Received: 23 November 2018 | Revised: 4 August 2019 | Accepted: 5 August 2019

DOI: 10.1111/mec.15222

ORIGINAL ARTICLE

MOLECULAR ECOLOGY WILEY

Using phylogeographic approaches to analyse the dispersal history, velocity and direction of viral lineages — Application to rabies virus spread in Iran

Simon Dellicour[1,2] | Cécile Troupin[3] | Fatemeh Jahanbakhsh[4] | Akram Salama[5] | Siamak Massoudi[6] | Madjid K. Moghaddam[6] | Guy Baele[1] | Philippe Lemey[1] | Alireza Gholami[4] | Hervé Bourhy[3]

[1]Department of Microbiology, Immunology and Transplantation, Rega Institute, Laboratory for Clinical and Epidemiological Virology, KU Leuven, Leuven, Belgium

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

[3]Unit Lyssavirus Epidemiology and Neuropathology, WHO Collaborating Centre for Reference and Research on Rabies, Institut Pasteur, Paris, France

[4]WHO Collaborating Centre for Reference and Research on Rabies, Pasteur Institute of Iran, Tehran, Iran

[5]Department of Medicine and Infectious Diseases, Faculty of Veterinary Medicine, University of Sadat City, Sadat City, Egypt

[6]Department of Environment, Wildlife Diseases Group, Wildlife Bureau, Tehran, Iran

Correspondence
Simon Dellicour, Department of Microbiology, Immunology and Transplantation, Rega Institute, Laboratory for Clinical and Epidemiological Virology, KU Leuven, Leuven, Belgium.
Email: simon.dellicour@kuleuven.be

Hervé Bourhy, Unit Lyssavirus Epidemiology and Neuropathology, WHO Collaborating Centre for Reference and Research on Rabies, Institut Pasteur, Paris, France.
Email: herve.bourhy@pasteur.fr

Present address
Cécile Troupin, Institut Pasteur de Guinée, BP 1147 Université Gamal Abdel Nasser de Conakry (UGANC), Conakry, Guinée

Funding information
Fonds National de la Recherche Scientifique; Fonds Wetenschappelijk Onderzoek; Internal Funds KU Leuven, Grant/Award Number: C14/18/094; European Union Seventh Framework Programme, Grant/Award Number: 278433; European Research Council; Wellcome Trust, Grant/Award Number: 206298/Z/17/Z and BSV3-0019; Bijzonder Onderzoeksfonds, Grant/Award Number: OT/14/115; Research Foundation—Flanders, Grant/Award Number: G066215N,

Abstract

Recent years have seen the extensive use of phylogeographic approaches to unveil the dispersal history of virus epidemics. Spatially explicit reconstructions of viral spread represent valuable sources of lineage movement data that can be exploited to investigate the impact of underlying environmental layers on the dispersal of pathogens. Here, we performed phylogeographic inference and applied different post hoc approaches to analyse a new and comprehensive data set of viral genomes to elucidate the dispersal history and dynamics of rabies virus (RABV) in Iran, which have remained largely unknown. We first analysed the association between environmental factors and variations in dispersal velocity among lineages. Second, we present, test and apply a new approach to study the link between environmental conditions and the dispersal direction of lineages. The statistical performance (power of detection, false-positive rate) of this new method was assessed using simulations. We performed phylogeographic analyses of RABV genomes, allowing us to describe the large diversity of RABV in Iran and to confirm the cocirculation of several clades in the country. Overall, we estimate a relatively high lineage dispersal velocity, similar to previous estimates for dog rabies virus spread in northern Africa. Finally, we highlight a tendency for RABV lineages to spread in accessible areas associated with high human population density. Our analytical workflow illustrates how phylogeographic approaches can be used to investigate the impact of environmental factors on several aspects of viral dispersal dynamics.

S. Dellicour and C. Troupin are equal contributions.

G0D5117N and G0B9317N; Agence Nationale de la Recherche

KEYWORDS
dogs, Iran, molecular epidemiology, phylogeography, rabies

1 | INTRODUCTION

RNA viruses are characterized by high rates of evolutionary change, which results from fast replication with error-prone RNA polymerases and a combination of natural selection and sometimes recombination (Holmes, 2004). The resulting high genetic variability of RNA viruses underpins their ability to adapt to changing environments (Kühnert, Wu, & Drummond, 2011), including their emergence and spread in new host species or ecological niches. In addition, spatial heterogeneity such as topographical features (e.g., rivers, mountains, deserts) and socio-economical characteristics (e.g., road networks, mode of commercial exchanges, levels of education and awareness) together with control measures can also affect viral spread by impeding or facilitating host movement and influencing host distributions, densities and susceptibility (Brunker et al., 2018; Dellicour et al., 2017). Therefore, the dynamic and inherently spatial dimension of epidemiological processes present unique challenges to studying and managing the spread of emerging and re-emerging infectious diseases.

The burgeoning field of landscape epidemiology aims at examining interactions between landscape heterogeneity and the associated environmental processes that drive the spread and persistence of diseases and the evolution of viruses (Grenfell, Bjørnstad, & Kappey, 2001; Keeling et al., 2001; Pybus, Tatem, & Lemey, 2015; Real & Biek, 2007). In this respect, the recent and rapidly increasing availability of viral genomic data provides an unprecedented opportunity to develop and apply new evolutionary approaches to explore how evolutionary and spatial processes give rise to geographical distributions of RNA viruses. In particular, phylogeographic analysis of genetic sequences sampled in two-dimensional space has emerged as a useful approach to study viral dispersal histories and dynamics in a spatially explicit context, that is without the need to delineate discrete sampling locations. The implementation of such a continuous phylogeographic model (Lemey, Rambaut, Welch, & Suchard, 2010; Pybus et al., 2012) allows inferring spatially and temporally referenced phylogenies while accommodating variation in dispersal velocity among branches (Baele, Dellicour, Suchard, Lemey, & Vrancken, 2018). These annotated trees can in turn be used to perform landscape phylogeographic analyses, that is to use phylogenetically informed movements to investigate the factors impacting virus dispersal, thereby opening up new opportunities to acquire a better understanding of how environmental conditions impact the spatial dynamics of rapidly evolving populations of viruses (Brunker et al., 2018; Dellicour, Vrancken, Trovão, Fargette, & Lemey, 2018). Recent years have seen the development of methods to investigate the impact of environmental factors on the lineage dispersal velocity of viruses (Dellicour, Rose, & Pybus, 2016; Jacquot, Nomikou, Palmarini, Mertens, & Biek, 2017) and the impact of landscape features acting as potential barriers that decrease the dispersal frequency between geographical areas (Dellicour, Baele, et al., 2018).

While analytical frameworks have been previously developed to analyse the association between environmental factors and lineage dispersal velocity and frequency, little attention has been paid on testing the tendency of lineages to remain in and/or disperse towards specific environmental conditions. Nonetheless, an environmental factor could not have any impact on the dispersal velocity of a virus spread but instead determines the probability of viral lineages to disperse towards given areas. Methods investigating the potential impact of external factors on the dispersal velocity and direction of lineages would thus represent complementary approaches that could be applied to identify the drivers of viral spreads.

Because of the impact on human and wildlife populations, rabies virus (RABV) spread represents an important study case for how understanding the impact of external factors may ultimately inform the prevention, prediction and control of the disease, especially when these analyses focus on the goal of rabies elimination (WHO, 2018). Rabies is a widespread zoonotic disease distributed worldwide and remains the disease with the highest case fatality rate in animals and humans (nearly 100% in dogs and in humans once symptoms develop) and an incidence in humans of approximately 59,000 cases per year (Hampson et al., 2015). In Africa and Asia, almost all human rabies cases are caused by infections with dog RABV, and the majority occurs due to lack of rabies vaccination in domestic dog populations (Dodet et al., 2008; Knobel et al., 2005). In all cases, transmission occurs through contact between infectious and susceptible hosts (through bites/scratches or through direct contact of mucosa with saliva from infected animals). Environmental heterogeneity and host distribution, density and contact rates all have a strong impact on the spread and maintenance of the disease. This has led to a growing demand for analytical tools to analyse spatially resolved genetic data together with epidemiological and environmental data.

Phylogeographic approaches have previously been used to study the dynamics and spread of RABV at both large (Biek, Henderson, Waller, Rupprecht, & Real, 2007; Horton et al., 2015; Kuzmina et al., 2013; Talbi et al., 2009; Troupin et al., 2016) and small geographical scales (Bourhy et al., 2016; Zinsstag et al., 2017). Here, we use continuous phylogeographic analyses as well as related post hoc approaches to study the dispersal history of RABV spread in Iran and to investigate the environmental factors impacting its dispersal dynamics. Iran is located in the centre of Eurasia and shares almost 5,500 kilometres of borders in total with eight countries. For centuries, rabies has been present in Iran and has been a notifiable disease since several decades (Baltazard & Ghodssi, 1954). Dogs (*Canis lupus familiaris*) are the primary source of human rabies in Iran, but many other rabid wild animals such as Blanford's foxes (*Vulpes cana*), golden jackal (*Canis aureus*),

mongooses (*Herpestes auropunctatus*, *Herpestes edwardsii*) and wolves (*Canis lupus*) are reported with high frequency throughout the country yearly, suggesting a complex epidemiological scenario (Janani et al., 2008; Picot et al., 2017; Seimenis, 2008). As a consequence, Iran is characterized by one of the highest annual rates of postexposure prophylaxis (PEP; 22/10000 by 2018) provided to exposed patients worldwide (Dehghani, Sharif, Madani, Kashani, & Sharif, 2016; Farahtaj, Fayaz, Howaizi, Biglari, & Gholami, 2014). Despite this high burden, epidemiological studies on rabies in Iran have so far been limited to epidemiological studies of animal bites (Charkazi et al., 2013; Dehghani et al., 2016; Feizhaddad, Kassiri, Lotfi, & Hoseini, 2014) and to small-scale molecular epidemiological investigations (Nadin-Davis, Simani, Armstrong, Fayaz, & Wandeler, 2003). More comprehensive data to define the distribution, reservoirs and dispersal dynamics of rabies in Iran are hence important in order to gain a better understanding of the role of environmental heterogeneity and potentially different animal host species in the maintenance and spread of the disease as well as the development of control measures.

The overall goal of the present study is to describe and apply a comprehensive workflow of landscape phylogeography, including a novel approach to investigate the impact of environmental factors on the dispersal direction of viral lineages. We apply this comprehensive workflow on a new data set of viral sequences to analyse the RABV dispersal in Iran. Specifically, we aim at (a) identifying the different RABV lineages spreading in Iran, (b) inferring the dispersal history of these distinct lineages, (c) comparing their dispersal velocity with other instances of RABV spread across the world, (d) investigating the impact of environmental factors on the dispersal velocity and direction of viral lineages in Iran, and (e) testing the performance of our new approach using simulations.

2 | MATERIALS AND METHODS

2.1 | Virus sampling

We analysed 101 nearly complete genome sequences from RABV isolates, collected in Iran between 2008 and 2015. Samples were obtained through a collaborative programme of passive public surveillance following protocols put in place by the Iranian Department of Environment and the Iranian Veterinary Organization, two institutions collaborating on the present study: brain samples of domestic, farm or wild animals suspected of rabies are collected across the entire countries by the Health Network of the Iranian Veterinary Organization and sent to the WHO Collaborating Center for References and Research on Rabies (at the Pasteur Institute of Iran in Teheran) for analyses. New samples analysed in the present study were selected at random but while trying to maximize the spatial coverage of the country. Furthermore, we also included a balanced proportion of nondomestic dog samples in our final data set (41%) in order to investigate, but also to avoid underestimating, the role of wildlife species in spreading the virus on the study area. The lower proportion of genomes originated from eastern regions of the country simply reflects the lower number of samples coming from these areas. All brain samples were sent to the reference laboratory along with a form that contains information on sampling origin, the GPS coordinates of which were later retrieved.

2.2 | RNA extraction and next-generation sequencing

Total RNA was extracted using TRIzol (Ambion) according to the manufacturer's instructions from brain samples. RNA was then reverse-transcribed using Superscript III reverse transcriptase with random hexamers (Invitrogen) according to the manufacturer's instructions. The complete viral genome (excluding the 3′ and 5′ extremities, corresponding to the leader and the trailer regions, respectively) of 101 new isolates was amplified with six overlapping PCR fragments by using the Phusion polymerase (Thermo Fisher) as previously described (Troupin et al., 2016). After electrophoresis, each PCR fragment was independently purified using the NucleoSpin Gel and PCR clean-up kit (Macherey-Nagel) and quantified using PicoGreen dsDNA quantification kit (Invitrogen). For each sample, all six PCR fragments were pooled with equimolar proportions to obtain 500 ng of dsDNA. dsDNA libraries were constructed using the Nextera XT kit (Illumina) and sequenced using a 2 × 150 nucleotides paired-end strategy on the NextSeq500 platform.

2.3 | Genome sequence analyses

All reads were preprocessed to remove low-quality or artefactual bases. Library adapters at 5′ and 3′ ends and base pairs with a Phred quality score <25 were trimmed using AlienTrimmer as implemented in Galaxy (Criscuolo & Brisse, 2013; https://research.pasteur.fr/en/tool/pasteur-galaxy-platform). Reads with length lower than 75 bp (base pairs) after these preprocessing steps or those containing >20% of bp with a Phred score of <25 were discarded. The filtered reads were mapped to specific complete genome sequences: isolates 91047FRA and NNV-RAB-H (with Accession nos. KX148127 and EF437215) for the cosmopolitan- and Arctic-related clade viruses, respectively. For that purpose, we used the CLC Genomics Assembly Cell in Galaxy (http://www.clcbio.com/products/clc-assembly-cell). The majority nucleotide (>50%) at each position with generally a minimum coverage of 200 was used to generate the consensus sequence. All consensus sequences were manually inspected for accuracy, such as the presence of intact open reading frames, using BIOEDIT (Hall, 1999). A multiple sequence alignment was constructed using CLUSTALW2 with default parameters (Larkin et al., 2007) implemented in Galaxy and manually adjusted when necessary. In addition to the 101 nearly full-length genome sequences generated in the present study, we also included in our data set 8 full-length genome sequences of Iranian isolates collected in Iran between 1974 and 1996, which were previously sequenced by Troupin et al. (2016). Sampling data and GenBank Accession nos. are summarized in Table S1, and a sampling map is displayed in Figure 1.

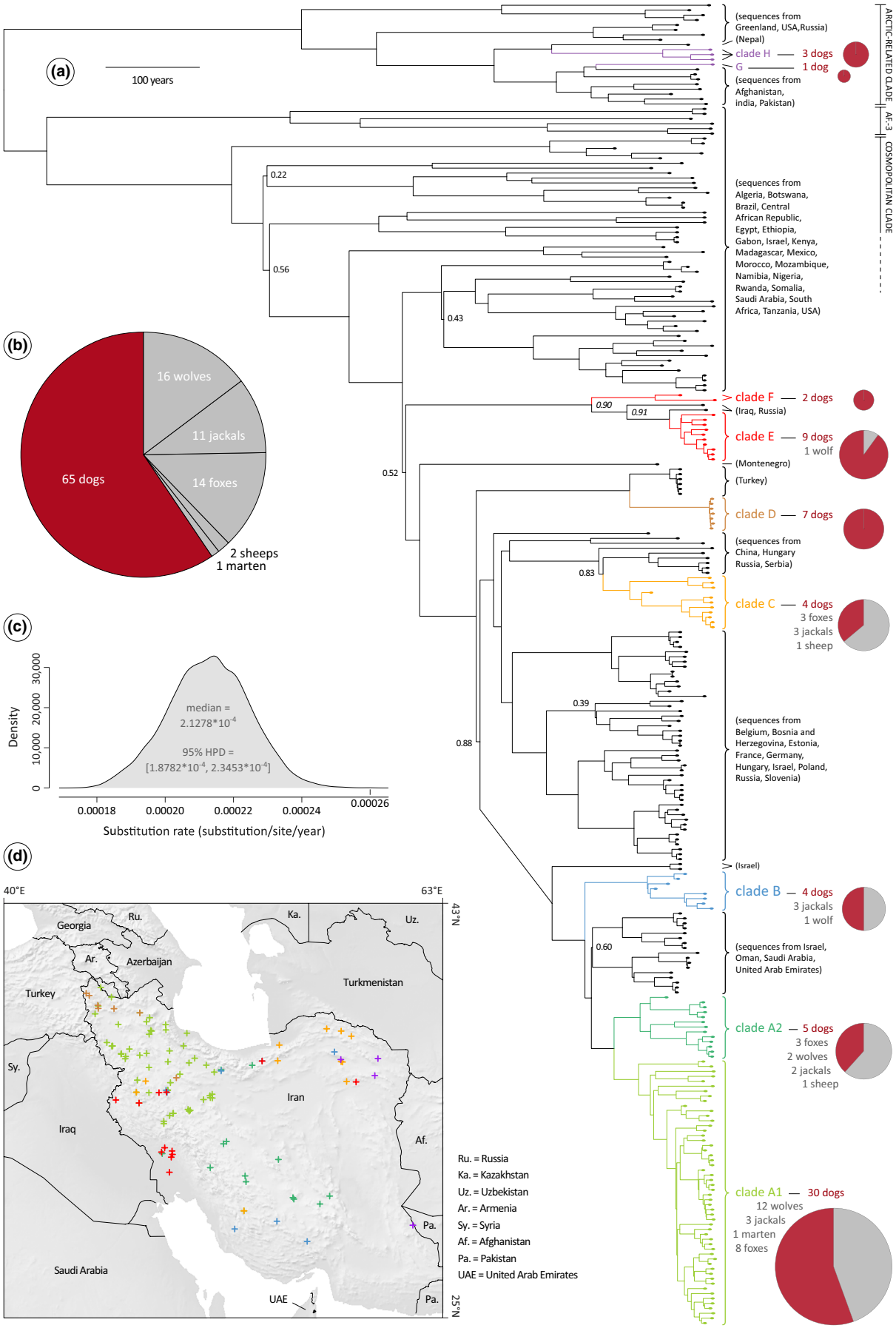
(a)
100 years
(sequences from Greenland, USA,Russia)
(Nepal)
clade H — 3 dogs
G — 1 dog
(sequences from Afghanistan, india, Pakistan)
ARCTIC-RELATED CLADE
AF-3
COSMOPOLITAN CLADE
0.22
0.56
0.43
(sequences from Algeria, Botswana, Brazil, Central African Republic, Egypt, Ethiopia, Gabon, Israel, Kenya, Madagascar, Mexico, Morocco, Mozambique, Namibia, Nigeria, Rwanda, Somalia, Saudi Arabia, South Africa, Tanzania, USA)
clade F — 2 dogs
0.90
0.91
(Iraq, Russia)
clade E — 9 dogs
1 wolf
(Montenegro)
0.52
(Turkey)
clade D — 7 dogs
(sequences from China, Hungary Russia, Serbia)
0.83
clade C — 4 dogs
3 foxes
3 jackals
1 sheep
0.39
0.88
(sequences from Belgium, Bosnia and Herzegovina, Estonia, France, Germany, Hungary, Israel, Poland, Russia, Slovenia)
(Israel)
clade B — 4 dogs
3 jackals
1 wolf
0.60
(sequences from Israel, Oman, Saudi Arabia, United Arab Emirates)
clade A2 — 5 dogs
3 foxes
2 wolves
2 jackals
1 sheep
clade A1 — 30 dogs
12 wolves
3 jackals
1 marten
8 foxes
(b)
16 wolves
11 jackals
14 foxes
65 dogs
2 sheeps
1 marten
(c)
Density
0
10,000
20,000
30,000
median = 2.1278*10^-4
95% HPD = [1.8782*10^-4, 2.3453*10^-4]
0.00018
0.00020
0.00022
0.00024
0.00026
Substitution rate (substitution/site/year)
(d)
40°E
63°E
43°N
25°N
Georgia
Ru.
Ka.
Uz.
Ar.
Azerbaijan
Turkey
Turkmenistan
Sy.
Iran
Iraq
Af.
Pa.
Saudi Arabia
UAE
Ru. = Russia
Ka. = Kazakhstan
Uz. = Uzbekistan
Ar. = Armenia
Sy. = Syria
Af. = Afghanistan
Pa. = Pakistan
UAE = United Arab Emirates

FIGURE 1 (a) Part of the maximum clade credibility (MCC) tree estimated from the discrete phylogeographic analysis performed to identify independent RABV introductions in Iran (see Figure S1 for the entire tree). Branch support is only reported for internal nodes associated with a posterior probability <0.95 (except for nodes connecting tip nodes for which lower support is not reported here). In addition, we indicate in italic the only two posterior probabilities lower than 0.95 associated with internal nodes for the ancestral states' reconstruction (involving only two possible discrete locations, i.e., "Iran" and "other"). For each clade, we also report the number of times an RABV sequence has been sampled in a given host, as well as the proportion of sequences sampled in dogs. "AF.3" refers to clade "Africa-3" (Troupin et al., 2016). (b) Total number of sequences sampled in each host. (c) Posterior distribution of substitution rate estimated from the same BEAST analysis. (d) Sampling map coloured by clades

2.4 | Discrete phylogeographic analysis

In order to identify the different RABV clades circulating and the potential introductions of the virus in Iran, we first performed a discrete phylogeographic analysis involving previously sequenced genomes that did not originate from Iran. This discrete phylogeographic analysis was performed in BEAST 1.10 (Lemey, Rambaut, Drummond, & Suchard, 2009; Suchard et al., 2018). For this analysis, we specified a simple constant population size coalescent prior, a GTR + I + Γ4 model of nucleotide substitution and a relaxed (uncorrelated log-normal) molecular clock. In addition to the new Iranian RABV genomes, we included a large number of non-Iranian genomes from a previous large-scale phylogenomic study (Troupin et al., 2016) as well as three full-length genomes of new isolates from Egypt (GenBank Accession nos. MK760768–MK60770). In line with the objective of this phylogeographic exploration, we only specified two possible location states: "Iran" and "other." In order to enhance computation speed, all BEAST analyses were performed using the BEAGLE library (Ayres et al., 2012).

2.5 | Continuous phylogeographic inference

The history of RABV lineage dispersal in Iran was inferred using the continuous phylogeographic method implemented in BEAST 1.10 (Lemey et al., 2010; Suchard et al., 2018). We performed a distinct continuous phylogeographic analysis for each Iranian clade identified by the discrete phylogeographic analysis. For these clade-specific analyses, we used a flexible Bayesian skygrid coalescent model (Gill et al., 2013) as well as a relaxed random walk diffusion model (RRW; Lemey et al., 2010). Because the low numbers of genomes involved in these analyses did not allow to adequately inform a molecular clock model, we used an informative prior on the substitution rate to obtain precise and realistic estimates of time-dependent estimates such as branch durations or the time of the most recent common ancestor (TMRCA; Jung et al., 2012). This informative prior was here set up as normal distribution, of which the mean and standard deviation were set according to posterior estimates of the overall discrete phylogeographic analysis. The MCMC (Markov chain Monte Carlo) analyses were run until adequate effective sample size (ESS) values were obtained (ESS >200; Drummond & Bouckaert, 2015): 100 million iterations for clades A, B and E–F, 500 million iterations for clade C, and 1 billion iterations for clade D. We used the program TRACER version 1.7 (Rambaut, Drummond, Xie, Baele, & Suchard, 2018) to examine ESS values, to determine the number of trees to discard as burn-in and to obtain highest posterior density (HPD) intervals for estimated parameters. Maximum clade credibility (MCC) trees were summarized using TREEANNOTATOR 1.10 (Suchard et al., 2018) and visualized with FIGTREE 1.7 (www.tree.bio.ed.ac.uk/software/figtree).

The spatio-temporal information contained in the inferred phylogenetic trees was then extracted using the "seraphim" R package (Dellicour, Rose, Faria, Lemey, & Pybus, 2016). After having discarded 10% of sampled trees as burn-in, we extracted the spatio-temporal information from the remaining subset of 900 trees sampled from the posterior distribution of trees inferred for each Iranian clade. After this extraction step, phylogenetic branches can be treated as conditionally independent movement vectors (Pybus et al., 2012). We also used the R package "seraphim" to estimate dispersal statistics based on these extracted movement vectors. We estimated the mean branch dispersal velocity, the weighted branch dispersal velocity, the mean diffusion coefficient as originally defined by Pybus et al. (2012) and the weighted diffusion coefficient as defined by Trovão et al. (2015). While the mean branch velocity and diffusion coefficient are estimates of the dispersal velocity and of the diffusion coefficient averaged over all tree branches respectively, their weighted average counterparts involve a weighting by branch time. As detailed in Dellicour et al. (2017), for a given tree, branches with short duration will have less of an impact on the weighted metrics, resulting in lower-variance estimates. Therefore, the weighted statistics are particularly useful when aiming to discriminate between different dispersal measures among data sets. On the other hand, the nonweighted metrics are useful when investigating the heterogeneity in lineage dispersal velocity or diffusion within a specific spread (see also the related "seraphim" tutorial for more details about the different dispersal metrics). All these dispersal metrics were summarized separately for each clade and for all the clades combined, that is collecting all the movements vectors extracted from the distinct continuous phylogeographic inferences.

2.6 | Investigating the impact of environmental factors on lineage dispersal velocity

Our analytical framework investigating the impact of environmental factors on lineage dispersal velocity comprised four distinct steps, which were previously applied in other studies (e.g., Dellicour, Rose, & Pybus, 2016; Laenen et al., 2016). The first step consisted in extracting the spatio-temporal information embedded in a collection of posterior trees obtained using continuous phylogeographic inference. Due to computational time limitations, we restricted this analysis to 100 trees sampled from each post-burn-in posterior distribution inferred using clade-specific phylogeographic inference.

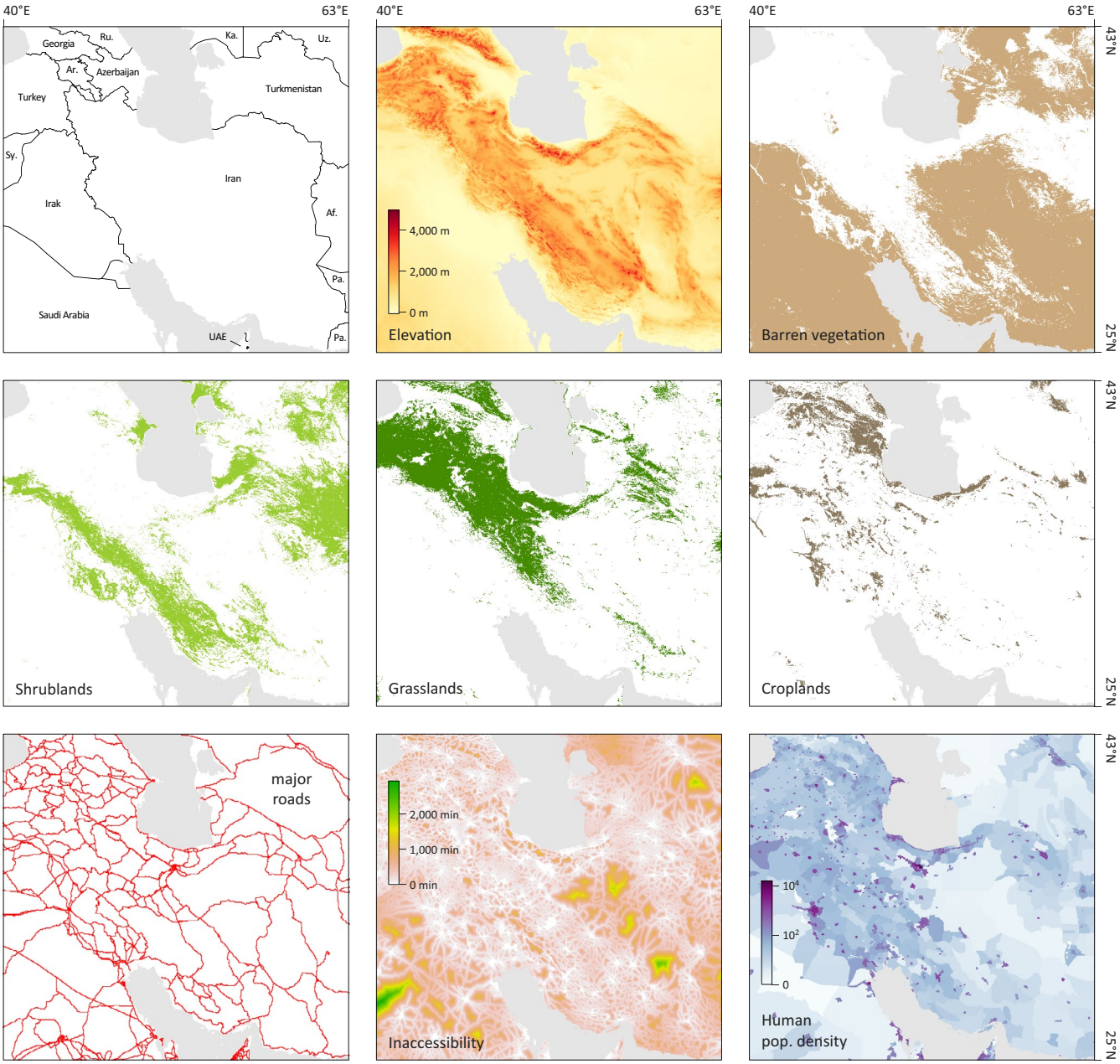

FIGURE 2 Environmental variables that were tested in the analysis of the RABV clades identified in Iran. The first map only displays the geopolitical context of the study area. On this map, "Ru." refers to Russia, "Ka." refers to Kazakhstan, "Uz." refers to Uzbekistan, "Ar." refers to Armenia, "Sy." refers to Syria, "Af." refers to Afghanistan, "Pa." refers to Pakistan and "UAE" refers to United Arab Emirates. The inaccessibility raster is in units of time (min) and indicates the time to travel to the nearest major city of at least 50,000 inhabitants

Specifically, each phylogenetic branch was considered a vector defined by its start and end locations (latitude and longitude), and its start and end dates (in decimal units). These phylogeny branches therefore represent conditionally independent viral lineage dispersal events (Pybus et al., 2012). Following this extraction step, all movement vectors extracted from the different Iranian clades were gathered for subsequent analyses.

In a second step, these movement vectors were assigned an environmental distance, that is a spatial distance that was weighted according to the values of an underlying environmental raster. In practice, we used two different path models to compute the environmental distances: (a) the least-cost path model, which uses a least-cost algorithm to determine the route taken between locations (Dijkstra, 1959), and (b) the Circuitscape path model, which uses circuit theory to accommodate uncertainty in the route taken (McRae, 2006). In this study, we investigated the impact of the following environmental factors (Figure 2): elevation, the most represented land cover variables on the study area (i.e., barren vegetation, shrublands, grasslands and croplands; land cover categorized according to the International Geosphere Biosphere Program, IGBP), major roads, human population density and inaccessibility (travel time to the nearest major city of >50,000 inhabitants; see Table S2 for the sources of the original raster

files). All factors were tested as potential conductance factors (i.e., factors facilitating movement) and as potential resistance factors (i.e., factors impeding movement). Further, several distinct rasters were generated by transforming original raster cell values with the following formula: $v_t = 1 + k^*(v_o/v_{max})$, where v_t and v_o are the transformed and original raster cell values, and v_{max} the maximum raster cell value recorded in the raster. The rescaling parameter k here allowed 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 environmental factor, we tested three different values for k (i.e., 10, 100 and 1,000).

In a third step, we estimated the correlation between branch durations and environmental distances with the statistic Q defined as the difference between two coefficients of determination (R^2): (a) R^2 obtained when branch durations are regressed against environmental distances computed on the environmental raster, and (b) R^2 obtained when branch durations are regressed against environmental distances computed on a null raster, that is an environmental raster with a value of "1" assigned to all the cells. 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 estimated per sampled tree, we thus obtained a distribution of Q values for each combination of environmental factor, k parameter value and path model. A variable can only be considered as potentially explanatory if both its distribution of regression coefficients and associated distribution of Q values are positive (Jacquot et al., 2017). In the final step, the statistical support of each Q distribution was evaluated against a null distribution generated by a randomization procedure and formalized as an approximated Bayes factor value (Dellicour et al., 2017).

Our framework relies on univariate testing of environmental factors (Dellicour, Rose, & Pybus, 2016) mostly because performing multivariate analyses would require dealing with multicollinearity issues among covariates, a notable limitation when co-analysing environmental distances computed on distinct environmental layers. Indeed, no matter the path model used to compute environmental distances between locations, such distances computed among the same locations but on different layers will remain more or less correlated because the distances inherently correlate with the geographical distances between locations (Dellicour, Vrancken, et al., 2018). Consequentially, the analytical strategy selected in the present workflow mainly served to analyse the different environmental factors independently, which were then compared in the context of a discussion. All the scripts related to this approach are available, along with tutorials and example files, in the R package "seraphim" (Dellicour, Rose, Faria, et al., 2016).

2.7 | Investigating the impact of environmental factors on lineage dispersal direction

We here introduce a new analytical framework to investigate the impact of environmental factors on the dispersal direction of viral lineages. Specifically, the idea is to test if lineages tended to remain in and/or tended to disperse towards particular environmental conditions. For instance, one may be interested in testing if a RABV lineage tended to disperse to areas associated with higher human population density, a variable that is at least partially correlated with domestic dog population density (Hampson et al., 2015). In practice, this framework consists of similar steps as the approach used to analyse the impact of environmental factors on lineage dispersal velocity (see above), but instead of comparing branch durations and associated environmental distances, we here focused on the environmental conditions at the locations of tree nodes.

For this purpose, we here propose to compute and test two distinct metrics: (a) E defined as the mean of the environmental values extracted at the nodes' positions, and (b) R defined as the proportion of branches for which the environmental value recorded at the oldest node position is higher than the environmental value recorded at the youngest node position. While E measures the tendency of tree nodes to remain located in lower/higher environmental values, R measures the tendency of lineages to disperse towards lower/higher environmental values. These two metrics are computed for each tree in the posterior sample, and we therefore obtain posterior distributions for E and R. Analogous to the final step of the first procedure described above, these posterior distributions were compared to null distributions of the same metrics computed after having randomized phylogenetic node positions within the study area, under the constraint that branch lengths, tree topology and root position remain unchanged. This approach only requires one randomization per sampled tree and leads to the approximation of a Bayes factor (BF) support for each metric. For a particular environmental factor e tested as a factor attracting lineages, the Bayes factor BF_e associated with the statistic E is approximated by the posterior odds that $E_{estimated} > E_{randomized}$ divided by the equivalent prior odds (the prior probability for $E_{estimated} > E_{randomized}$ is considered to be 0.5):

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

where p_e is the posterior probability that $E_{estimated} > E_{randomized}$, that is the frequency at which $E_{estimated} > E_{randomized}$ in the samples from the posterior distribution. The prior odds is 1 because we can assume an equal prior expectation for $E_{estimated}$ and $E_{randomized}$. The formal estimate of posterior predictive odds is analogous to computing Bayes factors in case two alternative hypotheses exist, such as for the inclusion of rate parameters or predictors in BSSVS procedures (Dellicour et al., 2017; Lemey et al., 2009). Alternatively, if the environmental factor was tested as a factor repulsing lineages, BF_e would be approximated by the posterior odds that $E_{estimated} < E_{randomized}$ divided by the equivalent prior odds. The same approach was used to approximate Bayes factor support for the R statistic. Whether the environmental factor is tested as a factor attracting or repulsing lineages, the posterior BF_e for R is approximated by the posterior odds that $R_{estimated} < R_{randomized}$ (attracting lineages) or that $R_{estimated} > R_{randomized}$ (repulsing lineages) divided by the equivalent prior odds. Following the same logic used for investigating the

impact of environmental factors on lineage dispersal velocity, each environmental factor was tested as a potential driver and as a potential impeder of virus dispersal. This new approach has been added to the R package "seraphim" (Dellicour, Rose, Faria, et al., 2016; evolve.zoo.ox.ac.uk/Evolve/Seraphim.html), along with a related tutorial. In addition to the RABV data set from Iran, we also used this new procedure to analyse a RABV data set from northern Africa previously published by Talbi et al. (2010) and also analysed by Dellicour et al. (2017) in a study focusing on lineage dispersal velocity. The re-analysis of this alternative domestic dog RABV data set was performed to allow a comparison of the environmental factors associated with the dispersal events of viral lineages in these two different regions.

2.8 | Assessing the performance of our approach to investigate lineage dispersal direction

The approach described above and developed to investigate the impact of environmental factors on lineage dispersal direction was tested on data sets simulated according to some control scenarios. To simulate these scenarios, we built on a procedure initially developed to simulate a forward-in-time relaxed random walk diffusion (RRW) process along branches of trees obtained by continuous phylogeographic inference conditional on the sampled precision matrix parameters and location at the root node (Dellicour, Baele, et al., 2018). We adapted this simulation approach in two distinct procedures to condition the simulation of each dispersal event along a phylogeny branch on the environmental values at the simulated node position. In the first simulation procedure, simulated node positions were more likely to fall in raster cells associated with higher (or alternatively lower) environmental values. In the second procedure, simulated node positions were more likely to fall in raster cells that maximize the positive (or alternatively negative) difference between the cell values at the ending (youngest) and starting (oldest) node positions. For a given branch, for which the position of the oldest node was already simulated (or fixed in the case of the root), the position of the youngest node was simulated 100 times and the environmental values below that nodes were recorded. In the first procedure, the probability to select one specific simulated position in the raster cell *i* was defined as follows:

$$p_i = v_i / v_{\text{tot}}$$

where v_i is the environmental value in cell *i*, and v_{tot} the sum of the raster cell values extracted at the 100 simulated node positions. Alternatively, in the second procedure, the probability p_i to select one specific simulated position in the raster cell *i* was defined as follows:

$$p_i = \begin{cases} e^{v_i} & \text{if } (v_i - v_0) < 0 \\ 1/v_i & \text{if } (v_i - v_0) > 0 \end{cases}$$

where v_0 is the environmental value in the raster cell of the oldest node. While the first procedure allows simulating dispersal scenarios in which lineages tend to *remain in*, and thus favour, particular environmental conditions, the second procedure allows to simulate scenarios where lineages tend to progressively *disperse towards* specific conditions. All these simulations were also 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 nonaccessible sea areas.

Specifically, we applied our two new methods described above (see "Investigating the impact of environmental factors on lineages dispersal direction") to test the impact of human population density on data sets consisting of 1,000 spatially annotated trees simulated under five scenarios: a scenario in which dispersal direction of lineages were not impacted by any environmental heterogeneity (scenario 1), scenarios in which lineages tended to *remain in* (scenarios 2–3, simulated with the first procedure) or to *disperse towards* (scenarios 4–5, simulated with the second procedure) areas of high human population density. The difference between scenarios 2 and 3, as well as between 4 and 5, is that the former ones are based on the original human population raster while the latter ones are based on a log-transformed version of that raster (which thus gives relatively less importance to highly populated areas). The analyses of spatially annotated trees simulated under scenario 1 aimed at assessing the absence of type I error (false positives), and simulations performed under scenarios 2–3 as well as under scenarios 4–5 allowed investigating the statistical power of the analytical frameworks based on the metrics *E* and *R* (and the associated randomization procedure), respectively.

3 | RESULTS

3.1 | Identification of independent introductions in Iran

The preliminary discrete phylogeographic analysis (Lemey et al., 2009) was based on 109 Iranian RABV genomes, of which 101 were sequenced in the context of this study, as well as 274 non-Iranian genomes from a previous large-scale phylogenomic study (Troupin et al., 2016). In this analysis, we only specified two possible location states, that is "Iran" and "other," to focus on the identification of Iranian introductions. This first analysis identifies independent introductions for one isolated sequence as well as for seven monophyletic clades whose tip nodes are all located in Iran (clades A–H in Figure 1). While the majority of Iranian RABV sequences have been sampled from dogs (60%; Figure 1), viral sequences have also been sampled from various other host species whose proportion varies between clades (Figure 1). Clade A, which combines clades A1 and A2, constitutes the major monophyletic clade of Iranian sequences. Along with clade B, they are both closely related to a clade of Middle Eastern sequences from Israel, Oman, Saudi Arabia and the United Arab Emirates, that are known to circulate in foxes. The proportion of positive samples originating from wildlife (fox, wolf, jackal, marten) among the 75 isolates belonging to these three clades amounts to 47%. By comparison, Iranian clade C is connected to two paraphyletic clades of sequences from eastern Europe, Russia and Asia, which are also known to circulate in foxes.

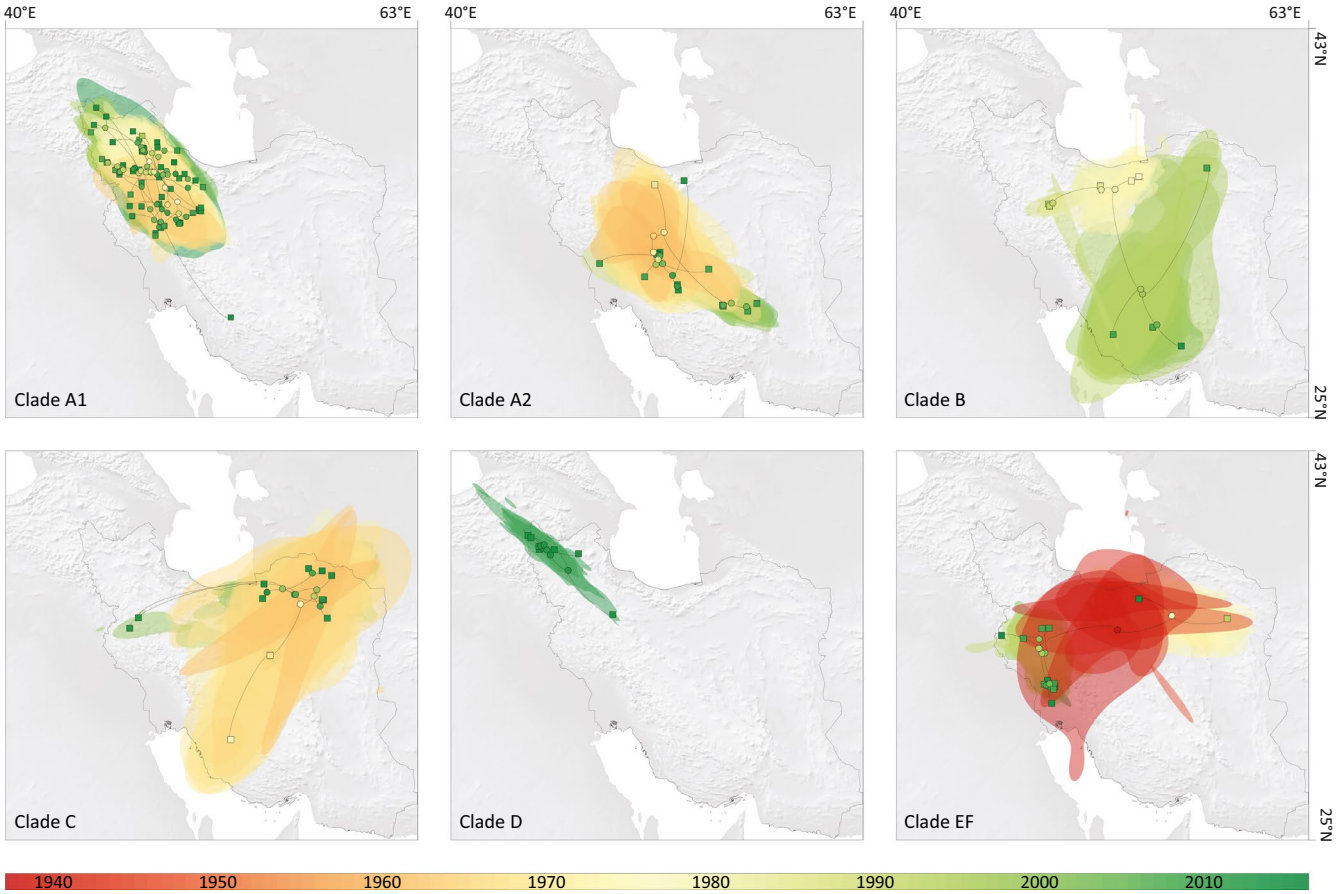

FIGURE 3 Reconstructed spatio-temporal diffusion for six monophyletic clades whose tip nodes are all located in Iran: maximum clade credibility (MCC) trees and 80% HPD regions based on 100 trees subsampled from the posterior distribution of each continuous phylogeographic analysis. Internal and tip nodes of the MCC trees are, respectively, displayed as dots and squares, and all nodes are coloured according to their time of occurrence. 80% HPD regions were computed for successive time layers and then superimposed using the same colour scale reflecting time. Only the borders of Iran are displayed on the map

Isolates collected from wildlife animals represent 55% (n = 11) of the sequences gathered in clade C. Clade D only includes isolates collected from dogs (n = 7) and its neighbouring clade exclusively consists of sequences sampled in Turkey. Clades E and F form, together with an Iraqi and a Russian sequence, another distinct monophyletic clade almost exclusively composed of dog samples (92%, n = 12). Finally, the more isolated Iranian clades G and H appear to be related to other sequences obtained from samples collected from Central Asia (Nepal, Afghanistan, India and Pakistan) and are exclusively composed of sequences collected from dog samples (n = 4). As these two clades only contained one and three sequences, respectively, they were not considered for the subsequent continuous phylogeographic analyses. Because this initial phylogenetic analysis was based on a large set of RABV sequences sampled across a range of multiple years, this also provided an opportunity to estimate the evolutionary rate to calibrate subsequent time-measured continuous phylogeographic analyses based on restricted numbers of sequences. From this initial analysis, we estimate a substitution rate of 2.13×10^{-4} substitutions per site per year (95% HPD [1.88, 2.35]; Figure 1), which is consistent with previous estimates obtained by Zhang et al. (2017).

3.2 | Continuous phylogeographic inferences

We performed continuous phylogeographic inference (Lemey et al., 2010) for the six distinct clades treated as potential separate introductions: A1, A2, B, C, D, as well as the combined Iranian clade formed by the closely related clades E and F (Figure 1). These inferences resulted in different phylogeographic reconstruction patterns (Figure 3). For instance, the continuous phylogeographic analyses of clades A1 and D clearly reveal that most of their dispersal has occurred in relatively restricted northwest regions of the country, which also correspond to the most populated areas. By contrast, clades A2, B and EF appear to be associated with more central and widespread distributions within the country. Finally, clade C presents a relatively widespread distribution mostly covering the northeast part of Iran.

3.3 | Comparative analysis of dispersal statistics

Spatio-temporal information contained in the inferred phylogenetic trees was subsequently used to estimate dispersal statistics such as branch dispersal velocities and diffusion coefficients. We summarize these estimates per clade (Table S3) but also for all clades considered

TABLE 1 RABV dispersal statistics estimated from continuous phylogeographic analyses performed on different data sets

	n	Mean branch velocity (km/year)	Weighted branch velocity (km/year)	Original diffusion coefficient (km^2/year)	Weighted diffusion coefficient (km^2/year)
RABV in Iran (present study)	105	55.5 [38.9, 142.4]	18.1 [16.3, 20.8]	2,676 [1,935, 5,066]	1,643 [1,356, 2,325]
Dog RABV in northern Africa	250	43.4 [30.9, 64.2]	19.5 [16.2, 23.5]	3,779 [2,444, 7,700]	1,511 [1,246, 1,936]
Raccoon RABV in North America	47	37.0 [22.3, 117.1]	11.8 [9.6, 13.3]	1,126 [744, 4931]	561 [454, 689]
Skunk RABV in North America	229	28.4 [20.1, 56.0]	9.4 [8.3, 10.6]	983 [633, 2,963]	579 [474, 675]
Bat RABV in Argentina	131	76.0 [60.9, 127.5]	34.7 [28.1, 41.6]	1,051 [720, 2,586]	721 [555, 929]
Bat RABV in eastern Brazil	41	37.4 [24.5, 148.6]	12.5 [7.8, 20.3]	615 [334.7, 3137]	273 [146.9, 423]
Bat RABV in Peru (lineage L1)	81	61.6 [34.9, 284.7]	21.8 [16.8, 28.2]	4,276 [2,166, 21,397]	2,056 [1,525, 2,800]
Bat RABV in Peru (lineage L3)	179	25.5 [17.5, 68.3]	8.7 [7.3, 10.2]	111 [64, 362]	61 [49, 76]

Note: For each statistic, we report both the median value and 95% HPD interval. In addition to the Iranian RABV data set introduced in the present study, we also report dispersal statistics estimated for several previously published RABV data sets: bat RABV data sets from Argentina (Torres et al., 2014), Brazil (Vieira, Pereira, Carnieli, Tavares, & Kotait, 2013) and Peru (Streicker et al., 2016), raccoon and skunk RABV data sets from North America (Biek et al., 2007; Kuzmina et al., 2013) and a dog RABV data set from northern Africa (Talbi et al., 2010). "*n*" indicates the number of sequences in each data set. See also Table S3 for separate estimates obtained from each RABV clade.

together, and we compare them with the same metrics previously reported for other RABV data sets (Table 1). First, the comparison between clades mainly highlights a much higher dispersal velocity and diffusion coefficient for clade D. This clade could correspond to a particularly rapid and recent RABV spread in the northwest region of the country. Sequences from that clade were exclusively sampled in domestic dogs, but this is also the case for clade E, which is associated with lower branch dispersal velocities. More generally, when comparing branch dispersal velocity and the host species composition, we were not able to identify any general trends (Table S3).

Second, the comparison with other RABV data sets reveals that dispersal statistics estimated for the RABV spread in Iran are highly similar with those estimated for the dog RABV dispersal in northern Africa, as inferred from the data set of Talbi et al. (2010). The similarity is particularly striking when comparing the "weighted" metrics, that is the weighted branch velocities and diffusion coefficients. The spread of dog RABV in Iran and in northern Africa are clearly associated with the highest values for the diffusion coefficients. Indeed, diffusion coefficient estimates are two to three times higher for these two data sets than for the raccoon and skunk RABV data sets, but also higher than for the bat-related data set from Argentina and Brazil. Only one of the bat-related data sets from Peru (lineage L1) is associated with branch velocity and diffusion coefficient values higher than those estimated for the present dog RABV data set from Iran.

3.4 | Impact of environmental factors on viral lineage dispersal

The analysis of lineage dispersal velocity further reveals non-negligible variability among all phylogenetic branches (CV = 2.82, 95% HPD [1.90, 7.53]). As a consequence, it is relevant to investigate if an environmental factor might explain the observed heterogeneity in dispersal velocity. To this end, we have used the analytical framework implemented in the R package "seraphim" (Dellicour, Rose, Faria, et al., 2016) to compare phylogenetic branch durations and a number of environmental distances. This analysis consists of using path models (Dijkstra, 1959; McRae, 2006) to compute, for each phylogenetic branch, environmental distances. These distances are computed on environmental rasters (i.e., grids) as well as on a so-called "null" raster with a value of "1" assigned to all accessible cells. The correlations between environmental distances and branch durations are then investigated to assess if heterogeneous environmental rasters could explain the differences in lineage dispersal velocity. The analysis reveals a relatively low correlation between branch durations and geographical distances, that is environmental distances computed on the "null" raster, no matter what path model is used to compute these distances (R^2 ~ 10%). Yet, none of the environmental distances computed on an environmental layer significantly increases this correlation with branch durations. Indeed, none of the tested environmental factors leads to a *Q* distribution with at least 90% positive values (Table S4), suggesting that they do not appropriately explain the RABV dispersal velocity better than the geographical distance factor alone. Because clade D is associated with a much higher branch dispersal velocity (Table S3), this analysis was also repeated after having discarded phylogeny branches belonging to that clade. However, discarding this potential outlier clade does not lead to overall different results, as none of the *Q* distributions tend to be clearly higher than zero (Table S5).

While the first analysis consists of testing the impact of environmental factors on dispersal velocity, we have also used a new approach that aims at testing the impact of such factors on the dispersal direction, that

TABLE 2 Impact of several environmental factors on the dispersal direction of RABV lineages in Iran and in northern Africa

	Testing the tendency of lineages to *remain in* specific environmental conditions (*E*)		Testing the tendency of lineages to *disperse towards* specific environmental conditions (*R*)	
Environmental factor	BF for factors treated as negative drivers	BF for factors treated as positive drivers	BF for factors treated as negative drivers	BF for factors treated as positive drivers
Data set: RABV in Iran				
Elevation	0.1	16.3	12.8	0.1
Barren vegetation	179.0	0.0	0.7	1.3
Shrublands	2.5	0.4	0.6	1.5
Grasslands	0.1	9.2	224.0	0.0
Croplands	0.1	13.5	3.5	0.3
Inaccessibility	>999	0.0	>999	0.0
Human pop. density	0.0	>999	0.0	>999
Data set: RABV in North Africa				
Elevation	>999	0.0	3.1	0.3
Barren vegetation	>999	0.0	0.0	27.1
Shrublands	25.5	0.0	0.4	2.4
Grasslands	6.1	0.2	0.7	1.4
Croplands	0.4	2.7	44.0	0.0
Inaccessibility	>999	0.0	41.9	0.0
Human pop. density	0.0	>999	0.0	127.6

Note: We report Bayes factor (BF) support obtained from analyses based on 900 trees sampled in each posterior distribution. Each environmental variable was treated as a positive and as a negative driver of the viral lineage dispersal. BF >3 and >20 are considered as "positive" and "strong" evidence, respectively (Kass & Raftery, 1995).

is on the tendency of viral lineages to remain in or disperse towards specific environmental conditions. In contrast to the analysis of the impact on dispersal velocity, the analysis of the impact on dispersal tendency reveals several environmental factors associated with Bayes factor values >20, which can be considered as a strong support (Table 2). Indeed, the analysis highlights that viral lineages tended to spread towards and remain in accessible areas associated with relatively high human population density. In addition, this analysis also underlines that lineages were less likely to spread towards grasslands and to occur in barren vegetation areas. The latter result related to barren vegetation is, however, a likely consequence of the nonuniform sampling across the study area (cfr the discussion). To put these results in perspective, we have also performed the same analyses on the dog RABV data set of northern Africa (Talbi et al., 2010). Interestingly, these analyses provide evidence for the same significant trends regarding the association between lineage dispersal direction and inaccessibility/human population density (Table 2). Several other factors are also highlighted when analysing this northern African data set: RABV lineages did not tend to occur in shrublands, barren vegetation and elevated areas, and did not tend to disperse towards croplands.

3.5 | Assessing the performance of our approach to investigate lineage dispersal direction

To test the statistical performance of our new approach focusing on lineage dispersal direction, we have analysed data sets simulated under different scenarios. The results are summarized in Table 3 and first indicate the absence of false positives when analysing data sets simulated on a uniform environmental layer, that is without any environmental impact on the dispersal direction (null model; BFs ~ 1). Furthermore, the analyses reveal a strong power of detection: approximated Bayes factor supports are >>20 when analysing the tendency of lineages to remain in or disperse towards populated areas, when analysing dispersal histories simulated under scenarios 2 (tendency to remain in highly populated areas) and 4 (tendency to progressively converge to highly populated areas), respectively (Table 3). However, when the tendency to remain in highly populated areas is simulated according to a log-transformed raster of human population density (scenario 3), we note a decrease in the statistical power of detection with approximated Bayes factor supports <10.

4 | DISCUSSION

Our discrete phylogeographic analysis confirms the co-occurrence of distinct lineages in Iran (Horton et al., 2015) and reveals the presence of at least eight RABV clades circulating in the region. As these lineages are likely to correspond to independent introductions of rabies in Iran, this highlights the importance of the geographical position of the Iranian region. Therefore, the Iranian region is everything but an area with isolated RABV spread, as previously reported for many parts of the world (Bourhy et al., 1999; Chen, Zou, Jin, & Ruan, 2015; Cliquet, Picard-Meyer, & Robardet, 2014; Horton et al.,

TABLE 3 Performances of the analytical workflow to test the impact of environmental factors on the dispersal direction of lineages

Different scenarios under which spatially annotated trees were simulated	Testing the tendency of lineages to *remain in* populated areas (*E*)	Testing the tendency of lineages to *disperse towards* populated areas (*R*)
1—No impact of human pop. density	1.0 (1.2)	1.1 (1.1)
2—Tendency to remain in populated area	999 (>999)	1.1 (1.1)
3—Tendency to remain in populated areas (log)	4.6 (8.1)	0.9 (0.9)
4—Tendency to disperse towards populated areas	0.4 (1.2)	499 (499)
5—Tendency to disperse towards populated areas (log)	0.4 (1.7)	>999 (>999)

Note: We report Bayes factor (BF) support obtained from analyses based on 1,000 spatially annotated trees simulated under different scenarios: dispersal direction of lineages was not impacted by any environmental heterogeneity (scenario 1), lineages tended to remain in (scenarios 2–3) or to disperse towards (scenarios 4–5) areas of high human population density. The difference between scenarios 2 and 3, as well as between 4 and 5, is that the former ones are based on the original human population raster and that the latter ones are based on a log-transformed version of that raster (which thus gives relatively less importance to highly populated areas). BF values reported between parentheses correspond to the support obtained when testing the log-transformed version of the human population density raster. BF >3 and >20 are considered as "positive" and "strong" evidence, respectively (Kass & Raftery, 1995).

2015; Pant et al., 2013; Talbi et al., 2009, 2010; Troupin et al., 2016). However, vaccination campaigns focusing on Iran could eliminate dog-mediated rabies in the country by restricting rabies cases to a few introductions from neighbouring countries that will not spread within vaccinated populations.

Further, the analysis of metadata associated with these Iranian clades suggests frequent RABV transmissions between dog and wildlife animal populations, and vice versa. Such a complex transmission pattern has already been identified in other regions such as Tanzania (Brunker et al., 2018) and in close countries such as Turkey (Horton et al., 2015; Marston et al., 2017). From an epidemiological point of view, the identification of this transmission pattern underlines the importance of the wildlife reservoir in the maintenance and circulation of the virus. It is widely accepted that host shifts of RABV and emergence of wildlife rabies from dog-adapted RABV occurred on many occasions (Troupin et al., 2016). Some of these shifts and emergence have been for instance described in Europe (Bourhy et al., 1999; McElhinney et al., 2011; Troupin et al., 2016), Taiwan (Lin et al., 2016) or more recently in Turkey (Marston et al., 2017; McElhinney et al., 2011). Understanding the key sources responsible for rabies epizootics and identifying host switches can have concrete implications on the implementation of rabies control measures in animals, as well as on the strategy of animal vaccination towards the ultimate goal of elimination (Fusaro et al., 2013; Un et al., 2012). In fact, rabies vaccination in wildlife requires specific strategies and stakeholder involvement (Freuling et al., 2013, Hsu et al., 2017, Müller et al., 2015, Wallace et al., 2018), which are different from those applied for infected dogs (Fahrion et al., 2017; Lembo, 2012). In Iran and the Middle East, rabies has been reported in many wild carnivore species such as foxes, golden jackals, wolves and martens (Janani et al., 2008; Picot et al., 2017; Seimenis, 2008). Our study clearly shows that some of the lineages circulating in Iran are more often found in wildlife species (in 50% of isolates from clades A1, A2, B and C altogether; $n = 86$) than those of the other clades (D, E, F, G and H), which were almost exclusively isolated from dogs (96%; $n = 25$). Consequently, a clear understanding of the role and the geographical distribution of the animal species potentially involved in the maintenance of the complex RABV epidemiological situation is crucial to improve the cost effectiveness of control measures as well as vaccination campaigns in low-income countries.

In this context, our spatially explicit continuous phylogeographic reconstructions have first allowed measuring a non-negligible dispersal velocity that appears to be highly similar to previously reported estimations for dog rabies spread in northern Africa (Dellicour et al., 2017; Talbi et al., 2010). One could therefore hypothesize that the dispersal velocity measured in both cases is intimately related to one of the main host species, that is dogs, whose distribution and individual movements are impacted by human activities. Yet, contrary to the study performed on the dog RABV data set from northern Africa (Dellicour et al., 2017), we do not detect any correlation between human-related factors (or any other tested factors) and lineage dispersal velocity. On the other hand, the analysis of the dispersal direction of lineages reveals the potential importance of two human-related factors, that is accessibility to major cities and human population densities for both study areas, that is northern Africa and Iran. Indeed, the analyses of both data sets revealed that lineages tended to remain in but also disperse towards human-populated and accessible areas. These two environmental factors are by definition correlated to each other, and it is thus not so surprising to detect the same association with lineage dispersal direction for both of them. These results underline the indirect importance of human-populated areas, which should be reasonably correlated with dog population density (Hampson et al., 2015), in attracting and further spreading the virus. It is also in line with results previously obtained by exploring the spatio-temporal circulation of dog RABV in a large African city (Bangui, the capital city of Central African Republic; Bourhy et al., 2016). This study revealed that, although dog RABV appears to be endemic in Bangui, its epidemiology is in fact shaped by the regular extinction of local chains of transmission coupled with the introduction of new lineages originating from outside the

city, generating successive waves of spread. In conclusion, populated areas represent strategic places for vaccination campaigns because they can act as crossroads and attractors of transmission chains. However, epidemiological surveillance and vaccination strategy should also consider their connection with less populated areas that can be responsible for RABV re-emergence. This latter aspect also highlights the complexity of implementing an efficient vaccination strategy allowing rabies eradication, which will require successive vaccination campaigns in the same areas (Rattanavipapong et al., 2018).

While the outcomes of continuous phylogeographic analyses can be exploited to investigate the impact of environmental factors on lineage dispersal velocity and direction, it is important to note that such analyses depend on inferred viral lineage movement and thus, to some extent, on the spatial distribution of the sampled sequences. Although a particular sampling will always affect the reconstructed dispersal history of viral lineages, continuous phylogeographic inference will still provide movement data that can inform on the dispersal dynamics of the virus. Indeed, even if connected through tree topologies, branches can be treated as distinct movement vectors informing on the mode and tempo (dispersal velocity, dispersal direction) of lineage dispersal events (Pybus et al., 2012). However, we acknowledge two potential impacts of sampling bias on these post hoc approaches. First, the effect of environmental conditions that are mostly represented in undersampled areas will be challenging to detect when testing their impact on lineage dispersal velocity or the tendency of lineages to disperse towards these particular conditions (based on the *R* metric and associated randomization procedure). Second, as currently implemented, sampling bias can directly impact the test on the tendency of lineages to remain in specific environmental conditions (based on the *E* metric and associated randomization procedure). This is the case for instance in the present sampling where barren vegetation areas are barely sampled but still, as expected, identified as a factor repulsing lineages. In conclusion, because of its stronger dependence on sampling heterogeneity, the latter method based on the metric *E* can be considered to be more descriptive or exploratory in nature rather than a proper statistical test and should hence be interpreted with caution.

Overall, our study demonstrates that continuous phylogeographic reconstructions represent a useful tool to describe but also to analyse the dispersal dynamic of virus spread. Indeed, phylogenetic trees inferred by continuous phylogeographic inference can first be exploited to estimate dispersal statistics, which mainly represent useful metrics in the context of a comparison between virus spread associated with different host species and/or environmental conditions, or even between different viruses. In addition, such annotated trees can be used to investigate the impact of environmental factors on the dispersal velocity and tendency, two different but important aspects to consider in order to elucidate the drivers behind an epidemic. However, these phylogeographic approaches come with relatively important limitations. First, and as discussed above, they remain dependent from the sampling pattern and quality. For instance, environmental conditions associated with the least sampled areas will be proportionally less investigated than in well sampled areas. Consequently, there is always a possible risk than an important environmental factor is not identified because of sampling bias. More generally, while the effects of sampling bias are relatively well known in the context of discrete phylogeographic inference (Baele et al., 2018; De Maio, Wu, O'Reilly, & Wilson, 2015), its implications for continuous phylogeographic are still unclear and should ideally be investigated in future studies. Second, these approaches require the availability of precise sampling data such as sampling dates but also sampling locations. While the availability of sampling dates is crucial to obtain temporal signal required to infer time-stamped phylogenies, sampling geographical coordinates are also compulsory for continuous phylogeographic inference. Precise sampling origins in particular are not frequently available along with publicly available sequences (directly in databases like GenBank or in associated publications). Increasing the availability of such metadata could open new opportunities for large-scale phylogeographic analyses with key benefits for epidemiologic investigations based on phylogeographic approaches.

ACKNOWLEDGEMENTS

We are grateful to three anonymous reviewers for all their constructive comments on an earlier version of the manuscript. SD is supported by the Fonds National de la Recherche Scientifique (FNRS, Belgium) and was previously funded by the Fonds Wetenschappelijk Onderzoek (FWO, Belgium). GB acknowledges support from the Interne Fondsen KU Leuven/Internal Funds KU Leuven under grant agreement C14/18/094. This work was supported by European Union Seventh Framework Programme PREDEMICS (grant number 278433). PL acknowledges funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 725422-ReservoirDOCS), from the Wellcome Trust Collaborative Award, 206298/Z/17/Z, from the Special Research Fund, KU Leuven ("Bijzonder Onderzoeksfonds", KU Leuven, OT/14/115), and from the Research Foundation—Flanders ("Fonds voor Wetenschappelijk Onderzoek—Vlaanderen", G066215N, G0D5117N and G0B9317N). HB acknowledges funding from the Agence Nationale de la Recherche (grant number BSV3-0019).

COMPETING INTERESTS

The authors declare no competing interest.

DATA AVAILABILITY STATEMENT

New RABV sequences from Iran and Egypt were deposited in GenBank under Accession nos. MK760667–MK760770.

ORCID

Simon Dellicour https://orcid.org/0000-0001-9558-1052
Guy Baele https://orcid.org/0000-0002-1915-7732
Philippe Lemey https://orcid.org/0000-0003-2826-5353
Hervé Bourhy https://orcid.org/0000-0002-2608-5589

REFERENCES

Ayres, D. L., Darling, A., Zwickl, D. J., Beerli, P., Holder, M. T., Lewis, P. O., … Suchard, M. A. (2012). BEAGLE: An application programming interface and high-performance computing library for statistical phylogenetics. *Systematic Biology*, *61*, 170–173.

Baele, G., Dellicour, S., Suchard, M. A., Lemey, P., & Vrancken, B. (2018). Recent advances in computational phylodynamics. *Current Opinion in Virology*, *31*, 24–32.

Baltazard, M., & Ghodssi, M. (1954). Prevention of human rabies; treatment of persons bitten by rabid wolves in Iran. *Bulletin of the World Health Organization*, *10*, 797–803.

Biek, R., Henderson, J. C., Waller, L. A., Rupprecht, C. E., & Real, L. A. (2007). A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. *Proceedings of the National Academy of Sciences of the USA*, *104*, 7993–7998.

Bourhy, H., Kissi, B., Audry, L., Smreczak, M., Sadkowska-Todys, M., Kulonen, K., … Zmudzinski, J. F. (1999). Ecology and evolution of rabies virus in Europe. *Journal of General Virology*, *80*, 2545–2557.

Bourhy, H., Nakouné, E., Hall, M., Nouvellet, P., Lepelletier, A., Talbi, C., … Rambaut, A. (2016). Revealing the micro-scale signature of endemic zoonotic disease transmission in an African urban setting. *PLoS Path*, *12*, e1005525.

Brunker, K., Lemey, P., Marston, D. A., Fooks, A. R., Lugelo, A., Ngeleja, C., … Biek, R. (2018). Landscape attributes governing local transmission of an endemic zoonosis: Rabies virus in domestic dogs. *Molecular Ecology*, *27*, 773–788.

Charkazi, A., Behnampour, N., Fathi, M., Esmaeili, A., Shahnazi, H., & Heshmati, H. (2013). Epidemiology of animal bite in Aq Qala city, northen of Iran. *Journal of Education and Health Promotion*, *2*(1), 13. https://doi.org/10.4103/2277-9531.112682

Chen, J., Zou, L., Jin, Z., & Ruan, S. (2015). Modeling the geographic spread of rabies in China. *PLoS Neglected Tropical Diseases*, *9*, e0003772.

Cliquet, F., Picard-Meyer, E., & Robardet, E. (2014). Rabies in Europe: What are the risks? *Expert Review of Anti-infective Therapy*, *12*, 905–908. https://doi.org/10.1586/14787210.2014.921570

Criscuolo, A., & Brisse, S. (2013). AlienTrimmer: A tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. *Genomics*, *102*, 500–506.

De Maio, N., Wu, C. H., O'Reilly, K. M., & Wilson, D. (2015). New routes to phylogeography: A Bayesian structured coalescent approximation. *PLoS Genetics*, *11*, e1005421.

Dehghani, R., Sharif, A., Madani, M., Kashani, H. H., & Sharif, M. R. (2016). Factors influencing animal bites in Iran: A descriptive study. *Osong Public Health and Research Perspectives*, *7*, 273–277.

Dellicour, S., Baele, G., Dudas, G., Faria, N. R., Pybus, O. G., Suchard, M. A., … Lemey, P. (2018). Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak. *Nature Communications*, *9*, 2222. https://doi.org/10.1038/s41467-018-03763-2

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

Dellicour, S., Rose, R., Faria, N. R., Vieira, L. F. P., Bourhy, H., Gilbert, M., … Pybus, O. G. (2017). Using viral gene sequences to compare and explain the heterogeneous spatial dynamics of virus epidemics. *Molecular Biology and Evolution*, *34*, 2563–2571.

Dellicour, S., Rose, R., & Pybus, O. G. (2016). Explaining the geographic spread of emerging epidemics: A framework for comparing viral phylogenies and environmental landscape data. *BMC Bioinformatics*, *17*, 1–12. https://doi.org/10.1186/s12859-016-0924-x

Dellicour, S., Vrancken, B., Trovão, N. S.,Fargette, D., & Lemey, P. (2018). On the importance of negative controls in viral landscape phylogeography. *Virus Evolution*, *4*, vey023.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, *1*, 269–271.

Dodet, B., Adjogoua, E. V., Aguemon, A. R., Amadou, O. H., Atipo, A. L., Baba, B. A., … Wateba, M. I. (2008). Fighting rabies in Africa: The Africa Rabies Expert Bureau (AfroREB). *Vaccine*, *26*, 6295–6298.

Drummond, A. J., & Bouckaert, R. R. (2015). *Bayesian evolutionary analysis with BEAST*. Cambridge, UK: Cambridge University Press.

Fahrion, A. S., Taylor, L. H., Torres, G., Müller, T., Dürr, S., Knopf, L., … Abela-Ridder, B. (2017). The road to dog rabies control and elimination – What keeps us from moving faster? *Frontiers in Public Health*, *5*, 103. https://doi.org/10.3389/fpubh.2017.00103

Farahtaj, F., Fayaz, A., Howaizi, N., Biglari, P., & Gholami, A. (2014). Human rabies in Iran. *Tropical Doctor*, *44*, 226–229.

Feizhaddad, M. H., Kassiri, H., Lotfi, M., & Hoseini, S. S. (2014). Epidemiology and public health aspects of animal biting in shush county, Khuzestan Province, Iran. *Archives of Clinical Infectious Diseases*, *9*, e18773.

Freuling, C. M., Hampson, K., Selhorst, T., Schröder, R., Meslin, F. X., Mettenleiter, T. C., & Müller, T. (2013). The elimination of fox rabies from Europe: Determinants of success and lessons for the future. *Philosophical Transactions of the Royal Society B*, *368*, 20120142.

Fusaro, A., Monne, I., Salomoni, A., Angot, A., Trolese, M., Ferrè, N., …De Benedictis, P. (2013). The introduction of fox rabies into Italy (2008–2011) was due to two viral genetic groups with distinct phylogeographic patterns. *Infection, Genetics and Evolution*, *17*, 202–209.

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B., & Suchard, M. A. (2013). Improving Bayesian population dynamics inference: A coalescent-based model for multiple loci. *Molecular Biology and Evolution*, *30*, 713–724.

Grenfell, B. T., Bjørnstad, O. N., & Kappey, J. (2001). Travelling waves and spatial hierarchies in measles epidemics. *Nature*, *414*, 716.

Hall, T. A. (1999). BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, *41*, 95–98.

Hampson, K., Coudeville, L., Lembo, T., Sambo, M., Kieffer, A., Attlan, M., … Dushoff, J. (2015). Estimating the global burden of endemic canine rabies. *PLoS Neglected Tropical Diseases*, *9*, e0003709.

Holmes, E. C. (2004). The phylogeography of human viruses. *Molecular Ecology*, *13*, 745–756.

Horton, D. L., McElhinney, L. M., Freuling, C. M., Marston, D. A., Banyard, A. C., Goharrriz, H., …Fooks, A. R. (2015). Complex epidemiology of a zoonotic disease in a culturally diverse region: Phylogeography of rabies virus in the Middle East. *PLoS Neglected Tropical Diseases*, *9*, e0003569.

Hsu, A.-P., Tseng, C.-H., Barrat, J., Lee, S.-H., Shih, Y.-H., Wasniewski, M., … & Tsai, H.-J. (2017). Safety, efficacy and immunogenicity evaluation of the SAG2 oral rabies vaccine in Formosan ferret badgers. *PLoS One*, *12*, e0184831.

Jacquot, M., Nomikou, K., Palmarini, M., Mertens, P., & Biek, R. (2017). 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.

Janani, A. R., Fayaz, A., Simani, S., Farahtaj, F., Eslami, N., Howaizi, N., … Sabetghadam, M. (2008). Epidemiology and control of rabies in Iran. *Developments in Biologicals*, *131*, 207–211.

Jung, M., Leye, N., Vidal, N., Fargette, D., Diop, H., Toure Kane, C., … Peeters, M. (2012). The origin and evolutionary history of HIV-1 Subtype C in Senegal. *PLoS ONE*, *7*, e33579.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795. https://doi.org/10.1080/01621459.1995.10476572

Keeling, M. J., Woolhouse, M. E. J., Shaw, D. J., Matthews, L., Chase-Topping, M., Haydon, D. T., … Grenfell, B. T. (2001). Dynamics of the 2001 UK foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape. *Science*, *294*, 813.

Knobel, D. L., Cleaveland, S., Coleman, P. G., Fèvre, E. M., Meltzer, M. I., Miranda, M. E. G., … Meslin, F. X. (2005). Re-evaluating the burden of rabies in Africa and Asia. *Bulletin of the World Health Organization*, *83*, 360–368.

Kühnert, D., Wu, C. H., & Drummond, A. J. (2011). Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. *Infection, Genetics and Evolution*, *11*, 1825–1841.

Kuzmina, N. A., Lemey, P., Kuzmin, I. V., Mayes, B. C., Ellison, J. A., Orciari, L. A., … Rupprecht, C. E. (2013). The phylogeography and spatio-temporal spread of south-central skunk rabies virus. *PLoS ONE*, *8*, e82348.

Laenen, L., Dellicour, S., Vergote, V., Nauwelaers, I., De Coster, S., Verbeeck, I., …Maes, P. (2016). Spatio-temporal analysis of Nova virus, a divergent hantavirus circulating in the European mole in Belgium. *Molecular Ecology*, *25*, 5994–6008.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., … Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, *23*, 2947–2948.

Lembo, T. & on behalf of the Partners for Rabies Prevention (2012). The blueprint for rabies prevention and control: A novel operational toolkit for rabies elimination. *PLoS Neglected Tropical Diseases*, *6*, e1388.

Lemey, P., Rambaut, A., Drummond, A. J., & Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS Computational Biology*, *5*, e1000520.

Lemey, P., Rambaut, A., Welch, J. J., & Suchard, M. A. (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology and Evolution*, *27*, 1877–1885.

Lin, Y. C., Chu, P. Y., Chang, M. Y., Hsiao, K. L., Lin, J. H., & Liu, H. F. (2016). Spatial temporal dynamics and molecular evolution of re-emerging rabies virus in Taiwan. *International Journal of Molecular Sciences*, *17*, 392.

Marston, D. A., Horton, D. L., Nunez, J., Ellis, R. J., Orton, R. J., Johnson, N., …Fooks, A. R. (2017). Genetic analysis of a rabies virus host shift event reveals within-host viral dynamics in a new host. *Virus Evolution*, *3*, vex038.

McElhinney, L. M., Marston, D. A., Freuling, C. M., Cragg, W., Stankov, S., Lalosević, D., …Fooks, A. R. (2011). Molecular diversity and evolutionary history of rabies virus strains circulating in the Balkans. *Journal of General Virology*, *92*, 2171–2180.

McRae, B. H. (2006). Isolation by resistance. *Evolution*, *60*, 1551–1561.

Müller, T., Freuling, C. M., Wysocki, P., Roumiantzeff, M., Freney, J., Mettenleiter, T. C., & Vos, A. (2015). Terrestrial rabies control in the European Union: Historical achievements and challenges ahead. *Veterinary Journal*, *203*, 10–17.

Nadin-Davis, S. A., Simani, S., Armstrong, J., Fayaz, A., & Wandeler, A. I. (2003). Molecular and antigenic characterization of rabies viruses from Iran identifies variants with distinct epidemiological origins. *Epidemiology and Infection*, *131*, 777–790.

Pant, G. R., Lavenir, R., Wong, F. Y. K., Certoma, A., Larrous, F., Bhatta, D. R., … Dacheux, L. (2013). Recent emergence and spread of an arctic-related phylogenetic lineage of rabies virus in Nepal. *PLoS Neglected Tropical Diseases*, *7*, e2560.

Picot, V., Rasuli, A., Abella-Rider, A., Saadatian-Elahi, M., Aikimbayev, A., Barkia, A., … Nel, L. (2017). The Middle East and Eastern Europe rabies Expert Bureau (MEEREB) third meeting: Lyon-France (7–8 April, 2015). *Journal of Infection and Public Health*, *10*, 695–701.

Pybus, O. G., Suchard, M. A., Lemey, P., Bernardin, F. J., Rambaut, A., Crawford, F. W., … Delwart, E. L. (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the National Academy of Sciences of the USA*, *109*, 15066–15071.

Pybus, O. G., Tatem, A. J., & Lemey, P. (2015). Virus evolution and transmission in an ever more connected world. *Proceedings of the Royal Society B: Biological Sciences*, *282*, 20142878.

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

Rattanavipapong, W., Thavorncharoensap, M., Youngkong, S., Genuino, A. J., Anothaisintawee, T., Chaikledkaew, U., & Meeyai, A. (2018). The impact of transmission dynamics of rabies control: Systematic review. *Vaccine*, in press, https://doi.org/10.1016/j.vaccine.2018.11.035

Real, L. A., & Biek, R. (2007). Spatial dynamics and genetics of infectious diseases on heterogeneous landscapes. *Journal of the Royal Society Interface*, *4*, 935–948.

Seimenis, A. (2008). The rabies situation in the Middle East. *Developments in Biologicals*, *131*, 43–53.

Streicker, D. G., Winternitzc, J. C., Satterfield, D. A., Condori-Condori, R. E., Broos, A., Tello, C., … Valderrama, W. (2016). Host-pathogen evolutionary signatures reveal dynamics and future invasions of vampire bat rabies. *Proceedings of the National Academy of Sciences of the USA*, *113*, 10926–10931.

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., & Drummond, A. J., &Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, *4*, vey016.

Talbi, C., Holmes, E. C., de Benedictis, P., Faye, O., Nakouné, E., Gamatié, D., … Bourhy, H. (2009). Evolutionary history and dynamics of dog rabies virus in western and central Africa. *Journal of General Virology*, *90*, 783–791.

Talbi, C., Lemey, P., Suchard, M. A., Abdelatif, E., Elharrak, M., Jalal, N., … Bourhy, H. (2010). Phylodynamics and human-mediated dispersal of a zoonotic virus. *PLoS Path*, *6*, e1001166.

Torres, C., Lema, C., Gury Dohmen, F., Beltran, F., Novaro, L., Russo, S., … Cisterna, D. M. (2014). Phylodynamics of vampire bat-transmitted rabies in Argentina. *Molecular Ecology*, *23*, 2340–2352.

Troupin, C., Dacheux, L., Tanguy, M., Sabeta, C., Blanc, H., Bouchier, C., … Bourhy, H. (2016). Large-scale phylogenomic analysis reveals the complex evolutionary history of rabies virus in multiple carnivore hosts. *PLoS Path*, *12*, e1006041.

Trovão, N. S., Baele, G., Vrancken, B., Bielejec, F., & Suchard, M. A.,Fargette, D., & Lemey, P. (2015). Host ecology determines the dispersal patterns of a plant virus. *Virus Evolution*, *1*(1), vev016.

Un, H., Eskiizmirliler, S., Unal, N., Freuling, C. M., Johnson, N., Fooks, A. R., … Aylan, O. (2012). Oral vaccination of foxes against rabies in Turkey between 2008 and 2010. *Berliner Und Munchener Tierarztliche Wochenschrift*, *125*, 203–208.

Vieira, L. F. P., Pereira, S. R. F. G., Carnieli, P. Jr, Tavares, L. C. B., & Kotait, I. (2013). Phylogeography of rabies virus isolated from herbivores and bats in the Espírito Santo State, Brazil. *Virus Genes*, *46*, 330–336.

Wallace, R. M., Lai, Y., Doty, J. B., Chen, C.-C., Vora, N. M., Blanton, J. D., … Pei, K. J. C. (2018). Initial pen and field assessment of baits to use in oral rabies vaccination of Formosan ferret-badgers in response to the re-emergence of rabies in Taiwan. *PLoS ONE*, *13*, e0189998.

WHO (2018). *0 by 30 our catalytic response*. Retrieved from http://www.who.int/rabies/United_against_Rabies

Zhang, Y., Vrancken, B., Feng, Y., Dellicour, S., Yang, Q., Yang, W., … Tian, H. (2017). Cross-border spread, lineage displacement and

evolutionary rate estimation of rabies virus in Yunnan Province, China. *Virology Journal*, *14*, 102.

Zinsstag, J., Lechenne, M., Laager, M., Mindekem, R., Naïssengar, S., Oussiguéré, A., … Chitnis, N. (2017). Vaccination of dogs in an African city interrupts rabies transmission and reduces human exposure. *Science Translational Medicine*, *9*, eaaf6984.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Dellicour S, Troupin C, Jahanbakhsh F, et al. Using phylogeographic approaches to analyse the dispersal history, velocity and direction of viral lineages — Application to rabies virus spread in Iran. *Mol Ecol*. 2019;28:4335–4350. https://doi.org/10.1111/mec.15222