# Persistent phylogeographic structure of an emerging virus on a homogeneous landscape

Trieste Musial[1,2*], Scott Duke-Sylvester[3], Rolan Davis[4], Roman Biek[5*], Leslie A. Real[2,6]

[1] Population Biology, Ecology, and Evolution Program, Emory University, Atlanta, Georgia, United States of America

[2] Center for Disease Ecology, Emory University, Atlanta, Georgia, United States of America

[3] Department of Biology, University of Louisiana at Lafayette, Lafayette, Louisiana, United States of America

[4] Rabies Laboratory, Veterinary Diagnostic Laboratory, Kansas State University, Manhattan, Kansas, United States of America

[5] Boyd Orr Centre for Population and Ecosystem Health, Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary, and Life Sciences, University of Glasgow, Glasgow, United Kingdom

[6] Department of Biology, Emory University, Atlanta, Georgia, United States of America

* Corresponding authors

E-mail: tmusial@emory.edu; roman.biek@glasgow.ac.uk

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27260v1 | CC BY 4.0 Open Access | rec: 7 Oct 2018, publ: 7 Oct 2018

## Abstract

Landscape composition and structure influence animal movement, which in turn can affect transmission of their diseases. Spatio-temporal variation in host diffusion, caused by landscape heterogeneity, is thus expected to generate corresponding phylogeographic patterns in the pathogen. However, establishing causative links between genetic structure in pathogen populations and environmental variation does require appropriate null models. Here, we present an empirical example of the emergence and multi-decade persistence of phylogeographic structure on a homogeneous landscape in a rapidly diversifying pathogen in the absence of any apparent landscape heterogeneity. By applying phylogeographic inference to 173 sequences of a raccoon-specific strain of rabies virus, we reconstruct patterns of the virus' evolution and diffusion on the Florida peninsula, USA, from its first emergence in the 1940's to the present. Consistent with a lack of significant landscape heterogeneity relevant to raccoon movement in Florida, we found that the speed of rabies virus diffusion was spatially homogeneous across the peninsula. In contrast, we document the emergence of strong phylogeographic structure in the virus, in the form of five monophyletic lineages that diverged during the early years of colonization and now each occupy a distinct sub-region of Florida. Based on samples taken over multiple decades, we show that the spatial distribution of these lineages has changed little over the past four decades. This phylogeographic stability allowed us to retrospectively identify a small set of counties within Florida as the likely source of the virus strain that seeded a much larger rabies outbreak in the northeastern USA in the 1970s. Our results provide a rare empirical demonstration that spatial genetic structure can arise and be maintained in the absence of landscape heterogeneity, which has wider implications for the interpretation of phylogeographic data and the reconstruction of historical colonization patterns from molecular data.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27260v1 | CC BY 4.0 Open Access | rec: 7 Oct 2018, publ: 7 Oct 2018

## Introduction

The geographic genetic organization of populations is a complex interaction between patterns of organismal movement, breeding structure, habitat heterogeneity, and dispersal and establishment.  On landscapes that are homogeneous with respect to individual movement probabilities, genetic organization traditionally is considered to be determined by patterns of local gene exchange and patterns of colonization by exogenous genotypes (Irwin 2002; Kuo & Avise 2005; Waters 2011; Waters *et al.* 2013; Wright 1943, 1951).  Spatial genetic patterns on these landscapes follow expectations of isolation-by-distance coupled to unpredictable stochastic migration events and stochastic ecological extinctions (Irwin 2002; Marske *et al.* 2013; Neigel & Avise 1993; Saunders *et al.* 1986).  In contrast, heterogeneous landscapes, with barriers to movement or gene exchange and/or characterized by areas of inhospitable terrain, tend to lead to more predictable spatial genetic relationships, delimited – and stabilized – by recognizable landscape features (Avise *et al.* 1987; Kuo & Avise 2005; Manel *et al.* 2003).  Significant spatial organization of genotypes independent of local gene exchange has long been considered *prima facie* evidence of landscape heterogeneity associated with restrictions on movement and establishment of genotypes distributed over space (Avise *et al.* 1987; Marske *et al.* 2013; Wright 1951).

Some research has indicated however that random local processes, as well as events during colonization and expansion, can also generate enduring spatial genetic patterns.  When this is the case, population spatial genetic structure may not be dictated solely by barrier arrangement.  On simulated homogeneous landscapes, significant "phylogeographic breaks", i.e. spatial partitioning of distinct phylogenetic lineages (Avise *et al.* 1987), can emerge independently of underlying geographic barriers within established, continuous populations where dispersal rates are low and/or populations sizes are small (Baptestini *et al.* 2013; Irwin 2002; Kuo & Avise 2005). Stochastic events will also shape the spatial arrangement of lineages that diverge during continuous range expansion, and these processes can generate lasting signatures that dominate long-term spatial genetic patterns (Excoffier & Ray 2008; Hallatschek *et al.* 2007; Waters *et al.* 2013). Phylogeographic breaks formed by stochastic processes contain no information about the spatial ecology of the lineage in question but can easily be mistaken for patterns generated by landscape-level processes (Crisp *et al.* 2011; Real *et al.* 2005a; Schwartz & McKelvey 2008). As of yet, the characterization of persistent genetic breaks on homogeneous landscapes has been limited to simulation studies (e.g., Baptestini *et al.* 2013; Ibrahim *et al.* 1996; Irwin 2002; Kuo & Avise 2005) and microbial *in vitro* experiments (Hallatschek *et al.* 2007; Hallatschek & Nelson 2008). Two key questions are thus: 1) under what conditions will these relationships form on natural landscapes? And 2) once formed, how stable are these spatial patterns across natural landscapes that are highly homogeneous? Here, we address these questions using a rapidly evolving zoonotic pathogen, rabies virus, as a model.

Terrestrial rabies, caused by an RNA virus, is often used as a model system for inferring the effects of landscape heterogeneity on phylogeographic patterns (e.g. Biek *et al.* 2007; Brunker *et al.* 2012; Szanto *et al.* 2011). The high mutation rate of rabies virus generates substitutions on a scale similar to the rate of ecological changes, making this and other RNA virus systems ideal for epidemiological reconstruction from genomic data (Biek *et al.* 2015; Drummond *et al.* 2003; Pybus & Rambaut 2009). Movement of rabies virus (like other directly-transmitted viruses) generally occurs across landscapes and distances that are amenable to independent host movement, though human-mediated long-distance translocation events are known to have facilitated various rabies epizootics and range expansions (Nettles *et al.* 1979;

Talbi *et al.* 2010; Wilson *et al.* 1997). Rabies lineages display strong regional compartmentalization (e.g., Biek *et al.* 2007; Kuzmina *et al.* 2013; Real *et al.* 2005a; Szanto *et al.* 2011), even in cases where a single strain is hosted by multiple species (Lembo *et al.* 2007; Nadin-Davis *et al.* 1994).

Although rabies virus can infect any mammal, different variants are usually maintained by a particular host species. One such host-specific rabies variant is raccoon rabies virus (RRV), which is distributed across much of eastern North America. Here we investigate the conditions under which phylogeographic breaks in RRV develop and persist on a homogeneous landscape, the Florida peninsula. RRV was first detected in Florida in 1947 (Bigler *et al.* 1973), and had covered the eastern U.S. within 50 years (Childs *et al.* 2000). The process of RRV expansion was distinctly different between its initial establishment in Florida and its later spread throughout the mid-Atlantic: In the southeastern US, RRV was detected only sporadically in the decades following its emergence in Florida and generated few detected, localized epizootic events (Bigler *et al.* 1973; Kappus *et al.* 1970; Scatterday *et al.* 1960). In fact, increased surveillance efforts during the decade following its detection in 1947 determined that RRV had already reached an enzootic state throughout peninsular Florida, having potentially gone undetected for much of its early expansion (Kappus *et al.* 1970; Scatterday *et al.* 1960). In contrast, the spread of RRV through the mid-Atlantic states, likely originating from infected raccoons translocated from southeastern states in the late 1970s (Nettles *et al.* 1979; Rupprecht & Smith 1994), proceeded rapidly and conspicuously, generating one of the largest wildlife epizootics in history (Childs *et al.* 2000). Mountains, rivers, and major water bodies frequently have been recognized as barriers to the movement of raccoons and, consequently, RRV (Biek *et al.* 2007; Smith *et al.* 2002; Wheeler & Waller 2008). Accordingly, phylogeographic analyses of RRV within its mid-

Atlantic range identified decreased viral velocity across mountain ranges (Biek *et al.* 2007). The landscape of Florida, in contrast, exhibits few features that might serve as barriers to raccoon movement. Florida's topography is relatively flat (with a max elevation of 105m above sea level), and there is an abundance of favorable and continuous raccoon habitat (including large swamps and long stretches of urbanized areas). Previous analyses based on microsatellite and mtDNA data found evidence for a single, well-mixed raccoon population that covers Florida, Georgia, Alabama, Tennessee, and South Carolina (Cullingham *et al.* 2008; Reeder-Carroll 2010), consistent with a lack of impediments to raccoon gene flow throughout this region.

The raccoon rabies system in Florida offers a chance to explore long-term evolutionary outcomes of invasion on a landscape devoid of spatial features that would be predicted to maintain phylogeographic patterns. On this landscape, one might expect to see little or no phylogeographic structure of viral strains and would instead predict lineages to spatially admix over time, eliminating early spatial genetic differentiation. We are testing this hypothesis by analyzing spatial evolutionary patterns of RRV expansion and lineage divergence in Florida. Utilizing novel methods to reconstruct spatial patterns of viral diffusion rates within our study area, we aimed to determine how RRV phylogeographic patterns arose and are maintained following emergence of the virus and how this affects the interpretation of genetic data in the context of RRV as well as biological invasion processes more generally.

## Materials and Methods

### Sample collection and preparation

We analyzed 173 brain tissue samples collected as part of ongoing rabies surveillance efforts by state and local public health departments in Florida (*n*=164) and Alabama (*n*=9) from 1982 to 2012. Sampling fell within three general time periods – 37 samples were collected from 1982 to 1988, 40 from 1998 to 2004, and 96 from 2009 to 2012. Sample locations from 2003 forward (*n*=103) were georeferenced to zip code; location data prior to 2003 were recorded as county centroid (Fig 1). Infection with the raccoon-specific rabies variant was initially confirmed using an indirect assay with monoclonal antibodies for the nucleocapsid protein (Smith *et al.* 1986). Total RNA was extracted from 50 – 100 mg of frozen brain tissue archived after post mortem analysis of rabid raccoons. RNA extraction was achieved using a hypotonic lysis buffer (Smith *et al.* 1991) followed by precipitation with TRIzol Reagent (Invitrogen, Carlsbad, CA, USA) and following the manufacturer's protocol. Pellets were stored at -20°C in DEPC-treated water until amplification.

RT-PCR was used to amplify the 5' portion of the rabies virus nucleoprotein (N) gene in all samples using forward primer 21G (5'-ATGTAACACCTCTACA-3') (Orciari *et al.* 2001) and reverse primer 304 (5'-ATGAGCAAGATCTTCGT-3'). Additionally, a portion of the glycoprotein (G) gene was amplified in a subset of 20 samples, using primers and conditions described in Biek *et al*. (2007). Our amplification procedure utilized the SuperScript III One-Step RT-PCR System with Platinum *Taq* High Fidelity (Invitrogen). Amplicons were examined by gel electrophoresis and samples with poor yield were reamplified prior to purification with the Promega Wizard PCR kit (Promega, Madison, WI) and sequencing on an ABI 377 DNA Sequencer (Applied Biosystems, Foster City, CA). Final sequences had no indels and were aligned manually and trimmed to final lengths of 591 nt at the N gene (*n*=173) and 1374 nt at the G gene (n=20). The N and G genes exhibit no significant incongruence (Biek *et al.* 2007) and were concatenated for our analyses.

**Phylogenetic analysis**

Potential evolutionary models were explored using the phyml package in jModelTest 2 (Darriba *et al.* 2012; Guindon & Gascuel 2003), and we used BEAST's (Drummond *et al.* 2012) path sampling procedure (Baele *et al.* 2012; Baele *et al.* 2013) to identify the best fit model; computations were performed using XSEDE cloud-computing resources via CIPRES (Miller *et al.* 2010; Towns *et al.* 2014). A strict molecular clock with gamma-distributed rate variation applied independently between codons at position 1+2 vs 3 ($HKY_{112}$ + $CP_{112}$ + $G_{112}$) (Hasegawa *et al.* 1985; Yang 1994) was clearly the best model for our data. We used BEAST to simultaneously estimate phylogenetic relationships and ancestral dispersal patterns from our samples. Initial analyses using marginal likelihood estimation (Baele *et al.* 2012; Baele *et al.* 2013), identified a strict clock combined with a codon-specific SDR06 model (Shapiro *et al.* 2005) as appropriate for our data. We modeled coalescence patterns with a Bayesian skyline demographic prior, applying a piecewise linear smoother (Drummond *et al.* 2005) after establishing its improved fit over the piecewise constant option. The molecular clock was parameterized with a normally distributed prior for the rate parameter with mean and variance set to the evolutionary rate estimates reported in Biek et al. (2007). Simultaneously, spatial locations at internal nodes were reconstructed as continuous traits at each tree in the posterior: Observed sample locations served as discrete priors for a Relaxed Random Walk model (RRW), which allows the variance in diffusion rate from node to descendent to vary from branch to branch within a topology (Lemey *et al.* 2009; Lemey *et al.* 2010). We tested a number of models for this diffusion process, and a gamma-distributed rate variation applied to an RRW was selected based on MLE. Because our samples were georeferenced to either a county or a zip code, our sample locations included 55 spatial duplicates, to which we added random spatial noise using a random "jitter" window of 0.5 before analysis with BEAST.

Finally, we ran our model for three independent runs, each of 300 million steps and sampling from the posterior distribution of trees every 30,000 steps, which yielded 10,000 posterior trees per run and high ESS values (1000+) for each estimated parameter within the runs.  Combining posterior outputs from the three runs caused ESS values to drop abruptly.). We determined this was caused by a monophyletic clade within our tree converging on different sister taxa between runs.  Since there were no practical differences between phylogenies holding this clade in different positions (the timing of its MRCA did not change, nor did its placement disrupt the branch lengths or topology of any other monophyly in our tree), we processed two more runs and ultimately combined the three that agreed.  We chose burn-in lengths based on posterior mixing patterns, discarding anywhere from 2,000 to 3,000 of the 10,000 trees per run before sampling from the remainder for combining in LogCombiner (Drummond & Rambaut 2007).  The Maximum Clade Credibility (MCC) tree drawn from the combined posteriors was summarized in TreeAnnotator.  For ease of comparisons we analyzed the N/G concatenated sequences using the same parameters and models as for the N data.

### Reconstructing spatial patterns of diffusion

Estimated ancestral locations and ages were used with sample locations and times to produce a two-dimensional surface model of viral diffusion rates.  Calculated simply as the geographic distance between a parent and daughter node and scaled by the time separating them, these diffusion rates are conditioned on phylogenetic history. To accommodate uncertainty in the estimated phylogeographic locations associated with nodes, as well as the phylogenetic dependence of inferred locations from neighboring nodes, 1000 trees of the N-gene drawn randomly from the posterior distributions logged by BEAST were used to build a data set of ancestral node heights and spatial coordinates.  For each tree, we determined temporal and geographic distances between parent and daughter nodes to obtain a branch-specific diffusion rate for each edge.  We assigned these diffusion rate estimates to point locations by generating

50 uniformly distributed (X, Y) coordinate pairs along the great-circle line connecting parent and daughter node. One intermediate point was randomly selected from this distribution and assigned a node height as a function of the height of the parent node, the distance of the intermediate from the parent, and the branch-specific dispersal rate. Applying this process to the sample of 1000 trees, yielded 344,000 georeferenced, time-stamped point estimates of viral diffusion rates.

Rate estimates and locations were used to model viral diffusion across Florida. A Generalized Additive Model (GAM – in R package mgcv (Wood 2011)) with penalized thin-plate regression smoothing was used to fit a model characterizing the relationship between diffusion rates and spatial coordinates. Predicted diffusion rates were then interpolated on a continuous grid of coordinates drawn from the rest of Florida. The optimal model function, a Gamma distribution with a logarithmic link, was selected by AIC.

## Simulations

As our genetic samples were drawn discontinuously in space and time, we examined the effect of sampling regime on our inference of the diffusion process. Simulated sequences were evolved on a grid that allowed equal probabilities of individuals diffusing to neighboring points consistent with a homogeneous landscape (see Duke-Sylvester, et al. (2013) for details), and sampled in spatiotemporal clusters representing different forms of sampling bias as well as the empirical sampling scheme used in our study (see SI).

## Phylogeographic analysis

Our analysis revealed the existence of five major RRV lineages in Florida that tended to cluster spatially (see Results). We used Delaunay triangulation of sample locations to visualize and track the distributions of these lineages over time. Delaunay triangulation is a method of calculating the two-dimensional geometric approximation of each point in a set, drawing polygons around each point such that every point in the set is closer to its own polygon than it is to any other point. We generated Delaunay diagrams for sample subsets corresponding to our three sampling time periods. For each period, we aggregated Delaunay cells by lineage membership and calculated the mean center of the resulting polygon, as well as the distance it had moved since the prior time slice.

### Reconstructing the source of mid-Atlantic RRV

The large RRV epizootic that emerged in the Virginias in the 1970's and rapidly spread throughout the mid-Atlantic states is thought to have been caused by the translocation of raccoons from Florida (Childs *et al.* 2000; Nettles *et al.* 1979; Rupprecht & Smith 1994). However, the origin and timing of this translocation event has never been identified. To infer the putative sequence of the virus that sparked the mid-Atlantic epizootic we applied marginal reconstruction in baseml (Yang 2007) to the maximum clade credibility tree of concatenated N and G sequences collected throughout the mid-Atlantic in an earlier study (Biek *et al.* 2007) to obtain the ancestral sequence at the root node. This sequence was incorporated into our phylogenetic analyses of the N-gene and the concatenated N/G sequence data sets from Florida.

## Results

### Phylogenetic analysis

Our 173 samples exhibited approximately 98% overall sequence identity and included 137 unique variants of the N gene. Phylogenetic analyses inferred five major clades, each supported by posterior probabilities of ~0.9 or more (Fig 1a). Relationships among these five clades were more difficult to resolve, with posterior probabilities of more ancestral nodes ranging from 0.3 to 0.7. The estimated rate of evolution was $3.1\text{x}10^{-4}$ substitutions/site/yr, yielding a date for the most recent common ancestor of 1944 (95% HPD=1933 to 1961). All five clades diverged rapidly in the years immediately following emergence of RRV; the 95% confidence intervals for the node height of all inferred lineage ancestors overlap (Table 1). We estimated that four clades were established by the late 1970s; the 95% HPD interval for the divergence time of clade IV extends to the early 1980s (Table 1, Fig 1a). All five clades were extant throughout the 30 years of our sample collection. Analysis of concatenated N and G gene sequences on a subset of our sample reaffirmed the patterns found with the N gene alone (Table 1 and SI 2).

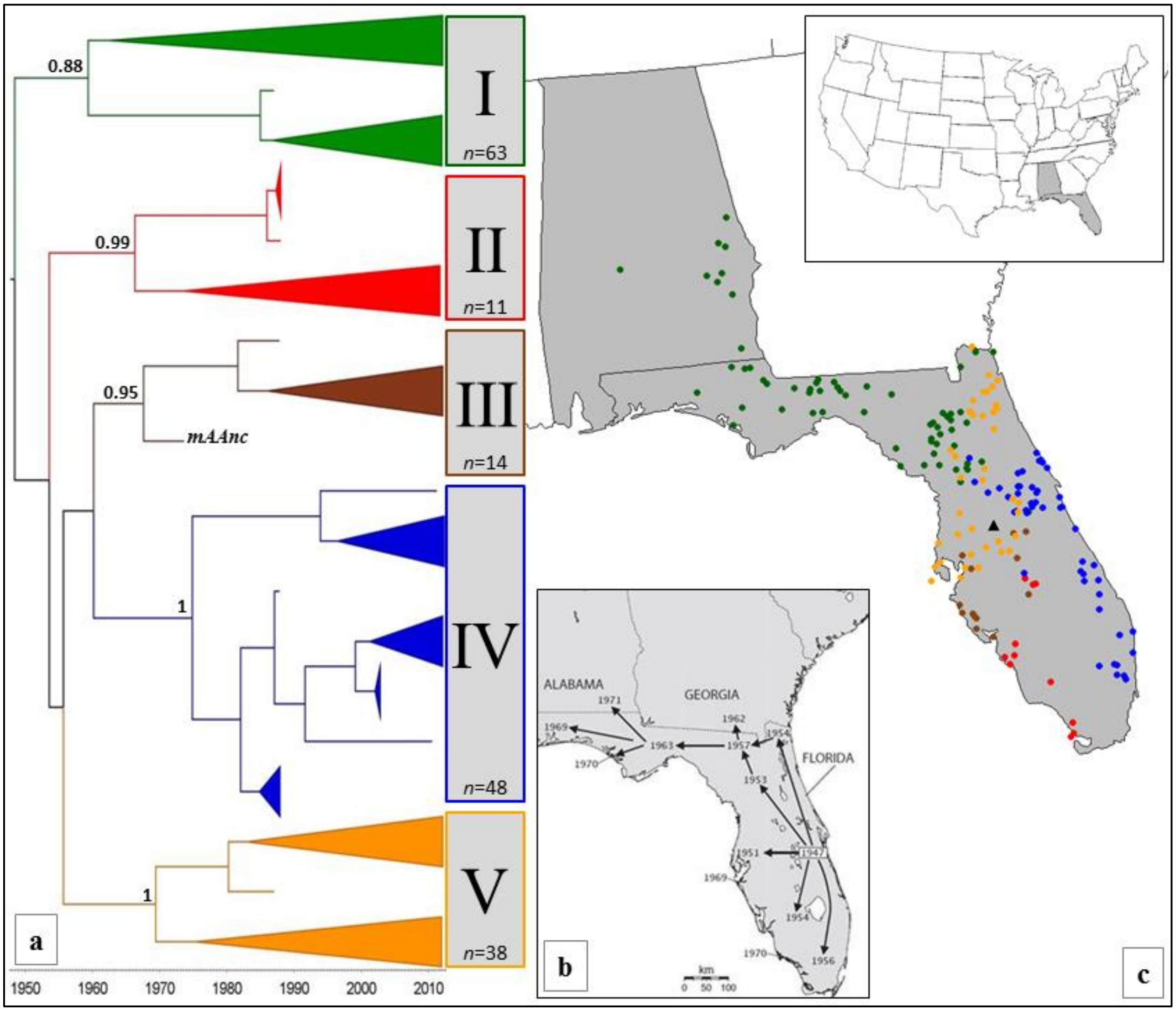


**Fig 1. Evolutionary relationships of raccoon rabies virus samples collected from southeastern US, with historic patterns of spread.** (**A**) Maximum clade credibility (MCC) tree from Bayesian coalescent analysis of N gene sequences, scaled to time. Node labels report posterior probabilities at each clade's ancestor. Groups within each clade are collapsed at posterior values <0.95. The sequence of the reconstructed ancestor of the mid-Atlantic RRV epizootic is labeled *mAAnc*. (**B**) RRV reported cases from the date and location of emergence, redrawn after Bigler et al. (1973). The first reported case is shown by a boxed date. (**C**) RRV positive samples sequenced at the N gene with their corresponding lineage assignments by color. The estimated tree root location is shown as a black triangle.

| | Estimated year of clade divergence [95% HPD] | |
|---|---|---|
| **Ancestor node** | **N-gene** | **N+G-genes** |
| Tree root | 1948 [1934, 1961] | 1957 [1940, 1969] |
| I | 1959 [1947, 1969] | 1962 [1947,1974] |
| II | 1966 [1954, 1976] | 1972 [1957, 1983] |
| III | 1968 [1959, 1973] | 1972 [1962, 1979] |
| IV | 1975 [1965, 1983] | 1985 [1977, 1991] |
| V | 1969 [1959, 1965] | 1971 [1957, 1981] |

**Table 1.** Median divergence times of each clade ancestor estimated using a time-scaled evolutionary rate applied to genetic data. All samples were sequenced at the nucleoprotein gene locus (N-gene), while a subset were analyzed at both the N and glycoprotein genes (N+G gene).

### Spatial patterns of diffusion

The spatiotemporal history of the N-gene phylogeny was reconstructed under a relaxed random walk model in continuous space (Lemey *et al.* 2010). Median viral diffusion rate estimated from posterior trees was 6.47 km/year (1$^{st}$ quartile=3.11, 3$^{rd}$ quartile=14.32), a value consistent with rates reported for RRV movement elsewhere (Biek *et al.* 2007; Lemey *et al.* 2010). Extracting the temporal and spatial locations of ancestral nodes from the posterior distribution of trees allowed us to interpolate a surface of parent-to-daughter diffusion rates across Florida (Fig 2), while accounting for uncertainty in the phylogeny and the coalescent model used to estimate those rates. The spatial distribution of estimated diffusion rates across the Florida peninsula did not reveal any obvious areas of heterogeneity, even under the relaxed random walk model, with the exception of minor clusters of elevated diffusion rates (Fig 2). These clusters corresponded to areas where we lacked samples. We confirmed using simulations that sequences evolved on a continuous landscape yielded similar spatial patterns of diffusion when sampled discontinuously (SI Fig 1c). We therefore propose that the “hotspots” of increased diffusion rates seen in the empirical data are likely the result of sampling gaps.

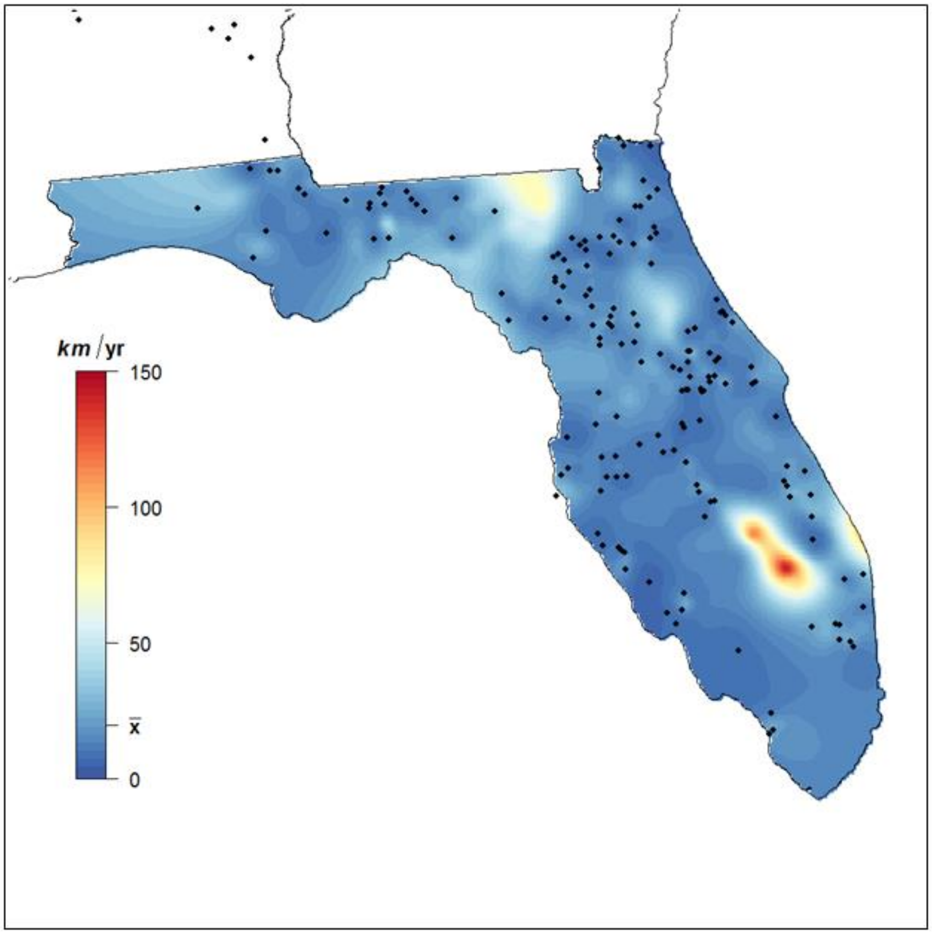


**Fig 2. Predicted averaged annual rate of raccoon rabies virus diffusion throughout the Florida peninsula.** Predicted diffusion was interpolated using spatial coordinates as predictor variables in a smoothed GAM of diffusion rates extracted from the posterior distribution of 1000 phylogenetic trees. The mean diffusion rate is indicated at x-bar. Increased diffusion rates are warmer in color. Sample locations are marked with black circles.

### Phylogeographic structure

The inferred clades exhibited strong geographic structure (Fig 1c), as our sample region can be partitioned into areas dominated by a certain clade. For example, clade IV is generally restricted to eastern Florida while sequences grouped in clade I are found in Alabama and the Florida panhandle (Fig 1c). All five clades can be found in the area around the estimated location of the tree's root (Fig 1c), positioned at a similar latitude but farther(?) to the west of the earliest reported case of RRV (Fig 1b). The geographic distribution of the five clades over time was remarkably stable, with mean centers during the second and third sampling period that remained within 100km to their estimated locations in the 1982 – 1988 sampling interval (Fig 3a, b). The geographic centers of clades I and V moved the least (<30 km total) (Fig 3b). Unlike clade I, however, clade V appears to have experienced some fragmentation by the expansion of clade IV (Fig 3a). The mean center of clade IV shifted a large distance from its estimated location, only to shift back the following sampling interval (Fig 3a, b). Only the mean centers of clades II and III exhibited repeated movement away from their respective locations in the 1982 – 1988 sampling interval (Fig 3a, b).

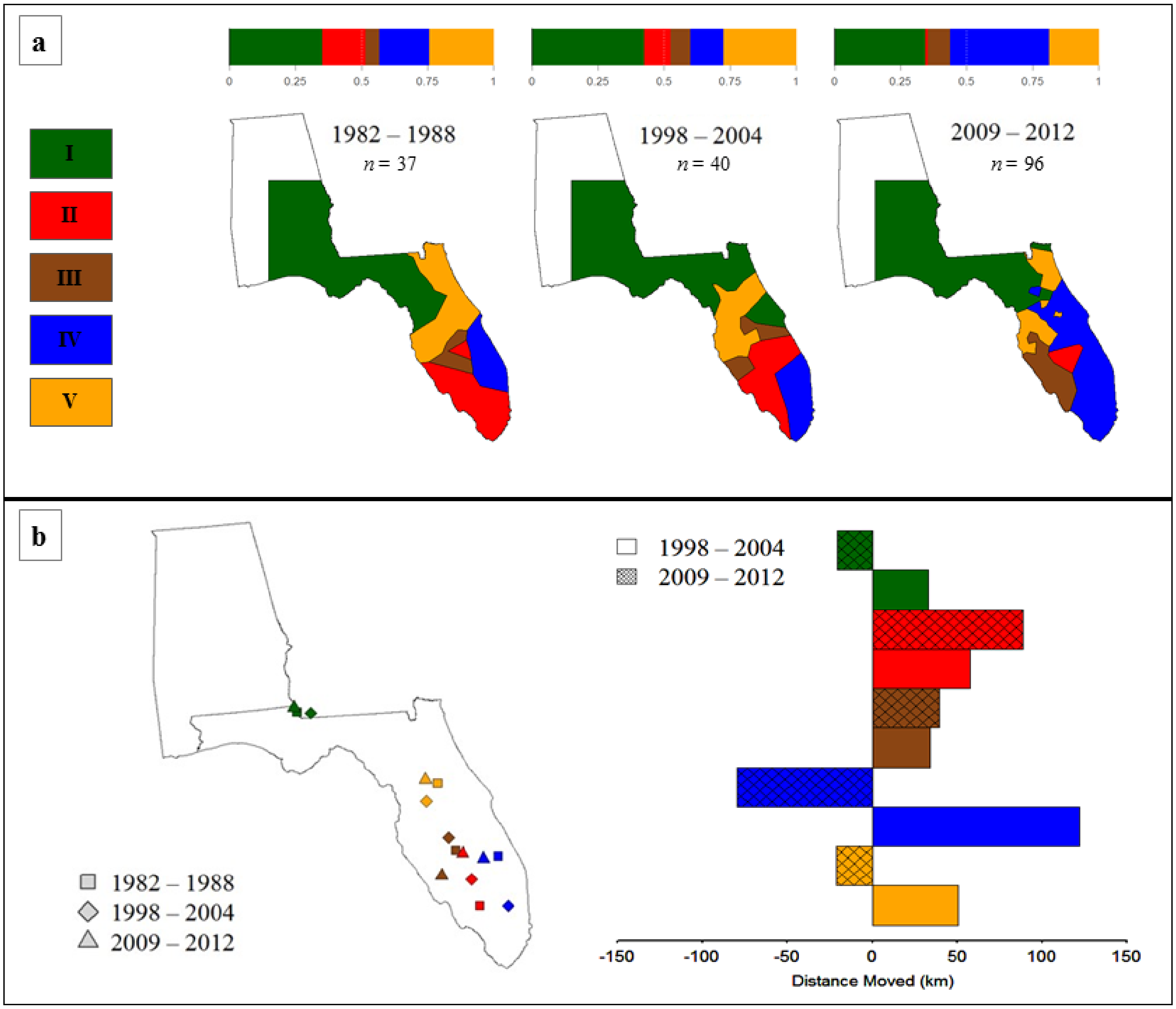


**Fig 3. Stable geographic distribution of raccoon rabies virus (RRV) lineages over time.** (**A**) Clade boundaries estimated from Delaunay triangulation at each general sampling interval. Colored squares indicate lineage number. The colored bars along the top indicate the proportion of samples assigned to a given lineage in each sampling interval. (**B**) Geometric mean center of each clade's triangulated area at each sampling interval, with the distances each mean center moved from the first sampling interval (1982 – 1988). Positive distances indicate a spatial shift away from the center's location in the first interval, negative distances are movements back toward that location.

## Mid-Atlantic origin

Included in our phylogenetic analyses was a viral sequence of the reconstructed ancestor of the mid-Atlantic RRV epizootic (Biek *et al.* 2007); the mid-Atlantic outbreak emerged in 1977 from a point far north of the RRV variant's original range in the southeastern states, likely due to translocation of an infected raccoon from Florida to West Virginia (Nettles *et al.* 1979; Rupprecht & Smith 1994). However, the exact spatial origin of the translocated raccoon has never been clarified. Our phylogenetic analysis placed the reconstructed ancestral sequence firmly within clade III, for both the N-gene-based phylogeny and the phylogeny constructed using both N and G genes (Figs 1a and SI 2). Both analyses indicated that this virus diverged early from the clade III ancestor. Clade III was one of the last clades to diverge and is restricted to a relatively small area in central western Florida. Our results thus suggest that the mid-Atlantic epizootic was caused by an infected raccoon originating from this area.

## Discussion

The results presented here provide evidence for the development and maintenance of spatial aggregates of rabies viral lineages on a landscape with few features that might impede host racoon movement. A key characteristic of rabies virus – indeed, of all rapidly-evolving RNA pathogens – is that there is a direct link between viral population dynamics and their molecular evolution, as mutations are fixed at the same tempo as population dynamics occur (Grenfell *et al.* 2004; Holmes 2004, 2009). This characteristic is central to the expectation that ecological processes are recorded in viral RNA genomes in near "real-time" (Drummond *et al.* 2003; Holmes 2008; Holmes 2009; Real *et al.* 2005a), yielding phylogenies that are rich in epidemiological information and predicted by viral ecology (Grenfell *et al.* 2004; Pybus *et al.* 2012; Real *et al.* 2005b). Without landscape features that significantly influence host movement, however, the spatial distribution and subsequent phylogeography of viral populations will be a result of host dynamics that occur at the local scale or that were written into the viral phylogeny during longer-term population processes such as colonization and epizootic expansion. We reconstructed evolutionary and ecological aspects of RRV back to the time and location of its original emergence in raccoons. The timescale of our inferred genealogy covers the known time period of RRV existence – the first reported case of rabies in raccoons occurred in 1947 (Bigler *et al.* 1973; Scatterday *et al.* 1960) (Fig 1b), which corresponds well with the estimated emergence date of our phylogenetic tree root. Reconstructed dispersal patterns suggest homogeneous viral diffusion across the state, a result that is consistent with the fact that the terrain of Florida presents few known barriers to raccoon movement, as well as with previous research supporting the presence of a single statewide raccoon population (Cullingham *et al.* 2008; Reeder-Carroll 2010). Early cases of raccoon rabies in Florida exhibited no spatial correlation – initial cases were sporadic and spatially disjoint – but RRV is believed to have covered the state by 1965 (Bigler *et al.* 1973). Localized outbreaks began to emerge in the late 1960s (Bigler *et al.* 1973; Kappus *et al.* 1970). We identified five distinct viral clades had diverged within ~20 years (by the early 1970s) following the mid-20th century emergence of RRV, with a phylogenetic structure that appears to be dominated by spatial radiation but in the absence of landscape-level effects. Despite the lack of landscape-level constraints on host movement, the distribution of the RRV lineages following their emergence has remained spatially compartmentalized and remarkably stable throughout the history of raccoon rabies in Florida.

Terrestrial rabies viral variants are frequently analyzed within the context of landscape heterogeneity and its effects on viral diffusion processes. In a few cases, phylogeographic structuring of terrestrial variants has been observed without the apparent influence of geographic barriers: Kuzmina et al. (2013) noted the presence of several spatially restricted, distinct phylogenetic lineages of skunk-specific rabies in Texas, remarking that there were no "obvious natural barriers restricting virus spread." An early investigation of fox rabies in Ontario, Canada suggested that viral lineages were segregated among distinct habitats (Nadin-Davis *et al.* 1999), however reanalysis of those data against a null model of isolation-by-distance found no evidence of genetic structuring due to ecological variables (Real *et al.* 2005a). There are many more examples of landscape heterogeneity driving the epidemiology and phylogeographic structure of terrestrial rabies (e.g., Biek *et al.* 2007; Cullingham *et al.* 2009; Smith *et al.* 2002; Talbi *et al.* 2010), which have identified a range of environmental predictors of rabies spread. However, our work shows that the existence of genealogical breaks in the virus phylogeny does not allow to conclude the existence of barriers to viral gene flow and transmission.

During population expansion, random events can generate spatial compartments of distinct evolutionary lineages (Cavalli-Sforza *et al.* 1993; Edmonds *et al.* 2004; Hallatschek *et al.* 2007; Waters *et al.* 2013). As new genetic lineages diverge and radiate outward, individuals located at their spatial centers are separated from sister clades by increasing distances, a phenomenon that should help solidify phylogeographic partitions as they form (Irwin 2002; Waters *et al.* 2013). Beyond colonization processes, it is also possible for phylogeographic structure to develop in established populations: Irwin (2002) noted that phylogeographic breaks emerged at random locations within simulated established and continuous populations modeled with low dispersal distances, even in the absence of physical barriers to gene flow. All the viral clades detected here diverged while rabies expanded through a novel host system, suggesting that colonization and expansion processes drove the phylogeographic patterns. Spatial compartmentalization of the five clades observed here remained stable throughout the history of RRV in our study area, with each clade restricted to distinct geographic areas that changed very little in their apparent arrangement. Our results suggest that the mean centers of each clade move very little – some shifting of clades' geographic centers was recorded throughout our sampling period, but the general locations of these centers were often preserved over multiple decades. Even clades experiencing fragmentation and/or contraction exhibited a constancy in mean center location, suggesting sampling variation as a possible cause rather than true shifts in distribution. This stability, as well as the absence of lineage turnover, suggest that spatial genetic patterns in rabies virus are preserved through time, long after the initial invasion process. It also highlights the overriding importance of local host movement processes, resulting in limited spatial admixture, in the maintenance of RRV. Interestingly, Hallatscheck *et al*. (2007) described similar spatial genetic partitions among expanding microbial colonies. The authors noted the inherent randomness to this process: The spatial arrangement of microbial diversity was ultimately determined by a few cells at the expanding edge contributing to the clear majority of future generations – and as these colonies were composed of genetically identical individuals expanding across a neutral landscape, it was not selection but random genetic drift that determined which microbial lines went on to dominate the gene pool (Hallatschek *et al.* 2007). Further, as the colonized range increases in size, core areas of distinct genetic lines are more likely to survive through time, buffered from turnover by average dispersal distance being less than the area of a genetic "sector" or spatial compartment (Hallatschek *et al.* 2007; Irwin 2002). Genetic lines or clades that are isolated from the expanding edge of the range, however, are consequently excluded from leaving much trace on the final phylogeny (Hallatschek *et al.* 2007). Using a measurably evolving pathogen, we are able to provide an empirical example of what has previously been demonstrated *in vitro* (Hallatschek *et al.* 2007) or using simulations (Irwin 2002).

A major goal of population genetics is to determine which of natural selection or neutral processes is driving a population's evolution. We find this system to be a clear example of neutral processes dominating population genetic structure. The stability of the phylogeographic patterns detected here enabled us to recover the potential source location of the mid-Atlantic rabies epizootic that began decades after RRV had reached an enzootic state in Florida. Crucially, we were able to identify this location even though our samples for this study were themselves collected long after RRV establishment in both Florida and in the mid-Atlantic states, as well as to reconstruct the source sequence of the mid-Atlantic epizootic. The clade from which the mid-Atlantic source appears to originate is not located near the expanding northern edge of RRV in the southeast. Instead, Clade III is surrounded by either water or by RRV sister clades. This event was enabled by human-mediated long-distance-dispersal (Nettles *et al.* 1979), without which we would predict that raccoon rabies would have continued its slow expansion up the east coast, and that Clade I – with exclusive access to the expanding edge of RRV in Florida – would have gone on to dominate future lineages.

RNA viruses in general have been identified as likely systems for evolutionary patterns driven by dispersal processes rather than selection, as they tend to cause acute infections that preclude co-divergence with their host species (Holmes 2004). Nevertheless, the rapid rate of rabies evolution has led to the expectation that its spatial genetic patterns will reflect its host population structure. Here, however, we find a significant and stable lack of spatial genetic congruence between RRV and its host, which displays no similar spatial compartmentalization (Cullingham *et al.* 2008; Reeder-Carroll 2010). Host contact rates and viral transmission dynamics are different, therefore, from those processes that drive host gene flow and migration patterns. Additionally, the ecological dynamics recoverable from this viral phylogeny my not have occurred on the same timescale as its sampling scheme. It is much more likely that the sweeping patterns set up during the disease's initial expansion are being recovered, while the stochastic and short-term dynamics of local movements are too transient to affect detectable phylogeographic changes (Carroll *et al.* 2007; Holmes 2004). Raccoon rabies control efforts that rely on landscape variables to assist in restraining viral movement (Elmore *et al.* 2017; Russell *et al.* 2005) will potentially not be applicable to systems such as this, and alternative methods of control will be necessary.

**Acknowledgements**

We thank the Florida Department of Health and the Center for Disease Control and Prevention for providing access to samples. We thank J. Caroline Henderson and Jennifer Snaman for their assistance in generating the sequence data. This research was supported by NIH grant no. RO1 AI047498.

## Data accessibility

The DNA sequences used for this paper have been submitted to Genbank and are available under accession nos. xxx-xxx.

AUTHOR CONTRIBUTIONS

TM, LAR and RB were involved in study design and concept. RD provided molecular data. TM carried out the analysis. SDS provided analytical tools and technical advice. TM wrote the manuscript with significant contributions from RB. All authors viewed and revised the final manuscript.

## References

Avise JC, Arnold J, Ball RM*, et al.* (1987) Intraspecific Phylogeography: The Mitochondrial DNA Bridge Between Population Genetics and Systematics. *Annual Review of Ecology and Systematics* **18**, 489-522.

Baele G, Lemey P, Bedford T*, et al.* (2012) Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular Biology and Evolution* **29**, 2157-2167.

Baele G, Li WL, Drummond AJ, Suchard MA, Lemey P (2013) Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. *Molecular Biology and Evolution* **30**, 239-243.

Baptestini EM, de Aguiar MA, Bar-Yam Y (2013) Conditions for neutral speciation via isolation by distance. *Journal of Theoretical Biology* **335**, 51-56.

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.

Biek R, Pybus OG, Lloyd-Smith JO, Didelot X (2015) Measurably evolving pathogens in the genomic era. *Trends in Ecology & Evolution* **30**, 306-313.

Bigler WJ, McLean RG, Trevino HA (1973) Epizootiologic aspects of raccoon rabies in Florida. *American Journal of Epidemiology* **98**, 326-335.

Brunker K, Hampson K, Horton DL, Biek R (2012) Integrating the landscape epidemiology and genetics of RNA viruses: rabies in domestic dogs as a model. *Parasitology* **139**, 1899-1913.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27260v1 | CC BY 4.0 Open Access | rec: 7 Oct 2018, publ: 7 Oct 2018

Carroll SP, Hendry AP, Reznick DN, Fox CW (2007) Evolution on ecological time-scales. *Functional Ecology* **21**, 387-393.

Cavalli-Sforza LL, Menozzi P, Piazza A (1993) Demic expansions and human evolution. *Science* **259**, 639-646.

Childs JE, Curns AT, Dey ME*, et al.* (2000) Predicting the local dynamics of epizootic rabies among raccoons in the United States. *Proc Natl Acad Sci U S A* **97**, 13666-13671.

Crisp MD, Trewick SA, Cook LG (2011) Hypothesis testing in biogeography. *Trends in Ecology & Evolution* **26**, 66-72.

Cullingham CI, Kyle CJ, Pond BA, Rees EE, White BN (2009) Differential permeability of rivers to raccoon gene flow corresponds to rabies incidence in Ontario, Canada. *Molecular Ecology* **18**, 43-53.

Cullingham CI, Kyle CJ, Pond BA, White BN (2008) Genetic structure of raccoons in eastern North America based on mtDNA: implications for subspecies designation and rabies disease dynamics. *Canadian Journal of Zoology* **86**, 947-958.

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **9**, 772.

Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG (2003) Measurably evolving populations. *Trends in Ecology & Evolution* **18**, 481-488.

Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7**, 214.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution* **22**, 1185-1192.

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* **29**, 1969-1973.

Duke-Sylvester SM, Biek R, Real LA (2013) Molecular evolutionary signatures reveal the role of host ecological dynamics in viral disease emergence and spread. *Philos Trans R Soc Lond B Biol Sci* **368**, 20120194.

Edmonds CA, Lillie AS, Cavalli-Sforza LL (2004) Mutations arising in the wave front of an expanding population. *Proc Natl Acad Sci U S A* **101**, 975-979.

Elmore SA, Chipman RB, Slate D*, et al.* (2017) Management and modeling approaches for controlling raccoon rabies: The road to elimination. *PLoS Negl Trop Dis* **11**, e0005249.

Excoffier L, Ray N (2008) Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution* **23**, 347-351.

Grenfell BT, Pybus OG, Gog JR*, et al.* (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327-332.

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52**, 696-704.

Hallatschek O, Hersen P, Ramanathan S, Nelson DR (2007) Genetic drift at expanding frontiers promotes gene segregation. *Proc Natl Acad Sci U S A* **104**, 19926-19930.

Hallatschek O, Nelson DR (2008) Gene surfing in expanding populations. *Theoretical Population Biology* **73**, 158-170.

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22**, 160-174.

Holmes EC (2004) The phylogeography of human viruses. *Molecular Ecology* **13**, 745-756.

Holmes EC (2008) Evolutionary history and phylogeography of human viruses. In: *Annual Review of Microbiology*, pp. 307-328.

Holmes EC (2009) *The evolution and emergence of RNA viruses* Oxford University Press.

Ibrahim KM, Nichols RA, Hewitt GM (1996) Spatial patterns of genetic variation generated by different forms of dispersal during range expansion. *Heredity* **77**, 282-291.

Irwin DE (2002) Phylogeographic breaks without geographic barriers to gene flow. *Evolution* **56**, 2383-2394.

Kappus KD, Bigler WJ, McLean RG, Trevino HA (1970) The raccoon an emerging rabies host. *Journal of Wildlife Diseases* **6**, 507-509.

Kuo CH, Avise JC (2005) Phylogeographic breaks in low-dispersal species: the emergence of concordance across gene trees. *Genetica* **124**, 179-186.

Kuzmina NA, Lemey P, Kuzmin IV*, et al.* (2013) The phylogeography and spatiotemporal spread of south-central skunk rabies virus. *PLOS ONE* **8**, e82348.

Lembo T, Haydon DT, Velasco-Villa A*, et al.* (2007) Molecular epidemiology identifies only a single rabies virus variant circulating in complex carnivore communities of the Serengeti. *Proc Biol Sci* **274**, 2123-2130.

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

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

Manel S, Schwartz MK, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution* **18**, 189-197.

Marske KA, Rahbek C, Nogués-Bravo D (2013) Phylogeography: spanning the ecology-evolution continuum. *Ecography* **36**, 1169-1181.

Miller MA, Pfeiffer W, Schwartz T (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees, 1-8.

Nadin-Davis SA, Casey GA, Wandeler AI (1994) A molecular epidemiological study of rabies virus in central Ontario and western Quebec. *Journal of General Virology* **75 ( Pt 10)**, 2575-2583.

Nadin-Davis SA, Sampath MI, Casey GA, Tinline RR, Wandeler AI (1999) Phylogeographic patterns exhibited by Ontario rabies virus variants. *Epidemiology and Infection* **123**, 325-336.

Neigel JE, Avise JC (1993) Application of a random walk model to geographic distributions of animal mitochondrial DNA variation. *Genetics* **135**, 1209-1220.

Nettles VF, Shaddock JH, Sikes RK, Reyes CR (1979) Rabies in translocated raccoons. *American Journal of Public Health* **69**, 601-602.

Orciari LA, Niezgoda M, Hanlon CA*, et al.* (2001) Rapid clearance of SAG-2 rabies virus from dogs after oral vaccination. *Vaccine* **19**, 4511-4518.

Pybus OG, Rambaut A (2009) Evolutionary analysis of the dynamics of viral infectious disease. *Nature Reviews: Genetics* **10**, 540-550.

Pybus OG, Suchard MA, Lemey P*, et al.* (2012) Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc Natl Acad Sci U S A* **109**, 15066-15071.

Real LA, Henderson JC, Biek R*, et al.* (2005a) Unifying the spatial population dynamics and molecular evolution of epidemic rabies virus. *Proc Natl Acad Sci U S A* **102**, 12107-12111.

Real LA, Russell C, Waller L, Smith D, Childs J (2005b) Spatial dynamics and molecular ecology of North American rabies. *Journal of Heredity* **96**, 253-260.

Reeder-Carroll SA (2010) *Population genetics of raccoons in the Eastern United States with implications for rabies transmission and spread* Ph.D., Emory University.

Rupprecht CE, Smith JS (1994) Raccoon rabies: the re-emergence of an epizootic in a densely populated area. *Seminars in Virology* **5**, 155-164.

Russell CA, Smith DL, Childs JE, Real LA (2005) Predictive spatial dynamics and strategic planning for raccoon rabies emergence in Ohio. *PLoS Biology* **3**, e88.

Saunders NC, Kessler LG, Avise JC (1986) Genetic Variation and Geographic Differentiation in Mitochondrial DNA of the Horseshoe Crab, LIMULUS POLYPHEMUS. *Genetics* **112**, 613-627.

Scatterday JE, Schneider NJ, Jennings WL, Lewis AL (1960) Sporadic animal rabies in Florida. *Public Health Reports* **75**, 945-953.

Schwartz MK, McKelvey KS (2008) Why sampling scheme matters: the effect of sampling scheme on landscape genetic results. *Conservation Genetics* **10**, 441-452.

Shapiro B, Rambaut A, Drummond AJ (2005) Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Molecular Biology and Evolution* **23**, 7-9.

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.

Smith JS, Fishbein DB, Rupprecht CE, Clark K (1991) Unexplained rabies in three immigrants in the United States. A virologic investigation. *N Engl J Med* **324**, 205-211.

Smith JS, Reid-Sanden FL, Roumillat LF*, et al.* (1986) Demonstration of antigenic variation among rabies virus isolates by using monoclonal antibodies to nucleocapsid proteins. *Journal of Clinical Microbiology* **24**, 573-580.

Szanto AG, Nadin-Davis SA, Rosatte RC, White BN (2011) Genetic tracking of the raccoon variant of rabies virus in eastern North America. *Epidemics* **3**, 76-87.

Talbi C, Lemey P, Suchard MA*, et al.* (2010) Phylodynamics and human-mediated dispersal of a zoonotic virus. *PLoS Pathog* **6**, e1001166.

Towns J, Cockerill T, Dahan M*, et al.* (2014) XSEDE: accelerating scientific discovery. *Computing in Science & Engineering* **16**, 62-74.

Waters JM (2011) Competitive exclusion: phylogeography's 'elephant in the room'? *Molecular Ecology* **20**, 4388-4394.

Waters JM, Fraser CI, Hewitt GM (2013) Founder takes all: density-dependent processes structure biodiversity. *Trends in Ecology & Evolution* **28**, 78-85.

Wheeler DC, Waller LA (2008) Mountains, valleys, and rivers: The transmission of raccoon rabies over a heterogeneous landscape. *J Agric Biol Environ Stat* **13**, 388-406.

Wilson ML, Bretsky PM, Cooper GH, Jr., *et al.* (1997) Emergence of raccoon rabies in Connecticut, 1991-1994: spatial and temporal characteristics of animal infection and human contact. *American Journal of Tropical Medicine and Hygiene* **57**, 457-463.

Wood SN (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**, 3-36.

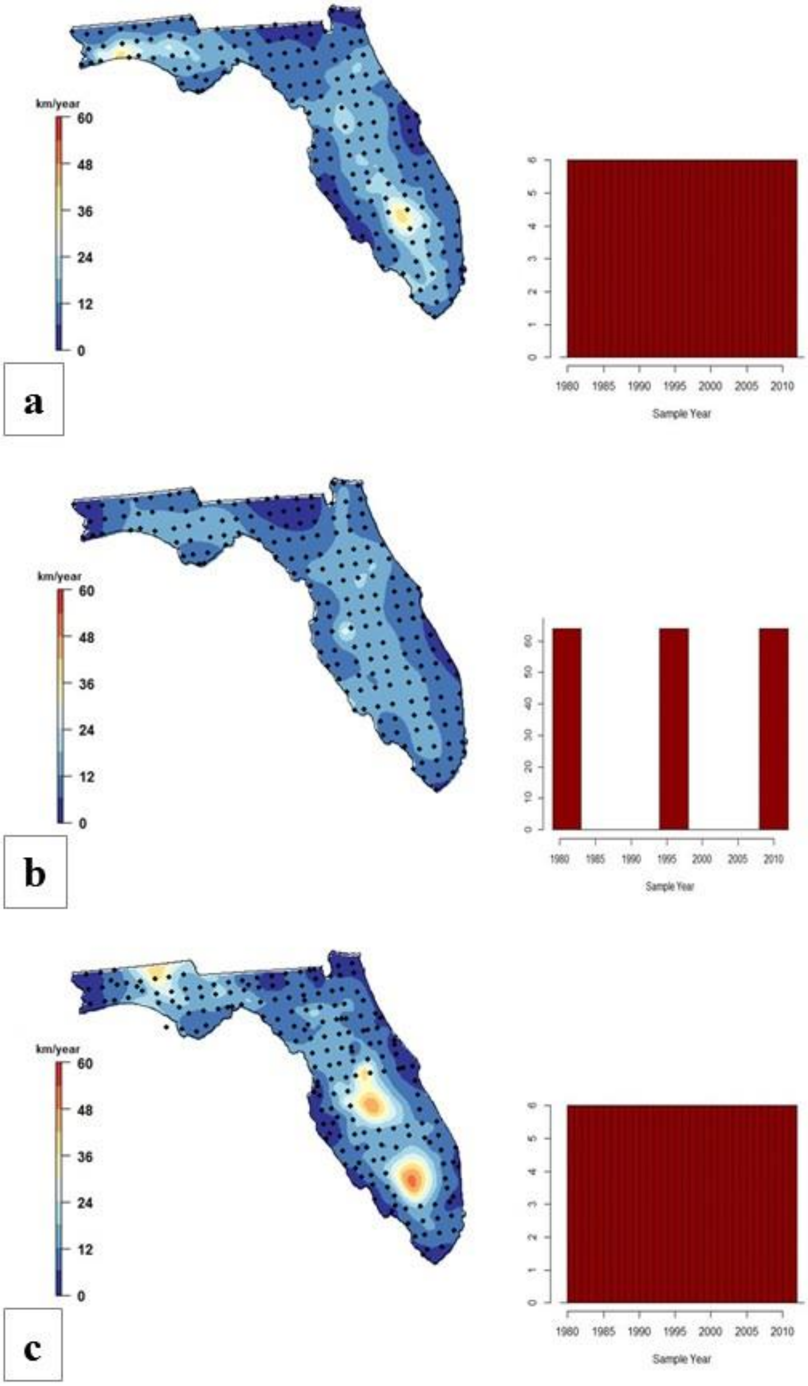
Wright S (1943) Isolation by Distance. *Genetics* **28**, 114-138.

Wright S (1951) The genetical structure of populations. *Ann Eugen* **15**, 323-354.

Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* **39**, 306-314.

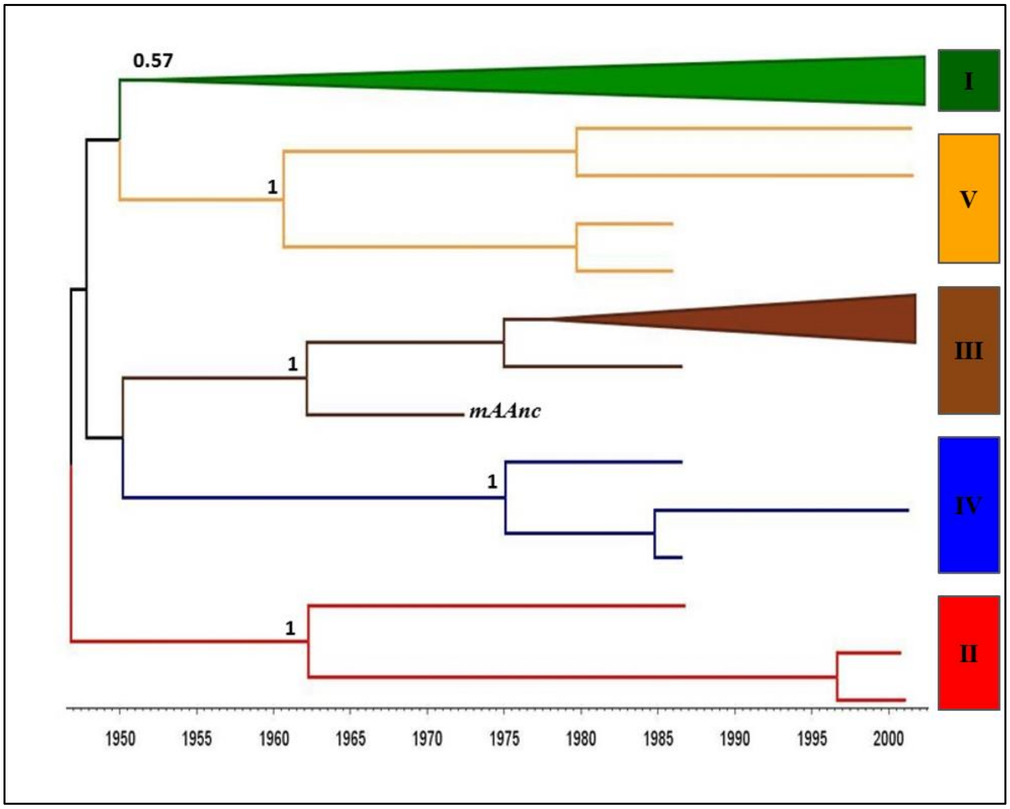
Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**, 1586-1591.

## Supplemental Information



***SI 1***

**SI Fig 1. Diffusion rates estimated from simulated sequence evolution on homogeneous landscapes and sampled under different schemes**. Host movement was modeled on a regular grid clipped to the shape of Florida, Georgia, and Alabama, with an index infection introduced at the site of the first reported case of RRV (Fig 1b) following five years of host movement. Host movement served as the backdrop for disease transmission and sequence evolution for a period of 75 years (see Duke-Sylvester et. al (2013) for details), after which sequences were sampled from host locations throughout Florida during the final 35 years of the model. Sampled sequences were analyzed in BEAST to estimate ancestral locations and diffusion rates for spatial interpolation of disease diffusion rates using a GAM. Sequences were collected (**A**) uniformly through space and time, (**B**) uniformly through space but discontinuously through time, and (**C**) continuously through time but discontinuously in space.



***SI 2***

**SI Fig 2. Evolutionary relationships of raccoon rabies virus among 20 samples using concatenate N and G gene sequences (1965bp)**. Maximum clade credibility (MCC) tree scaled to time, with posterior probabilities of each clade's ancestor at nodes (relationships are collapsed where posterior values are less than 0.95). The sequence of the reconstructed ancestor of the mid-Atlantic racoon rabies epizootic is indicated at individual *mAAnc*.