by *Culex pipiens* mosquitoes. *PLoS Pathogens* **4**, e1000092 (2008).

40. DeFelice, N. B. *et al.* Use of temperature to improve West Nile virus forecasts. *PLoS Computational Biology* **14**, e1006047 (2018).

41. Morin, C. W. & Comrie, A. C. Regional and seasonal response of a West Nile virus vector to climate change. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 15620–15625 (2013).

42. Samy, A. M. *et al.* Climate change influences on the global potential distribution of the mosquito *Culex quinquefasciatus*, vector of West Nile virus and lymphatic filariasis. *PLoS One* **11**, e0163863 (2016).

43. Shocket, M. S. *et al.* Transmission of West Nile virus and other temperate mosquito-borne viruses occurs at lower environmental temperatures than tropical mosquito-borne diseases. *bioRxiv* 597898 (2019).

44. DeFelice, N. B. *et al.* Modeling and surveillance of reporting delays of mosquitoes and humans infected with West Nile virus and associations with accuracy of West Nile virus forecasts reporting delays for West Nile virus and accuracy of West Nile virus forecasts reporting delays for West Nile virus and accuracy of West Nile virus forecasts. *JAMA Network Open* **2**, e193175 (2019).

45. Gardy, J. L. & Loman, N. J. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nature Reviews Genetics* **19**, 9–20 (2018).

46. Grubaugh, N. D. *et al.* Tracking virus outbreaks in the twenty-first century. *Nature Microbiology* **4**, 10–19 (2019).

47. Dellicour, S. *et al.* Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak. *Nature Communications* **9**, 2222 (2018).

48. Katoh, K. & Standley, D. M. Mafft multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* **30**, 772–780 (2013).

49. Larsson, A. AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278 (2014).

50. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).

51. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using beast 1.10. *Virus Evolution* **4**, vey016 (2018).

52. Ayres, D. L. *et al.* BEAGLE 3: Improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. *Systematic Biology* (2019).

53. Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences* **17**, 57–86 (1986).

54. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biology* **4**, 699–710 (2006).

55. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology* **67**, 901–904 (2018).

56. Fisher, A. A., Ji, X., Lemey, P. & Suchard, M. Relaxed random walks at scale. *arXiv* 1906.04834 (2019).

57. Lemey, P. *et al.* Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS Pathogens* **10**, e1003932 (2014).

58. Bedford, T. *et al.* Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature* **523**, 217 (2015).

59. Dellicour, S., Rose, R., Faria, N. R., Lemey, P. & Pybus, O. G. SERAPHIM: Studying environmental rasters and phylogenetically informed movements. *Bioinformatics* **32**, 3204–3206 (2016).

60. Suchard, M. A., Weiss, R. E. & Sinsheimer, J. S. Models for estimating bayes factors with applications to phylogeny and tests of monophyly. *Biometrics* **61**, 665–673 (2005).

61. Sullivan, B. L. *et al.* eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation* **142**, 2282–2292 (2009).

Acknowledgments We are grateful to Frank La Sorte for sharing their estimated flyway grids. 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). SD is supported by the *Fonds National de la Recherche Scientifique* (FNRS, Belgium) and was previously funded by the *Fonds Wetenschappelijk Onderzoek* (FWO, Belgium). LdP and OGP are supported by the European Research Council under the European Commission Seventh Framework Programme (grant agreement no. 614725-PATHPHYLODYN) and by the Oxford Martin School. PL acknowledges support by the Research Foundation - Flanders (*Fonds voor Wetenschappelijk Onderzoek - Vlaanderen*, G066215N, G0D5117N and G0B9317N). SL, BV and PB are funded by the *Fonds Wetenschappelijk Onderzoek* (FWO, Belgium). The Artic Network receives funding from the Welcome Trust through project 206298/Z/17/Z. MAS is partially supported by NSF grant DMS 1264153 and NIH grants R01 AI107034 and U19 AI135995. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.