PROCEEDINGS OF THE ROYAL SOCIETY B

rspb.royalsocietypublishing.org

Research

Cite this article: Mollentze N, Nel LH, Townsend S, le Roux K, Hampson K, Haydon DT, Soubeyrand S. 2014 A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proc. R. Soc. B* **281**: 20133251.
http://dx.doi.org/10.1098/rspb.2013.3251

Received: 20 December 2013
Accepted: 3 February 2014

Subject Areas:
ecology, health and disease and epidemiology, microbiology

Keywords:
endemic disease, rabies virus, spatial epidemiology, transmission trees, surveillance

Author for correspondence:
Daniel T. Haydon
e-mail: daniel.haydon@glasgow.ac.uk

[†]Present address: Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow, Glasgow G12 8QQ, UK.

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2013.3251 or via http://rspb.royalsocietypublishing.org.

Royal Society Publishing

A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data

Nardus Mollentze[1,†], Louis H. Nel[1], Sunny Townsend[2], Kevin le Roux[3], Katie Hampson[2], Daniel T. Haydon[2] and Samuel Soubeyrand[4]

[1]Department of Microbiology and Plant Pathology, University of Pretoria, Pretoria 0002, South Africa
[2]Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow, Glasgow G12 8QQ, UK
[3]Directorate of Veterinary Services, KwaZulu Natal Department of Agriculture and Environmental Affairs, Pietermaritzburg 3202, South Africa
[4]INRA, UR546 Biostatistics and Spatial Processes, 84914 Avignon CEDEX 9, France

We describe a statistical framework for reconstructing the sequence of transmission events between observed cases of an endemic infectious disease using genetic, temporal and spatial information. Previous approaches to reconstructing transmission trees have assumed all infections in the study area originated from a single introduction and that a large fraction of cases were observed. There are as yet no approaches appropriate for endemic situations in which a disease is already well established in a host population and in which there may be multiple origins of infection, or that can enumerate unobserved infections missing from the sample. Our proposed framework addresses these shortcomings, enabling reconstruction of partially observed transmission trees and estimating the number of cases missing from the sample. Analyses of simulated datasets show the method to be accurate in identifying direct transmissions, while introductions and transmissions via one or more unsampled intermediate cases could be identified at high to moderate levels of case detection. When applied to partial genome sequences of rabies virus sampled from an endemic region of South Africa, our method reveals several distinct transmission cycles with little contact between them, and direct transmission over long distances suggesting significant anthropogenic influence in the movement of infected dogs.

1. Introduction

Understanding the spatial aspects of disease transmission is increasingly recognized as an essential component of successful control strategies [1,2]. However, disease transmission is usually a highly elusive event and reconstructing 'who-infected-whom' in outbreaks of infectious disease remains a challenging problem. The advent of high volume and more affordable pathogen genome sequencing to complement conventional space-time incidence data promises a step-change in our ability to understand transmission at the population level. Yet, progress will only be made with advances in statistical methodology to accompany this ever increasing access to genetic and other data.

Two different but complementary approaches that use spatial, temporal and pathogen genetic information to reconstruct the dynamics of epidemics have been developed in recent years. The first approach is based on coalescent models that assume some form of population dynamic model to relate the demography of the pathogen to its evolution, while implementing a diffusion model to account for the movement of the pathogen over geographical space [3]. These models can be used to estimate various parameters of interest, such as the rate of spatial spread of the pathogen [4] and the rate of evolution over time [5]. This approach has the advantage that it is relatively robust to the intensity of

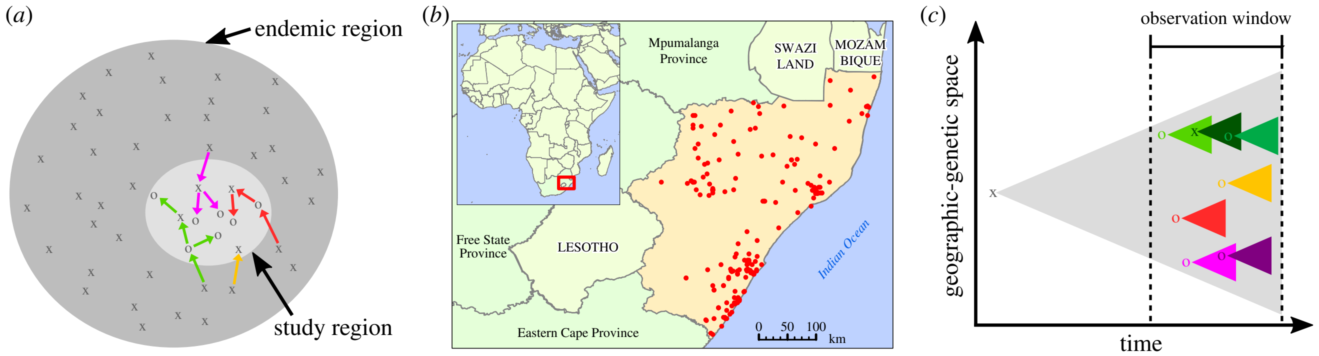


Figure 1. Modelling the transmission of endemic diseases. (*a*) All cases in the study region are in some way related both genetically and spatially because they form part of a larger epidemic that originated from a single progenitor. This, along with the fact that some cases go undetected, makes determining dependence among transmission chains difficult. Letters O represent sampled cases, while X represent unsampled cases. (*b*) Map of the KwaZulu Natal province of South Africa, showing the locations of the 176 cases used to infer the transmission tree (see also the electronic supplementary material, table S1). (*c*) Pathogens radiate both in terms of genetic diversity and in terms of the geographic space invaded. Triangles represent possible locations in the geographic–genetic space to which cases can move and evolve, with the grey triangle showing the radiation of the entire epidemic, which can also be viewed as the indirect radiation of the index case (represented by a black X) through its descendants. In the relatively short observation window, three types of relationships are apparent: direct transmissions (purple), introductions, which will be more closely related to the common ancestor of all sampled cases than to any other cases (red and yellow), and indirect transmissions via unsampled intermediary cases (green).

epidemiological sampling, but because such models do not have an explicit epidemiological formulation, the inferences cannot easily be related to real epidemiological processes. The second approach is based on spatial epidemiological models of transmission and simple models of genetic drift and directly reconstructs the transmission tree reflecting ‘who-infected-whom’, thus explicitly recognizing the host population structure and the epidemiological processes that govern the interaction of host and pathogen. In this approach, an epidemiological model of disease progression in individuals is used to estimate probability distributions for possible dates of infection and the infectious period of all cases. When coupled with a model of spatial diffusion and a model of the accumulation of point mutations over time, the probability of any two cases A and B being causally related can be calculated based on the likelihood that case A was infectious and case B was infected during the same time window, the probability that the pathogen could have dispersed from the geographical location at which case A was observed to the location at which B was observed in the time between observations, and the probability that the pathogen genetic sequence from case A could have mutated to the sequence from case B in the time between observations. This approach enables inferences to be made about epidemiological processes [6], the transmission tree [6,7], the mechanism of transmission [8] and the rate of evolution ‘per transmission event’ [9]. More recently, the two approaches have been combined, using a coalescent model to account for the influence of intra-host population dynamics on the structure of pathogen genetic data while reconstructing the transmission tree, thus addressing an important source of inaccuracy at high sampling intensities [10]. However, current transmission tree-based methods cannot handle large numbers of missing infections, and therefore require a high proportion of infected hosts from the outbreak to be present within the sample.

In general, these techniques have been applied to epidemics, and to data that are assumed to arise from a single introduction to the region under study (thus making its structure monophyletic). When pathogens are sampled from infected hosts in an endemic context (i.e. where the pathogen is stably maintained in an area in the absence of introductions from outside of that area), the epidemiological situation is potentially more complex. In this context, the connection between cases applies at two scales (figure 1*a*). At the scale of the entire endemic region, all cases may be related in some way (through the global transmission tree), leading to genetic relatedness and spatial autocorrelation between sampled cases. However, in a given study region (even one that has been exhaustively sampled), only some cases will be directly related through chains of transmission, and many chains of transmission may exist that are only indirectly related to each other by virtue of sharing a common ancestor outside the sampled area. The sample of pathogens within the study area is therefore polyphyletic. The picture is further complicated because surveillance is unlikely to be exhaustive, and therefore the sampling will be incomplete. Undetected or unsampled cases will reduce the detectable correlation between cases that are nevertheless causally related. If we hope to use genetic data to understand the detailed transmission biology of endemic pathogens, the challenge will be to develop algorithms that can accommodate the polyphyletic nature of pathogen population structure, and account for and make inferences regarding the unobserved and unsampled infections.

Here, we describe the extension of a spatial-genetic SEIR (susceptible/exposed/infectious/removed) model of transmission to accommodate the complexities inherent to polyphyletic and partially sampled outbreak data containing space, time and genetic information. In addition, we infer the infected host population size over the study period and region by developing a mark–recapture method applied to the virus lineages occurring in the transmission tree, thus providing upper and lower estimates of the number of undetected or unsampled cases. We test this technique by analysing various simulated scenarios, before applying it to endemic rabies virus in a province of South Africa (figure 1*b*), and show how it can be used to better understand the spatial epidemiology of the virus. Such knowledge is crucial for advancing the effectiveness of large-scale vaccination campaigns—some of which have been in place for decades, but have failed to eliminate the disease in question.

Rabies is a complex disease endemic to much of the developing world [11]. The mutation rates of RNA viruses are so high that population genetic and epidemiological processes occur on similar timescales, and spatial expansion and epidemiology leave a discernible fingerprint on the genetic structure of these viruses [12,13]. Rabies virus is typically transmitted by direct contact through biting [14]. However, the epidemiological dynamics of rabies are complicated by two factors. First, rabies has a highly variable incubation period [15,16] and second, rabies has a very large host range that includes all mammals, many of which would play no part in the onward transmission of the virus [14]. In southern Africa, two distinct genetic variants of rabies virus are known to circulate—one among members of the Canidae, including domestic dogs (*Canis lupus familiaris*), and the other among several members of the Herpestidae [17]. Nevertheless, the majority of infections in humans are associated with rabid domestic dogs [11,18], and it is in dogs that the disease must be controlled if the burden on humans is to be reduced [19].

2. Material and methods

(a) Data collection

In the KwaZulu Natal province of South Africa (KZN), suspected cases are primarily collected through a network of state and private veterinarians. Further cases are collected by travelling vaccination teams of a Bill and Melinda Gates Foundation-sponsored rabies elimination project active throughout KZN. All cases testing positive for rabies virus by the gold-standard fluorescent antibody test [20] between 1 March 2010 and 8 June 2011 were selected for analysis ($n = 195$; electronic supplementary material, table S1). Five cases were negative by polymerase chain reaction (PCR; see below) after multiple attempts and were excluded from further analysis. One sequence, from an unrecorded wildlife species, matched the herpestid variant of rabies virus by BLAST [21] and was also excluded. A further 13 cases lacked GPS coordinates and were therefore excluded from the transmission tree reconstruction.

(b) RT-PCR and sequencing

RNA was extracted from original brain material using TRIzol reagent (Invitrogen). Reverse transcription (RT)-PCR and sequencing were performed as described in the electronic supplementary material. Consensus sequences were aligned using the FFT-NS-i algorithm of MAFFT v. 6 [22]. Sequences were trimmed to equal length (760 nucleotides, encompassing the last 224 nucleotides of the glycoprotein gene, the G-L intergenic region and 118 nucleotides of the polymerase gene, based on the genome of the Pasteur rabies virus strain [23]). The overall mean distance between sequences in the alignment was calculated using MEGA v. 5 [24].

(c) Transmission tree reconstruction

The transmission trees linking cases were reconstructed using the trimmed alignment described above, which was realigned with MAFFT after exclusion of 13 cases lacking GPS coordinates (electronic supplementary material, table S1).

The core algorithm used here is a generalization of the algorithm of Morelli *et al.* [7] to allow its application to any directly transmitted disease. We start with an epidemiological model in which any susceptible host i becomes infected at time T_i^{inf}. Following an incubation period $\mathbf{L}_i$, it becomes infectious for time-period $\mathbf{D}_i$ before dying. Both $\mathbf{L}_i$ and $\mathbf{D}_i$ are random variables with informative prior distributions based on contact tracing data from Tanzania [15]. From this data, it is possible to calculate the probability of a transmission from any host j to any host i based on the probability of j being infectious at the time of i's infection, if we assume the known observation date occurred shortly after the end of the infectious period [7].

However, this forms only part of the probability of transmission between hosts. The spatial component of the likelihood equation was modified to accommodate a wide variety of spatial transmission patterns by replacing the exponential transmission kernel used in [7] with the exponential-power spatial transmission kernel described by [25]. This kernel is often used in dispersal studies and can take a variety of shapes, making it well suited to a range of endemic situations where often little is known regarding spatial transmission patterns. We also replaced the simplified substitution model of [7] with the Kimura three-parameter model [26].

(d) Extension to polyphyletic transmission trees

In a partially sampled outbreak, any given infected host which was sampled might have been infected by: (i) another sampled host (through direct transmission), (ii) an unsampled host which had been infected directly or indirectly by a sampled infected host (termed 'indirect transmission' here) or (iii) an unsampled host which has no ancestors within the sample, i.e. transmission from an exogenous source (figure 1*a*,*c*). The model of [7] allows for only a single virus introduction (i.e. a single 'exogenous' transmission) followed by direct transmissions for the rest of the outbreak. We extended this model by allowing multiple unobserved cases to arise anywhere in both space and time within the set of inferred transmissions.

The likelihood equation of [7] models the spatial radiation and genetic evolution of cases over time to determine the likelihood of various parameters at any point in time and thus calculate the probability of different transmissions. In our model, this is equivalent to the approach taken for direct transmissions, where each sampled infected host species able to transmit the virus can be a source of infection. These are modelled by the probability distribution $\mathcal{P}_{\mathrm{direct}}$, defined over the geographical–genetic space and evolving in time (represented by coloured cones in figure 1*c*). $\mathcal{P}_{\mathrm{direct}}$ is dependent on the infection time of the host (estimated as described above), its incubation duration (estimated), its removal or observation time (observed), a spatial dispersal kernel (estimated) and substitution rates for the sequence evolution (estimated).

Each sampled infected host which can spread the disease can also be an indirect source of observed infections after its removal, as a consequence of unsampled intermediate hosts: case A (sampled) infects B (unsampled) which infects C (unsampled) which infects D (sampled). As these unsampled cases extend the influence of case A in both geographical and genetic space, their effect can be modelled by allowing observed cases to continue moving and evolving after their death. This is represented by probability distribution $\mathcal{P}_{\mathrm{indirect}}$, again defined over the geographical–genetic space and evolving in time and depending on the same parameters as $\mathcal{P}_{\mathrm{direct}}$. The spatial influence contributed by unsampled cases is harder to determine. We considered two different specifications for the dispersal kernel governing indirect transmissions ($\mathcal{K}_{\mathrm{indirect}}$). In the first specification, we conservatively assume that $\mathcal{K}_{\mathrm{indirect}}$ is the same as the spatial dispersal kernel used for the direct transmissions, thus allowing only movement over transmission distances observed for (single) direct transmissions. In this scenario, infections occurring after the death of the source host are attributed to unsampled intermediate hosts. However, this does not adequately accommodate a scenario encompassing multiple unsampled intermediate cases, where greater geographical distances between the indirectly connected cases would be possible. We therefore also considered a more liberal specification, where $\mathcal{K}_{\mathrm{indirect}}$ is a uniform distribution over the whole study region, thus allowing unsampled intermediate hosts to carry the virus to any location within the sampled region. These two scenarios form extremes between which the true process can reasonably be expected to occur.

In a similar vein, the source of exogenous transmissions can be modelled as a probability distribution $\mathcal{P}_{\mathrm{exo}}$ defined over the geographical–genetic space and evolving in time (represented by the grey cone in figure 1*c*). $\mathcal{P}_{\text{exo}}$ can be completely specified based on an ancestral virus sequence (determined *a priori* through ancestral state reconstruction, in our case using the FastML server under the generalized time reversible model [27]), a time for the ancestral sequence and the same substitution rates as above (both of which are co-estimated with the transmission tree). The ancestral sequence and the sampled infected hosts generate a mixture $\mathcal{M}$ of spatio-temporal-genetic distributions ($\mathcal{P}_{\text{exo}}$, $\mathcal{P}_{\text{direct}}$ and $\mathcal{P}_{\text{indirect}}$) from which the infection events are drawn. Estimating the source that infected a given host involves assessing in which component of the mixture model $\mathcal{M}$ the infection of the host arose.

Conceptually, however, the source of both types of transmissions involving unobserved ancestors (indirect and exogenous) can be modelled in the same way—as being external to the sampled dataset, meaning the transmissions arise in $\mathcal{P}_{\text{exo}}$. Thus, to reduce complexity and computation time, we distinguished only between direct and 'unsampled' sources in the primary Markov chain Monte Carlo (MCMC) sampling procedure (only $\mathcal{P}_{\text{direct}}$ and $\mathcal{P}_{\text{exo}}$ were used to define $\mathcal{M}$), with a post-processing algorithm to distinguish between indirect and true exogenous transmissions. In the previously described monophyletic model [7], the posterior distributions of the incubation and infectious period durations can be deformed by indirect links between cases. We used narrow priors for the parameters governing these distributions, essentially forcing a decision between direct transmission or linkage to an exogenous source in the first step. To distinguish between exogenous and indirect transmissions, the post-processing analysis applies a Metropolis–Hastings update to the 'unsampled' transmission links determined by the MCMC algorithm, which involves comparing the probability that the transmission was really from an exogenous source (based on $\mathcal{P}_{\text{exo}}$, as described above) with the probability that it was merely indirect (based on $\mathcal{P}_{\text{indirect}}$). This post-processing was applied under both the conservative and liberal specifications of the spatial transmission kernel ($\mathcal{K}_{\text{indirect}}$) described above.

(e) Population size estimation

To determine the true number of cases represented by indirect links, we developed a mark–recapture technique applied to the virus lineages identified in the previous analysis. If we split the transmission tree dataset into two parts based on the sampling times of the hosts, any host sampled in the second time-period is considered as recaptured if it was directly or indirectly infected by a host observed in the first part of the dataset. Although the full transmission tree is not known, the previous analysis provides a sample of its posterior distribution. For each element of this sample, the number of recaptured virus lineages can be calculated, generating a posterior distribution of the number of recaptured virus lineages. With this distribution, one can determine the posterior distribution of the population size using a mark–recapture analysis, which takes into account uncertainty regarding changes in the population size from the first to the second time-period.

(f) Simulations

The accuracy of the method was assessed using 100 simulated datasets from each of six scenarios (i.e. 600 simulations in total). The first four scenarios were used to investigate overall accuracy and the effect of sampling rate on the reconstruction method with high (three-quarters of all cases), moderate (two-thirds of all cases), intermediate (one-half of all cases) and low (one-quarter of all cases) detection rates, respectively. A further two scenarios were used to test the sensitivity of the method to small and large misspecifications of epidemiological parameters. The simulation model was based on the probability distributions and specifications described above and in the electronic supplementary material, but contained a more realistic specification for the external source of infection. While the inference model assumes a single external source with a constant infection strength (constant in both space and time), the simulation model allows for multiple sources of novel lineages, occurring both inside and outside the sampling region, with infection strengths that are localized in time and space. The simulated epidemics were initiated from a single point in time and space outside the sampling period and region and allowed to progress until a set number of hosts had been infected. Only data from one-third of the region and time-period affected by the simulated epidemic were retained and subsampled with the detection rates above determining the probability of a case being retained.

A more formal description of the model, inference procedures and simulations described here can be found in the electronic supplementary material.

3. Results and discussion

Reconstructions of 600 simulated outbreaks reveal that the method described here accurately recovers most parameters regardless of sampling intensity or model misspecification (electronic supplementary material, table S2). As can be expected, reconstruction of transmission events is sensitive to the informative priors used for the incubation and infectious periods (electronic supplementary material, table S3). This limits the suitability of the approach to diseases where the epidemiology is reasonably well known. The reconstruction of direct transmissions remains fairly accurate regardless of sampling intensity (mean posterior probability of true transmission events more than 0.73; electronic supplementary material, table S3) and actually increases in accuracy when sampling intensity decreases. Reconstruction of transmissions involving unobserved cases is moderately accurate at high sampling intensities, but becomes increasingly unreliable when 50% or fewer of the cases in the sampling region are sampled. At these sampling intensities, the post-processing algorithm cannot accurately distinguish between indirect and exogenous connections, which in turn also leads to a significant underestimation of the total number of cases (electronic supplementary material, table S4). At high to moderate sampling intensities (three-quarter to two-thirds of all cases in the sampled area), however, the 95% posterior interval (PI) inferred for the total population size covers the true value in more than 97% of cases under both the conservative and liberal specifications of the model.

Between 1 March 2010 and 8 June 2011, 195 rabies virus-positive cases were detected in KZN. The majority of these cases occurred close to densely populated areas, often in the peri-urban townships surrounding cities and large towns (electronic supplementary material, figure S1). A 760 nucleotide fragment spanning the highly variable G-L intergenic region was sequenced from 190 of these samples (electronic supplementary material, table S1). Despite the small spatial and temporal scale, the overall mean distance between the 189 canid-associated rabies virus sequences generated was 8.42 nucleotides. However, many clusters of identical sequences exist, and the phylogenetic divergence was not sufficient to generate a well-resolved phylogeny (electronic supplementary material, figure S2).

The transmission trees linking cases were estimated using 176 canid-associated rabies cases for which detailed epidemiological data were available (electronic supplementary material, table S1). When considering only direct transmissions, there were several independent chains of transmission and many

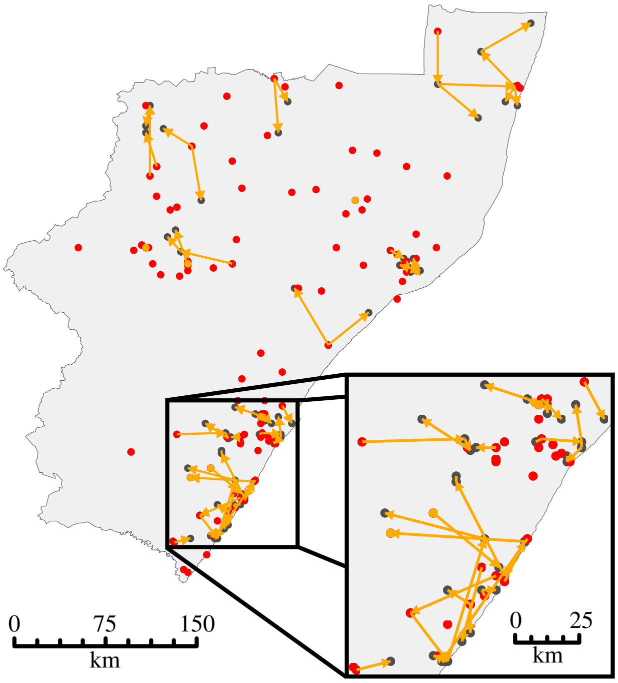


Figure 2. Transmission trees showing the direct pairwise transmissions with highest posterior probabilities. Transmission links between cases are represented by orange arrows. Red dots represent cases for which no direct ancestor was detected and black dots represent all other cases. The inset shows an enlarged view of connections in the southern coast of KZN, where the majority of cases were detected.

transmissions inferred to have taken place over long distances (figure 2 and electronic supplementary material, figure S3). The mean distance between the most probable directly connected cases was 14.9 km (0.025- and 0.975-quantiles: 0.0 and 56.1 km; electronic supplementary material, figure S3). This was despite the use of narrow prior distributions for the parameters governing the durations of infections, which would tend to minimize the distance between directly connected cases in favour of indirect or exogenous connections instead. Occasional long-distance transmissions in this region, particularly along the major highways that follow the KZN coast, have been identified before (based on phylogenetic patterns) and were ascribed to motorized transportation of dogs [28]. Road distances have also been shown to be a better predictor of rabies dissemination than absolute distances in northern Africa [29]. The long distances and short time-periods between cases in the transmission tree (electronic supplementary material, figures S3 and S4) provide further evidence for motorized transportation of infected dogs, but such transmissions were not restricted to any one area and instead appear to be a common feature of the epidemiology of rabies in this area. This might be owing to the high prevalence of circular human migration and migrant labour in many parts of KZN, with migrants visiting their rural households (and, it would seem, taking their dogs with them) on a regular basis [30].

The majority of cases could not be linked through direct transmissions—69 (95% PI: 60–79) direct transmissions were identified, while unsampled sources were the most likely link for the remaining 107 (95% PI: 97–117) cases (electronic supplementary material, figure S5). The conservative specification of the post-processing algorithm identified a further 37 (95% PI: 27–47) indirect transmission links over the 15 month study period, while the liberal version of the algorithm identified 67 (95% PI: 57–78) indirect transmissions (figure 3). Sixteen cases were assigned different indirect ancestors by the two specifications, while a further 35 were interpreted as having an exogenous source by the conservative specification, but were assigned indirect ancestors by the liberal specification. There are no obvious similarities between cases assigned different ancestors by the two specifications, with no evidence of either phylogenetic clustering (assessed using Moran's I to measure autocorrelation to inverse phylogenetic distances between cases, p-value of 0.16 when the null hypothesis is no clustering) or spatial clustering (assessed using a spatial scan statistic with a null hypothesis that there is no more clustering among cases interpreted differently than among cases in general; p-value of 0.69 for the best supported cluster) [31–33]. The same was true for cases interpreted as having an exogenous source by one specification but not the other, with no evidence of either phylogenetic (p-value = 0.86) or spatial clustering (p-value = 0.08 for the best supported cluster).

When considering both direct and indirect connections, there are many separate, unjoined transmission trees (electronic supplementary material, figure S6). For the most probable connections under both the conservative and liberal specifications of our algorithm, these transmission trees can be grouped into eight distinct spatial clusters. Transmission between different spatial clusters was rare—we detected only one such transmission with the conservative specification of the algorithm, and 10 such transmissions with the liberal specification. In addition, such transmissions do not appear to seed substantial additional numbers of cases, as only one instance of onward spread in the new cluster was detected under either specification, causing just one additional case in both instances. Interestingly, four of the inter-cluster transmissions identified under the liberal specification involved transmission from one cluster to another and then back to more-or-less the same location, before onward transmission in the original cluster, further supporting the hypothesis of migrants moving dogs back-and-forth between their urban and rural homes.

To gain a better understanding of the surveillance failures leading to the high number of indirect connections detected, we estimated the true number of cases occurring in the study area. This yielded a posterior median estimate of 389 cases (95% PI: 260–881) using the conservative specification of the post-processing algorithm, and 195 cases (95% PI: 182–298) using the liberal specification, over the 15 month study period (electronic supplementary material, figure S7). Our analyses of simulated datasets show that this mark–recapture approach is only accurate at fairly high sampling intensities, owing to difficulties in distinguishing between indirect and exogenous transmissions, and we note that the 95% PI of the number of recaptured lineages under the conservative specification is fairly wide (electronic supplementary material, figure S8). However, direct transmissions are accurately identified regardless of sampling intensity (electronic supplementary material, table S2), and in this dataset the conservative algorithm identified almost all infections involving unsampled individuals as exogenous transmissions, while the liberal algorithm identified most of these infections as indirect transmissions. Thus, the conservative algorithm minimized the number of recaptured lineages, while the liberal algorithm maximized it, which means the inferred population sizes can be interpreted as a lower and upper bound of the true value. As the herpestid-associated genetic variant of rabies virus is rare in KZN, the five cases which could not be sequenced were most likely

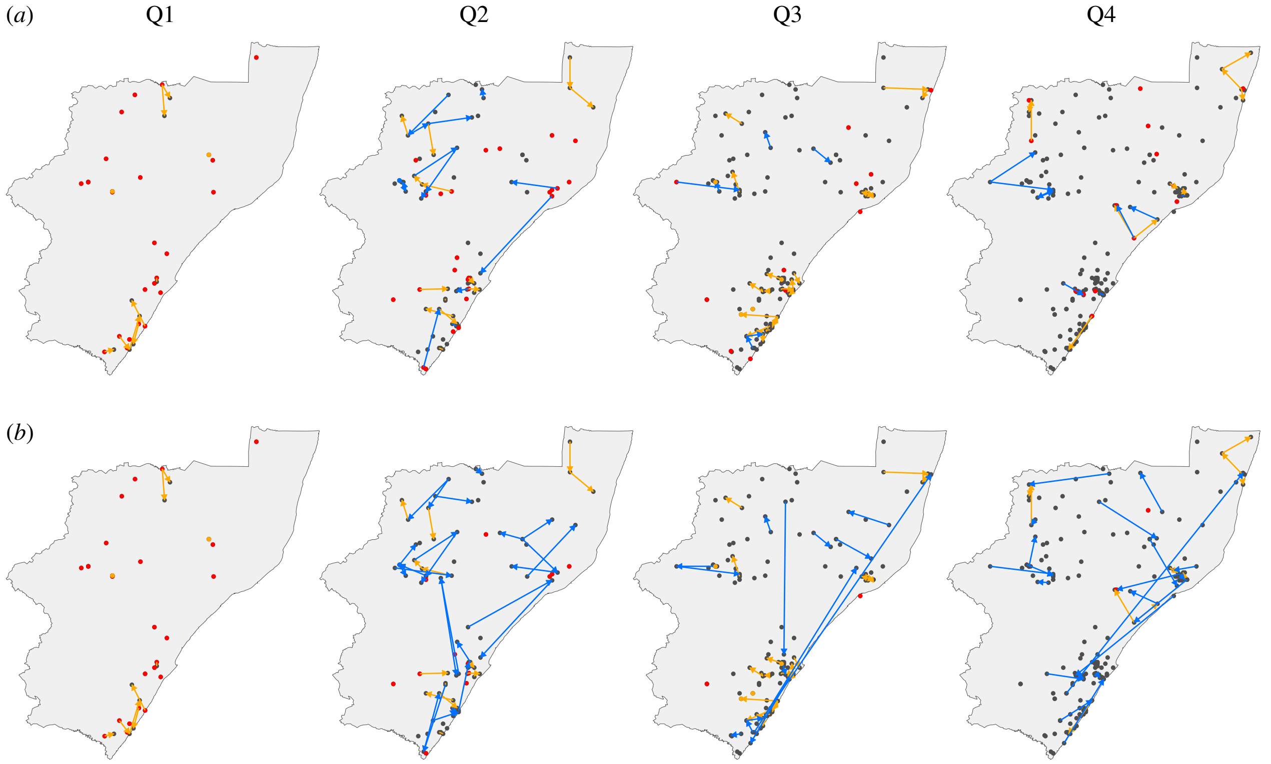


Figure 3. Pairwise transmissions with the highest posterior probabilities in each quarter of the sampled period, including indirect transmissions. Black dots represent all cases since the start of the sampling period, while red dots represent cases appearing in that quarter that have an exogenous source. Orange arrows represent direct transmission events. Blue arrows represent indirect transmissions inferred using the conservative (*a*) and liberal specification (*b*) of the post-processing algorithm. Note that detected cases (black dots) are displayed cumulatively. Q1–Q4: first to fourth quarter of the sampling period.

representatives of the canid-associated variant. Thus, surveillance detected 194 cases of infection with the canid-associated variant, or between 49.87 and 99.49% of all canid-associated cases (based on the posterior medians of the conservative and liberal specifications, respectively). Such high detection rates are exceptional for rabies [34] and need further confirmation by contact tracing. However, surveillance effort (measured as the number of samples submitted per month) was fairly constant over the study period while incidence concurrently declined, suggesting that the ongoing intensive control programme is effectively driving rabies towards elimination, which could account for the low total number of cases inferred from this analysis. The areas where cases are still being missed can be deduced from our identification of indirect links (figure 3), providing a powerful tool for improving detection rates which would be particularly important if rabies is indeed close to being eliminated in this province.

4. Conclusion

To successfully control rabies and other endemic diseases in a changing landscape, a detailed understanding of its spatial epidemiology is required. The method described here allows for the detailed reconstruction of the transmission events of endemic infectious diseases, providing information that can be used both in designing more efficient control strategies and to measure and improve the quality of surveillance programmes. Importantly, key parameters could be recovered accurately regardless of sampling intensity.

The long distances characterizing many internal transmissions point to a significant anthropogenic influence on the epidemiology of rabies in KZN, the causes of which require further study. Despite these long-distance transmissions, clear spatial groupings could be discerned (electronic supplementary material, figure S6). In addition, the frequent long-distance transmissions cause most of these spatial clusters to consist of a relatively small core area and numerous surrounding cases (figure 3). Thus, identifying the connections of surrounding cases to specific clusters enables more directed vaccination, where targeting the smaller core areas would allow control of rabies over large areas. Identifying the spatial scale at which independent control strategies can be applied means it is possible to replace the thin spread of limited resources across the province with intense, focused campaigns that move across the province on an annual basis. Also crucial to the success of any disease elimination effort is effective surveillance. By identifying the true state of surveillance as well as the areas where cases are being missed from existing, routinely collected data, the method described here can be used as a starting point to investigate the causes of poor surveillance in specific parts of the region of concern.

By applying the methods described here to data from multiple years, important information will be revealed about how to iteratively improve surveillance and adapt rabies control strategies by identifying areas to be prioritized during annual vaccination campaigns. In addition, these methods can easily be adapted to other endemic diseases, and the high mutation rate of other RNA viruses makes them ideal candidates for this approach. Particularly encouraging is the fact that the small genome region sequenced here provided sufficient resolution for this analysis, making the generation of adequate data for large numbers of cases feasible even in resource-poor areas.

Acknowledgements. The authors wish to thank the staff of the Allerton Veterinary Laboratory of the KZN Department of Agriculture and

Environmental Affairs for providing samples, and Roman Biek and Joseph Hughes for critical discussions. They also wish to thank two anonymous reviewers, whose contributions improved the manuscript.

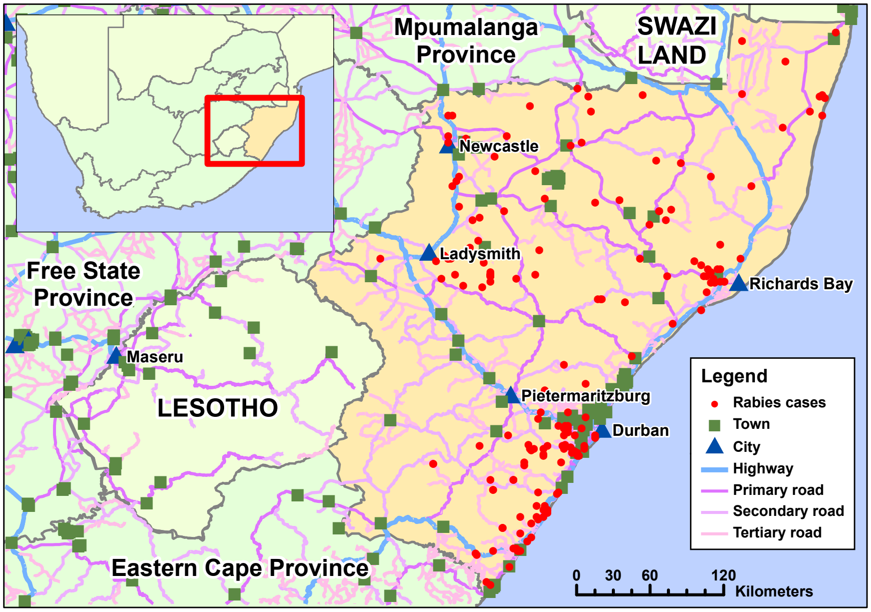
Data accessibility. Genetic data: GenBank accessions nos. KC660160–KC660352.

Funding statement. N.M. was financially supported by the NRF (grant no. 74606), the PRF (grant no. 11/85) and the University of Pretoria Postgraduate Study Abroad Programme. K.H. was supported by the Wellcome Trust (grant no. 095787/Z/11/Z). D.T.H. and S.T. were financially supported by the UK Medical Research Council (MRC, grant no. G0901135). S.S. was financially supported by the ANR (grant EMILE). The sequencing was financially supported by the MRC (grant no. G0901135). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

1. Ferguson NM, Keeling MJ, Edmunds WJ, Gani R, Grenfell BT, Anderson RM, Leach S. 2003 Planning for smallpox outbreaks. *Nature* **425**, 681–685. (doi:10.1038/nature02007)
2. Keeling MJ, Woolhouse MEJ, May RM, Davies G, Grenfell BT. 2003 Modelling vaccination strategies against foot-and-mouth disease. *Nature* **421**, 136–142. (doi:10.1038/nature01343)
3. Lemey P, Rambaut A, Welch JJ, Suchard MA. 2010 Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* **27**, 1877–1885. (doi:10.1093/molbev/msq067)
4. Pybus OG *et al.* 2012 Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc. Natl Acad. Sci. USA* **109**, 15 066–15 071. (doi:10.1073/pnas.1206598109)
5. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. 2002 Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161**, 1307–1320.
6. Ypma RJF, Bataille AMA, Stegeman A, Koch G, Wallinga J, Van Ballegooijen WM. 2012 Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proc. R. Soc. B* **279**, 444–450. (doi:10.1098/rspb.2011.0913)
7. Morelli MJ, Thébaud G, Chadœuf J, King DP, Haydon DT, Soubeyrand S. 2012 A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Comput. Biol.* **8**, e1002768. (doi:10.1371/journal.pcbi.1002768)
8. Ypma RJF, Jonges M, Bataille A, Stegeman A, Koch G, Van Boven M, Koopmans M, Van Ballegooijen WM, Wallinga J. 2013 Genetic data provide evidence for wind-mediated transmission of highly pathogenic avian influenza. *J. Infect. Dis.* **207**, 730–735. (doi:10.1093/infdis/jis757)
9. Cottam EM, Thébaud G, Wadsworth J, Gloster J, Mansley L, Paton DJ, King DP, Haydon DT. 2008 Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc. R. Soc. B* **275**, 887–895. (doi:10.1098/rspb.2007.1442)
10. Ypma RJF, Van Ballegooijen WM, Wallinga J. 2013 Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics* **195**, 1055–1062. (doi:10.1534/genetics.113.154856)
11. World Health Organization. 2002 *World survey of rabies number 35 for the year 1999*. Geneva, Switzerland: World Health Organization.
12. Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, Holmes EC. 2004 Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327–332. (doi:10.1126/science.1090727)
13. 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. USA* **104**, 7993–7998. (doi:10.1073/pnas.0700741104)
14. Rupprecht CE, Hanlon CA, Hemachudha T. 2002 Rabies re-examined. *Lancet Infect. Dis.* **2**, 327–343. (doi:10.1016/S1473-3099(02)00287-6)
15. Hampson K, Dushoff J, Cleaveland S, Haydon DT, Kaare M, Packer C, Dobson A. 2009 Transmission dynamics and prospects for the elimination of canine rabies. *PLoS Biol.* **7**, e1000053. (doi:10.1371/journal.pbio.1000053)
16. Charlton KM, Nadin-Davis S, Casey GA, Wandeler AI. 1997 The long incubation period in rabies: delayed progression of infection in muscle at the site of exposure. *Acta Neuropathol.* **94**, 73–77. (doi:10.1007/s004010050674)
17. Nel LH, Sabeta CT, Von Teichman B, Jaftha JB, Rupprecht CE, Bingham J. 2005 Mongoose rabies in southern Africa: a re-evaluation based on molecular epidemiology. *Virus Res.* **109**, 165–173. (doi:10.1016/j.virusres.2004.12.003)
18. Cleaveland S, Kaare M, Knobel D, Laurenson MK. 2006 Canine vaccination: providing broader benefits for disease control. *Vet. Microbiol.* **117**, 43–50. (doi:10.1016/j.vetmic.2006.04.009)
19. Lembo T *et al.* 2011 Renewed global partnerships and redesigned roadmaps for rabies prevention and control. *Vet. Med. Int.* **2011**, 1–18. (doi:10.4061/2011/923149)
20. Dean DJ, Abelseth MK, Atanasiu P. 1996 The fluorescent antibody test. In *Laboratory techniques in rabies* (eds FX Meslin, MM Kaplan, H Koprowski), pp. 88–95, 4th edn. Geneva, Switzerland: World Health Organization.
21. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990 Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410. (doi:10.1016/S0022-2836(05)80360-2)
22. Katoh K, Toh H. 2008 Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinf.* **9**, 286–298. (doi:10.1093/bib/bbn013)
23. Tordo N, Poch O, Ermine A, Keith G, Rougeon F. 1988 Completion of the rabies virus genome sequence determination: highly conserved domains among the L (polymerase) proteins of unsegmented negative-strand RNA viruses. *Virology* **165**, 565–576. (doi:10.1016/0042-6822(88)90600-9)
24. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011 MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739. (doi:10.1093/molbev/msr121)
25. Austerlitz F, Dick CW, Dutech C, Klein EK, Oddou-Muratorio S, Smouse PE, Sork VL. 2004 Using genetic markers to estimate the pollen dispersal curve. *Mol. Ecol.* **13**, 937–954. (doi:10.1111/j.1365-294X.2004.02100.x)
26. Kimura M. 1981 Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl Acad. Sci. USA* **78**, 454–458. (doi:10.1073/pnas.78.1.454)
27. Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, Pupko T. 2012 FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res.* **40**, W580–W584. (doi:10.1093/nar/gks498)
28. Coetzee P, Nel LH. 2007 Emerging epidemic dog rabies in coastal South Africa: a molecular epidemiological analysis. *Virus Res.* **126**, 186–195. (doi:10.1016/j.virusres.2007.02.020)
29. Talbi C *et al.* 2010 Phylodynamics and human-mediated dispersal of a zoonotic virus. *PLoS Pathog.* **6**, e1001166. (doi:10.1371/journal.ppat.1001166)
30. Posel D, Marx C. 2013 Circular migration: a view from destination households in two urban informal settlements in South Africa. *J. Dev. Stud.* **49**, 819–831. (doi:10.1080/00220388.2013.766717)
31. Paradis E, Claude J, Strimmer K. 2004 APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290. (doi:10.1093/bioinformatics/btg412)
32. Gittleman JL, Kot M. 1990 Adaptation: statistics and a null model for estimating phylogenetic effects. *Syst. Biol.* **39**, 227–241. (doi:10.2307/2992183)
33. Kulldorff M, Nagarwalla N. 1995 Spatial disease clusters: detection and inference. *Stat. Med.* **14**, 799–810. (doi:10.1002/sim.4780140809)
34. Townsend SE, Lembo T, Cleaveland S, Meslin FX, Miranda ME, Putra AAG, Haydon DT, Hampson K. 2013 Surveillance guidelines for disease elimination: a case study of canine rabies. *Comp. Immunol. Microbiol. Infect. Dis.* **36**, 249–261. (doi:10.1016/j.cimid.2012.10.008)

Supplementary figures and tables



[H]

Figure S1. Detailed map of KwaZulu Natal showing the cases detected between 1 March 2010 and 8 June 2011 in the context of major roads, towns and cities. Note that in addition to the 180 cases shown, a further 15 cases were detected for which coordinates were not recorded (table S1). Road and town data © OpenStreetMap contributors.

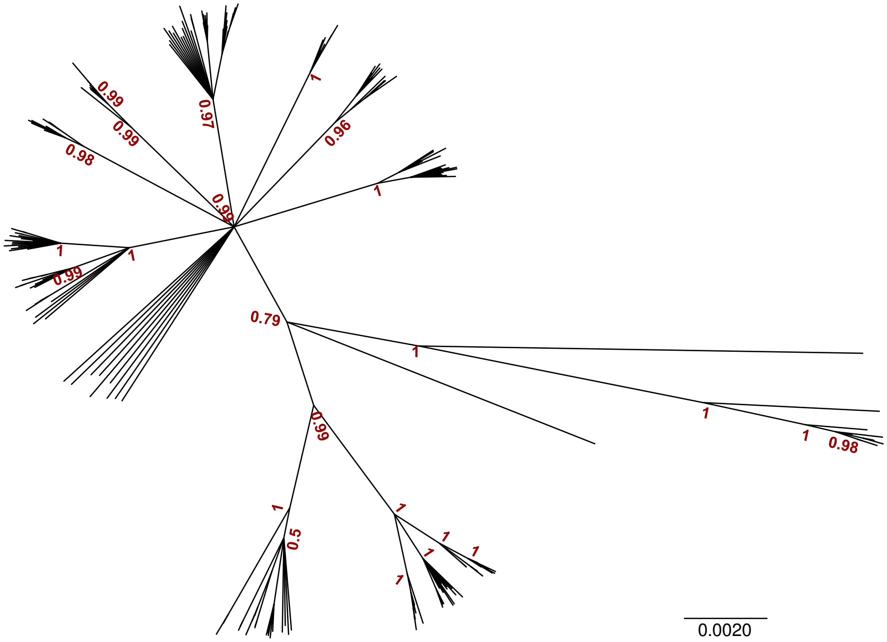


Figure S2. Unrooted consensus phylogeny of all cases sequenced in this study (including those lacking coordinates), inferred using Beast version 1.7.3 under an exponential growth coalescent model and assuming a strict molecular clock. Branch lengths are in number of substitutions per site, as indicated by the scale bar, while red labels at key nodes indicate their posterior probability. The tree shown is a majority consensus phylogeny, in which nodes with posterior probabilities lower than 0.5 are collapsed.

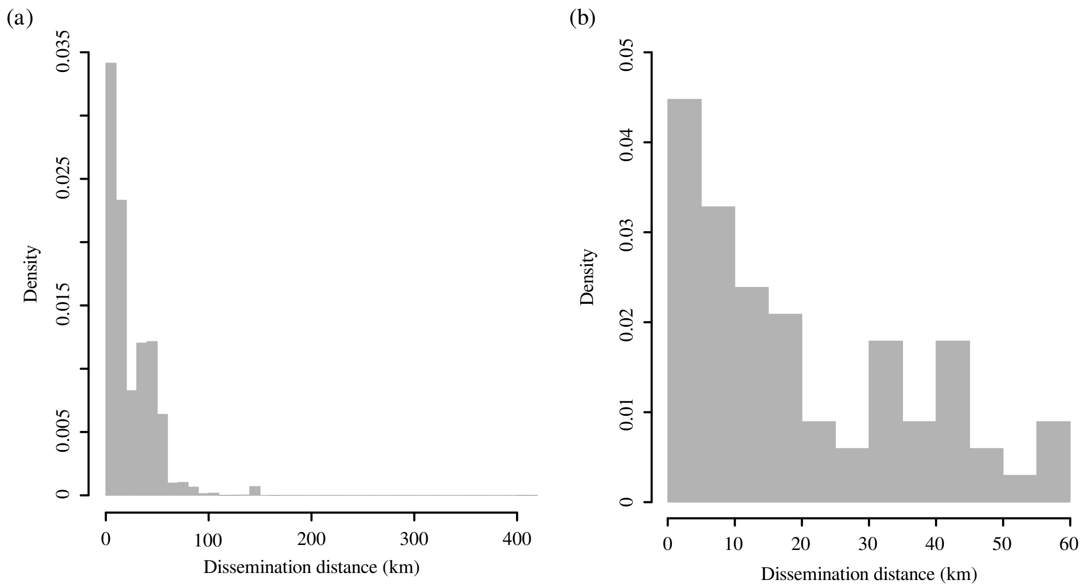


Figure S3. Posterior distributions of the transmission distances between directly connected cases. (*a*): Transmission distances between all *a posteriori* directly connected cases. (*b*): Distances between connected cases corresponding to the direct transmission links with the highest posterior probabilities (i.e. only the most probable links).

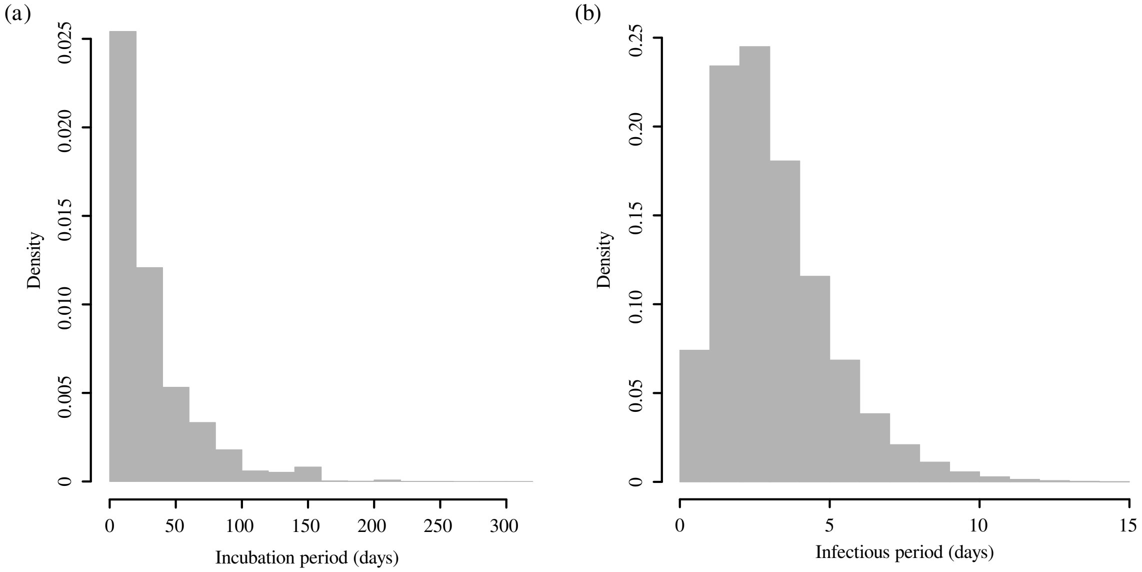


Figure S4. Posterior distributions of the incubation period (*a*) and infectious period (*b*) of directly connected cases. These distributions were obtained by aggregating the respective posterior distributions of all cases responsible for onward transmission through direct connections in the transmission tree.

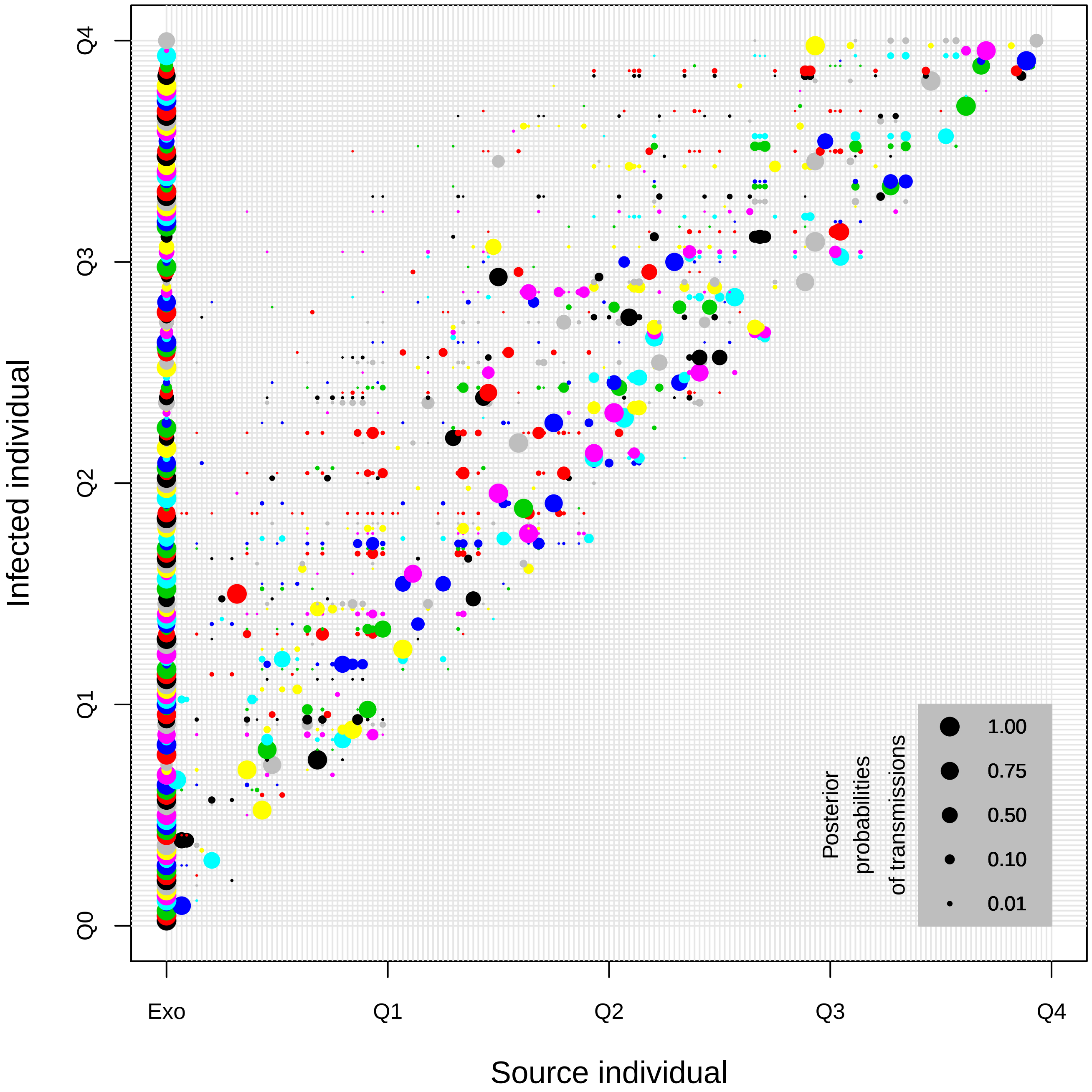


Figure S5. Graphical representation of the posterior distribution of inferred sources from the main inference algorithm (which allows only direct and exogenous sources). Individuals are arranged by observation date on both axes, with each infected individual (horizontal rows) indicated in an alternating colour for clarity. Q0 indicates the start of the sampling period, while Q1–Q4 indicate the ends of quarters of the sampling period. "Exo" indicates infection from an external source, encompassing indirect transmissions and introductions from outside the dataset.

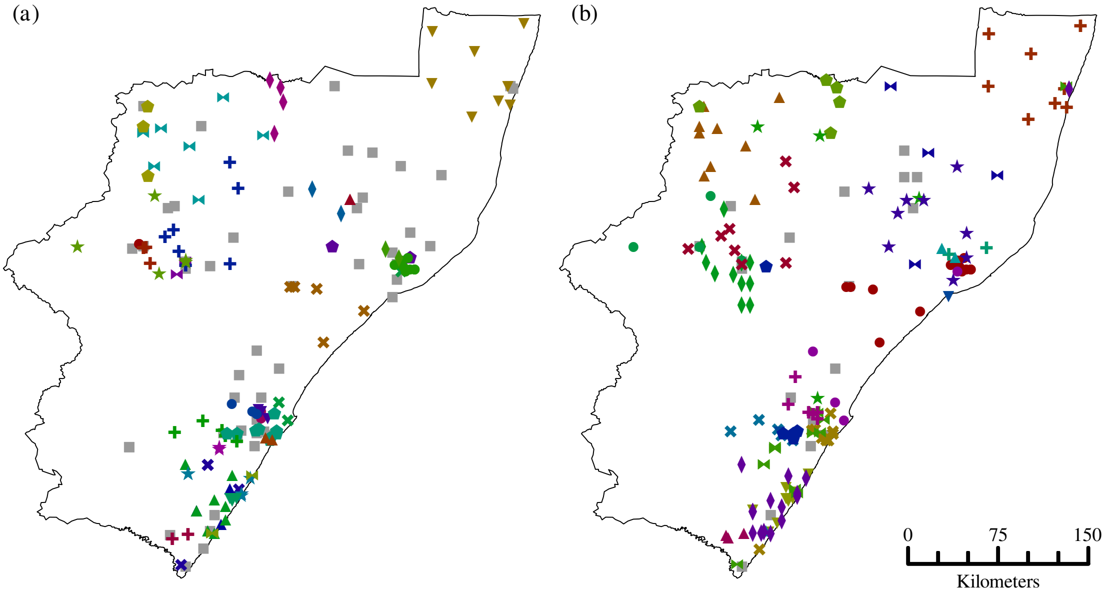


Figure S6. Cases belonging to independent transmission trees (indicated by shapes and colours) and completely unconnected cases (indicated by grey squares) when considering the most probable direct and indirect connections between cases. Indirect connections were determined by a conservative (*a*) and liberal (*b*) specification of the post-processing algorithm.

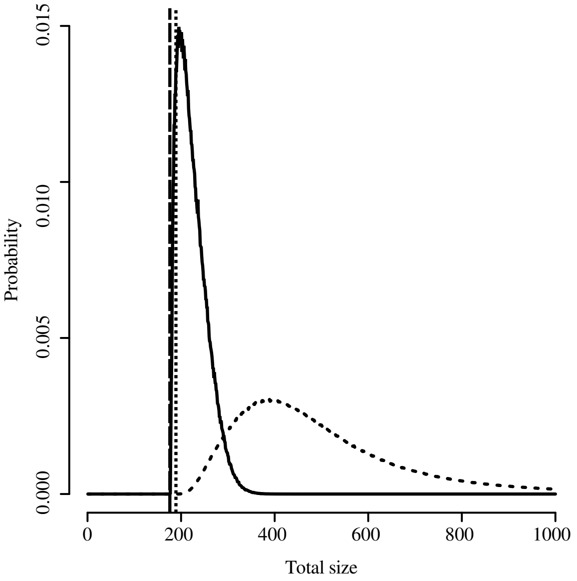


Figure S7. Estimated total number of animals infected by the canid-associated variant of rabies virus over the sampled period. Curves show the posterior distribution of cases under the conservative (dotted curve) and liberal (solid curve) specification of the transmission kernel ($\mathcal{K}_{indirect}$). The dashed vertical line shows the number of cases included in the analysis, while the dotted vertical line shows the number of cases detected by surveillance (including 5 which could not be confirmed as belonging to the canid-associated genetic variant).

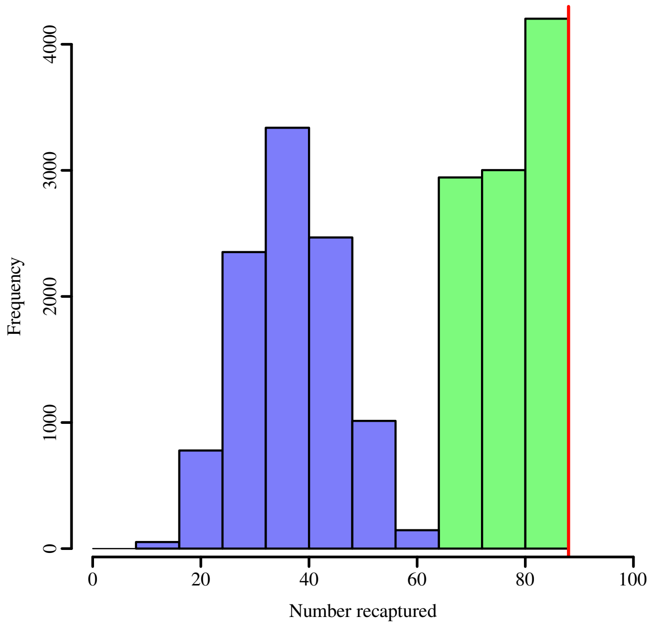


Figure S8. Posterior distributions of the number of recaptured virus lineages in the second half of the data series ($\mathcal{R}$ in figure S10). In total, 88 cases were detected during this period ($\mathcal{C}$ in figure S10), as indicated by the vertical red line. The histogram indicated in blue corresponds to data from the conservative specification of the post-processing algorithm, while the light green histogram corresponds to data from the liberal specification.

Table S1. Rabies cases detected in KwaZulu Natal between 1 March 2010 and 8 June 2011

Case number	Date	Host species	Latitude (degrees)	Longitude (degrees)	Accession number
10/129	2010/03/03	Unspecified caprine species	-30.62	30.45	KC660293
10/128	2010/03/05	*Canis lupus familiaris*	-30.78	30.13	KC660323
10/146	2010/03/12	*Canis lupus familiaris*	-27.35	30.87	KC660234
10/155	2010/03/19	*Canis lupus familiaris*	-27.75	30.90	KC660179
10/157	2010/03/20	*Bos taurus*	-28.60	29.92	KC660255
10/153	2010/03/21	*Canis lupus familiaris*	-28.60	29.92	KC660207
10/154	2010/03/21	*Canis lupus familiaris*	-29.98	30.65	KC660233
10/164	2010/03/22	*Bos taurus*	-28.30	30.15	KC660254
10/167	2010/03/22	*Canis lupus familiaris*	-28.25	31.47	KC660224
10/173	2010/03/23	*Bos taurus*	-28.73	30.23	KC660183
10/168	2010/03/25	*Canis lupus familiaris*	-29.38	30.77	KC660240
10/175	2010/03/26	*Bos taurus*	-29.52	30.93	KC660227
10/188	2010/03/29	*Canis lupus familiaris*	-28.25	31.47	KC660196
10/183	2010/04/02	*Canis lupus familiaris*	-27.00	32.08	KC660341
10/184	2010/04/02	*Canis lupus familiaris*	-30.02	30.85	KC660282
10/212	2010/04/10	*Canis lupus familiaris*	-28.72	30.23	KC660230
10/202	2010/04/11	*Canis lupus familiaris*	-27.52	30.97	KC660201
10/203	2010/04/12	*Canis lupus familiaris*	-27.70	30.35	KC660167
10/195	2010/04/13	*Canis lupus familiaris*	-30.45	30.65	KC660329
10/205	2010/04/16	*Canis lupus familiaris*	-29.83	30.80	KC660244
10/209	2010/04/16	*Canis lupus familiaris*	-30.58	30.32	KC660285
10/207	2010/04/19	*Canis lupus familiaris*	-28.72	30.23	KC660229
10/220	2010/04/22	*Canis lupus familiaris*	-30.12	30.48	KC660327
10/229	2010/04/24	*Canis lupus familiaris*	-30.73	30.42	KC660351
10/233	2010/04/27	*Canis lupus familiaris*	-28.32	31.52	KC660222
10/236[2]	2010/05/02	*Canis lupus familiaris*	N.R.	N.R.	KC660235

N.R: not recorded; N.S: not sequenced (RT-PCR unsuccessful)

[1] Herpestid-associated variant of RABV (excluded from analysis)

[2] Excluded from analysis (RT-PCR unsuccessful or coordinates not recorded)

Table S1 – continued from previous page

Case number	Date	Species	Latitude	Longitude	Accession number
10/237	2010/05/03	*Canis lupus familiaris*	-29.90	30.77	KC660328
10/245[2]	2010/05/04	*Canis lupus familiaris*	N.R.	N.R.	KC660166
10/252	2010/05/12	*Canis lupus familiaris*	-27.48	30.52	KC660163
10/261	2010/05/16	*Canis lupus familiaris*	-28.62	29.83	KC660228
10/277	2010/05/17	*Canis lupus familiaris*	-30.75	30.25	KC660322
10/268	2010/05/18	*Canis lupus familiaris*	-30.42	30.57	KC660305
10/267	2010/05/19	*Canis lupus familiaris*	-29.88	30.80	KC660243
10/269[2]	2010/05/19	*Canis lupus familiaris*	N.R.	N.R.	KC660265
10/272	2010/05/19	*Canis lupus familiaris*	-28.72	30.23	KC660275
10/274	2010/05/23	*Canis lupus familiaris*	-30.68	30.50	KC660299
10/271	2010/05/24	*Canis lupus familiaris*	-28.73	31.53	KC660209
10/278	2010/05/26	*Canis lupus familiaris*	-30.75	30.45	KC660303
10/286[1]	2010/06/08	Unspecified wildlife species	-28.03	29.98	KC660352
10/294	2010/06/22	*Canis lupus familiaris*	-28.54	30.59	KC660185
10/295	2010/06/22	*Canis lupus familiaris*	-30.32	30.58	KC660297
10/305	2010/06/29	*Canis lupus familiaris*	-30.23	30.23	KC660315
10/307	2010/06/29	*Canis lupus familiaris*	-28.58	29.89	KC660211
10/308	2010/06/30	*Canis lupus familiaris*	-28.58	29.89	KC660226
10/309	2010/06/30	*Canis lupus familiaris*	-28.72	30.56	KC660216
10/310	2010/07/01	*Canis lupus familiaris*	-29.85	30.77	KC660246
10/314	2010/07/05	*Canis lupus familiaris*	-28.63	30.18	KC660184
10/317	2010/07/06	*Canis lupus familiaris*	-28.07	32.15	KC660238
10/318	2010/07/06	*Canis lupus familiaris*	-28.73	31.82	KC660248
10/321	2010/07/06	*Canis lupus familiaris*	-27.41	30.95	KC660221
10/324	2010/07/07	*Canis lupus familiaris*	-29.83	30.78	KC660326
10/325	2010/07/07	*Bos taurus*	-28.19	31.00	KC660218
10/327	2010/07/09	*Canis lupus familiaris*	-29.98	30.15	KC660311
10/331	2010/07/16	*Canis lupus familiaris*	-28.68	31.92	KC660194
10/342	2010/07/23	*Bos taurus*	-30.10	30.75	KC660325
10/347	2010/07/28	*Canis lupus familiaris*	-30.73	30.40	KC660298
10/351	2010/07/30	*Canis lupus familiaris*	-29.90	30.85	KC660318
10/352	2010/07/30	*Canis lupus familiaris*	-29.73	30.60	KC660200
10/353	2010/07/30	*Bos taurus*	-30.10	30.48	KC660320
10/355	2010/08/02	*Canis lupus familiaris*	-31.00	30.23	KC660324
10/366	2010/08/06	*Canis lupus familiaris*	-28.81	30.17	KC660161
10/367	2010/08/06	*Canis lupus familiaris*	-28.70	31.85	KC660273
10/369	2010/08/11	*Canis lupus familiaris*	-28.60	31.33	KC660217
10/370	2010/08/11	*Canis lupus familiaris*	-28.52	30.08	KC660165
10/376	2010/08/16	*Canis lupus familiaris*	-28.78	31.85	KC660219
10/379[2]	2010/08/17	*Canis lupus familiaris*	N.R.	N.R.	KC660269
10/385	2010/08/23	*Canis lupus familiaris*	-28.32	30.10	KC660262
10/387	2010/08/24	*Canis lupus familiaris*	-30.45	30.62	KC660287
10/392	2010/08/25	*Canis lupus familiaris*	-29.77	30.93	KC660350
10/393	2010/08/25	*Canis lupus familiaris*	-30.47	30.65	KC660291
10/395	2010/08/27	*Canis lupus familiaris*	-27.39	32.08	KC660340
10/397	2010/08/30	*Canis lupus familiaris*	-30.98	30.20	KC660331
10/400	2010/08/31	*Canis lupus familiaris*	-29.85	30.80	KC660317
10/401	2010/09/01	*Canis lupus familiaris*	-28.60	32.07	KC660241
10/410	2010/09/06	*Canis lupus familiaris*	-29.97	30.51	KC660295
10/420	2010/09/13	*Canis lupus familiaris*	-28.00	30.00	KC660195
10/425	2010/09/14	*Canis lupus familiaris*	-27.85	30.26	KC660225
10/440	2010/09/22	*Canis lupus familiaris*	-28.17	31.18	KC660223
10/441	2010/09/23	*Canis lupus familiaris*	-28.72	29.97	KC660278
10/445	2010/09/27	*Canis lupus familiaris*	-28.77	30.23	KC660172

N.R: not recorded; N.S: not sequenced (RT-PCR unsuccessful)

[1] Herpestid-associated variant of RABV (excluded from analysis)

[2] Excluded from analysis (RT-PCR unsuccessful or coordinates not recorded)

Table S1 – continued from previous page

Case number	Date	Species	Latitude	Longitude	Accession number
10/447	2010/09/27	*Bos taurus*	-28.60	29.94	KC660181
10/452	2010/09/29	*Canis lupus familiaris*	-28.25	30.33	KC660162
10/453[2]	2010/09/29	*Canis lupus familiaris*	N.R.	N.R.	KC660236
10/456	2010/09/29	*Canis lupus familiaris*	-30.23	30.40	KC660330
10/458	2010/09/30	Unspecified caprine species	-30.11	29.81	KC660334
10/459	2010/09/30	*Canis lupus familiaris*	-30.42	30.63	KC660319
10/460	2010/09/30	*Canis lupus familiaris*	-27.77	30.82	KC660212
10/461	2010/09/30	*Canis lupus familiaris*	-27.75	29.92	KC660164
10/462	2010/09/30	*Canis lupus familiaris*	-28.16	30.63	KC660176
10/463	2010/09/30	*Canis lupus familiaris*	-29.57	30.63	KC660245
10/472	2010/10/05	*Canis lupus familiaris*	-29.97	30.78	KC660231
10/479	2010/10/07	*Canis lupus familiaris*	-27.64	32.38	KC660348
10/481	2010/10/08	*Canis lupus familiaris*	-28.00	31.85	KC660247
10/485	2010/10/12	*Canis lupus familiaris*	-29.85	30.77	KC660250
10/488	2010/10/12	*Canis lupus familiaris*	-28.75	30.42	KC660215
10/492	2010/10/14	*Canis lupus familiaris*	-28.73	30.23	KC660252
10/495	2010/10/15	*Canis lupus familiaris*	-30.52	30.58	KC660313
10/496[2]	2010/10/18	*Canis lupus familiaris*	-28.73	30.23	N.S.
10/498	2010/10/19	*Canis lupus familiaris*	-30.00	30.62	KC660281
10/504	2010/10/21	*Canis lupus familiaris*	-29.98	30.76	KC660177
10/506	2010/10/21	*Canis lupus familiaris*	-29.98	30.92	KC660204
10/508	2010/10/22	*Canis lupus familiaris*	-27.90	31.63	KC660173
10/509	2010/10/25	*Canis lupus familiaris*	-30.01	30.53	KC660294
10/515	2010/10/26	*Canis lupus familiaris*	-28.75	31.87	KC660272
10/517	2010/10/27	*Canis lupus familiaris*	-28.60	29.92	KC660180
10/518	2010/10/27	*Canis lupus familiaris*	-28.98	31.78	KC660263
10/524	2010/10/28	*Canis lupus familiaris*	-30.50	30.57	KC660314
10/525	2010/10/28	*Bos taurus*	-30.23	30.40	KC660321
10/527[2]	2010/11/01	*Canis lupus familiaris*	-29.92	31.00	N.S.
10/528	2010/11/01	*Canis lupus familiaris*	-29.85	30.90	KC660214
10/530	2010/11/01	*Canis lupus familiaris*	-29.85	30.77	KC660178
10/531	2010/11/03	*Canis lupus familiaris*	-30.33	30.72	KC660304
10/532	2010/11/03	*Canis lupus familiaris*	-30.58	30.32	KC660300
10/533	2010/11/03	*Canis lupus familiaris*	-30.58	30.32	KC660283
10/536	2010/11/04	*Canis lupus familiaris*	-28.47	30.14	KC660169
10/541	2010/11/09	*Canis lupus familiaris*	-29.83	30.73	KC660168
10/552	2010/11/12	*Canis lupus familiaris*	-30.00	30.92	KC660199
10/557	2010/11/13	*Canis lupus familiaris*	-30.45	30.65	KC660296
10/560[2]	2010/11/16	*Canis lupus familiaris*	N.R.	N.R.	KC660242
10/562	2010/11/17	*Canis lupus familiaris*	-30.30	30.25	KC660308
10/564	2010/11/17	Unspecified jackal species	-28.50	31.92	KC660259
10/572	2010/11/17	*Canis lupus familiaris*	-28.60	29.92	KC660174
10/579[2]	2010/11/18	*Canis lupus familiaris*	N.R.	N.R.	KC660258
10/583	2010/11/19	*Canis lupus familiaris*	-29.99	30.82	KC660316
10/593	2010/11/24	*Bos taurus*	-30.11	29.81	KC660335
10/600	2010/11/26	*Canis lupus familiaris*	-28.60	29.42	KC660276
10/608	2010/11/30	*Canis lupus familiaris*	-28.73	31.80	KC660274
10/613	2010/12/02	*Canis lupus familiaris*	-28.77	31.95	KC660271
10/614	2010/12/02	*Canis lupus familiaris*	-28.77	31.95	KC660191
10/624	2010/12/06	Unspecified caprine species	-27.96	30.56	KC660253
10/633	2010/12/08	*Canis lupus familiaris*	-30.00	30.55	KC660280
10/635	2010/12/09	*Bos taurus*	-30.77	30.12	KC660332
10/641	2010/12/13	*Equus ferus caballus*	-29.78	30.58	KC660251
10/644	2010/12/14	*Canis lupus familiaris*	-28.35	31.40	KC660190

N.R: not recorded; N.S: not sequenced (RT-PCR unsuccessful)

[1] Herpestid-associated variant of RABV (excluded from analysis)

[2] Excluded from analysis (RT-PCR unsuccessful or coordinates not recorded)

Table S1 – continued from previous page

Case number	Date	Species	Latitude	Longitude	Accession number
10/655	2010/12/20	*Canis lupus familiaris*	-30.30	30.25	KC660292
10/656	2010/12/21	*Canis lupus familiaris*	-30.05	30.87	KC660267
10/659	2010/12/21	*Canis lupus familiaris*	-27.72	30.05	KC660171
11/02	2011/01/04	*Canis lupus familiaris*	-30.03	30.87	KC660232
11/15[2]	2011/01/07	*Canis lupus familiaris*	N.R.	N.R.	KC660312
11/16[2]	2011/01/10	*Canis lupus familiaris*	N.R.	N.R.	KC660188
11/28	2011/01/14	*Canis lupus familiaris*	-29.90	30.36	KC660309
11/29	2011/01/14	*Canis lupus familiaris*	-27.41	32.65	KC660342
11/51	2011/01/24	*Canis lupus familiaris*	-28.62	31.73	KC660337
11/54	2011/01/24	*Canis lupus familiaris*	-30.50	30.62	KC660302
11/55	2011/01/24	*Canis lupus familiaris*	-30.65	30.53	KC660307
11/56	2011/01/24	*Canis lupus familiaris*	-30.55	30.53	KC660286
11/61	2011/01/26	*Canis lupus familiaris*	-29.90	31.00	KC660239
11/62	2011/01/26	*Canis lupus familiaris*	-27.52	32.58	KC660347
11/63	2011/01/27	*Canis lupus familiaris*	-28.78	31.88	KC660197
11/66	2011/01/28	*Canis lupus familiaris*	-30.50	30.45	KC660284
11/67[2]	2011/01/28	*Canis lupus familiaris*	N.R.	N.R.	KC660339
11/84	2011/02/02	*Canis lupus familiaris*	-29.73	30.80	KC660249
11/96	2011/02/08	*Canis lupus familiaris*	-27.42	32.69	KC660345
11/99	2011/02/09	*Canis lupus familiaris*	-30.87	30.37	KC660213
11/100[2]	2011/02/11	*Canis lupus familiaris*	N.R.	N.R.	KC660288
11/101	2011/02/14	*Canis lupus familiaris*	-28.70	30.23	KC660277
11/108	2011/02/15	*Canis lupus familiaris*	-28.85	31.82	KC660260
11/115	2011/02/16	*Canis lupus familiaris*	-29.32	31.27	KC660256
11/120	2011/02/17	*Bos taurus*	-27.40	32.67	KC660349
11/121	2011/02/17	*Canis lupus familiaris*	-28.80	30.03	KC660210
11/124[2]	2011/02/21	*Canis lupus familiaris*	N.R.	N.R.	N.S.
11/127	2011/02/23	*Canis lupus familiaris*	-29.08	31.57	KC660192
11/129[2]	2011/02/24	*Canis lupus familiaris*	-28.75	29.87	N.S.
11/144	2011/03/07	*Bos taurus*	-28.90	31.02	KC660270
11/178	2011/03/22	*Canis lupus familiaris*	-28.90	31.04	KC660338
11/181	2011/03/23	*Canis lupus familiaris*	-27.40	31.35	KC660279
11/185	2011/03/24	*Canis lupus familiaris*	-30.04	30.62	KC660187
11/186	2011/03/25	*Canis lupus familiaris*	-27.15	32.40	KC660343
11/188	2011/03/28	*Canis lupus familiaris*	-30.00	30.52	KC660264
11/191	2011/03/29	*Canis lupus familiaris*	-28.77	31.92	KC660261
11/195	2011/03/29	*Canis lupus familiaris*	-30.05	30.62	KC660310
11/240	2011/04/05	*Canis lupus familiaris*	-28.68	31.90	KC660198
11/241	2011/04/05	*Canis lupus familiaris*	-28.68	31.83	KC660189
11/203	2011/04/07	*Canis lupus familiaris*	-28.65	31.78	KC660336
11/208	2011/04/08	*Canis lupus familiaris*	-28.07	29.95	KC660170
11/209	2011/04/11	*Bos taurus*	-28.22	30.00	KC660206
11/212	2011/04/14	*Bos taurus*	-28.72	30.23	KC660175
11/217	2011/04/18	*Canis lupus familiaris*	-30.32	30.73	KC660301
11/221[2]	2011/04/19	*Canis lupus familiaris*	N.R.	N.R.	KC660205
11/224	2011/04/20	*Canis lupus familiaris*	-27.55	29.95	KC660160
11/232	2011/04/28	*Bos taurus*	-27.88	31.45	KC660266
11/251	2011/05/06	*Canis lupus familiaris*	-28.24	31.56	KC660182
11/252	2011/05/09	*Bos taurus*	-27.55	29.92	KC660220
11/257	2011/05/09	*Canis lupus familiaris*	-30.00	30.77	KC660333
11/262	2011/05/11	*Canis lupus familiaris*	-26.94	32.77	KC660344
11/268	2011/05/16	*Canis lupus familiaris*	-30.03	30.83	KC660186
11/270	2011/05/17	*Canis lupus familiaris*	-30.05	30.88	KC660202
11/272	2011/05/17	*Canis lupus familiaris*	-30.75	30.43	KC660290

N.R: not recorded; N.S: not sequenced (RT-PCR unsuccessful)

[1] Herpestid-associated variant of RABV (excluded from analysis)

[2] Excluded from analysis (RT-PCR unsuccessful or coordinates not recorded)

Table S1 – continued from previous page

Case number	Date	Species	Latitude	Longitude	Accession number
11/273	2011/05/17	*Canis lupus familiaris*	-30.75	30.44	KC660289
11/284	2011/05/27	*Canis lupus familiaris*	-28.90	31.05	KC660193
11/296	2011/05/31	*Canis lupus familiaris*	-27.70	29.92	KC660268
11/300[2]	2011/06/01	*Canis lupus familiaris*	N.R.	N.R.	KC660306
11/305	2011/06/06	*Canis lupus familiaris*	-27.55	32.67	KC660346
11/306[2]	2011/06/07	*Canis lupus familiaris*	N.R.	N.R.	N.S.
11/308	2011/06/08	Unspecified caprine species	-28.92	31.22	KC660257

N.R: not recorded; N.S: not sequenced (RT-PCR unsuccessful)

[1] Herpestid-associated variant of RABV (excluded from analysis)

[2] Excluded from analysis (RT-PCR unsuccessful or coordinates not recorded)

Supplementary text

1. Genetic data

1.1 RT-PCR and sequencing

TRIzol-extracted RNA was eluted in 50 µl nuclease free water. Next, RT-PCR was performed using primers GL4614F (5'-GATTTTGTAGAGGTT CACC-3') and GL5632R (5'-GACCTGGAGCAATTGTCTG-3'), amplifying a 1019 nucleotide region from position 4614 to position 5632 on the Pasteur rabies virus genome [1]. For reverse transcription, 2 pmol GL4614F was incubated with 5 µl of a 1:4 or 1:9 dilution of RNA at 70°C for 5 minutes. After a further 5 minutes in an ice bath, 4.1 µl nuclease-free water, 4 µl Improm-II reaction buffer (Promega), 3 mM $MgCl_2$, 0.5 mM dNTP mix (Roche), 20 units Protector RNase inhibitor (Roche) and 1 unit Improm-II reverse transcriptase (Promega) was added. This was followed by incubation at 25°C for 5 minutes, 42°C for 60 minutes and finally 70°C for 15 minutes. The entire reaction mixture was used for PCR by adding 10 pmol GL4614F, 12.5 pmol GL5632R, 10 µl DreamTaq buffer (Fermentas), 1.25 units DreamTaq polymerase (Fermentas) and 67.5 µl nuclease free water. This reaction was incubated at 94°C for 1 minute, followed by 40 cycles of 94°C for 30 seconds, 56°C for 30 seconds and 72°C for 90 seconds, before a final incubation step of 72°C for 7 minutes. PCR products were purified from a 1% agarose gel containing 0.0001 mg/ml Ethidium Bromide using the Wizard SV gel and PCR cleanup system (Promega). Amplicons were sequenced in both directions by dye-terminator chemistry on an ABI 3100 or 3500xL sequencer using the BigDye v3.1 cycle sequencing kit (Applied Biosystems).

1.2 Phylogenetic analysis

A Bayesian phylogeny was constructed with Beast version 1.7.3 [2] using Kimura's 3-parameter nucleotide substitution model [7] and assuming a strict molecular clock with a broad substitution rate prior (defined by a uniform distribution between 1×10^{-5} and 1×10^{-2}). The best-fitting demographic model was determined to be one describing exponential growth using both path sampling and stepping stone sampling of individual Markov chains of 50 million iterations each, saving every 5000th step [3]. This model was used to construct two Markov chains of 50 million steps each, saving every 10 000th step. The resulting estimates were checked for convergence and the posterior estimates of trees were combined after a burn-in of 10% of each chain, and then summarised as a majority consensus tree using Dendroscope version 3.2.2 [4].

2. Transmission model including direct and external transmissions

We extended the model of Morelli *et al.* (2012) [5] to allow for the complexities inherent to polyphyletic epidemics and endemic diseases. Their model reconstructs transmission trees from a combination of temporal, spatial and genetic data, but allows for only a single introduction, followed by direct transmission for the remainder of the epidemic. This means that missing cases will deform the distributions describing the incubation and infectious periods, while the polyphyletic origins of many datasets will not be adequately reflected. These factors make it unsuitable for application to endemic diseases. In what follows, we present only the changes to the model of Morelli *et al.*, using the same notation for clarity (a summary of the notation used here is provided in table S5 at the end of this document).

2.1 Transmission tree and infection times

Concerning the transmission dynamics, we added into the model of Morelli *et al.* (i) the possibility of multiple introductions (i.e. transmissions from sources external to the dataset) and (ii) the existence of two classes of hosts, namely those that can spread the disease and those that are definite dead-end hosts. The joint conditional probability distribution of the transmission tree J and the infection times $\mathbf{T}^{inf}$ given incubation periods $\mathbf{L}$, infectious periods $\mathbf{D}$, observation times $\mathbf{T}^{obs}$ (which are assumed to coincide with the times of deaths), host locations $\mathbf{X}$ and capacities to spread the disease $\mathbf{C}$ (see the definition of $\mathbf{C}$ below) can be written:

$$\begin{aligned} p(J, \mathbf{T}^{\mathit{inf}} | \mathbf{L}, \mathbf{D}, \theta, \mathbf{T}^{\mathit{obs}}, \mathbf{X}, \mathbf{C}) &= p\left(J(1), T_1^{\mathit{inf}} \mid \mathbf{L}, \mathbf{D}, \theta, \mathbf{T}^{\mathit{obs}}, \mathbf{X}, \mathbf{C}\right) \\ &\times \prod_{i=2}^{I} p\left(J(i), T_i^{\mathit{inf}} \mid J\{1:(i-1)\}, T_{1:(i-1)}^{\mathit{inf}}, \mathbf{L}, \mathbf{D}, \theta, \mathbf{T}^{\mathit{obs}}, \mathbf{X}, \mathbf{C}\right), \end{aligned} \tag{1}$$

where θ is a vector of all unknown parameters, I is the total number of hosts, the index i is sorted with respect to increasing infection times T_i^{inf}, $J\{1:(i-1)\} = (J(1), \ldots, J(i-1))$ and $T_{1:(i-1)}^{\mathit{inf}} = (T_1^{\mathit{inf}}, \ldots, T_{i-1}^{\mathit{inf}})$. Note that in contrast to the model of Morelli *et al.* (2012), there is no delay here between the time when a host is observed as infected and its death. The model assumes that observation of infection leads to the animal being euthanised (or that cases are detected immediately upon death).

Each host has the same chance $(1/I)$ to be infected first (by an external source $J(1) = 0$), and its infection time is assumed to be less than or equal to the first observation time ($\min\{\mathbf{T}^{\mathit{obs}}\}$):

$$p\left(J(1), T_1^{\mathit{inf}} \mid \mathbf{L}, \mathbf{D}, \theta, \mathbf{T}^{\mathit{obs}}, \mathbf{X}, \mathbf{C}\right) = \frac{1}{I} \times \mathbf{1}(T_1^{\mathit{inf}} \leq \min\{\mathbf{T}^{\mathit{obs}}\}),$$

where $\mathbf{1}$ is the indicator function. Subsequent infections (i.e. for $i > 1$) occur with the following probabilities:

$$\begin{aligned} & p\left(J(i), T_i^{\mathit{inf}} \mid J\{1:(i-1)\}, T_{1:(i-1)}^{\mathit{inf}}, \mathbf{L}, \mathbf{D}, \theta, \mathbf{T}^{\mathit{obs}}, \mathbf{X}, \mathbf{C}\right) \\ &= \exp\left(-\alpha_0(T_i^{\mathit{inf}} - T_1^{\mathit{inf}}) - \int_{T_1^{\mathit{inf}}}^{T_i^{\mathit{inf}}} \sum_{j=1}^{i-1} \alpha_1 C_j \mathbf{1}(T_j^{\mathit{inf}} + L_j \leq t \leq T_j^{\mathit{obs}}) f_{\alpha_2}(||X_i - X_j||)dt\right) \\ &\quad \times \left(\alpha_0 \mathbf{1}\{J(i) = 0\} + \alpha_1 C_{J(i)} \mathbf{1}(T_{J(i)}^{\mathit{inf}} + L_{J(i)} \leq T_i^{\mathit{inf}} \leq T_{J(i)}^{\mathit{obs}}) f_{\alpha_2}(||X_i - X_{J(i)}||) \mathbf{1}\{J(i) > 0\}\right) \end{aligned}$$

where the exponential term is the probability that host i has not been infected between times T_1^{inf} and T_i^{inf}, and the second term is the probability density that host i has been infected by $J(i)$ at time T_i^{inf}. Here, if $J(i) > 0$ the source is observed, while the source is external to the dataset (an introduction) if $J(i) = 0$. α_0 is the infection strength of the external source, assumed to be constant in time and space, α_1 is the infection strength of an observed source, and α_2 is a vector of the parameters defining the shape of the spatial transmission kernel f_{α_2} (see below). $||\cdot||$ is a geographic distance (the great-circle distance in this study). The binary variable C_j indicates whether j can be contagious ($C_j = 1$) or not ($C_j = 0$). In this study, dogs and jackals can spread the disease, whereas livestock (cattle, goats and sheep) are dead-end hosts.

2.2 Transmission kernel

We used the exponential-power transmission kernel [6], which is a two-parameter kernel (scale parameter $\alpha_{2,1}$ and shape parameter $\alpha_{2,2}$) and has the advantage of including fat-tailed kernels (when $\alpha_{2,2} < 1$), thin-tailed kernels (when $\alpha_{2,2} > 1$), the exponential kernel ($\alpha_{2,2} = 1$) used in Morelli *et al.* (2012), and the normal kernel ($\alpha_{2,2} = 2$). Thus, the exponential-power kernel, which is often used in dispersal studies, is a very general kernel, making it well suited to a range of endemic situations where often very little is known regarding spatial transmission patterns. The exponential-power kernel satisfies, for all distances $r \geq 0$:

$$f_{\alpha_2}(r) = \frac{\alpha_{2,2}}{2\pi(\alpha_{2,1})^2 \Gamma(2/\alpha_{2,2})} \exp\left\{-\left(\frac{r}{\alpha_{2,1}}\right)^{\alpha_{2,2}}\right\}.$$

2.3 Nucleotide substitutions

To take into account heterogeneity in the rates of different types of nucleotide substitutions, we replaced the Jukes-Cantor substitution model used in Morelli *et al.* (2012) with the 3-parameter Kimura model [7]. In this model, the substitution rates are different for transitions (U ↔ C and A ↔ G), transversions of type 1 (U ↔ A and C ↔ G) and transversions of type 2 (U ↔ G and A ↔ C) [7]. Therefore, the numbers of transitions, type-1 transversions, type-2 transversions and unchanged bases over a time lag Δ are distributed according to a multinomial distribution, say $P_{\mu,s}(\cdot \mid \Delta)$, with size equal to the length s of the observed sequence fragment and with the following vector of probabilities:

$$\frac{1}{4}\big(1 - e_1 - e_2 + e_3,\ 1 - e_1 + e_2 - e_3,\ 1 + e_1 - e_2 - e_3,\ 1 + e_1 + e_2 + e_3\big),$$

where $e_1 = \exp\{-2(\mu_1 + \mu_2)\Delta\}$, $e_2 = \exp\{-2(\mu_1 + \mu_3)\Delta\}$, $e_3 = \exp\{-2(\mu_2 + \mu_3)\Delta\}$, and μ_1, μ_2 and μ_3 are the genetic substitution rates per nucleotide per day, for transitions, type-1 transversions and type-2 transversions, respectively. In what follows, we use the notation $\mu = (\mu_1, \mu_2, \mu_3)$ for simplicity.

2.4 Observed and unobserved pathogen sequences

For the purposes of this discussion, we use the terms "sequence" and "sequence fragment" interchangeably. Here, we are interested in the conditional distribution $p_{\mu,s}(\mathbf{S}^{obs} \mid J, \mathbf{T}^{obs}, \mathbf{T}^{inf}, S_{\text{MRCA}}, T_{\text{MRCA}})$ of observed pathogen sequences $\mathbf{S}^{obs}$ given the transmission tree J, infection times $\mathbf{T}^{inf}$, observation times $\mathbf{T}^{obs}$, and the genetic sequence S_{MRCA} and time T_{MRCA} of the most recent common ancestor (MRCA) of the observed pathogens. This probability distribution, which depends on the nucleotide substitution rates μ and the length s of the sequence fragment that is observed, can be written as a sum over the unknown transmitted genetic sequences S_i at time T_i^{inf} (for all i such that $J(i) > 0$):

$$\begin{aligned} &p_{\mu,s}(\mathbf{S}^{obs} \mid J, \mathbf{T}^{obs}, \mathbf{T}^{inf}, S_{\text{MRCA}}, T_{\text{MRCA}}) \\ &= \sum_{\substack{\{S_i \in \mathbb{S}:\ i=1,\ldots,I \\ \text{and } J(i)>0\}}} \left\{ \left(\prod_{i=1}^{I} P_{\mu,s}\{M(S_i^{obs}, S^{\dagger}_{\text{prec}(i,obs)}) \mid \Delta = T_i^{obs} - T^{\dagger}_{\text{prec}(i,obs)}\} \right) \right. \\ &\qquad \left. \times \left(\prod_{\substack{i=1 \\ J(i)>0}}^{I} P_{\mu,s}\{M(S_i, S^{*}_{\text{prec}(i,inf)}) \mid \Delta = T_i^{inf} - T^{*}_{\text{prec}(i,inf)}\} \right) \right\}. \end{aligned} \tag{2}$$

In equation (2), $\mathbb{S}$ is the set of all possible sequences (the size of $\mathbb{S}$ is 4^s, where s is the length of the sequence fragment), $M(S', S)$ is the vector of the numbers of transitions, type-1 transversions, type-2 transversions and unchanged bases between S and S', and $P_{\mu,s}\{M(S', S) \mid \Delta = T' - T\}$ is the probability given by the multinomial distribution in Section 2.3.

The subscript $\text{prec}(i, obs)$ can take two types of values. First case: if i was infected from outside the dataset ($J(i) = 0$) and i did not infect any other observed host, then $\text{prec}(i, obs) = \text{MRCA}$, which means $S^{\dagger}_{\text{prec}(i,obs)} = S_{\text{MRCA}}$ and $T^{\dagger}_{\text{prec}(i,obs)} = T_{\text{MRCA}}$. Second case: if $J(i) > 0$ or if i infected another observed host, then $\text{prec}(i, obs)$ denotes the host whose node of infection belongs to the tree path from the root of the tree to the observation of i (at time T_i^{obs}) and whose infection immediately precedes the observation of i (i.e. $S^{\dagger}_{\text{prec}(i,obs)}$ is the transmitted sequence $S_{\text{prec}(i,obs)}$ at the infection time $T^{\dagger}_{\text{prec}(i,obs)} = T^{inf}_{\text{prec}(i,obs)}$). The node of infection of a given host k is defined as the point on the tree at which "the branch leading to the observation of k" and "the branch leading to the observation of the infecting host $J(k)$" diverged. The tree path from one point of the tree to another is defined as the most direct line on the graph connecting the two points. In the second case, if $J(i) > 0$ and i did not infect any other host, then $\text{prec}(i, obs)$ is i itself.

The subscript $\text{prec}(i, inf)$ can also take two types of values. First case: if $J(J(i)) = 0$ and $J(i)$ did not infect any other observed host before the infection of i at T_i^{inf}, then $\text{prec}(i, inf) = \text{MRCA}$, which means $S^{*}_{\text{prec}(i,inf)} = S_{\text{MRCA}}$ and $T^{*}_{\text{prec}(i,inf)} = T_{\text{MRCA}}$. Second case: if $J(J(i)) > 0$ or if $J(i)$ infected another observed host before the infection of i at T_i^{inf}, then $\text{prec}(i, inf)$ denotes the host whose node of infection belongs to the tree path from the root of the tree to the infection of i (at time T_i^{inf}) and whose infection is just preceding the infection of i (i.e. $S^{*}_{\text{prec}(i,inf)}$ is the transmitted sequence $S_{\text{prec}(i,inf)}$ at the infection time $T^{*}_{\text{prec}(i,inf)} = T^{inf}_{\text{prec}(i,inf)}$).

In other words, the first series of factors in equation (2) accounts for the probabilities of the number of substitutions between an observed sequence and the immediately preceding unobserved, transmitted sequence or the sequence of the MRCA. The second series of factors accounts for the probabilities of the number of substitutions between each transmitted sequence and the transmitted sequence immediately preceding it in time or the sequence of the MRCA.

2.5 Prior distributions

Independent prior distributions were used for all parameters.

Independent vague exponential priors with a mean value of 100 were chosen for the parameter α_0 relating to the infection strength of external sources and parameter α_1 relating to the infection strength of observed sources.

A vague exponential prior with mean value of 100 was chosen for the scale parameter $\alpha_{2,1}$ of the transmission kernel, while a gamma prior distribution with mean $a_{2,2}^{(1)} = 1$ and standard deviation $a_{2,2}^{(2)} = 1$ was specified for the shape parameter $\alpha_{2,2}$ of the transmission kernel. Doing so allows classical kernels to have a non-negligible weight, in particular the thin-tailed normal kernel ($\alpha_{2,2} = 2$), the exponential kernel ($\alpha_{2,2} = 1$) and the fat-tailed kernel corresponding to $\alpha_{2,2} = 0.5$.

For the parameters governing incubation and infectious periods, we used very narrow prior distributions centered around values matching distributions fitted to contact tracing data from rural Tanzania [8]. Narrow prior distributions were used to prevent the inference algorithm from creating direct connections by extending the incubation and infectious periods when one or more intermediate cases have not been sampled.

Incubation periods were modelled by independent gamma distributions (which were the best fitting distributions for both the incubation and infectious periods in [8]) with mean parameter β_1 and standard deviation β_2. A narrow gamma prior with mean $b_1^{(1)} = 22.1$ days and standard deviation $b_1^{(2)} = 0.01$ was specified for β_1. A narrow gamma prior with mean $b_2^{(1)} = 21.2$ days and standard deviation $b_2^{(2)} = 0.01$ was specified for β_2.

Infectious periods were modeled by independent gamma distributions with mean parameter δ_1 and standard deviation δ_2. A narrow gamma prior with mean $d_1^{(1)} = 3.1$ days and standard deviation $d_1^{(2)} = 0.01$ was specified for δ_1. A narrow gamma prior with mean $d_2^{(1)} = 1.8$ days and standard deviation $d_2^{(2)} = 0.01$ was specified for δ_2.

Independent exponential prior distributions with a mean parameter $m = 2 \times 10^{-6}$ substitutions per nucleotide per day were used for the substitution rates μ_1, μ_2 and μ_3.

A vague normal prior distribution with mean $t_{\text{MRCA}}^{(1)} = -5500$ days and standard deviation $t_{\text{MRCA}}^{(2)} = 10000$ days was specified for T_{MRCA}.

3. Inference

In this study, we estimate the same parameters and latent variables as in Morelli *et al.* (2012) [5], and additionally also the time T_{MRCA} of the MRCA and the substitution rates μ_1, μ_2 and μ_3. We modified the pseudo-distribution for the genetic data introduced by Morelli *et al.* specifically to take into account the possibility for multiple introductions and used interacting Markov chain Monte Carlo (MCMC) instead of simple MCMC to allow us to reduce computation time by exploiting multiple computation cores.

3.1 Pseudo-distribution for the observed pathogen sequences

The sequence S_{MRCA} was reconstructed *a priori* using the FastML web server with a general time reversible (GTR) substitution model [9, 10] and was considered as fixed. In future developments of the method, this sequence could be estimated within our inference algorithm.

To reduce the complexity of the inference algorithm, we used a conditional pseudo-distribution of $\mathbf{S}^{obs}$, noted $\tilde{p}_{\mu,s}(\mathbf{S}^{obs} \mid J, \mathbf{T}^{obs}, \mathbf{T}^{inf}, S_{\text{MRCA}}, T_{\text{MRCA}})$, instead of the exact conditional distribution $p_{\mu,s}(\mathbf{S}^{obs} \mid J, \mathbf{T}^{obs}, \mathbf{T}^{inf}, S_{\text{MRCA}}, T_{\text{MRCA}})$. The conditional pseudo-distribution does not depend on the extra latent vectors $\{S_i : i = 1, \dots, I,\ J(i) > 0\}$ appearing in equation (2) – see Morelli *et al.* [5] for details.

With index i being sorted with respect to increasing infection times T_i^{inf}, the distribution $p_{\mu,s}(\mathbf{S}^{obs} \mid J, \mathbf{T}^{obs}, \mathbf{T}^{inf}, S_{\text{MRCA}}, T_{\text{MRCA}})$ can be written as:

$$\begin{aligned} p_{\mu,s}(\mathbf{S}^{obs} \mid J, \mathbf{T}^{obs}, \mathbf{T}^{inf}, S_{\text{MRCA}}, T_{\text{MRCA}}) &= p_{\mu,s}(S_1^{obs} \mid J, \mathbf{T}^{obs}, \mathbf{T}^{inf}, S_{\text{MRCA}}, T_{\text{MRCA}}) \\ &\times \prod_{i=2}^{I} p_{\mu,s}(S_i^{obs} \mid S_{1:(i-1)}^{obs}, J, \mathbf{T}^{obs}, \mathbf{T}^{inf}, S_{\text{MRCA}}, T_{\text{MRCA}}), \end{aligned} \tag{3}$$

where $S_{1:(i-1)}^{obs}$ is the set of observed sequences from hosts $1, \dots, i-1$. For the first infected host,

$$p_{\mu,s}(S_1^{obs} \mid J, \mathbf{T}^{obs}, \mathbf{T}^{inf}, S_{\text{MRCA}}, T_{\text{MRCA}}) = P_{\mu,s}\{M(S_1^{obs}, S_{\text{MRCA}}) \mid \Delta = T_1^{obs} - T_{\text{MRCA}}\}.$$

For the other hosts infected by external sources (i.e. $i > 1$ such that $J(i) = 0$),

$$p_{\mu,s}(S_i^{obs} \mid J, \mathbf{T}^{obs}, \mathbf{T}^{inf}, S_{\text{MRCA}}, T_{\text{MRCA}}) = P_{\mu,s}\{M(S_i^{obs}, S_{\text{MRCA}}) \mid \Delta = T_i^{obs} - T_{\text{MRCA}}\}.$$

For hosts not directly infected by external sources (i.e. $i > 1$ such that $J(i) > 0$), we replaced the conditional probability $p_{m,s}(S_i^{obs} \mid S_{1:(i-1)}^{obs}, J, \mathbf{T}^{obs}, \mathbf{T}^{inf}, S_{\text{MRCA}}, T_{\text{MRCA}})$ of S_i^{obs} given sequences S_j^{obs} ($j = 1, \dots, i-1$) by the product, up to a power, of the conditional probabilities of S_i^{obs} given *each* sequence S_j^{obs} such that $j \in 1, \dots, i-1$ and j is in the transmission chain leading to i (the latter condition is mathematically written: $\exists n \in \mathbb{N}^*, J^n(i) = j$):

$$\left(\prod_{\substack{j=1 \\ \exists n \in \mathbb{N}^*, J^n(i) = j}}^{i-1} P_{m,s}\{M(S_i^{obs}, S_j^{obs}) \mid \Delta = \mid T_i^{obs} - T_{\text{div}(i,j)}^{inf} \mid + \mid T_j^{obs} - T_{\text{div}(i,j)}^{inf} \mid\} \right)^{1/\eta_i},$$

where η_i is the number of terms in the product, $T_{\text{div}(i,j)}^{obs}$ denotes the infection time at which the chain of infection leading to i and the chain of infection leading to j diverged ($T_{\text{div}(i,j)}^{inf}$ is one of the latent variables in $\mathbf{T}^{inf}$, also called "augmented data") and $\Delta = \mid T_i^{obs} - T_{\text{div}(i,j)}^{inf} \mid + \mid T_j^{obs} - T_{\text{div}(i,j)}^{inf} \mid$ is the evolutionary duration separating the observation of S_i^{obs} and S_j^{obs}. The use of the power $1/\eta_i$ is a way to get a quantity homogeneous to a single probability and not to a product of probabilities whatever the length of the transmission chain leading to i. This means the hosts have similar weights in the pseudo-distribution given below.

Thus, the conditional pseudo-distribution of $\mathbf{S}^{obs}$ satisfies:

$$\begin{aligned} &\tilde{p}_{\mu,s}(\mathbf{S}^{obs} \mid J, \mathbf{T}^{obs}, \mathbf{T}^{inf}, S_{\text{MRCA}}, T_{\text{MRCA}}) \\ &= \prod_{\substack{i=1 \\ J(i)=0}}^{I} P_{\mu,s}\{M(S_i^{obs}, S_{\text{MRCA}}) \mid \Delta = T_i^{obs} - T_{\text{MRCA}}\} \\ &\times \prod_{\substack{i=1 \\ J(i)>0}}^{I} \left(\prod_{\substack{j=1 \\ \exists n \in \mathbb{N}^*, J^n(i) = j}}^{i-1} P_{m,s}\{M(S_i^{obs}, S_j^{obs}) \mid \Delta = \mid T_i^{obs} - T_{\text{div}(i,j)}^{inf} \mid + \mid T_j^{obs} - T_{\text{div}(i,j)}^{inf} \mid\} \right)^{1/\eta_i}. \end{aligned} \tag{4}$$

3.2 Interacting MCMC

The MCMC algorithm of Morelli *et al.* can be directly adapted to the new model developed for endemic diseases. However, to decrease the computation time, we ran 20 interacting Markov chains in parallel. The chains were independently run except every 2000 iterations when importance sampling among the 20 current chain states was performed [11]. Following a burn-in of 5000 iterations for each of the 20 chains, we sampled every 250th iteration to gain a total posterior sample size of 10^4. For our study, the parallel algorithm took 10 days using a computation cluster equipped with Intel Xeon X5690 processors (3.46GHz and below).

3.3 Indirect transmissions

As described in the main text, indirect transmissions were inferred using a post-processing analysis of the output provided by the algorithm described above. Sources of indirect transmission were represented by extending the infectious potentials of infectious hosts after their deaths, as illustrated by figure S9.

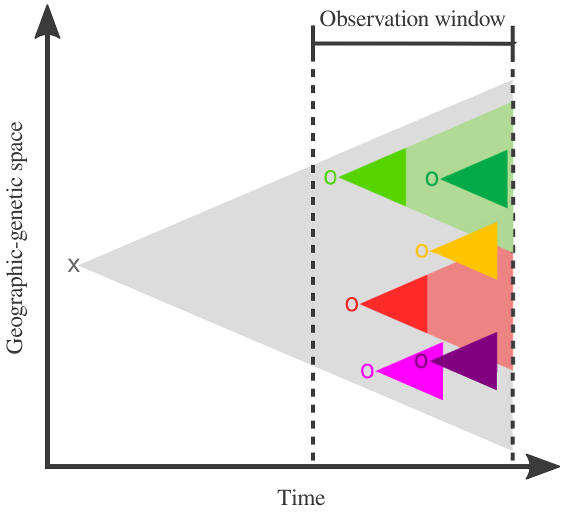


Figure S9. Inferring indirect transmissions. The unsampled index case is represented by an x, while o's represent observed cases. The influence of cases in both the genetic and geographic space will be extended by cases further down the transmission chain. Thus, the effect of unsampled cases can be modelled by allowing cases to continue moving and evolving after the death of the host involved, as illustrated by the green and red cones. In this way, we can detect the indirect causal connection between the light and dark green cases in the figure caused by an unsampled intermediate case whose host was infected by the light green case and which then went on to infect the host of the dark green case.

The algorithm described below aims to determine if hosts originally considered as infected by external sources might have been infected by these sources of indirect transmission. Although including this reconstruction in the main inference algorithm would be more statistically rigorous, this task would require more computational effort. We will test this possibility in future work.

Indirect transmissions determined with genetic data only (liberal specification). The output of the inference algorithm is a table of 10^4 rows corresponding to the sampled states of the MCMC chains. In this table each column corresponds to a given unknown parameter or latent variable. We introduce the index b that indicates which state of the posterior sample is considered. For example, $J^{(b)}(i)$ is the source of host i at the bth state of the posterior sample.

- For each $b = 1, \ldots, 10^4$:
  - Initialize the vector $\tilde{J}^{(b)}$ by setting $\tilde{J}^{(b)} = J^{(b)}$;
  - For each host i such that $J^{(b)}(i) = 0$:
    - Compute the value $r_i^{(b)}$ of the pseudo-distribution of the genetic data (equation (4)) with $(J, \mathbf{T}^{\mathit{inf}}, T_{\text{TMRCA}}, \mu) = (J^{(b)}, \mathbf{T}^{\mathit{inf}(b)}, T_{\text{TMRCA}}^{(b)}, \mu^{(b)})$;
    - For each host $j \neq i$ satisfying $C_j = 1$ and $T_j^{\mathit{inf}} + L_j \leq T_i^{\mathit{inf}}$ (j is a possible indirect source since we ignore its death time T_j^{obs}):
      - Generate a new transmission tree $J^{(b,*)}$ such that $J^{(b,*)}$ coincides with $J^{(b)}$ except $J^{(b,*)}(i) = j$;
      - Compute the value $r_i^{(b,*)}$ of the pseudo-distribution of the genetic data (equation (4)) with $(J, \mathbf{T}^{\mathit{inf}}, T_{\text{TMRCA}}, \mu) = (J^{(b,*)}, \mathbf{T}^{\mathit{inf}(b)}, T_{\text{TMRCA}}^{(b)}, \mu^{(b)})$;
    - Select j such that $r_i^{(b,*)}$ is maximum and set $\tilde{J}^{(b)}(i) = J^{(b,*)}(j)$ with probability $\min(1, r_i^{(b,*)}/r_i^{(b)})$.

If $\tilde{J}^{(b)}(i)$ is changed from zero to j, then i is considered as indirectly infected by j in state b. Therefore, we obtain a posterior distribution of the transmission tree including and differentiating direct, indirect and external transmissions.

Indirect transmissions determined with genetic and spatial data (conservative specification). In the conservative specification, both genetic and spatial information is taken into account when inferring indirect transmissions, which restricts the spatial influence of undetected cases. In this paragraph, we extend the formula of the distribution of the transmission tree and the infection times (equation (1)) by differentiating the death of host j at time T_j^{obs} and the end of its potential influence at time T_j^{end}. Between times T_j^{obs} and T_j^{end}, j cannot infect any other host, but unobserved hosts infected by j before T_j^{obs} might go on to infect other hosts. Thus, the distribution of the transmission tree and the infection times given in equation (1) is replaced by:

$$\begin{aligned} p^{\mathit{indirect}}(J, \mathbf{T}^{\mathit{inf}} \mid \mathbf{L}, \mathbf{D}, \theta, \mathbf{T}^{\mathit{obs}}, \mathbf{T}^{\mathit{end}}, \mathbf{X}, \mathbf{C}) &= \frac{1}{I} \times \mathbf{1}(T_1^{\mathit{inf}} \leq \min\{\mathbf{T}^{\mathit{obs}}\}) \\ \times \prod_{i=2}^{I} \exp\left(-\alpha_0(T_i^{\mathit{inf}} - T_1^{\mathit{inf}}) - \int_{T_1^{\mathit{inf}}}^{T_i^{\mathit{inf}}} \sum_{j=1}^{i-1} \alpha_1 C_j \mathbf{1}(T_j^{\mathit{inf}} + L_j \leq t \leq T_j^{\mathit{end}}) f_{\alpha_2}(\|X_i - X_j\|) dt\right) \\ \times \left(\alpha_0 \mathbf{1}\{J(i) = 0\} + \alpha_1 C_{J(i)} \mathbf{1}(T_{J(i)}^{\mathit{inf}} + L_{J(i)} \leq T_i^{\mathit{inf}} \leq T_{J(i)}^{\mathit{end}}) f_{\alpha_2}(\|X_i - X_{J(i)}\|) \mathbf{1}\{J(i) > 0\}\right). \end{aligned} \tag{5}$$

The difference between equations (1) and (5) is the time interval during which j is considered as infectious ($[T_j^{\mathit{inf}} + L_j, T_j^{\mathit{obs}}]$ in equation (1) and $[T_j^{\mathit{inf}} + L_j, T_j^{\mathit{end}}]$ in equation (5)).

The algorithm described in the paragraph above is replaced by the following one:

- For each $b = 1, \ldots, 10^4$:
  - Initialize the vector $\tilde{J}^{(b)}$ by setting $\tilde{J}^{(b)} = J^{(b)}$;
  - For each host i such that $J^{(b)}(i) = 0$:
    - Compute the value $r_i^{(b)}$ of the pseudo-distribution of the genetic data (equation (4)) with $(J, \mathbf{T}^{\mathit{inf}}, T_{\text{TMRCA}}, \mu) = (J^{(b)}, \mathbf{T}^{\mathit{inf}(b)}, T_{\text{TMRCA}}^{(b)}, \mu^{(b)})$;
    - Compute the value $q_i^{(b)}$ of the distribution of the transmission tree and the infection times (equation (1)) with $(J, \mathbf{T}^{\mathit{inf}}, \mathbf{L}, \mathbf{D}, \theta) = (J^{(b)}, \mathbf{T}^{\mathit{inf}(b)}, \mathbf{L}^{(b)}, \mathbf{D}^{(b)}, \theta^{(b)})$;
    - For each host $j \neq i$ satisfying $C_j = 1$ and $T_j^{\mathit{inf}} + L_j \leq T_i^{\mathit{inf}}$ (j is a possible indirect source since we ignore its death time T_j^{obs}):
      - Generate a new transmission tree $J^{(b,*)}$ such that $J^{(b,*)}$ coincides with $J^{(b)}$ except $J^{(b,*)}(i) = j$;
      - Generate an end-time vector $\mathbf{T}^{\mathit{end}(b,*)}$ such that $\mathbf{T}^{\mathit{end}(b,*)}$ coincides with $\mathbf{T}^{\mathit{obs}}$ except $T_j^{\mathit{end}(b,*)} = T_i^{\mathit{inf}} + 1$ (the end time of j is fixed at the infection time of i plus one day such that j may indirectly infect i);
      - Compute the value $r_i^{(b,*)}$ of the pseudo-distribution of the genetic data (equation (4)) with $(J, \mathbf{T}^{\mathit{inf}}, T_{\text{TMRCA}}, \mu) = (J^{(b,*)}, \mathbf{T}^{\mathit{inf}(b)}, T_{\text{TMRCA}}^{(b)}, \mu^{(b)})$;
      - Compute the value $q_i^{(b,*)}$ of the new distribution of the transmission tree and the infection times (equation (5)) with $(J, \mathbf{T}^{\mathit{inf}}, \mathbf{L}, \mathbf{D}, \theta, \mathbf{T}^{\mathit{end}}) = (J^{(b,*)}, \mathbf{T}^{\mathit{inf}(b)}, \mathbf{L}^{(b)}, \mathbf{D}^{(b)}, \theta^{(b)}, \mathbf{T}^{\mathit{end}(b,*)})$;
    - Select j such that the product $r_i^{(b,*)} \times q_i^{(b,*)}$ is maximum and set $\tilde{J}^{(b)}(i) = J^{(b,*)}(j)$ with probability $\min(1, r_i^{(b,*)} \times q_i^{(b,*)} / \{r_i^{(b)} \times q_i^{(b)}\})$.

As above, if $\tilde{J}^{(b)}(i)$ is changed from zero to j, then i is considered as indirectly infected by j in state b. Therefore, we obtain a posterior distribution of the transmission tree including and differentiating direct, indirect and external transmissions.

4. Population size

Consider a set of infected hosts observed at times $\mathbf{T}^{\mathit{obs}}$ and linked through a transmission tree $\tilde{J}$ including direct and indirect transmissions as well as transmissions from external sources. This set of hosts is a fraction of the population of infected hosts in the study area and period; let $\mathcal{N}$ be the size of this population that we aim to infer.

The set of observed hosts is split into two halves based on the observation times of the infected hosts (note that one of the groups will have one more host than the other if the total number of observed hosts is odd), and the following variables are introduced: $\mathcal{M}$ is the number of *marked* virus lineages during the first period of time; it is equal to the number of hosts in the first half of the data set. $\mathcal{C}$ is the number of *captured* virus lineages during the second period of time; it is equal to the number of hosts in the second half of the data set. $\mathcal{R}$ is the number of *recaptured* virus lineages during the second period of time; it is equal to the number of hosts in the second half of the data set that were infected directly or indirectly by hosts in the first half of the data set. $\mathcal{R}$ is directly computed from the transmission tree $\tilde{J}$. The conditional probability $P(\mathcal{N} \mid \mathcal{M}, \mathcal{C}, \mathcal{R})$ of the population size $\mathcal{N}$ satisfies:

$$P(\mathcal{N} \mid \mathcal{M}, \mathcal{C}, \mathcal{R}) = \sum_{\mathcal{N}_2=0}^{\mathcal{N}} P(\mathcal{N}, \mathcal{N}_2 \mid \mathcal{M}, \mathcal{C}, \mathcal{R}),$$

where $\mathcal{N}_2$ is the size of the population in the second period of time. Using Bayes' theorem, the conditional joint probability of $(\mathcal{N}, \mathcal{N}_2)$ satisfies:

$$P(\mathcal{N}, \mathcal{N}_2 \mid \mathcal{M}, \mathcal{C}, \mathcal{R}) = \frac{P(\mathcal{R} \mid \mathcal{N}, \mathcal{N}_2, \mathcal{M}, \mathcal{C}) P(\mathcal{N}_2 \mid \mathcal{N}, \mathcal{M}, \mathcal{C}) P(\mathcal{N} \mid \mathcal{M}, \mathcal{C})}{P(\mathcal{R} \mid \mathcal{M}, \mathcal{C})}.$$

In what follows, the terms of the numerator of the previous fraction are provided. Figure S10 provides a graphical explanation of the variables considered here.

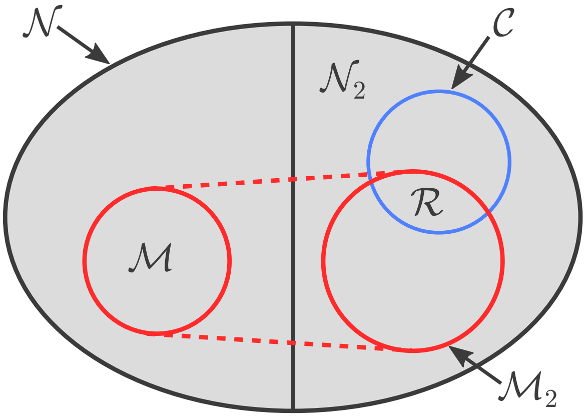


Figure S10. A graphical representation of the parameters considered in the mark-recapture-based calculation of the true number of cases affecting the study area. $\mathcal{N}$ is the total number of infected hosts, which we wish to infer, while $\mathcal{N}_2$ is the number of infected hosts in the second half of the sampling period. $\mathcal{M}$ is the number of hosts from the first half of the sampling period that are marked (i.e. all detected cases), which increases or decreases to $\mathcal{M}_2$, the number of hosts in the second half of the sampling period which are infected directly or indirectly by the $\mathcal{M}$ marked hosts. $\mathcal{C}$ is the number infected hosts captured (detected) in the second time period, while $\mathcal{R}$ is the number of marked hosts that are recaptured.

Expression of $P(\mathcal{R} \mid \mathcal{N}, \mathcal{N}_2, \mathcal{M}, \mathcal{C})$. Let $\mathcal{M}_2$ be the number of hosts (observed or not) corresponding to the second period of time and infected, directly or indirectly, by the $\mathcal{M}$ hosts marked during the first period of time ($\mathcal{M}_2$ is unknown). The conditional probability of $\mathcal{R}$ can be written as follows:

$$P(\mathcal{R} \mid \mathcal{N}, \mathcal{N}_2, \mathcal{M}, \mathcal{C}) = \sum_{\mathcal{M}_2=0}^{\mathcal{N}_2} P(\mathcal{R} \mid \mathcal{N}, \mathcal{N}_2, \mathcal{M}, \mathcal{M}_2, \mathcal{C}) P(\mathcal{M}_2 \mid \mathcal{N}, \mathcal{N}_2, \mathcal{M}, \mathcal{C}),$$

where, assuming that observed hosts are independently sampled,

$$\mathcal{R} \mid \mathcal{N}, \mathcal{N}_2, \mathcal{M}, \mathcal{M}_2, \mathcal{C} \sim \text{Binomial}\left(\mathcal{C}, \frac{\mathcal{M}_2}{\mathcal{N}_2}\right)$$
$$\mathcal{M}_2 \mid \mathcal{N}, \mathcal{N}_2, \mathcal{M}, \mathcal{C} \sim \text{Binomial}\left(\mathcal{N}_2, \frac{\mathcal{M}}{\mathcal{N} - \mathcal{N}_2}\right).$$

Expression of $P(\mathcal{N}_2 \mid \mathcal{N}, \mathcal{M}, \mathcal{C})$. Let z be the multiplication coefficient between the size of the population corresponding to the first period of time and the size of the population corresponding to the second period of time (z is unknown). By assuming that z does not depend on $(\mathcal{N}, \mathcal{M}, \mathcal{C})$, the conditional probability of $\mathcal{N}_2$ can be written as follows:

$$P(\mathcal{N}_2 \mid \mathcal{N}, \mathcal{M}, \mathcal{C}) = \int_z P(\mathcal{N}_2 \mid z, \mathcal{N}, \mathcal{M}, \mathcal{C}) f(z) dz, \tag{6}$$

where f is the probability density function of z that is assumed to be the uniform density over the interval $[0.5, 1.5]$ and

$$P(\mathcal{N}_2 \mid z, \mathcal{N}, \mathcal{M}, \mathcal{C}) = P(\mathcal{N}_2 \mid z, \mathcal{N}, \mathcal{N}_2 < \mathcal{N} - \mathcal{M}, \mathcal{N}_2 \geq \mathcal{C})$$
$$\mathcal{N}_2 \mid z, \mathcal{N} \sim \text{Binomial}\left(\mathcal{N}, \frac{z}{1+z}\right).$$

Therefore, in equation (6),

$$z \sim \text{Uniform}([0.5; 1.5])$$
$$\mathcal{N}_2 \mid z, \mathcal{N}, \mathcal{M}, \mathcal{C} \sim \text{Truncated Binomial}\left(\mathcal{N}, \frac{z}{1+z}, \mathcal{C}, \mathcal{N} - \mathcal{M} - 1\right),$$

where the two latter arguments in the truncated binomial distribution are the lower and upper truncation bounds.

Expression of $P(\mathcal{N} \mid \mathcal{M}, \mathcal{C})$. $\mathcal{N}$ is assumed to be uniform over the set of values $\{0, \ldots, \mathcal{N}^{max}\}$ (we used $\mathcal{N}^{max} = 10^3$) and, consequently,

$$\mathcal{N} \mid \mathcal{M}, \mathcal{C} \sim \text{Uniform}(\{\mathcal{M} + \mathcal{C}, \ldots, \mathcal{N}^{max}\}).$$

Importance sampling. Therefore, the conditional distribution of $\mathcal{N}$ given observed variables $(\mathcal{M}, \mathcal{C}, \mathcal{R})$ satisfies:

$$P(\mathcal{N} \mid \mathcal{M}, \mathcal{C}, \mathcal{R}) = \frac{1}{P(\mathcal{R} \mid \mathcal{M}, \mathcal{C})} \sum_{\mathcal{N}_2=0}^{\mathcal{N}} \sum_{\mathcal{M}_2=0}^{\mathcal{N}_2} \int_z \Big(P(\mathcal{R} \mid \mathcal{N}, \mathcal{N}_2, \mathcal{M}, \mathcal{M}_2, \mathcal{C}) P(\mathcal{M}_2 \mid \mathcal{N}, \mathcal{N}_2, \mathcal{M}, \mathcal{C}) \\ \times P(\mathcal{N}_2 \mid z, \mathcal{N}, \mathcal{M}, \mathcal{C}) f(z) P(\mathcal{N} \mid \mathcal{M}, \mathcal{C}) \Big) dz.$$

where all the terms in the multiple sum have explicit expressions. The multiple sum can be assessed with importance sampling [11] by simulating z, $\mathcal{N}_2$ and $\mathcal{M}_2$ under their conditional distributions (we used 10^3 particles in the importance sampling). Then, the domain of $\mathcal{N}$ being finite, a simple average is performed to obtain $P(\mathcal{N} \mid \mathcal{M}, \mathcal{C}, \mathcal{R})$ for all integer values of $\mathcal{N}$ between 0 and $\mathcal{N}^{max}$.

Final estimate of $P(\mathcal{N} \mid \mathcal{M}, \mathcal{C}, \mathcal{R})$. The previous paragraph shows how to compute the conditional distribution of $\mathcal{N}$ given the variables $(\mathcal{M}, \mathcal{C}, \mathcal{R})$. In our study, the number $\mathcal{R}$ of recaptured lineages is merely inferred and we only have access to its posterior distribution (for each chain state $b \in \{1, \dots, 10^4\}$, we can determine the number $\mathcal{R}^{(b)}$ of hosts in the second half of the data set that were infected directly or indirectly by hosts in the first half of the data set). Consequently, we integrated the conditional distribution $P(\mathcal{N} \mid \mathcal{M}, \mathcal{C}, \mathcal{R})$ of the population size with respect to the posterior sample of $\mathcal{R}$ provided by the inference algorithm described in Section 3.3:

$$\hat{P}(\mathcal{N} \mid \mathcal{M}, \mathcal{C}, \text{data}) = \frac{1}{10^4} \sum_{b=1}^{10^4} P(\mathcal{N} \mid \mathcal{M}, \mathcal{C}, \mathcal{R}^{(b)}),$$

where $P(\mathcal{N} \mid \mathcal{M}, \mathcal{C}, \mathcal{R}^{(b)})$ is estimated with the importance sampling procedure cited above.

5. Estimation of R_e

5.1 Method

Due to incomplete sampling as well as limited sampling in both time and space, calculating the effective number of secondary cases per infection, R_e, is subject to various sources of bias. We cannot detect transmission to secondary cases occurring after the sampling period (bias 1), occurring outside the sampled region (bias 2) or occurring within the sampling period and within the sampling region but remaining unobserved (bias 3).

Bias 1 was taken into account by restricting the calculation of R_e to the first half of observed infected hosts whose secondary cases are likely to be observed in the sampling period. Bias 2 was taken into account by restricting the calculation of R_e to the hosts distant from the western and northern borders of the study region whose secondary cases are likely to be observed in the sampling region (only source hosts observed at longitudes above 30.5 degrees and latitudes below -28.0 degrees were considered). This temporally and spatially restricted set of hosts is denoted by $\mathcal{I}_{inner} \subset \{1, \dots, I\}$ and the number of hosts in this set is denoted by $|\mathcal{I}_{inner}|$. Using these two restrictions for source hosts, the following posterior sample based on inferred direct transmissions gives an initial assessment for R_e:

$$\left\{ R^{(b)}_{e,\,direct} = \frac{1}{|\mathcal{I}_{inner}|} \sum_{j \in \mathcal{I}_{inner}} \sum_{i=1}^{I} \mathbf{1}(J^{(b)}(i) = j) : b = 1, \dots, 10^4 \right\}, \tag{7}$$

where $\sum_{i=1}^{I} \mathbf{1}(J^{(b)}(i) = j)$ is the number of direct secondary cases for j at the b-th saved iteration of the MCMC algorithm. This posterior sample, illustrated in figure S11 (*a*), can be interpreted as a lower bound of R_e.

Transmission to unsampled cases (bias 3) was taken into account by multiplying $R^{(b)}_{e,\,direct}$ by a correcting value that was obtained as follows. R_e is the sum of $R^{(b)}_{e,\,direct}$ plus a term composed of unobserved secondary cases, say $R^{(b)}_{e,\,unobserved}$: $R_e = R^{(b)}_{e,\,direct} + R^{(b)}_{e,\,unobserved}$. The ratio $R^{(b)}_{e,direct}/R_e$ is equal to the probability of "observing a secondary infected host" that is approximately equal to the ratio between the total number of sampled cases I and the expectation of the population size in the study region and period, i.e. $\sum_{\mathcal{N}=0}^{\infty} \mathcal{N} P(\mathcal{N} \mid \mathcal{M}, \mathcal{C}, \mathcal{R}^{(b)})$, where $P(\mathcal{N} \mid \mathcal{M}, \mathcal{C}, \mathcal{R}^{(b)})$ depends on the reconstruction specification and the iteration b of the MCMC algorithm (see section 4 of this supplementary text):

$$\frac{R^{(b)}_{e,\,direct}}{R_e} = Pr(\text{observing a secondary infected host}) \approx \frac{I}{\sum_{\mathcal{N}=0}^{\infty} \mathcal{N} P(\mathcal{N} \mid \mathcal{M}, \mathcal{C}, \mathcal{R}^{(b)})}.$$

Thus, transmission to unsampled cases (bias 3) was taken into account by multiplying $R^{(b)}_{e,\,direct}$ by the correcting value $(1/I) \sum_{\mathcal{N}=0}^{\infty} \mathcal{N} P(\mathcal{N} \mid \mathcal{M}, \mathcal{C}, \mathcal{R}^{(b)})$, and the following posterior sample gives two further assessments for R_e (one for each specification of the postprocessing algorithm):

$$\left\{ R^{(b)}_{e,\,indirect} = \frac{R^{(b)}_{e,\,direct}}{I} \sum_{\mathcal{N}=0}^{\infty} \mathcal{N} P(\mathcal{N} \mid \mathcal{M}, \mathcal{C}, \mathcal{R}^{(b)}) : b = 1, \dots, 10^4 \right\}, \tag{8}$$

Figure S11 shows these posterior samples as obtained with the conservative (*b*) and liberal (*c*) specifications of the reconstruction of indirect transmissions.

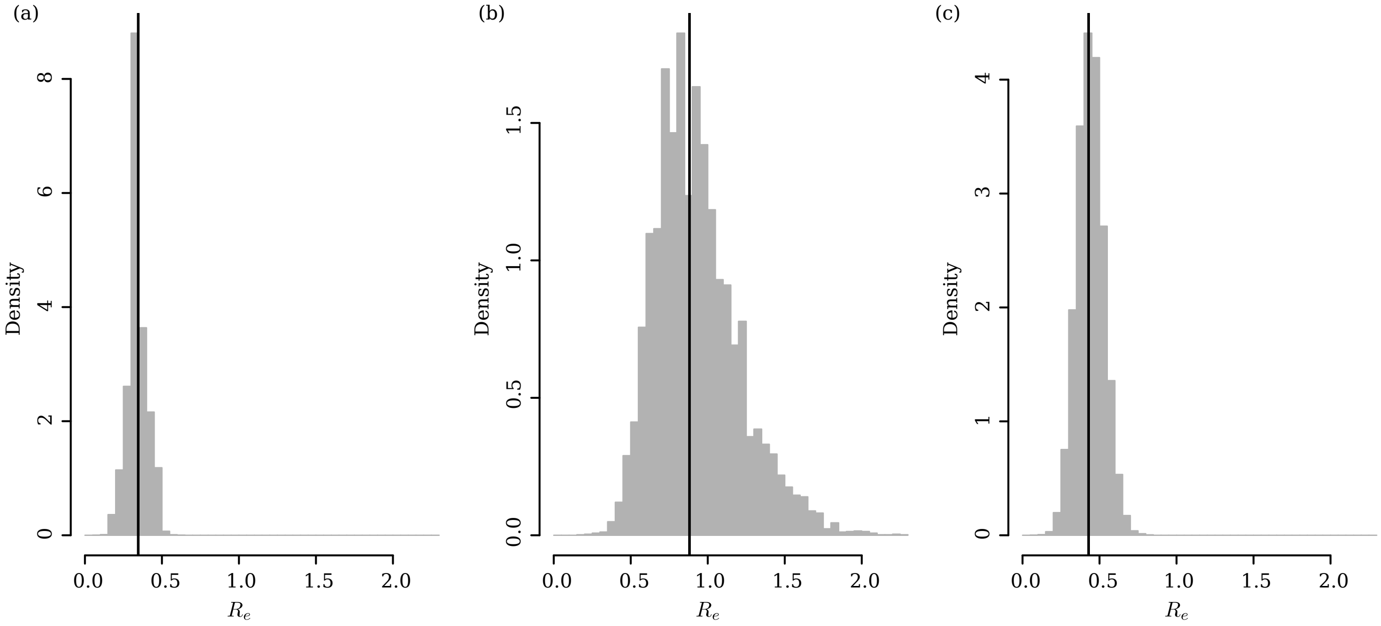


Figure S11. Posterior distributions of R_e, the effective number of secondary cases per infection. Considering only direct connections gives the lower limit of R_e (*a*). Estimates of R_e when including the probability of missing secondary cases are shown in (*b*) and (*c*) for the conservative and liberal specifications respectively. Vertical black lines indicate the mean in each plot.

5.2 Results

When considering only direct connections, the approach described above gave a median estimate of 0.35 (95%-PI: 0.23-0.47) (this value can be seen as a lower limit of R_e), and 0.88 (95%-PI: 0.50-1.58) and 0.43 (95%-PI: 0.27-0.62) when also accounting for the risk of missing cases calculated by inclusion of indirect connections, as determined under the conservative and liberal post-processing algorithms respectively, to the mark recapture approach (figure S11). Given the continued transmission of rabies virus in the study area throughout and after the study period, the conservative algorithm provides more plausible results for this dataset, with values above 1 in the 95% posterior interval. This result is consistent with estimates of R_0 for dog rabies outbreaks around the world, which are generally only barely above the threshold of one required for epidemic expansion [8]. The fact that our estimate does not rule out values below one could be due to: (i) the fact that the R_0 of dog rabies epidemics is generally close to one and our estimate is imprecise, (ii) the fact that in KZN, R_e will be influenced by active control programmes, and it may be that Re has dropped below one at a provincial scale, or (iii) the fact that the sources of bias mentioned above are not sufficiently compensated for in our approach.

6. Simulations

6.1 Method

We tested the accuracy of the method using 100 simulated datasets from each of six scenarios. Scenarios 1 to 4 were used to investigate overall accuracy and the effect of sampling rate on the reconstruction method, representing high (3/4 of all cases), moderate (2/3 of all cases), intermediate (1/2 of all cases) and low (1/4 of all cases) detection rates respectively. Scenarios 5 and 6 were used to test the sensitivity of the method to small and large misspecifications of epidemiological parameters.

The simulation model is based on the probability distributions and specifications given in section 2.1 of this supplementary text. The single difference between the simulation model and the model used for inference concerns infection from external sources. In the inference model, there is a single external source with an infection strength that is constant in time and space. In the simulation model, the external sources are handled in a different manner (and more realistic manner for our case study): the external sources are infectious hosts within or outside the sampling region generating infectious risks that are local in time and space.

For each simulation of scenario 1, the epidemic is initiated at time zero with one infected host localized at the origin $(0, 0)$ and 119 susceptible hosts uniformly and randomly localized in the $[0.0, 0.3] \times [0.0, 0.1]$ rectangle. At time zero, the genetic sequence of the virus in the infected host consists purely of adenine bases. The incubation duration parameters are $\beta_1 = 50$ (mean) and $\beta_2 = 5$ (standard deviation). The infectious duration parameters are $\delta_1 = 10$ (mean) and $\delta_2 = 1$ (standard deviation). The infection strength of each infectious host is 10^6. The dispersal parameters are $\alpha_{2,1} = 0.25$ (scale) and $\alpha_{2,2} = 1$ (shape). The substitution rates are $\mu_1 = \mu_2 = \mu_3 = 2 \times 10^{-5}$. The observed hosts form a sample of the 120 hosts: the sampling is uniform among the hosts with x-coordinate larger than 0.2 and the sampling effort is 3/4 (3/4 of cases with x-coordinate larger than 0.2 are observed). For genetic data, a sequence fragment of length 800 is observed.

Figure S12 presents an example of a simulation under scenario 1. For this scenario and those below, the prior distributions were the same as for the inference in the real case study, but we changed the values of the parameters: (i) informative priors for incubation and infectious durations parameterized by $(b_1^1, b_1^2, b_2^1, b_2^2) = (50, 0.01, 5, 0.01)$ and $(d_1^1, d_1^2, d_2^1, d_2^2) = (10, 0.01, 1, 0.01)$; (ii) vague exponential priors with mean 10^6 for strengths of the external source and the observed sources; (iii) exponential prior with mean value 1.0 for the scale dispersal parameter and prior parameters $a_{2,1}^{(1)} = 1$ and $a_{2,2}^{(2)} = 1$ for the shape dispersal parameter; (iv) exponential prior distributions with mean $m = 2 \times 10^{-5}$ for the substitution rates; (v) normal prior with mean $t_{MRCA}^{(1)} = 0$ and standard deviation $t_{MRCA}^{(2)} = 10$ for T_{MRCA}.

In scenario 2 (moderate detection rate), the sampling effort is 2/3 instead of 3/4. To ensure that the same *number* of observed hosts were sampled, the spatial domain was extended to the $[0.0, 0.3\sqrt{135/120}] \times [0.0, 0.1\sqrt{135/120}]$ rectangle and the total number of hosts in this rectangle was increased to 135 instead of 120.

In scenario 3 (intermediate detection rate), the sampling effort is 1/2. To ensure that the same number of observed hosts were sampled, the spatial domain was extended to the $[0.0, 0.3\sqrt{1.5}] \times [0.0, 0.1\sqrt{1.5}]$ rectangle and the total number of hosts in this rectangle was increased to 180.

In scenario 4 (low detection rate), the sampling effort is 1/4. To ensure that the same number of observed hosts were sampled, the spatial domain was extended to the $[0.0, 0.3\sqrt{3}] \times [0.0, 0.1\sqrt{3}]$ rectangle and the total number of hosts in this rectangle was increased to 360.

In scenario 5 (small misspecification of the model), we changed the mean parameter values of the incubation and infectious durations used for the simulation: $\beta_1 = 45$ instead of 50 and $\delta_1 = 12$ instead of 10, but we left the strongly informative priors of the reconstruction algorithm unchanged, with means of 50 and 10 for the incubation and infectious durations, respectively.

In scenario 6 (large misspecification of the model), we changed the mean parameter values of the incubation and infectious durations used for the simulation: $\beta_1 = 30$ instead of 50 and $\delta_1 = 15$ instead of 10, but again left the strongly informative priors of the reconstruction algorithm unchanged.

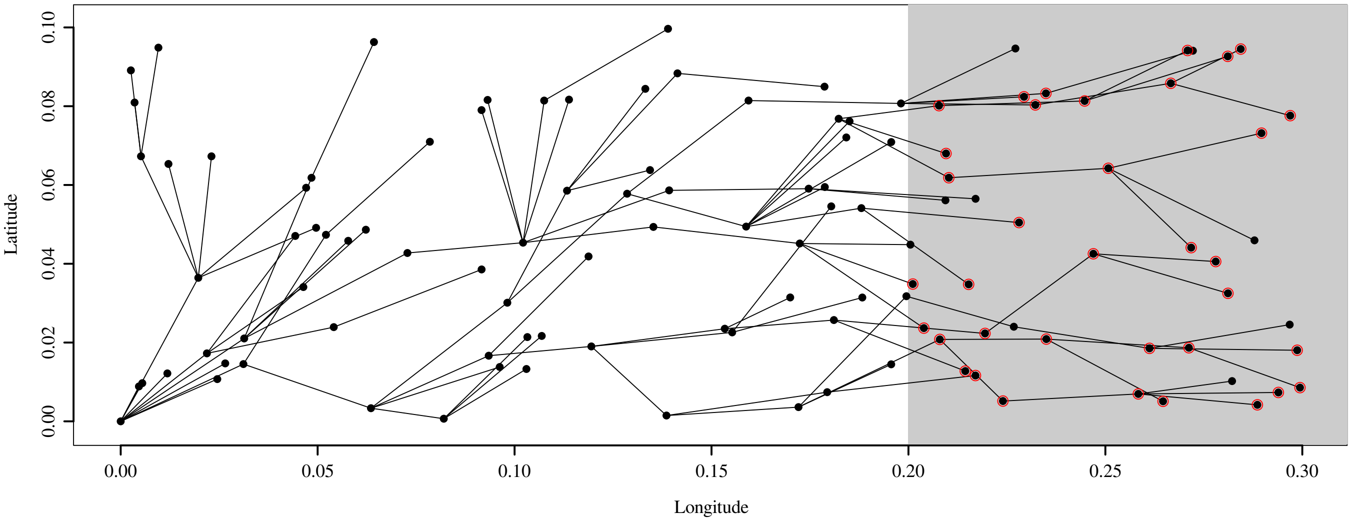


Figure S12. Example of a simulated epidemic under scenario 1. Black dots represent hosts, while black segments represent transmissions. Samples are take from only a small region, indicated in grey, and in this area, not all cases are detected or sampled. In simulation scenario 1, cases in the sampling area have a probability of 3/4 of being sampled. Sampled hosts are indicated in red.

6.2 Results

Substitution parameters and the date of the MRCA were appropriately estimated whatever the simulation scenario (rates of coverage by the 95%-posterior intervals are high – see table S2, while the lengths of these intervals are moderate – data not shown). The dispersal parameters were less accurately estimated (see rates of coverage of $\alpha_{2,1}$ and $\alpha_{2,2}$ by their 95%-posterior intervals in table S2). In these simulations, inaccuracies are observed when the priors for incubation and infectiousness parameters are biased (scenarios 5 and 6) and, surprisingly, when the sampling rate is high (scenarios 1 and 2). The latter case can be explained by the reduced spatial extent of the sampled domain. In all simulation scenarios, both the number of sampled individuals and the spatial density of all cases (sampled and unsampled) were kept constant, which required an increase in the sampled spatial domain. Thus, at lower sampling rates, the range of dispersal distances that can be observed is larger and, consequently, the estimation of dispersal parameters is improved even if the number of direct connections is reduced. Other simulation results are discussed in the main text.

Table S2. Performance of the transmission tree reconstruction approach for the estimation of key parameters. For each parameter, the mean (and standard deviation) is reported based on 100 simulations for each simulation type.

Parameter	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6
$\alpha_{2,1}$	0.61 (0.49)	0.66 (0.48)	0.71 (0.45)	0.85 (0.36)	0.59 (0.49)	0.64 (0.48)
$\alpha_{2,2}$	0.66 (0.47)	0.68 (0.47)	0.78 (0.42)	0.91 (0.29)	0.62 (0.49)	0.91 (0.28)
μ_1	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
μ_2	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
μ_3	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
T_{MRCA}	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)

Table S3. Performance of the approach for the estimation of various transmission events. In each case the mean (and standard deviation) of the posterior probabilities of true transmission events is reported based on 100 simulations for each simulation type.

Transmission type	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6
Direct connection between two observed cases	0.73 (0.12)	0.73 (0.10)	0.78 (0.13)	0.82 (0.12)	0.60 (0.15)	0.03 (0.03)
Infection of an observed case by an unobserved[1] case	0.64 (0.19)	0.62 (0.17)	0.54 (0.15)	0.57 (0.15)	0.73 (0.15)	0.76 (0.17)
Infection of an observed case by an exogenous[2] case (conservative specification)	0.66 (0.22)	0.58 (0.20)	0.45 (0.18)	0.33 (0.14)	0.72 (0.18)	0.85 (0.17)
Infection of an observed case by an exogenous[2] case (liberal specification)	0.65 (0.22)	0.57 (0.20)	0.44 (0.19)	0.29 (0.12)	0.72 (0.18)	0.85 (0.15)
Direct or indirect infection of an observed case by another observed case (conservative specification)	0.63 (0.15)	0.62 (0.13)	0.57 (0.17)	0.40 (0.18)	0.54 (0.15)	0.08 (0.07)
Direct or indirect infection of an observed case by another observed case (liberal specification)	0.64 (0.14)	0.63 (0.12)	0.59 (0.15)	0.42 (0.17)	0.53 (0.14)	0.07 (0.07)

[1] An unobserved case can be inside or outside the study region

[2] An exogenous case is an unobserved case *outside* the study region

Table S4. Performance of the approach for the estimation of the total number $\mathcal{N}$ of infected cases in the sampling region. For each parameter, the mean (and standard deviation) is reported based on 100 simulations for each simulation type.

Parameter	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6
True value of $\mathcal{N}$	39.5 (5.3)	44.2 (6.0)	59.4 (7.0)	120.4 (8.5)	38.9 (4.8)	39.9 (5.1)
Rate of coverage of the true value by the 95%-posterior interval (conservative specification)	1.00 (0.00)	0.98 (0.14)	0.15 (0.36)	0.16 (0.37)	1.00 (0.00)	0.97 (0.18)
Length of the 95% posterior interval (conservative specification)	31.9 (83.0)	30.0 (63.3)	41.3 (107.1)	89.5 (160.3)	57.9 (165.3)	148.2 (228.7)
Rate of coverage of the true value by the 95%-posterior interval (liberal specification)	1.00 (0.00)	0.97 (0.17)	0.05 (0.21)	0.00 (0.00)	1.00 (0.00)	0.98 (0.15)
Length of the 95% posterior interval (liberal specification)	22.2 (4.6)	21.5 (1.2)	21.5 (1.2)	21.7 (1.4)	22.5 (5.8)	56.3 (85.4)

7. Posterior distributions

Figure S13 to S19 show the posterior distributions of model parameters and other relevant quantities.

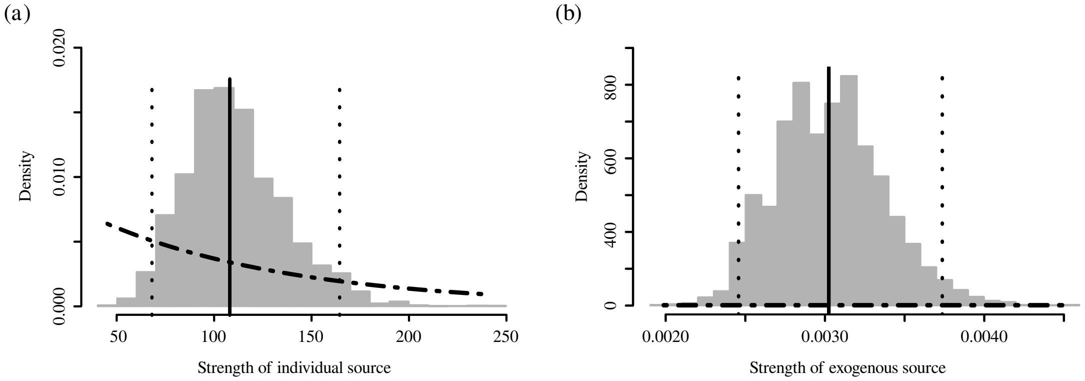


Figure S13. Posterior distributions of α_1, the strength of observed source individuals (*a*) and α_0, the strength of the exogenous source, which is assumed to be constant in time (*b*). Solid lines indicate posterior medians, dotted lines show the 95% posterior intervals, and dashed curves represent prior distributions.

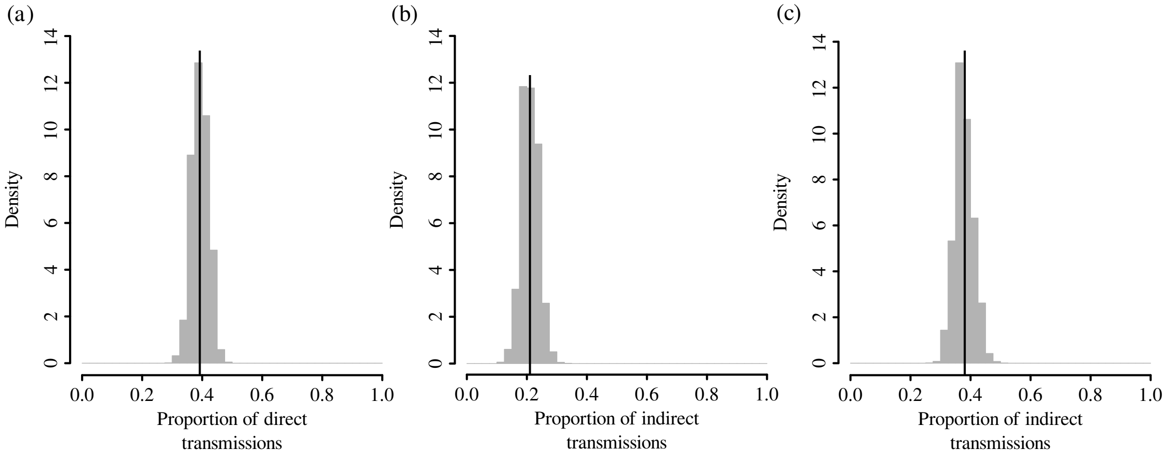


Figure S14. The proportion of transmissions inferred as direct (*a*) and indirect (*b* and *c*) in the posterior distribution of transmission trees. The remaining cases are connected to an exogenous source occurring either before the sampling period or outside the sampled region. The proportion of indirect transmissions inferred using the conservative specification is shown in (*b*), while (*c*) shows the proportion inferred under the liberal specification.

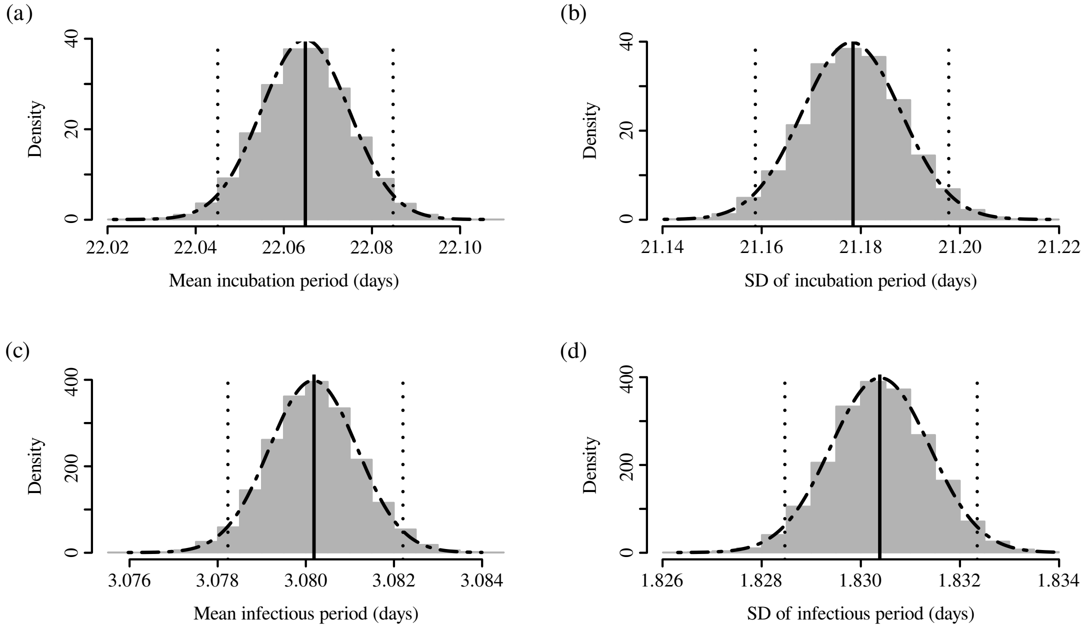


Figure S15. Posterior distributions of the mean and standard deviation of the incubation (*a* and *b*) and infectious (*c* and *d*) periods for all direct connections. The solid lines indicate posterior medians, while the dotted, vertical lines show the 95% posterior intervals. Dashed curves represent the prior distributions for each parameter.

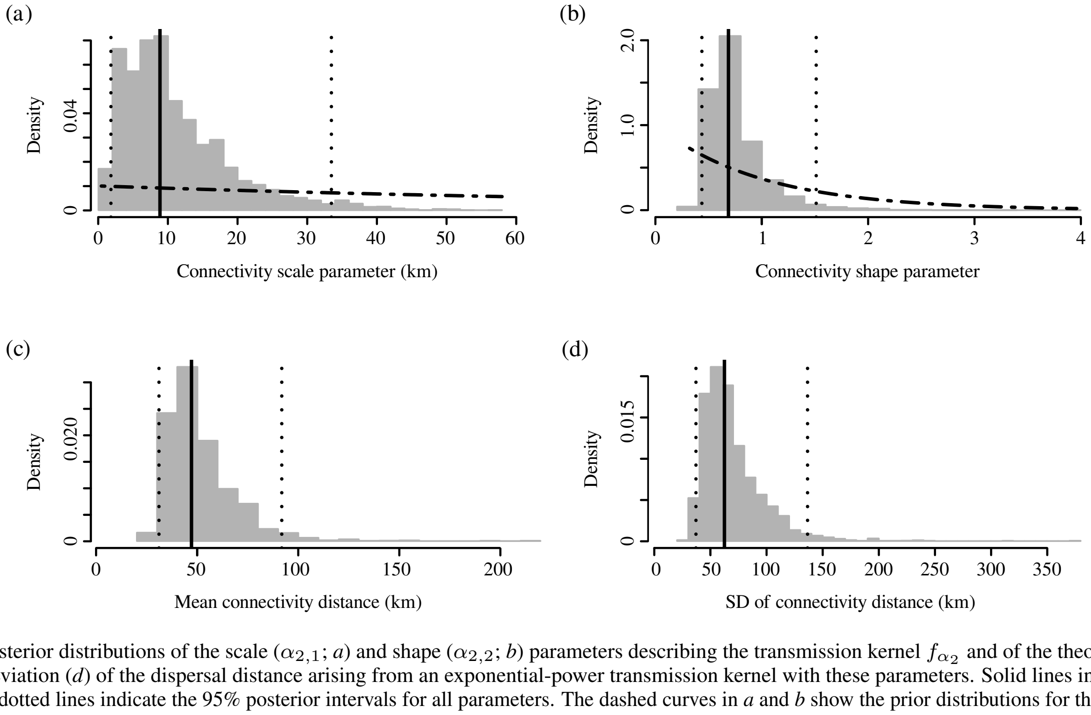


Figure S16. Posterior distributions of the scale ($\alpha_{2,1}$; *a*) and shape ($\alpha_{2,2}$; *b*) parameters describing the transmission kernel f_{α_2} and of the theoretical mean (*c*) and standard deviation (*d*) of the dispersal distance arising from an exponential-power transmission kernel with these parameters. Solid lines indicate posterior medians, while dotted lines indicate the 95% posterior intervals for all parameters. The dashed curves in *a* and *b* show the prior distributions for these parameters.

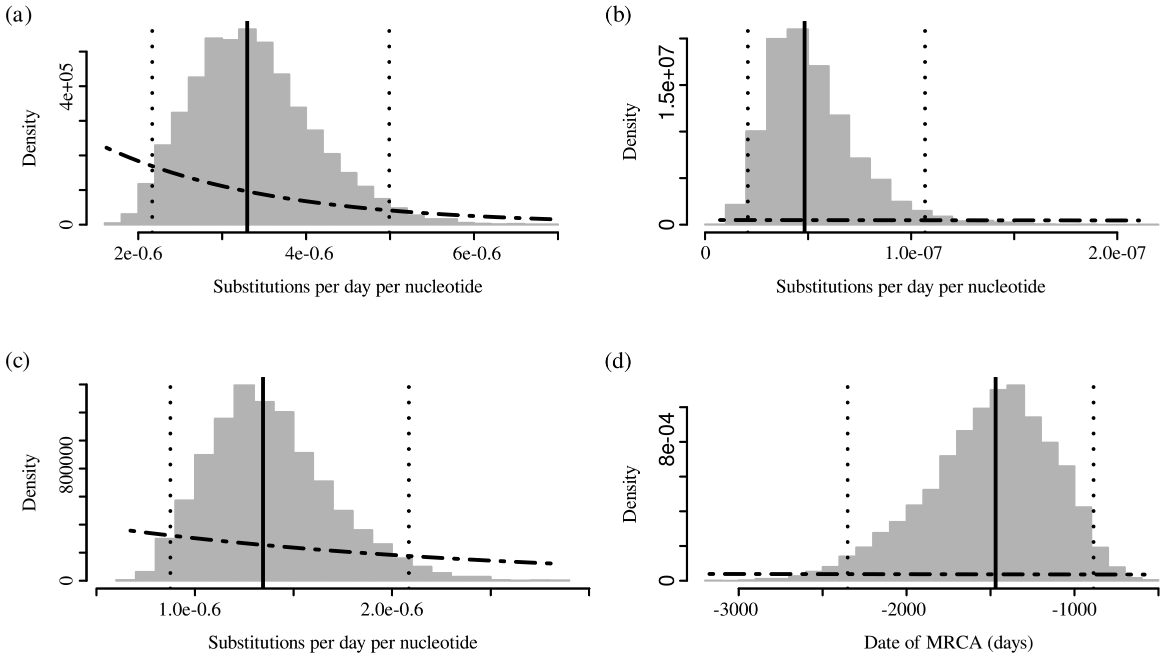


Figure S17. Posterior distributions of the rate of transitions (μ_1; *a*), rate of type 1 transversions (μ_2; *b*), rate of type 2 transversions (μ_3; *c*) and T_{MRCA} (*d*). Solid lines indicate posterior medians, while the dotted lines show the 95% posterior interval for each parameter. Dashed curves indicate the prior distributions of each parameter.

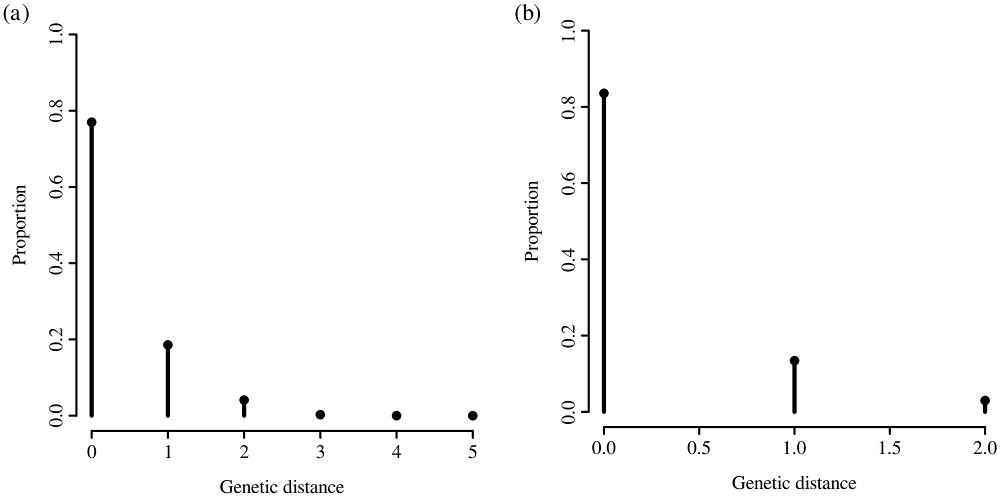


Figure S18. Posterior distributions of the number of genetic differences between directly connected cases. Genetic distances between all *a posteriori* directly connected cases are shown in (*a*), while (*b*) shows the genetic distances between connected cases corresponding to the direct transmission links with the highest posterior probabilities (i.e. only the most probable links).

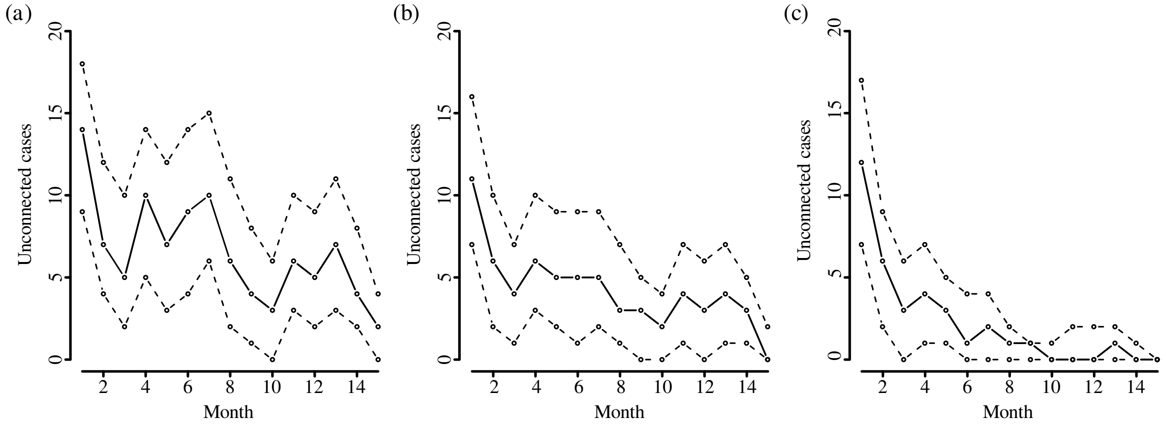


Figure S19. Number of cases with no detectable source in the dataset per month of the sampling period. (*a*): Number of cases infected by an exogenous source under the main inference algorithm allowing only direct or exogenous connections. (*b*): Cases still infected by an exogenous source after applying the post-processing algorithm using the conservative specification, and (*c*): after applying the post-processing algorithm using the liberal specification. Solid lines indicate the median of the posterior distributions, while dashed lines show the 95% posterior intervals. These figures show that the high numbers of cases inferred as stemming from exogenous transmissions (i.e. introductions) early in the data set are most likely the result of dogs already being infected by the time we started sampling. The number of cases with exogenous sources stabilises at a posterior median of 2.27 (95%-PI:0.6-5.1) and 0.2 (95%-PI: 0.0-1.3) per month for the conservative and liberal algorithms respectively.

8. Notation

Table S5. Summary of the notation used in this document

Symbol	Description
$J\{1:i-1\}$	The transmission tree ($J\{1:i-1\} = (J(1), ..., J(i-1))$, a vector of source-descendant pairs, where $J(i)$ is the source of J)
$\mathbf{T}^{obs}$	Vector of observation times
$\mathbf{T}^{inf}$	Vector of infection times
$\mathbf{L}$	Vector of incubation periods
$\mathbf{D}$	Vector of infectious periods
$\mathbf{X}$	Vector of observed host locations
C_i	Capacity of host i to spread disease ($C_i = 0$ for dead-end hosts; $C_i = 1$ for hosts able to transmit the disease)
I	Total number of hosts
f_{α_2}	The spatial transmission kernel
α_0	Infection strength of the exogenous source of infections
α_1	Infection strength of observed hosts
α_2	Set of parameters describing the shape of the spatial transmission kernel ($\alpha_2 = (\alpha_{2,1}, \alpha_{2,2})$)
$\alpha_{2,1}$	Scale parameter of the exponential-power transmission kernel
$\alpha_{2,1}$	Shape parameter of the exponential-power transmission kernel
$\lvert\lvert \cdot \rvert\rvert$	Geographic distance between two cases
β_1	Mean of the incubation period
β_2	Standard deviation of the incubation period
δ_1	Mean of the infectious period
δ_2	Standard deviation of the infectious period
μ	Parameter vector of nucleotide substitution rates ($\mu = (\mu_1, \mu_2, \mu_3)$)
μ_1	Rate of transitions per nucleotide per day
μ_2	Rate of type-1 transversions per nucleotide per day
μ_3	Rate of type-2 transversions per nucleotide per day
T_{MRCA}	Estimated date of detection of the most recent common ancestor of all cases in the dataset
S_{MRCA}	Nucleotide sequence of the most recent common ancestor of all cases in the dataset (determined *a priori*)
$M(S', S)$	Vector of the number of transitions, type-1 and type-2 transversions and unchanged bases between S' and S
θ	Vector of all unknown parameters

References

1. Tordo, N., Poch, O., Ermine, A., Keith, G. & Rougeon, F. 1988 Completion of the rabies virus genome sequence determination: highly conserved domains among the L (polymerase) proteins of unsegmented negative-strand RNA viruses. *Virology* **165**, 565–576.

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

3. Baele G, Lemey P, Bedford T , Rambaut A, Suchard MA, Alekseyenko AV (2012) Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol. Biol. Evol.* 29:2157–2167.

4. Huson DH, Scornavacca C (2012) Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.* 61:1061–1067.

5. Morelli MJ, Thébaud G, Chadœuf J, King DP, Haydon DT, Soubeyrand S (2012) A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Comput. Biol.* 8:e1002768.

6. Austerlitz F, Dick CW, Dutech C, Klein EK, Oddou-Muratorio S, Smouse, PE (2004) Using genetic markers to estimate the pollen dispersal curve. *Mol. Ecol.* 13:937–954.

7. Kimura M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA* 78:454–458.

8. Hampson K, Dushoff J, Cleaveland S, Haydon DT, Kaare M, Packer C, Dobson A (2009) Transmission dynamics and prospects for the elimination of canine rabies. *PLoS Biol.* 7:e1000053.

9. Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, Pupko T (2012) FastML: A web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res.* 40:W580–W584.

10. Rodriguez F, Oliver JL, Marin A, Medina JR (1990) The general stochastic model of nucleotide substitution. *J. Theor. Biol.* 142:485–501.

11. Parent E, Bernier J (2007) *Le raisonnement bayésien: modélisation et inférence* (Springer-Verlag, Paris).