Molecular evolutionary signatures reveal the role of host ecological dynamics in viral disease emergence and spread

Scott M. Duke-Sylvester1,2, Roman Biek4,5 and Leslie A. Real2,3,6

1Department of Biology, University of Louisiana at Lafayette, Lafayette, LA 70506, USA
2Center for Disease Ecology, and 3Department of Biology, Emory University, Atlanta, GA 30322, USA
4MRC-University of Glasgow Centre for Virus Research and Institute of Biodiversity, Animal Health and
5Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow,
Glasgow G12 8QQ, UK
6Fogarty International Center, National Institutes of Health, Bethesda, MD 20892, USA

RNA viruses account for numerous emerging and perennial infectious diseases, and are characterized by rapid rates of molecular evolution. The ecological dynamics of most emerging RNA viruses are still poorly understood and difficult to ascertain. The availability of genome sequence data for many RNA viruses, in principle, could be used to infer ecological dynamics if changes in population numbers produced a lasting signature within the pattern of genome evolution. As a result, the rapidly emerging phylogeographic structure of a pathogen, shaped by the rise and fall in the number of infections and their spatial distribution, could be used as a surrogate for direct ecological assessments. Based on rabies virus as our example, we use a model combining ecological and evolutionary processes to test whether variation in the rate of host movement results in predictive diagnostic patterns of pathogen genetic structure. We identify several linearizable relationships between host dispersal rate and measures of phylogenetic structure suggesting genetic information can be used to directly infer ecological process. We also find phylogenetic structure may be more revealing than demography for certain ecological processes. Our approach extends the reach of current analytic frameworks for infectious disease dynamics by linking phylogeography back to underlying ecological processes.

1. Introduction

Ecological and evolutionary processes are inevitably and intrinsically linked [1]. However, the importance of coupled ecological–evolutionary interaction is often underappreciated owing to a general perception that ecological and evolutionary processes operate on vastly different time-scales. Evolutionary processes are often thought of as occurring over centuries to millennia while ecological dynamics are measured in months or decades. The discordance of time-scale leads to a perception that ecological and evolutionary processes can be treated independently. However, theoretical results that embrace both sets of processes consistently point towards the importance of ecological and evolutionary interactions even when rates are dissimilar [2–5]. Several recent studies also suggest that comparable ecological and evolutionary rates for plants and animals may be more common than previously assumed [6–9].

Interactions between ecological and evolutionary processes are nowhere more evident than in the dynamics of infectious diseases. Pathogens can evolve rapidly under strong selective pressure to develop resistance to antimicrobial drugs and to avoid host immune system defences [10,11]. Accelerated rates of molecular substitution and neutral evolution are also observed, and are a result of several features of pathogen biology that enhance the emergence of new genotypes [12,13]. Many pathogens are characterized by generation times measurable in hours or days and by high rates of population growth and turnover. In addition, a rapid succession of population bottlenecks as
infection spreads to new hosts and limited gene flow among demes can promote local genetic differentiation. Rates of molecular evolution are also accelerated, because of limited fidelity during genome replication. This situation is most prominent among single-stranded RNA viruses such as hepatitis C, measles and influenza. Such viruses lack a mechanism for correcting errors during replication, resulting in evolutionary rates ranging from $10^{-2}$ to $10^{-4}$ substitution per site per year, several orders of magnitude faster than observed rates in plants and animals [12,13].

Over the course of an epidemic, a rapid evolutionary rate will result in the emergence of multiple viral genotypes [14,15]. A phylogenetic tree inferred from these genotypes traces the pattern of ancestry among viral lineages and transmission between hosts. The structure of the phylogeny, and the timing and frequency of bifurcations are correlated with transmission events and the effective number of infections ($N_e$) [16–18]. Through the application of coalescent methods, such as skyline plots, the demographic history of a viral population can be reconstructed from the pathogen's phylogeny, linking genetic and ecological dynamics [18–20]. As an infection expands, the virus' genome is passed from infected to susceptible hosts, analogous to offspring receiving a genome from their parents. As a consequence, the skyline plot and the viral phylogeny are a reflection of the infection dynamics in the host population.

The linkage between ecological and evolutionary dynamics occurring on the same time-scale suggests the possibility that the evolutionary history of a virus can be used to distinguish among different ecological processes governing the spread of disease [17]. We suggest that different ecological processes will result in characteristic differences in evolutionary history making the former recoverable as distinct 'signatures' from virus sequence data. We also suggest these signatures may persist over time so that past ecological processes can be identified from viral sequence data collected years or decades after the fact.

One candidate process for establishing linkages between ecological and behavioural dynamics and the formation of lasting evolutionary genetic signatures is the dispersal of both host and pathogen. At present, we lack general approaches for relating spatial-genetic patterns seen in rapidly evolving pathogens back to underlying ecological and epidemiological processes [21]. Opportunities for spatial diffusion vary greatly between different types of virus, such that some are characterized by predominantly local contact (e.g. measles [22]), whereas long-distance translocation (LDT) has been highly influential in the spread of many other viral diseases of public health concern (e.g. monkeypox, severe acute respiratory syndrome, influenza, HIV and West Nile virus [23–25]).

LDT facilitates the establishment of new centres of infection that are far from existing infected populations and can contribute to the rapid spread of disease. Individuals in many wildlife species are capable of traversing long distances, and in the process transporting their diseases [26,27]. LDT can also be the result of deliberate or unintentional human actions, as illustrated by the global dissemination of many human and animal diseases over the last few centuries [23,24,28].

We hypothesize that the presence of LDT during an epidemic, as a specific mechanism of movement that adds to the otherwise local processes of dispersal, will produce recognizable patterns in the viral phylogeny. Differences in phylogenetic branching patterns are expected as a result of the increased growth afforded by LDT. In the absence of LDT the advancing wave of infection is expected to travel at a constant velocity, $v$ [29,30]. As a result we expect the area, and number of infected hosts, to grow in proportion to the square of time ($N_e \tau \propto \pi(vt)^2$). On the other hand, we expect growth with LDT to be exponential. This follows from the exponential growth in the number of new foci of infection that emerge early in the epidemic. Considering a landscape composed of $N$ patches that are either infected or uninfected: at any time, the number of already infected hosts dispersing by LDT is proportional to the number of infected patches. LDT allows each disperser to potentially reach a large number of patches, and earlier in the epidemic most of these will still be uninfected. As a result, the rate of change in the number of infected hosts is also proportional to the current number of infected patches. That is $dN/dt = \lambda N$ where $\lambda$ is the growth rate in the number of infected patches. The net effect in this case is sustained exponential growth of the epidemic ($N_e \tau = Ce^{\lambda t}$).

Are these two cases of epidemic expansion, with and without LDT, distinguishable from their resulting viral phylogenies?

We also expect there to be detectable differences in the spatial arrangement of viral lineages associated with changes in the mode or rate of dispersal. One manifestation of these differences is a decline in isolation by distance (IBD). IBD is the amount of genetic variance between lineages that can be attributed to the geographical distance that separates those lineages [31]. Moreover, as the geographical distance between lineages increases, the genetic distance between lineages is expected to increase. Dispersal, on the other hand, is expected to decrease IBD because it tends to genetically homogenize populations and reduce the amount of between group variation. What remains unclear, however, is the specific form of the relationship between rates of movement and changes in IBD. We further expect the area occupied by a viral lineage, as well as its spatial overlap with other lineages, to increase with the rate of LDT.

Here, we explore the interaction between the molecular evolution of a virus and ecological processes that govern its spread, and how these interactions manifest themselves in the structure of the viral phylogeny. Our results are based on a spatially explicit SIR-type model [32] that includes the molecular evolution of the virus that occurs concurrently with the spread of infection. For illustrative purposes, we focus on the spatial spread of raccoon rabies and the concurrent molecular evolution of the rabies virus. Raccoon rabies is an excellent case study for exploring our hypothesis. The rabies virus genome consists of a single strand of RNA, eliminating the complicating effects of reassortment [33], shows little or no recombination, and has a rapid evolutionary rate (approx. $3 \times 10^{-4}$ substitution per site per year) [34]. In addition, the rabies strain associated with raccoons is thought to be exclusively maintained and dispersed by this host species [33]. We compare the phylogenetic structure of the rabies virus that emerges under different estimated rates of LDT [35] with the following research aims: first, we use simulated data based on reasonable parameter values to determine whether the presence of LDT during an epidemic can be reliably detected from genetic data collected retrospectively. Second, we test whether evidence for LDT can be detected in empirical genetic data from a large raccoon rabies epidemic in the USA. Finally, we assess the spatial-genetic patterns predicted to arise from different levels of
emerge in the viral phylogeny can be attributed to changes in molecular evolution of the virus along the branches and nodes from one host to another. Finally, our model simulates the infectious pathogen. The second part of our model tracks population that emerge under the spread and persistence of LTD. We discuss our findings in the context of a rising need for inference methods producing epidemiological insights from pathogen sequence data.

2. Material and methods

Our model is a discrete stochastic model of the spatial dynamics of the host population that emerge under the spread and persistence of an infectious pathogen. The second part of our model tracks the bifurcating tree of infections created as infection is spread from one host to another. Finally, our model simulates the molecular evolution of the virus along the branches and nodes of the infection tree. We use our model to simulate viral phylogeny that emerges under six different rates of LTD (\( \phi_{\text{LDT}} \)). The equations for our ODE model are

\[
\begin{align*}
S_i' &= aN_i - bNiSi - \beta_iE_i - (\phi + \phi_{\text{LDT}})Si + \sum_{j} (\delta_{ij} + \phi_{\text{LDT}}k_{ij})S_j, \\
E_i' &= \beta_iE_i - bNiE_i - \alpha_iE_i - (\phi + \phi_{\text{LDT}})E_i + \sum_{j} (\delta_{ij} + \phi_{\text{LDT}}k_{ij})E_j, \\
I_i' &= \alpha_iE_i - \beta_iI_i - \alpha_iI_i - (\phi + \phi_{\text{LDT}})I_i + \sum_{j} (\delta_{ij} + \phi_{\text{LDT}}k_{ij})I_j,
\end{align*}
\]

where \( S_i, E_i, \) and \( I_i \) are the number of susceptible, exposed, and infected individuals at location \( i \), respectively. The total population size in location \( i \) is \( N_i \).

In the absence of infection, population growth is logistic with an intrinsic rate of increase given by \( a \), and a density-dependent mortality rate, \( b \). This results in a stable disease-free equilibrium at \( (S_0, E_0, I_0) = (a/b, 0, 0) \) in all locations \( i \). Our estimate for the intrinsic rate of increase, \( a \), is 3.67, \( 10^{-3} \) day\(^{-1} \), and is based on a mean birth rate of 1.33 kits per year [37–39]. Our estimate for the density-dependent death rate \( b = 2.29375 \times 10^{-7} \) day\(^{-1} \) yields 16,000 raccoons per county, equivalent to a density of 10 raccoons km\(^{-2} \) [40,41]. Our estimated values for these and all other model parameters are summarized in table 1.

The infection of susceptible individuals occurs at a per capita rate of \( \beta_i \) where \( \beta \) is the transmission coefficient. Values for \( \beta \) in epidemiological models are often difficult to estimate, because infectious contacts are rarely observed [48]. Given the other model parameters, our estimate for \( \beta \) reproduces the observed 48-month cycle in infectious raccoons [42], and observed net reproductive rate, \( R_0 \), for raccoon rabies of approximately 1.2 [34]. The transmission coefficient \( \beta \) must also be scaled to match the spatial resolution of the model since the spread of infection between individuals in our model is a density-dependent process. In estimating values for the transmission rate and emigration rates, we applied a heuristic algorithm that evaluated \( \beta \) in increments of \( 10^{-5} \), while holding other model parameters fixed to find a value that minimized the absolute difference between the value of a statistic for the model and observed data.

Following infection, there is a latency period during which the virus spreads within a host from the point of infection, through peripheral nerve tissue and into the central nervous system [33]. This is included in the model by the term \( aE_i \) in equations (2.3) and (2.4), where 1/\( \sigma \) is the expected length of the latency period. The length of the latency period in raccoons is highly variable and ranges from 18 to 107 days with a mean of approximately 1/\( \sigma \) = 50 days [43–45]. The invasion of the central nervous system is rapidly followed by the migration of virus into the salivary glands and the shedding of virus in the saliva. At this point hosts are infectious, and this phase corresponds with the onset of the end-stage symptoms typical of rabies (e.g. hyper-aggression, hydrophobia, foaming saliva) [33]. Death follows quickly after the onset of symptoms and is included in the model through the term \( al_i \), where 1/\( \alpha \) is the

### Table 1. Parameter description for ODE model (1–5) and the derived stochastic model and the estimated parameter values used in our results.

<table>
<thead>
<tr>
<th>parameter</th>
<th>description</th>
<th>value</th>
<th>references</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a )</td>
<td>per capita birth rate</td>
<td>( 3.67 \times 10^{-3} ) day(^{-1} )</td>
<td>[37–39]</td>
</tr>
<tr>
<td>( b )</td>
<td>non-disease induced, density-dependent death rate</td>
<td>( 2.293 \times 10^{-2} ) day(^{-1} )</td>
<td>[40,41]</td>
</tr>
<tr>
<td>( \beta )</td>
<td>contact rate</td>
<td>( 1 \times 10^{-8} ) (animal days(^{-1} ))</td>
<td>[34,42]</td>
</tr>
<tr>
<td>1/( \sigma )</td>
<td>latency period</td>
<td>50 days</td>
<td>[43–45]</td>
</tr>
<tr>
<td>1/( \alpha )</td>
<td>infectious period</td>
<td>14 days</td>
<td>[33,43,46]</td>
</tr>
<tr>
<td>( \phi )</td>
<td>local movement rate</td>
<td>( 6 \times 10^{-6} ) day(^{-1} )</td>
<td>[34,47]</td>
</tr>
<tr>
<td>( \phi_{\text{LDT}} )</td>
<td>long-distance translocation rate</td>
<td>( 6 \times 10^{-2} ) day(^{-1} )</td>
<td>[35]</td>
</tr>
<tr>
<td>( k_{ij} )</td>
<td>fraction of emigrants from ( j ) to ( i )</td>
<td>[ \begin{array}{ll} 0.125 &amp;</td>
<td>\text{row}_i -</td>
</tr>
</tbody>
</table>
life expectancy of an infectious raccoon. The length of this period is extremely short, on the order of 14 days [33].

Our model includes both local dispersal and LTD of raccoons. Local dispersal is the movement of raccoons over the landscape in search of food, mates and shelter. Local dispersal of hosts is included in the model equations, where the probability distribution of events that may occur at each step. Our equations accumulate independently (governed by a common evolutionary process) and the substitutions that characterize each viral population and genetic distances to assess the magnitude of IBD. For each replicate run, we extracted the infection tree that connected 100 infected raccoons through a skyline plot analysis [20]. For each replicate run, we extracted the infection tree that connected 100 infected raccoons selected at random, and simulated the molecular evolution of the virus among the branches and nodes of the infection tree. Once a new host has been infected, the processes of mutation and fixation occur independently in the two viral populations, and the substitutions that characterize each viral population accumulate independently (governed by a common evolutionary model and rate). The accumulated differences in the sequence are the basis for the molecular diversification of the viral phylogeny. Further, because of the rapid evolutionary rate of RNA viruses, the structure of viral phylogeny closely follows the structure of the infection tree (see the electronic supplementary material, figure S2).

(d) Skyline plot analysis

We measure the overall growth of an epidemic based on the rate of increase in the effective number of infections \(N_e \tau\) revealed through a skyline plot analysis [20]. For each replicate run, we extracted the infection tree that connected 100 infected raccoons selected at random, and simulated the molecular evolution of the virus among the branches and nodes of the infection tree to produce a viral phylogeny. We then computed a generalized skyline plot for each viral phylogeny using the R package ape [51] to estimate the change over time in the effective number of infections \(N_e \tau\). Next, we used maximum-likelihood to fit both a time-squared \((N_e \tau = \pi(t + d)^2)\) and exponential \((N_e \tau = Re^{\alpha})\) model of growth to the log-transformed time-series of \(N_e \tau\) values produced by the skyline analysis. We used \(\Delta AIC\) to determine which of the two growth models was a better fit.

(e) Spatial analysis

For each rate of LTD, we correlated the pairwise geographical and genetic distances to assess the magnitude of IBD. For each correlation we use the slope \(m_{\text{linear}}\) and the correlation coefficient \(r_{\text{linear}}\) of IBD to measure the strength of the relationship between geographical and genetic distance. The slope \(m_{\text{linear}}\) of the probability mass function. At each step in the model, the expected time, \(1/r\), and the new probability mass function are recomputed from current values for the rates of each event type. The time to the next event is drawn at random from an exponential distribution with mean \(1/r\). Time is updated to the next event time and the type of event that occurs is determined from the probability mass function. The process repeats until the desired simulation length is reached.
correlation indicates the rate at which the genetic distance between samples increases with increasing geographical distance. The IBD correlation coefficient ($R_{IBD}^2$) is a measure of the amount of variation accounted for by a linear relationship. The slope and correlation coefficient together reflect the strength of the correlation and the degree of IBD.

The spatial spread of infection causes viral genotypes to occur in clusters that partition the landscape into patches (see the electronic supplementary material, figure S3) [34,52]. We computed the total area and the area of overlap of the geographical ranges encompassed by clusters of related viral genotypes. The area and overlap of genetic clusters explicitly includes the hierarchical details of the viral phylogeny. A single convex polygon (i.e. one with no concave features) that just encompasses a set of related viral genotypes can be used to define the area occupied by a viral lineage. The area of such a polygon is expected to increase as the rate of LDT is increased because related viral genotypes are more likely to be separated at higher rates of dispersal. The aggregation of genetic lineages into groups contains an element of arbitrariness. We address this issue by defining the endemic phase groups in terms of transition time $t_T$, at which the epidemiological dynamics shift from epidemic to endemic. Specifically, we are interested in those groups of viruses where all isolates share a common ancestor ‘a’ that is more recent than $t_T$ and that maximize the number of isolates within each group.

(f) Sensitivity analysis
We performed a sensitivity analysis of our model to see if our results were robust to changes in the model parameters. We performed an additional 12 runs in which we either increased or decreased by 10 per cent the value of each of the six model parameters ($a$, $b$, $\beta$, $\sigma$, $\alpha$, $\phi$). We also considered an increased spatial domain for our model (50 30 cells) to explore the possibility that our results were scale-dependent or influenced by close proximity of the landscape boundaries along the long axis of the model domain.

In contrast to our model, real-world landscapes are spatially heterogeneous which can introduce noise to biological processes that makes detecting patterns difficult. For example, random variations in habitat quality may result in spatial variation in the demography of the host population and may alter the pattern and tempo of an epidemic. We addressed the question whether such spatial heterogeneity could obscure the phylogenetic signatures associated with different rates of LDT by making the local carrying capacity of each cell and the rate of movement between cells spatially random variables.

To vary carrying capacity, we assigned a random value to the per capita birth rate, $a$, to each location in our simulated landscape. Random values were drawn from a lognormal distribution with a mean equal to the baseline birth rate ($a_b$), to each location in our simulated landscape. The variance for the distribution was chosen to place half of distribution area between $\pm$10% of the mean value and half outside these limits (see the electronic supplementary material, figure S4). This resulted in the carrying capacity ranging between 7.5 and 13.3 raccoons km$^{-2}$ at 95 per cent of locations. Spatial variation in the dispersal rates was created using the same method of assigning a random deviate to each location from a lognormal distribution centred at the mean local dispersal rate ($\phi = 6 \times 10^{-1}$ day$^{-1}$).

For each replicate run of the model we created a new random landscape that was held constant over the course of each replicate. We considered scenarios that included stochastic variation in the carrying capacity only, variation in the dispersal rate only and a scenario in which spatial variation in both demographic parameters was combined.

3. Results
We find that there are unique signatures embedded within the structure of the viral phylogeny produced by different rates of LDT. Skyline plots reveal the effects of LDT as changes in the growth rate of the effective number of infections ($N_e$; figure 1a). As we hypothesized, there are discernible and diagnostic differences in the pattern of growth characteristics of the epidemic with and without LDT. Our
Table 2. Percentage of replicates in which the epidemic growth is better described by either a time-squared or exponential model of growth. ΔAIC values comparing maximum-likelihood fits of an exponential model of epidemic growth and a time-squared model of growth. For each comparison we computed the difference between the AIC for the exponential model and the time-squared model (ΔAIC = AIC_{exponential} − AIC_{time-squared}). (Smaller AIC values are associated with a better fit.) Positive ΔAIC values indicate a smaller AIC value for the time-squared and a better fit by the time-squared model of growth, while negative ΔAIC indicate a better fit by the exponential model of growth. (Absolute number of replicates better described by each growth model are shown parenthetically.)

<table>
<thead>
<tr>
<th>LDT rate (Δφ_{LDT})</th>
<th>N replicatesa</th>
<th>time-squared growth</th>
<th>exponential growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>40</td>
<td>72.5% (29)</td>
<td>27.5% (11)</td>
</tr>
<tr>
<td>3.75 × 10^{-8}</td>
<td>40</td>
<td>20% (8)</td>
<td>80% (32)</td>
</tr>
<tr>
<td>7.5 × 10^{-8}</td>
<td>38</td>
<td>5.3% (2)</td>
<td>94.7% (36)</td>
</tr>
<tr>
<td>1.5 × 10^{-7}</td>
<td>37</td>
<td>18.9% (7)</td>
<td>81.1% (30)</td>
</tr>
<tr>
<td>3.0 × 10^{-7}</td>
<td>36</td>
<td>8.3% (3)</td>
<td>91.7% (33)</td>
</tr>
<tr>
<td>6.0 × 10^{-7}</td>
<td>38</td>
<td>10.5% (4)</td>
<td>89.5% (34)</td>
</tr>
</tbody>
</table>

aThe number of replicate runs varies because the model is stochastic, and in some runs the epidemic goes extinct before reaching the end of the 20 year simulation.

results, based on the sign of ΔAIC and summarized in table 2, support our hypothesis that epidemics characterized by local dispersal (Δφ_{LDT} = 0) should follow a time-squared pattern of growth while exponential growth will arise when spatial spread includes LDT. In 72.5 per cent of epidemics without LDT the time-squared model is a better fit to the skyline data (N_{t,τ}). On the other hand, exponential growth is a better fit in 80−94.7% of epidemics when LDT is included as a mode of host dispersal (Δφ_{LDT} ≠ 0). Increasing rates of LDT also produce differences in the phylogenetic structures that can be revealed as increasing rates of growth in the skyline analysis. Higher rates of LDT result in faster epidemic growth rates with a linear relationship between the rate of exponential growth and the non-zero rate of LDT (Δφ_{LDT} ≠ 0). Increasing rates of LDT result in higher coefficient of determination (R²) and slopes that were not significantly different from zero. However, the lack of correspondence was in large part due to demographic stochasticity to which the skyline plots are not sensitive. On the other hand, we found a strong correspondence when we used the rate of exponential growth to compare the demographic data to the skyline reconstructions. A linear regression of the rates of exponential growth from both sources indicates a strong linear relationship between the two (figure 2).

We also found that the demographic data was not as reliable a source as the skyline plots for detecting the differences between time-squared and exponential growth. For each replicate run, we computed the total number of infected hosts, I_{total}, from the spatial dynamics. We then fitted models that assumed either time-squared or exponential growth during the period of spatial spread to the time-series of I_{total}. Maximum likelihood was used to estimate values for the growth model parameters and AIC was used to compare the relative fit of each model with the demographic data. Unlike our results for the skyline plot (table 2), the time-squared growth model was not the better fit to the demographic data when Δφ_{LDT} = 0 and instead the exponential model was a better fit to I_{total} both with and without LDT (see the electronic supplementary material, table S4). The difficulty in detecting time-squared growth arises because the demographic data is noisy as a result of demographic stochasticity. This tends to mask differences between time-squared and exponential growth.

We apply our model to explore the mode of epidemic growth exhibited by the mid-Atlantic raccoon rabies epidemic reported by Biek et al. [34]. Based on their analysis, Biek et al. [34] divided the raccoon rabies epidemic into three phases of epidemic expansion corresponding to different physical features of the landscape over which the epidemic spread. We have re-examined these three phases and performed a maximum-likelihood fit of a time-squared...
versus exponential model to the growth in each phase (phase 1: 1975 through 1982, figure 3a; phase 2: 1986 through 1989, figure 3b; phase 3: 1989 through 1992, figure 3c). As in the classification derived from our model, these results suggest that LDT of infected hosts was not only occurring, but also had a pronounced effect during the mid-Atlantic spread of raccoon rabies. While we are not claiming this as definitive proof of a role of LDT, the exponential growth observed in the phylogenetic data is consistent with an analysis of the demographic data (figure 1b), with the subscript added in this figure to make clear distinction with the growth rate estimated from the demographic data (\( \lambda_{\text{demographic}} \)).

**4. Discussion**

Our results suggest that differences in fundamental ecological processes, represented in our analysis by differences in the magnitude of LTD, may result in detectable differences in the topology and geographical pattern of pathogen phylogenies. These differences in genetic effects leave a persisting ‘signature’ indicative of specific ecological processes. Even when samples collected from an epidemic reflect many local processes interacting over space, different rates of LDT leave a detectable signature within a viral phylogeny.

Our model includes the demographic stochasticity of the host’s birth/death process as well as the disease cycle of rabies from transmission, to incubation, infectiousness and finally death. We have also considered the effects of spatial heterogeneity in the demography of the host species. Despite the considerable noise these processes are expected to add, we are able to reveal several differences in the phylogeny of the virus produced by LTD. Using skyline plots to examine the sequence of coalescent events in the viral phylogeny, we can compare the rate of growth of the epidemic for different movement scenarios. In accordance with our expectations, we find that epidemic growth without LTD follows a time-squared function, whereas growth with LTD is exponential. Further, we find an increasing linear relationship between the rate of exponential growth and the log-transformed rate of LTD.

Our comparison of the skyline plots to the epidemic dynamics indicates that the skyline analysis can provide a...
reliable estimate of the true demographic changes in the host population. We found that the rate of exponential growth estimate from $N_e$ over $\tau$ was in close agreement with the rate estimate taken directly from the demographic data (figure 2).

Our results also point to the value of combining phylogenetic analysis with analysis of demographic data collected from an epidemic. We found that the mode of epidemic growth may be easier to detect from the signatures embedded in the viral phylogeny than from the analysis of the demographic data. Time-squared and exponential growth were distinguishable from the pattern of the viral phylogeny (table 1), but this was not the case using only demographic data (see the electronic supplementary material, table S4). Demographic stochasticity in the dynamics of an epidemic introduce noise into the demographic data which may tend to obscure differences between ecological processes such as the model of dispersal. On the other hand, while the rate of rabies virus evolution is fast, it is still not capable of tracking day-to-day changes in the size of an epidemic, and so may be providing a smoother picture of epidemic growth.

In an earlier analysis, solely based on ecological data, we established the importance of LDT during the spread of raccoon rabies in Connecticut over a five-year period of expansion [35]. Mathematical models simulating the dynamics of spread that included LDT as a distinct process result in a better agreement with the observed pattern of spread relative to models that did not include LDT. This earlier analysis used models and data calibrated at the spatial scale of townships, with a mean area of 85 km$^2$. Our current analysis, however, suggests that the effects of LDT are preserved across spatial scales. The spatial resolution of our analysis of the entire eastern seaboard epidemic is at the level of counties, with a mean area of 1600 km$^2$. The combined analysis reveals a process of expansion characterized by foci of secondary infection acting in advance of the encroaching epidemic wave at both local and regional spatial scales.

Our model results are also consistent with results suggesting that there may have been considerable heterogeneity in the rate of rabies spread in the eastern USA. Using a model that allowed for overdispersed variation in the rate of spread along different branches in the phylogeny (a relaxed random walk), Lemey et al. [54] document a better fit to the genetic data than a time-homogeneous Brownian diffusion model. As an important distinction, the approach by Lemey et al. is focused on revealing and quantifying heterogeneity in the spatial diffusion process from geo-referenced samples without being tied to any particular ecological mechanism. Variation in branch-specific diffusion rates in their case may thus occur for a variety of reasons, including variation in the rate of local spread, the existence of dispersal barriers as well as some contribution of long-distance dispersal. Here, we suggest a complementary approach in which genetic data is analysed with a model that makes explicit the contribution of different modes of dispersal and can be used to make predictions about spatial patterns of growth.

Our analysis based on skyline plots does not include spatial location information from our spatially explicit model. However, our results suggest that phylogenetic patterns associated with spatial processes can be detected even when reliable spatial data are not available. Collection of spatial data that span the period of an epidemic has been an important asset in reconstructing the geographical spread of several infectious diseases including foot and mouth disease [55], and rabies [34,50,54]. However, the collection of accurate and complete public health data is a significant challenge in many parts of the world [56–59]. Our results suggest that analysis of viral sequence data may be useful in overcoming some of these data deficiencies.

Our results are based on viral sequences collected 20 years after the start of the epidemic, when the disease is well into its

Figure 3. Growth in the effective number of infections ($N_e\tau$) during the three phases of the mid-Atlantic raccoon rabies epidemic [34]. In each panel the circles show the growth in $N_e\tau$ obtained from a Bayesian skyline plot analysis of 44 rabies sequences [34], the red line shows the fit of an exponential growth model and the blue line shows a time-squared model of growth. For each phase of the epidemic the exponential model of growth is a better fit (a) $\Delta\text{AIC} = -65.2$; (b) $\Delta\text{AIC} = -6.6$; (c) $\Delta\text{AIC} = -9.6$, suggesting that long-distance translocation facilitated the spread of raccoons in the mid-Atlantic epidemic.
endemic phase. We chose this retrospective sampling regime because long delays between when an epidemic starts and when it is first sampled are common in infectious disease studies. Wildlife diseases are notoriously difficult to detect because of insufficient surveillance [60]. Even in human populations, epidemics can go unnoticed for prolonged periods of time as highlighted by the discovery of HIV in the early 1980s nearly 30 years after the virus’ suspected emergence in Africa [61,62]. The late sampling date allows for prolonged action of stochastic processes, such as genetic drift and small numbers of infected hosts, which tend to obscure the effect ecological processes have on the viral phylogeny. Despite this, we find that differences in the mode and tempo of dispersal can be recovered from the viral phylogeny (figures 1, 3 and 4) decades later. We expect alternative sampling procedures that include collection of some or all samples at times closer to the start of an epidemic to provide better estimates of past ecological processes.

Our results also indicate that at the spatial scale of US counties used in our model, the ability to detect the effects of LTD are limited if sampling were to occur immediately after the start of an epidemic. The skyline plots for different rates of LTD in figure 1a overlap and are essentially indistinguishable in the period immediately following the epidemic. As a result, it would be impossible to detect differences associated with different rates of LTD. This outcome is expected because early in the epidemic there will have been few or no opportunities for LTD events larger than a single county to occur. In such fortunate circumstances where an epidemic is identified before it has a chance to spread, direct monitoring of the epidemic would be a more efficacious approach to studying the disease dynamics. In addition, highly accurate georeferenced genotype data may allow an earlier assessment of the role of LTD we observe in our model. However, the chance to observe an epidemic from its very beginning is likely to be the exception. For many epidemics, in both human and wildlife populations, techniques and theory can help reveal unobserved epidemic dynamics.

As long as mutational events occur on similar timescales as the generation time and the timescales of spatial movement of the pathogen, this should result in the emergence of phylogeographic structure that can yield insights into the ecological processes that shape the spread of disease, here specifically dispersal. Systems this may apply to are other viruses in terrestrial wildlife hosts, such as hanta- or arenaviruses (potentially long infectious periods offset by limited dispersal ability in the host) or canine distemper, as well as certain plant viruses.

The authors thank Vijay Panjetti and Derek Johnson for their comments. This research was supported by NIH grant no. ROI AI047498 to L.A.R., and the RAPIDD Programme of the Science and Technology Directorate, The Department of Homeland Security and The Fogarty International Center, National Institutes of Health and the Louisiana Board of Regents grant no. LEQSF(2011-14)-RD-A-27 to S.M.D. Computer simulations were performed on the Emory High Performance Computer Cluster and the Louisiana Optical Network Initiative computers.

References


