# Using Viral Gene Sequences to Compare and Explain the Heterogeneous Spatial Dynamics of Virus Epidemics

Simon Dellicour,*[,1] Rebecca Rose,[2] Nuno Rodrigues Faria,[3] Luiz Fernando Pereira Vieira,[4] Hervé Bourhy,[5] Marius Gilbert,[6] Philippe Lemey,[1] and Oliver G. Pybus[3]

[1]Clinical and Epidemiological Virology, Department of Microbiology and Immunology, Rega Institute, KU Leuven—University of Leuven, Leuven, Belgium

[2]BioInfoExperts LLC, Thibodaux, LA

[3]Department of Zoology, University of Oxford, Oxford, United Kingdom

[4]Department of Laboratorial Diagnosis, Institute of Agricultural and Forest Defense of Espírito Santo (IDAF), Vitoria, Brazil

[5]Institut Pasteur, Lyssavirus Dynamics and Host Adaptation Unit, WHO Collaborating Centre for Reference and Research on Rabies, Paris, France

[6]Spatial Epidemiology Lab (SpELL), Université Libre de Bruxelles, Brussels, Belgium

***Corresponding author:** E-mail: simon.dellicour@kuleuven.be.

**Associate editor:** Claus Wilke

## Abstract

**Rabies is an important zoonotic disease distributed worldwide. A key question in rabies epidemiology is the identification of factors that impact virus dispersion. Here we apply new analytical methods, based on phylogeographic reconstructions of viral lineage movement, to undertake a comparative evolutionary-epidemiological study of the spatial dynamics of rabies virus (RABV) epidemics in different hosts and habitats. We compiled RABV data sets from skunk, raccoon, bat and domestic dog populations in order to investigate the viral diffusivity of different RABV epidemics, and to detect and compare the environmental factors that impact the velocity of viral spread in continuous spatial landscapes. We build on a recently developed statistical framework that uses spatially- and temporally-referenced phylogenies. We estimate several spatial statistics of virus spread, which reveal a higher diffusivity of RABV in domestic dogs compared with RABV in other mammals. This finding is explained by subsequent analyses of environmental heterogeneity, which indicate that factors relating to human geography play a significant role in RABV dispersion in domestic dogs. More generally, our results suggest that human-related factors are important worldwide in explaining RABV dispersion in terrestrial host species. Our study shows that phylogenetically informed viral movements can be used to elucidate the factors that impact virus dispersal, opening new opportunities for a better understanding of the impact of host species and environmental conditions on the spatial dynamics of rapidly evolving populations.**

***Key words:*** **phylodynamics, viral phylogeography, molecular epidemiology, relaxed random walk, RABV, spread.**

## Introduction

Understanding the processes that cause variation in the rate of biological invasions (which includes both emerging pathogens and ecologically invasive species) remains a key question in spatial population biology. In addition to host behavior and environmental factors, the spatial dynamics of an emerging infectious disease will also be affected by stochastic variation in the invasion process itself (Melbourne and Hastings 2009). One way to evaluate this variation is to directly compare different spatial dynamics caused by the same pathogen or invasive species. Although large data sets of spatio-temporal incidence can be used to make such comparisons (Mundt et al. 2009), such data can be difficult and time consuming to collect. Through phylogeographic approaches, gene sequence data may provide a practical alternative to direct observation for the investigation of spatial processes in ecology and epidemiology.

Phylogeographic analyses based on viral gene sequences are increasingly used to gain insight into spatial epidemic processes. Phylogeography is the study of shared ancestry, inferred using phylogenetic methods, in a geographic and temporal context (Silva et al. 2013). Phylogeographic analysis of animal and human viruses can provide valuable insights into the spatial dissemination of outbreaks and epidemics (Holmes 2004; Faria et al. 2011). Complementing traditional epidemiology, evolutionary analyses can reconstruct the spatio-temporal history of an epidemic even when surveillance records are scarce or absent (Dellicour, Rose, and Pybus 2016). Further, phylogeographic approaches can infer linkages among infections and reveal rare but influential founder







Article

events that may not be evident otherwise (Pybus et al. 2012). The rapidly increasing availability of viral genomic data provides an opportunity to develop and apply new evolutionary approaches that can quantify the spatial dynamics of virus. Here we use new phylogeographic techniques to investigate whether host species and environmental factors can explain the spatial dispersion of rabies virus. Answering questions such as these will help to better predict infectious disease emergence from animal reservoirs.

The rabies virus (RABV, genus Lyssavirus, family Rhabdoviridae) is an RNA virus whose genome comprises ~12,000 nucleotides and contains five genes. RABV is the etiological cause of rabies, a fatal infection of the central nervous system usually transmitted through bites by infected animals. Various mammal species including dogs, foxes, raccoons, skunks, mongooses and bats, act as both reservoirs and vectors of the virus. In 1886, Louis Pasteur developed a rabies vaccine that has since allowed the development of prevention strategies and postexposure prophylaxis (WHO 2005; Warrell and Warrell 2004). Despite the availability of this vaccine, RABV still causes approximately 59,000 human deaths per year, and remains one of the most virulent diseases of animals and humans (Hampson et al. 2015). Almost all human deaths are caused by infections with dog RABV (Bourhy et al. 2008) and the majority occur due to the lack of ready accessibility to rabies vaccine and immunoglobulins (Knobel et al. 2005; Dodet et al. 2008). In recent years, a number of RABV epidemics have been studied using phylogeographic approaches (Carnieli et al. 2011; Hayman et al. 2011; Picard-Meyer et al. 2012; Piñero et al. 2012; Fusaro et al. 2013; Seetahal et al. 2013; Tohma et al. 2014; Zieger et al. 2014; Horton et al. 2015). Phylogeography has been used to understand the spread of rabies virus at both large (Biek et al. 2007; Talbi et al. 2009; Kuzmina et al. 2013; Troupin et al. 2016) and small geographic scales (Bourhy et al. 2016). However, these efforts have used a variety of different analysis techniques and each stands in isolation. It is therefore important to undertake formal comparative analyses in order to elucidate the differential impact of host species and environmental conditions on RABV spread.

Here, we perform a comparative analysis of the spatial genetics of different outbreaks of the same virus, with the aim of understanding to what extent different host species and environmental variables affect spatial dynamics. We analyze five instances of RABV spread in different mammalian populations and different locations using a common statistical framework. Our approach computes statistics relating to spatial dissemination and allows the identification of specific environmental factors that impact viral lineage dispersion velocity. We selected five publicly available viral genetic data sets associated with distinct epidemics: (1) RABV in North American skunks (Kuzmina et al. 2013), (2) RABV in North American raccoons (Biek et al. 2007), (3) RABV in North African domestic dogs (Talbi et al. 2010), (4) RABV in vampire bat populations in eastern Argentina (Torres et al. 2014) and (5) RABV in vampire bats in eastern Brazil (Vieira et al. 2013). These data sets are named hereafter as the "skunk", "raccoon", "dog", "bat-1", and "bat-2" RABV data sets, respectively. The data sets all contain acceptably long nucleotide sequences (800–3,000 nt) that are spatially distributed over broad geographic ranges and are associated with precise information on the dates and locations of sampling, and on host species.

Through a detailed comparative analysis, we aim to (1) reconstruct the spatio-temporal epidemic history of each outbreak, (2) estimate and compare spatial diffusion coefficients for each instance of RABV spread, and (3) use statistical tests to identify which environmental factors determine RABV spread, and whether those factors are shared or vary among independent outbreaks.

## New Approaches

We present a phylogeographic approach in continuous space to address epidemiological questions in a quantitative framework. We aim to identify the environmental factors impacting the spatial dynamics of different epidemics of the same virus (the rabies virus, RABV) in different host species (domestic dog, skunk, raccoon, and vampire bats) and different geographic contexts. To directly compare epidemics, we apply the same analysis procedure to multiple independent RABV data sets. Our method improves on that described in Dellicour, Rose, and Pybus (2016) which is available in the R package SERAPHIM (Dellicour et al. 2016; see the related manual and new tutorials available within the package for further details). An important feature of the modification used here is the ability to approximate Bayes factor support values for each environmental factor. This improvement enables a more straightforward interpretation of the statistical results. Further, because the new approach now requires only one randomization per sampled tree, we also significantly decrease computation time. An increase in computational efficiency makes more practical large-scale comparative studies, such as those reported here.

## Results

We employ an analytical procedure that comprises five distinct steps, described in detail in the "Materials and Methods" section. The workflow can be summarized as follows: (1) we first extract information contained in spatially- and temporally-referenced phylogenies, which are inferred using a Bayesian continuous phylogeographic method implemented in BEAST (Lemey et al. 2010). In order to take into account uncertainty in phylogenetic estimation, we use a sample of 100 trees from the posterior distribution of phylogenies for each data set. The velocity, distance and duration of spatial movement along each phylogeny branch in each tree are represented by a vector. (2) Second, these vectors are used to estimate and compare spatial diffusion coefficients and other spatial summary statistics for each RABV host species. (3) Third, a series of environmental factors are investigated. Each factor (e.g., elevation) is described by a raster that defines the spatial heterogeneity of that variable. This raster is used to calculate an "environmental distance" for each phylogeny branch, which represents both the actual distance travelled and the degree to which that environmental factor facilitated or impeded lineage movement. (4) Correlations between

phylogenetic branch durations and the environmental distances are then estimated. (5) Finally, the statistical support for these correlations is evaluated against a null distribution generated by a randomization procedure and formalized as Bayes factor support.

A visual comparison of the general trends in spatial diffusion history for the five RABV epidemics analyzed here is shown in figure 1. Each is inferred from a set of phylogeographic molecular clock trees whose nodes all have defined locations and durations. These reconstructions reveal two main diffusion patterns: (1) continual range expansion, as observed in the skunk and raccoon data sets, and the bat-2 data set, and (2) viral lineage diffusion within an endemic area, which is apparent for the dog and bat-1 RABV data sets. For the two latter cases, uncertainty in the estimated locations of the most ancestral nodes (colored in red) is large and encompasses a considerable fraction of entire sampling area. Furthermore, plots for each data set of the maximal spatial epidemic wavefront through time (supplementary fig. S1, Supplementary Material online) indicate that the bat-2 RABV spread in Brazil is associated with the slowest wavefront velocity (∼4 km/year), followed by the skunk RABV spread (∼6 km/year). This result contrasts with the higher wavefront velocity estimated for bat-1 RABV in Argentina (∼24 km/year). Finally, dog and raccoon RABV display similar wavefront velocities (∼15–22 and ∼20 km/year, respectively).

In order to compare the RABV diffusivity associated with each host species, we estimated spatial diffusion coefficients for each data set using two alternative statistics: the original ($D_{original}$; Pybus et al. 2012) and weighted ($D_{weighted}$; Trovão et al. 2015) estimators. While $D_{original}$ is an estimate of the average diffusion coefficient associated with each branch in the tree, $D_{weighted}$ is a weighted average across the tree. Consequently, values of $D_{original}$ are approximately twice as large as those for $D_{weighted}$ (supplementary fig. S2, Supplementary Material online). In addition, it is interesting to note that the mean and coefficient of variation of the diffusion coefficient values tend to be positively correlated for the $D_{original}$ metric, but for not for the $D_{weighted}$ statistic (supplementary fig. S2, Supplementary Material online). While $D_{original}$ is calculated in a manner similar to an arithmetic mean of the branch-specific values, $D_{weighted}$ is calculated in a manner similar to a weighted mean. As a consequence, $D_{weighted}$ is less sensitive to extreme values on short branches while $D_{original}$ will be more affected by the among-branch variation this imposes (see the "Materials and Methods" section for detail). Despite these differences, both statistics clearly show that the dog RABV epidemic in North Africa is characterized by the highest diffusivity ($D_{weighted} = \sim 1300$ km/year$^2$; fig. 2). In contrast, estimated diffusion coefficients for the skunk, raccoon and bat RABV are lower and more similar to each other ($D_{weighted} = \sim 550$–$700$ km/year$^2$; fig. 2). In addition, if we compare the variation in diffusion coefficients among lineages then we observe greater among-branch variation for the dog RABV and smaller among-branch variation for the raccoon RABV (fig. 2).

For each RABV data set, several environmental factors were tested as potential correlates of virus dispersal. As an illustration of the environmental data used, the nine factors evaluated for the spread of dog RABV in North Africa are shown in supplementary figure S3, Supplementary Material online. Randomization tests were performed to assess the level of significance of the correlations between phylogeny branch durations and "environmental distances". Results of the randomization tests are reported in supplementary table S1, Supplementary Material online, and the most important results are gathered in table 1 for comparison. We report a Bayes factor (BF) value for each combination of data set, path model, and environmental factor that was tested. Following the interpretation of BF values defined by Jeffreys (1961), BF values higher than 10 and $10^{3/2}$ are considered as "strong" and "very strong" evidence for a correlation between dispersal durations and environmentally scaled distances.

For the bat RABV data sets, none of the environmental factors were significantly associated with viral lineage movement, under any model/parameter combination (BFs $<10$; see supplementary table S1, Supplementary Material online). For the skunk RABV data set, we found strong evidence (BF $>10$) for a correlation between dispersal durations and environmental distances, for the croplands and human population density factors (when treated as conductance factors with the least-cost path model). For the same data set, we also find a very strong effect (BF $>10^{3/2}$) of the human population density raster using the random walk path model. For the raccoon RABV data set, both the accessibility to nearest major cities (time travel to nearest major cities of $>50{,}000$ inhabitants, hereafter referred as "inaccessibility") and urban areas are identified as important predictors with BFs $>10$ or higher (table 1) when respectively treated as resistance and conductance factors by the least-cost path model. For the raccoon RABV data set, there is also strong evidence that elevation (treated as a resistance factor) is associated with rabies lineage dissemination. Specifically, the BF value for that factor is $>10^{3/2}$ under the least-cost path model and $>10$ under the random walk path model. Finally, randomization tests performed on the dog RABV data set highlight several environmental factors that are associated with viral lineage spread (table 1), which are: inaccessibility (BF $>10$, resistance factor; least-cost path model), grasslands (BF $>10$ or higher; conductance factor; least-cost path model), urban areas (BF $>10^{3/2}$; conductance factor; least-cost path model), human population density (BF $>10$; conductance factor; least-cost path model) and elevation (BF $>10$ or higher, resistance factor; least-cost path). Human population density is thus identified as an important factor for both the skunk and the dog data sets, and inaccessibility, urban areas and elevation for both the raccoon and dog data sets.

## Discussion

In this study, we show that the patterns of spread in five RABV data sets are characterized by different dynamics of spatial dispersal. Skunks and raccoons constitute ∼33% and 27% of rabid animals documented in United States in 2011,

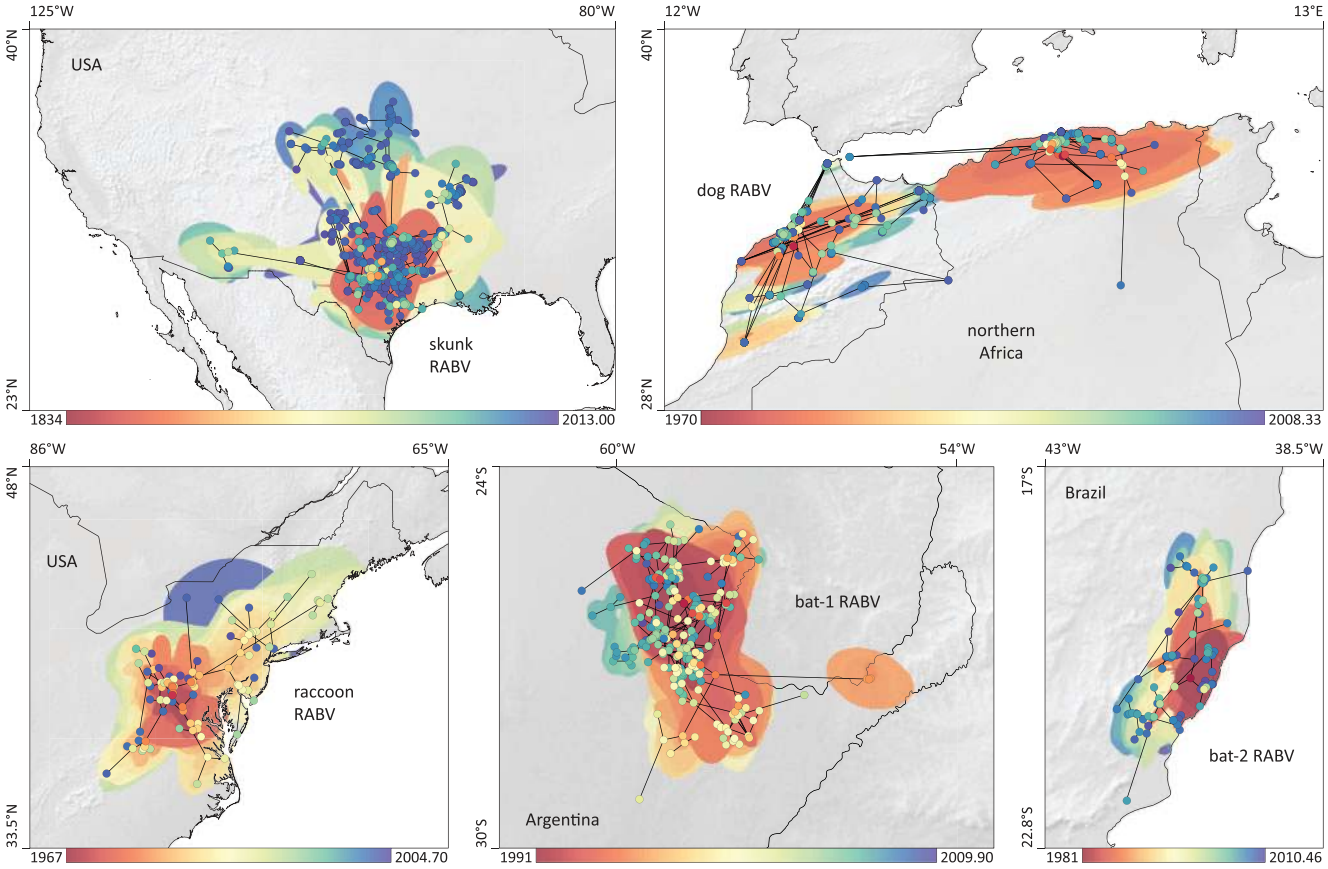


**Fig. 1.** Reconstructed spatio-temporal diffusion of five RABV data sets: mapped consensus trees and 95% HPD regions based on 100 trees subsampled from the post burn-in posterior distribution. Nodes of the consensus trees are coloured according to time, using a scale ranging from red (MRCA) to blue (most recent sampling time). 95% HPD regions were computed for a series of time points and then superimposed using the same red-to-blue temporal colour scale.

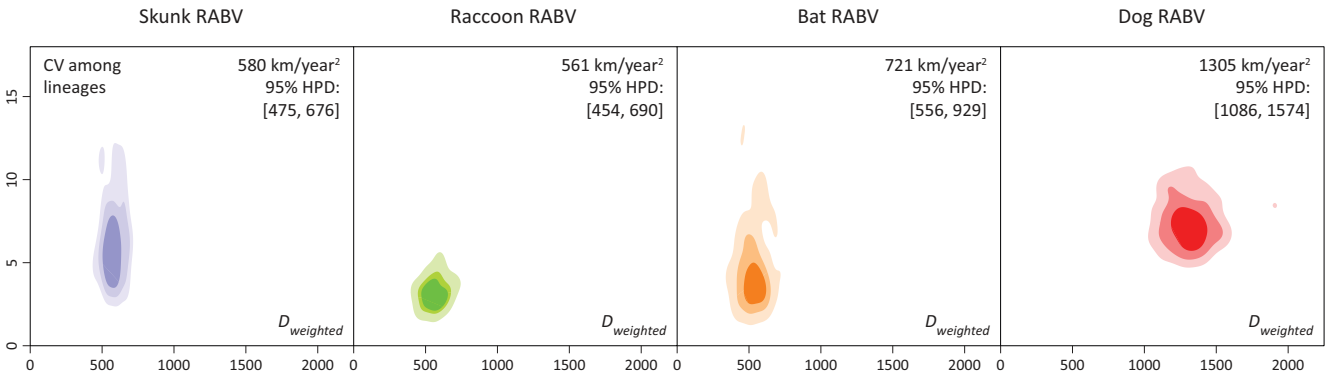


**Fig. 2.** Kernel density estimates of the diffusion coefficient parameters obtained using the $D_{weighted}$ statistic, for each data set. Plots show the mean diffusion coefficient among branches (*x* axis) versus the coefficient of variation "CV" of that value among branches (*y* axis). In each case, the three contours show, in shades of decreasing darkness, the 50%, 75%, and 95% highest posterior density regions via kernel density estimation. We also report the median value and 95% HPD interval of this statistic.

respectively (Blanton et al. 2012), and the skunk rabies virus is one of the most broadly distributed terrestrial viral lineages in North America (Kuzmina et al. 2013). For the skunk RABV dataset, our results suggest that croplands and human population density are associated with a greater degree of spatial epidemic spread. Although Kuzmina et al. (2013) proposed the importance of deserts and mountains as important barriers to dispersal of skunk RABV, we found no significant link between branch velocity and the "barren vegetation" and "elevation" environmental layers for the same dataset.

In contrast, as previously highlighted by Biek et al. (2007) and Dellicour, Rose, and Pybus (2016), there is strong evidence that, for the raccoon RABV data set, elevation has a significant negative impact on viral spatial spread. We also find some evidence that human-geographic variables have an impact on the lineage velocity of rabies viruses in the raccoon

**Table 1.** Selected Results of Randomization Tests: Bayes Factors (BF) of model combinations for which the Estimated BF Value Is >10 for One Path Model.

| RABV Data Set | Environmental Factor | *k* | Least-Cost Path Model | | Random Walk Path Model | |
|---|---|---|---|---|---|---|
| | | | Conductance | Resistance | Conductance | Resistance |
| Skunk RABV | Croplands | 1000 | 13.29 | 0.16 | 4.26 | 0.23 |
| | Human pop. density | 100 | 4.26 | 0.37 | 32.33 | 0.30 |
| | Human pop. density | 1000 | 13.29 | 0.43 | *49.00* | 0.33 |
| Dog RABV | Inaccessibility | 10 | 1.08 | 15.67 | 0.15 | 0.96 |
| | Inaccessibility | 100 | 0.35 | 13.29 | 0.12 | 2.70 |
| | Inaccessibility | 1000 | 0.20 | 10.11 | 0.11 | 3.55 |
| | Grasslands | 100 | 13.29 | 1.50 | 2.70 | 0.49 |
| | Grasslands | 1000 | 32.33 | 0.72 | 24.00 | 0.32 |
| | Urban areas | 10 | 32.33 | 3.35 | 0.52 | 0.00 |
| | Elevation | 10 | 0.32 | *49.00* | 0.05 | 9.00 |
| | Elevation | 100 | 0.11 | 5.67 | 0.14 | 11.50 |
| | Elevation | 1000 | 0.19 | 4.56 | 0.28 | 10.11 |
| | Human pop. density | 10 | 15.67 | 0.39 | 2.23 | 0.22 |
| Raccoon RABV | Inaccessibility | 100 | 0.28 | 32.33 | 0.43 | 2.45 |
| | Inaccessibility | 1000 | 0.30 | 15.67 | 0.30 | 2.45 |
| | Urban areas | 10 | 19.00 | 0.35 | 1.56 | 0.32 |
| | Urban areas | 100 | 24.00 | 0.30 | 1.63 | 0.28 |
| | Urban areas | 1000 | 24.00 | 0.35 | 1.38 | 0.32 |
| | Elevation | 10 | 0.09 | *49.00* | 0.09 | 19.00 |
| | Elevation | 100 | 0.16 | *49.00* | 0.35 | 24.00 |
| | Elevation | 1000 | 0.20 | *49.00* | 0.47 | 19.00 |
| Bat-1 RABV | — | — | — | — | — | — |
| Bat-2 RABV | — | — | — | — | — | — |

NOTE.—According to Jeffreys (1961), BF's >10 and $>10^{3/2}$ (31.62; in italics) are respectively considered as "strong" and "very strong" evidence of statistical significance, that is, of a significant correlation between environmental distance and dispersal duration.

data set (i.e., the inaccessibility and urban areas rasters treated as resistance and conductance factors, respectively). However, we do not find any significant impact of rivers acting as barriers. This differs from the results reported by Smith et al. (2002) in a more localized study of raccoon RABV in Connecticut. Smith et al.'s (2002) study, which was not based on viral genetic sequences but instead used a more traditional model based on epidemiological records, identified a slower local spread of RABV associated with river crossing events.

As dogs are terrestrial animals, their natural locomotion alone is unlikely to explain the high diffusivity of dog RABV identified here. The movement of domestic dogs is likely to be affected by human activity. Correspondingly, we identified three human-related environmental rasters as important factors for dog RABV spread: inaccessibility (treated as resistance factor), urban areas, and human population density (both treated as a conductance factor). Hence the dispersal of RABV within the North African domestic dog population appears to be shaped by human-based connectivity and mobility. Based on a discrete phylogeographic analysis and marginal likelihood estimations, Talbi et al. (2010) suggested that the spatial dynamics of RABV in North African dogs was best described by pairwise road distances between sampling locations. Our study statistically supports this hypothesis and further highlights the importance of human geography in explaining the spatial dynamics of domestic dog RABV. More generally, the three factors correlated with dog RABV spread are also identified as potential determinants of the spatial dynamics of RABV in skunks (human population density) and raccoons (inaccessibility, urban areas). Thus human-related factors appear important across different continents and host species in explaining RABV dispersion in terrestrial host species. In bats, however, we do not find any evidence that human factors have an impact on viral lineage spread. As discussed below, this could be because bats are nonterrestrial hosts whose dispersal remains largely unaffected by landscape features shaped by human activities.

In Latin America, the common vampire bat (*Desmodus rotundus*) is the most important source of human and animal rabies (Benavides et al. 2016). This species is an important RABV reservoir host and every year cattle and horses die from rabies transmitted by this haemotophagous bat species (Vieira et al. 2013; Torres et al. 2014). Benavides et al. (2016) report that, in these regions, rabies virus invasions form wavefronts that can advance towards large and unvaccinated livestock populations that are bitten by bats. Because of their aerial locomotion, it is unsurprising that we find higher diffusivity of RABV in bats than for RABV in skunks and raccoons. However, our study failed to identify strong support (BF >10) for environmental factors that might be driving viral spread in bats. There are several potential explanations for this. Firstly, the complexity of RABV circulation in noeotropical bat communities (de Thoisy et al. 2016) may obscure the impact of any individual environmental factor on RABV dispersion velocity. Secondly, as mentioned earlier, landscape features may have comparably less impact on the dispersal of a nonterrestrial species. This would mean that there would be no environmental factor that appropriately explains the RABV dispersion time, other than geographic distance alone. Thirdly, it is possible that the factors we tested did not

contain the ones that are relevant to the ecology of RABV in bats. Finally, sampling bias or sampling from a restricted area within a wider region of bat dispersal may compromise the statistical power necessary to identify relevant factors.

While sampling bias imposes a general limitation on phylogeographic analyses, its impact is best characterized for the discrete phylogeographic method (Lemey et al. 2009; De Maio et al. 2015; Baele et al. 2017). Indeed, with this method, over or under sampling can directly affect estimates of model parameters (i.e., transition rates in and out of locations) and hence can affect ancestral reconstruction. In the continuous phylogeographic method (Lemey et al. 2010), the relationship between sampling density in specific areas and model parameters (i.e., the variance co-variance matrix of the diffusion process) is, however, less straightforward, and this should be addressed in more detail in future investigations. In addition to sampling heterogeneity, we also need to consider the impact of incomplete spatial coverage. If we fail to include a clade or lineage from an unsampled area, then we simply do not test the impact of the environment in that area on the diffusion process. The conclusions we draw relate to how environmental factors shape viral dispersal, and they therefore pertain only to the area from which we were able to sample. Sparse, incomplete and poorly representative sampling is still expected to result in ancestral reconstructions that may not capture the underlying dispersal pattern well. While our approach is conditional on these reconstructions, this will primarily impact the statistical power in detecting relevant environmental factors.

According to the distribution of rabies cases in United States, published by the CDC (Centers for Disease Control and Prevention, www.cdc.gov), the sequences in our skunk and raccoon RABV data sets provide relevant coverage of the occurrence records. Similarly, given that human population density acts as a good proxy of domestic dog presence in northern Africa, viral sequence sampling for that data set appears to cover the host distribution in that region. Bat RABV sampling, on the other hand, is harder to evaluate and is more localized within the broader vampire bat distribution in South America. This may explain, at least partially, the absence of strong support (i.e., Bayes factors $>10$) for environmental factors impacting RABV spread in bat populations (in line with the abovementioned point about undersampling leading to reduced statistical power). Therefore, we cannot exclude the possibility that environmental variables act as significant factors outside the relatively restricted sampling areas for bat RABV.

Phylogeographic analyses of genetic sequences sampled in two dimensional space (so-called "continuous" phylogeography), enables detailed comparisons of the spatial dynamics of different populations, or, in this case, different strains of the same virus spreading in different regions or host species. By inferring evolutionary relationships among sampled individuals, phylogenetic analysis (coupled with longitude and latitude coordinates) can quantify virus spatial spread. Phylogenetic branches from spatially- and temporally referenced trees can be treated as movement vectors (Pybus et al. 2012) and collections of such vectors can be used to estimate dispersal statistics and to test the association between dispersal velocity and environmental factors. Comparative application of these methods can reveal how environmental factors impact the dispersal dynamics of different epidemics. This could become a useful addition to pre-existing quantitative tools with applications to other emerging infectious diseases that infect animals and humans. Such approaches could lead to a better understanding of pathogen spread and could ultimately inform the prevention, prediction and control of emerging and zoonotic infectious diseases. The fact that we can detect different associations between the environment and viral dispersal for terrestrial and nonterrestrial host species means that, in principle, it may be possible to use the spatial dynamics of lineages to infer the likely host species of a newly discovered pathogen whose reservoir or source population is unknown.

## Materials and Methods

A common analysis workflow was applied to each RABV data set and involved five steps, outlined below. All analytical steps were performed with R functions available in an updated version of the package SERAPHIM (Dellicour et al. 2016; see the related tutorials within the package for further practical details about these scripts).

### Step 1

The history of lineage dispersal was recovered from phylogenies that were generated using a phylogeographic model, such that the trees have branches that represent time and have tips and internal nodes that each have a defined location. In this study, such trees were reconstructed using a phylogenetic relaxed random walk (RRW) diffusion model implemented in BEAST (Lemey et al. 2010).

It is important that a RRW is used, and not a strict Brownian motion model, because it is impossible to test whether environmental factors correlate with branch diffusion velocities if those velocities do not vary among phylogeny branches.

For each data set, we used the nucleotide substitution model, molecular clock model, coalescent prior, and RRW model that were used in the original study, when available (supplementary table S2, Supplementary Material online). Prior to phylogeographic analyses, we assessed the phylogenetic temporal signal within each dataset using regressions of root-to-tip genetic distances against sequence sampling times. Analyses used maximum likelihood trees inferred with PhyML (Guindon et al. 2010) and the correlation and determination coefficient ($R^2$) of the regression were estimated with TempEst (Rambaut et al. 2016). The $P$-values were calculated using the approach of Murray et al. (2016) and based on 1,000 random permutations of the sequence sampling dates (Navascués et al. 2010). Root-to-tip regression results are reported in supplementary table S3, Supplementary Material online, and confirm the presence of significant temporal signal ($P$-value $< 0.05$) in all data sets reported here. For the dog RABV data set, a single phylogeographic reconstruction resulted in unacceptably high uncertainty of the root location estimate (results not shown),

which is not unusual for estimates of continuous traits at deep phylogenetic nodes (Schluter et al. 1997). Consequently, for the dog RABV data set, we performed separate analyses for the two major country-specific clades as originally identified in Talbi et al. (2010). For the present study, we selected a subset of 100 trees sampled at regular intervals from the posterior distribution of trees (after burn-in had been removed) and extracted the spatio-temporal information embedded in these trees using the "treeExtractions" function of the R package SERAPHIM (Dellicour et al. 2016). Specifically, each phylogenetic branch was considered a vector defined by its start and end location (latitude and longitude), and its start and end dates (in decimal units). Each phylogeny branch therefore represents a conditionally independent viral lineage dispersal event (Pybus et al. 2012). Using these extracted vectors we generated a graphical representation of the inferred spatio-temporal spread of each data set using the "spreadGraphic" function of the R package SERAPHIM (Dellicour et al. 2016). In addition, for each RABV data set, we also generated spatially referenced maximum clade credibility (MCC) consensus trees using TreeAnnotator 1.8.3 (Drummond et al. 2012).

## Step 2

Statistics of spatial dispersion for each data set were calculated from the information extracted in step 1. We estimated the spatial diffusion coefficient using two different approaches, $D_{original}$ and $D_{weighted}$. $D_{original}$ is an estimation of the average diffusion coefficient associated with each branch in the tree (Pybus et al. 2012), whereas $D_{weighted}$ is a weighted average across the tree of the ability to diffuse (Trovão et al. 2015).

$$D_{original} = \frac{1}{n}\sum_{i=1}^{n}\frac{d_i^2}{4t_i} \text{ and } D_{weighted} = \frac{\sum_{i=1}^{n} d_i^2}{\sum_{i=1}^{n} 4t_i} \quad (1)$$

In equations (1), $d_i$ and $t_i$ are, respectively, the geographic distance travelled (great-circle distance in km) and the time elapsed (in years) on each phylogeny branch. For a given tree, branches with short duration will have less of an impact on $D_{weighted}$ than on $D_{original}$, and therefore also on the resulting variance among $D_{weighted}$ values across all trees. While both statistics are measures of diffusivity, $D_{weighted}$ can potentially allow a better discrimination among epidemics with different diffusivity because it is associated with a smaller variance. In addition to the diffusion coefficients, we also generated graphs that display, for each analysis, the temporal evolution of the epidemic wavefront distance from the estimated root location.

## Step 3

Each of the vectors obtained in step 1 were assigned an "environmental distance", that is, a metric that was weighted according to the values of an environmental variable at each location (Dellicour, Rose, and Pybus 2016). We used two distinct path models to compute the environmental distance allocated to each phylogeny branch for a given environmental raster: (1) the *least-cost path* model, which uses a least-cost algorithm to determine the route taken between the start and end points (Dijkstra 1959), and (2) the *random walk path* model, which uses circuit theory to accommodate uncertainty in the route taken (McRae 2006; McRae et al. 2008). Note that for these path models, each environmental raster must be considered twice, once as a conductance factor (i.e., as a variable that facilitates movement) and once as a resistance factor (i.e., impedes movement).

The spatial heterogeneity of each environmental variable was defined using rasters. We investigated whether the following environmental rasters could explain variation in dissemination among rabies virus lineages: elevation, human population density, inaccessibility (time travel to nearest major cities of >50,000 inhabitants), major roads, main rivers and the key land cover variables for each study area (e.g., "barren vegetation", "croplands", "forests", "grasslands", "savannas", "shrublands", "urban areas", "wetlands"; land cover categorized according to the International Geosphere Biosphere Programme, IGBP). For RABV spread in Argentina and western Brazil, we also included in the analysis the cattle population density raster. The sources of each of the original raster files used in this study are listed in supplementary table S4, Supplementary Material online. Note that we did not systematically test every variable for each RABV spread. Some environmental factors were discarded for specific data sets because they were absent or virtually absent within the area under investigation. Population density rasters (human and cattle population densities) were also log-transformed in order to avoid providing an excessive weight to a few areas associated with high values. Furthermore, several transformed conductance and resistance values $\nu_C$ and $\nu_R$ were generated for each original raster, $\nu_C$ and $\nu_R = 1 + k^*\nu$, where $\nu$ is the original raster cell value. This transformation was not applied to raster cells with no data, which mostly correspond to bodies of water. The parameter $k$ allows us to define and test different strengths of raster cell conductance or resistance, relative to the conductance/resistance of a cell with a minimum value set to "1." We tested three different values for $k$: 10, 100 and 1,000. A detailed list of the different combinations of environmental rasters, associated $k$ values and path models tested for each spread is available in supplementary table S1, Supplementary Material online.

## Step 4

The correlation between the duration of each phylogeny branch, that is, branch length in units of time, and its weighted distances (see Step 3) was estimated for each of the 100 posterior trees in the five data sets (Dellicour, Rose, and Pybus 2016). Specifically, we estimated the statistic $Q = (R^2_{env} - R^2_{null})$, where $R^2_{env}$ is the coefficient of determination obtained when branch durations are regressed against environmental distances computed on the environmental raster, and $R^2_{null}$ is the coefficient of determination obtained when branch durations are regressed against environmental distances computed on a "null" raster, that is, the environmental raster with a value of "1" assigned to all the cells (except cells with no original data). The $Q$ statistic (previously

referred as "*D*" in Dellicour, Rose, and Pybus 2016) therefore represents how much variation in lineage movement is explained when spatial heterogeneity in the environmental variable is taken into account, above and beyond that explained by distance alone (Dellicour, Rose, and Pybus 2016). Therefore, when $Q > 0$, distances weighted according to a heterogeneous environmental raster are correlated more strongly with branch duration than distances computed on a "null" raster (which represents geographical distance alone). Since one *Q* value was calculated per sampled posterior tree, we then obtained 100 *Q* values for each combination of environmental factor, *k* parameter value and path model.

### Step 5

The statistical significance of the observed *Q* values was tested against a null model of no impact of the environmental factor. To generate an appropriate null distribution for *Q*, we used a randomization procedure implemented by Dellicour, Rose, and Pybus (2016) and Dellicour et al. (2016): phylogenetic node positions were randomized within the study area, under the constraint that branch lengths, tree topology and root position are unchanged. Each sampled posterior tree was randomized once to generate the equivalent value under the null hypothesis; this results in a null distribution of *Q* values that can be compared directly to the posterior distributions of observed *Q* values. We approximate Bayes factor (BF) support for the environmental factors through the ratio of the posterior odds over the prior odds for $Q_{observed} > Q_{randomized}$:

$$\mathrm{BF} = \frac{p_e}{1 - p_e} \Big/ \frac{0.5}{1 - 0.5} \quad (2)$$

where $p_e$ is the posterior probability that $Q_{observed} > Q_{randomized}$, that is, the frequency at which $Q_{observed} > Q_{randomized}$ in the samples from the posterior distribution. The prior odds is 1 because we have an equal prior expectation for $Q_{observed}$ and $Q_{randomized}$. The formal estimate of posterior predictive odds is analogous to computing BFs in situations in which two alternative hypotheses exist, e.g. the inclusion of rate parameters or predictors in BSSVS procedures (Bayesian stochastic search variable selection; see equation [6] in Lemey et al. 2009). As described in Jeffreys (1961), BF values higher than 10 and $10^{3/2}$ (31.62) are respectively considered as "strong" and "very strong" evidences of the statistical significance of $Q_{observed}$.

## Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

## Acknowledgments

We thank two anonymous reviewers as well as Bram Vrancken for their helpful comments on this study. S.D. is a postdoctoral research fellow funded by the Fonds Wetenschappelijk Onderzoek (FWO, Flanders, Belgium) and previously funded by the Wiener-Anspach Foundation. R.R. received funding from the Medical Research Council under a Methodology Research Fellowship grant agreement no. 99204. N.R.F. is funded by a Sir Henry Dale Fellowship (Wellcome Trust—Royal Society Grant 204311/Z/16/Z). MG is a Senior Research Associate of the Fonds National *de la Recherche Scientifique* (FNRS, Brussels, Belgium). HB and PL acknowledge funding by European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 278433-PREDEMICS. PL also acknowledges funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) under ERC Grant agreement no. 260864-ViralPhylogeography. O.G.P. received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 614725-PATHPHYLODYN.

## References

Baele G, Suchard MA, Rambaut A, Lemey P. 2017. Emerging concepts of data integration in pathogen phylodynamics. *Syst Biol*. 66(1):e47–e65.

Benavides JA, Valderrama W, Streicker DG. 2016. Spatial expansions and travelling waves of rabies in vampire bats. *Proc R Soc B* 283:1–9.

Biek R, Henderson JC, Waller LA, Rupprecht CE, Real LA. 2007. A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. *Proc Natl Acad Sci U S A*. 104:7993–7998.

Blanton JD, Dyer J, McBrayer J, Rupprecht CE. 2012. Rabies surveillance in the United States during 2011. *J Am Vet Med Assoc*. 241:712–722.

Bourhy H, Nakouné E, Hall M, Nouvellet P, Lepelletier A, Talbi C, Watier L, Holmes EC, Cauchemez S, Lemey P, et al. 2016. Revealing the micro-scale signature of endemic zoonotic disease transmission in an African urban setting. *PLoS Pathog*. 12(4):e1005525.

Bourhy H, Reynes JM, Dunham EJ, Dacheux L, Larrous F, Huong VTQ, Xu G, Yan J, Miranda MEG, Holmes EC. 2008. The origin and phylogeography of dog rabies virus. *J Gen Virol*. 89:2673–2681.

Carnieli P Jr, de Novaes Oliveira R, Macedo CI, Castilho JG. 2011. Phylogeography of rabies virus isolated from dogs in Brazil between 1985 and 2006. *Arch Virol*. 156:1007–1012.

de Thoisy B, Bourhy H, Delaval M, Pontier D, Dacheux L, Darcissac E, Donato D, Guidez A, Larrous F, Lavenir R, et al. 2016. Bioecological drivers of rabies virus circulation in a neotropical bat community. *PLoS Negl Trop Dis*. 10(1):e0004378.

Dellicour S, Rose R, Faria NR, Lemey P, Pybus OG. 2016. SERAPHIM: studying environmental rasters and phylogenetically-informed movements. *Bioinformatics* 32(20):3204–3206.

Dellicour S, Rose R, Pybus OG. 2016. Explaining the geographic spread of emerging epidemics: a framework for comparing viral phylogenies and environmental landscape data. *BMC Bioinform*. 17:1–12.

De Maio N, Wu C-H, O'Reilly KM, Wilson D. 2015. New routes to phylogeography: a Bayesian structured coalescent approximation. *PLoS Genet*. 11:e1005421.

Dijkstra EW. 1959. A note on two problems in connexion with graphs. *Numer Math*. 1:269–271.

Dodet B, Adjogoua EV, Aguemon AR, Amadou OH, Atipo AL, Baba BA, Bara Ada S, Boumandouki P, Bourhy H, Diallo MK, et al. 2008. Fighting rabies in Africa: The Africa Rabies Expert Bureau (AfroREB). *Vaccine* 26:6295–6298.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol*. 29:1969–1973.

Faria NR, Suchard MA, Rambaut A, Lemey P. 2011. Toward a quantitative understanding of viral phylogeography. *Curr Opin Virol*. 1:423–429.

Fusaro A, Monne I, Salomoni A, Angot A, Trolese M, Ferrè N, Mutinelli F, Holmes EC, Capua I, Lemey P, et al. 2013. The introduction of fox rabies into Italy (2008-2011) was due to two viral genetic groups with distinct phylogeographic patterns. *Infect Genet Evol*. 17:202–209.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 59(3):307–321.

Hampson K, Coudeville L, Lembo T, Sambo M, Kieffer A, Attlan M, Barrat J, Blanton JD, Briggs DJ, Cleaveland S, et al. 2015. Estimating the global burden of endemic canine rabies. *PLoS Negl Trop Dis*. 9(5):e0003709.

Hayman DTS, Johnson N, Horton DL, Hedge J, Wakeley PR, Banyard AC, Zhang S, Alhassan A, Fooks AR. 2011. Evolutionary history of rabies in Ghana. *PLoS Negl Trop Dis*. 5(4):e1001.

Holmes EC. 2004. The phylogeography of human viruses. *Mol Ecol*. 13(4):745–756.

Horton DL, McElhinney LM, Freuling CM, Marston DA, Banyard AC, Goharrriz H, Wise E, Breed AC, Saturday G, Kolodziejek J, et al. 2015. Complex epidemiology of a zoonotic disease in a culturally diverse region: phylogeography of rabies virus in the Middle East. *PLoS Negl Trop Dis*. 9(3):e0003569.

Jeffreys H. 1961. Theory of probability. 3rd ed. Oxford: Oxford University Press.

Knobel DL, Cleaveland S, Coleman PG, Fèvre EM, Meltzer MI, Miranda MEG, Shaw A, Zinsstag J, Meslin F-X. 2005. Re-evaluating the burden of rabies in Africa and Asia. *Bull World Health Organ*. 83:360–368.

Kuzmina NA, Lemey P, Kuzmin IV, Mayes BC, Ellison JA, Orciari LA, Hightower D, Taylor ST, Rupprecht CE. 2013. The phylogeography and spatiotemporal spread of South-Central skunk rabies virus. *PLoS One* 8(12):e82348.

Lemey P, Rambaut A, Drummond AJ, Suchard MA. 2009. Bayesian phylogeography finds its roots. *PLoS Comput Biol*. 5(9):e1000520.

Lemey P, Rambaut A, Welch JJ, Suchard MA. 2010. Phylogeography takes a relaxed random walk in continuous space and time. *Mol Biol Evol*. 27:1877–1885.

McRae BH. 2006. Isolation by resistance. *Evolution* 60:1551–1561.

McRae BH, Dickson BG, Keitt TH, Shah VB. 2008. Using circuit theory to model connectivity in ecology, evolution, and conservation. *Ecology* 89:2712–2724.

Melbourne BA, Hastings A. 2009. Highly variable spread rates in replicated biological invasions: fundamental limits to predictability. *Science* 325:1536–1539.

Mundt CC, Sackett KE, Wallace LD, Cowger C, Dudley JP. 2009. Long-distance dispersal and accelerating waves of disease: empirical relationships. *Am Nat*. 173:456–466.

Murray GGR, Wang F, Harrison EM, Paterson GK, Mather AE, Harris SR, Holmes MA, Rambaut A, Welch JJ. 2016. The effect of genetic structure on molecular dating and tests for temporal signal. *Methods Ecol Evol*. 7:80–89.

Navascués M, Depaulis F, Emerson BC. 2010. Combining contemporary and ancient DNA in population genetic and phylogeographical studies. *Mol Ecol Res*. 10:760–772.

Picard-Meyer E, Robardet E, Moroz D, Trotsenko Z, Drozhzhe Z, Biarnais M, Solodchuk V, Smreczak M, Cliquet F. 2012. Molecular epidemiology of rabies in Ukraine. *Arch Virol*. 157:1689–1698.

Piñero C, Dohmen F, Beltran F, Martinez L, Novaro L, Russo S, Palacios G, Cisterna DM. 2012. High diversity of rabies viruses associated with insectivorous bats in Argentina: presence of several independent enzootics. *PLoS Negl Trop Dis*. 6(5):e1635.

Pybus OG, Suchard MA, Lemey P, Bernardin FJ, Rambaut A, Crawford FW, Gray RR, Arinaminpathy N, Stramer SL, Busch MP, Delwart EL. 2012. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc Natl Acad Sci U S A*. 109:15066–15071.

Rambaut A, Lam TT, Max Carvalho L, Pybus OG. 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol*. 2(1):vew007.

Schluter D, Price T, Mooers AØ, Ludwig D. 1997. Likelihood of ancestor states in adaptive radiation. *Evolution* 51:1699–1711.

Seetahal JFR, Velasco-Villa A, Allicock OM, Adesiyun AA, Bissessar J, Amour K, Phillip-Hosein A, Marston DA, McElhinney LM, Shi M, et al. 2013. Evolutionary history and phylogeography of rabies viruses associated with outbreaks in Trinidad. *PLoS Negl Trop Dis*. 7(8):e2365.

Silva SR, Katz ISS, Mori E, Carnieli P, Vieira LFP, Batista HBCR, Chaves LB, Scheffer KC. 2013. Biotechnology advances: a perspective on the diagnosis and research of rabies virus. *Biologicals* 41:217–223.

Smith DL, Lucey B, Waller LA, Childs JE, Real LA. 2002. Predicting the spatial dynamics of rabies epidemics on heterogeneous landscapes. *Proc Natl Acad Sci U S A*. 99:3668–3672.

Talbi C, Holmes EC, de Benedictis P, Faye O, Nakouné E, Gamatié D, Diarra A, Elmamy BO, Sow A, Adjogoua EV, et al. 2009. Evolutionary history and dynamics of dog rabies virus in western and central Africa. *J Gen Virol*. 90:783–791.

Talbi C, Lemey P, Suchard MA, Abdelatif E, Elharrak M, Jalal N, Faouzi A, Echevarría JE, Morón SV, Rambaut A, et al. 2010. Phylodynamics and human-mediated dispersal of a zoonotic virus. *PLoS Pathog*. 6(10):e1001166.

Tohma K, Saito M, Kamigaki T, Tuason LT, Demetria CS, Orbina JRC, Manalo DL, Miranda ME, Noguchi A, Inoue S, et al. 2014. Phylogeographic analysis of rabies viruses in the Philippines. *Infect Genet Evol*. 23:86–94.

Torres C, Lema C, Gury Dohmen F, Beltran F, Novaro L, Russo S, Freire MC, Velasco-Villa A, Mbayed VA, Cisterna DM. 2014. Phylodynamics of vampire bat-transmitted rabies in Argentina. *Mol Ecol*. 23:2340–2352.

Troupin C, Dacheux L, Tanguy M, Sabeta C, Blanc H, Bouchier C, Vignuzzi M, Duchene S, Holmes EC, Bourhy H. 2016. Large-scale phylogenomic analysis reveals the complex evolutionary history of rabies virus in multiple carnivore hosts. *PLoS Pathog*. 12:e1006041.

Trovão NS, Suchard MA, Baele G, Gilbert M, Lemey P. 2015. Bayesian inference reveals host-specific contributions to the epidemic expansion of Influenza A H5N1. *Mol Biol Evol*. 32(12):3264–3275.

Vieira LFP, Pereira SRFG, Carnieli P Jr, Tavares LCB, Kotait I. 2013. Phylogeography of rabies virus isolated from herbivores and bats in the Espírito Santo State, Brazil. *Virus Genes* 46:330–336.

Warrell MJ, Warrell DA. 2004. Rabies and other lyssavirus diseases. *Lancet* 363:959–969.

World Health Organization. 2005. WHO expert consultation on rabies. First report, Geneva: Technical Report Series. 931.

Zieger U, Marston DA, Sharma R, Chikweto A, Tiwari K, Sayyid M, Louison B, Goharriz H, Voller K, Breed AC, et al. 2014. The phylogeography of rabies in Grenada, West Indies, and implications for control. *PLoS Negl Trop Dis*. 8(10):e3251.