Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak

Simon Dellicour[1,*], Guy Baele[1], Gytis Dudas[2], Nuno R. Faria[3], Oliver G. Pybus[3], Marc A. Suchard[4,5,6], Andrew Rambaut[7,8], and Philippe Lemey[1]

[1]Rega Institute for Medical Research, Clinical and Epidemiological Virology, Department of Microbiology and Immunology, KU Leuven–University of Leuven, Minderbroedersstaat 10, 3000 Leuven, Belgium

[2]Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, 98109, Seattle, WA, USA

[3]Department of Zoology, University of Oxford, Oxford OX1 3PS, United Kingdom

[4]Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles, CA, USA

[5]Department of Biomathematics David Geffen School of Medicine at UCLA, University of California, Los Angeles, CA, USA

[6]Department of Human Genetics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, CA, USA

[7]Institute of Evolutionary Biology, University of Edinburgh, King's Buildings, Edinburgh, EH9 3FL, UK

[8]Fogarty International Center, National Institutes of Health, Bethesda, MD, USA

∗Corresponding author (simon.dellicour@kuleuven.be)

The recent Ebola virus (EBOV) outbreak in West Africa witnessed considerable efforts to obtain viral genomic data as the epidemic was unfolding. Such data are critical for the investigation of viral molecular epidemiology and can complement contact tracing by public health agencies. Analysing the accumulated EBOV genetic data can also deliver important insights into epidemic dynamics, as demonstrated by a recent viral genome study that revealed a metapopulation pattern of spread. Although metapopulation dynamics were critical for connecting rural and urban areas during the epidemic, the implications for specific intervention scenarios remain unclear. Here, we address this question using a collection of phylodynamic approaches. We show that long-distance dispersal events (between administrative areas >250 km apart) were not crucial for epidemic expansion and that preventing viral lineage movement to single locations would, in most cases, have had little impact. In addition, urban areas – specifically those encompassing the three capital cities and their suburbs – represented major 'transit centers' for transmission chains, but preventing viral lineage movement to all three simultaneously would have only contained epidemic size to about one third. Using continuous phylogeographic reconstructions we estimate a distance kernel for EBOV spread and reveal considerable heterogeneity in dispersal velocity through time. We also show that announcements of border closures were followed by a significant but transient effect on international virus dispersal. Our study illustrates how phylodynamic analyses can answer specific epidemiological and epidemic control questions and can be used to quantify the hypothetical impact of different intervention strategies as well as the impact of barriers on dispersal frequency.

Keywords: Ebola virus, West Africa, phylogeography, spatial epidemiology, intervention strategies, borders impact.

The recent Ebola virus (EBOV) epidemic in West Africa emerged around end of 2013 in the prefecture of Guéckédou in Guinea[1] and caused at least 11,310 deaths among 28,616 recorded cases in Guinea, Sierra Leone and Liberia[2]. It has been argued that the West African EBOV epidemic illustrates problems in the early detection of, and rapid response to, infectious disease outbreaks of public health importance[3]. In general, inadequate surveillance and/or reporting systems can cause delayed detection and underreporting. In addition, misdiagnosis and a lack of real-time data sharing impedes understanding of the magnitude of an epidemic[3]. The use of rapid point-of-care (POC) diagnostic tests to improve the speed of diagnosis can improve outbreak control, and the promise of effective assistance can help to overcome reluctance to report disease events[3].

Various reasons may explain the slow initial response to the West African EBOV epidemic, including poor public health infrastructure and local unfamiliarity with Ebola virus disease (EVD), as well as a lack of preparedness by the international community. This highlights the need for effective systems to detect and stop infectious disease threats and for rapidly expanding international response capacity when confronted with such threats[4]. Improving infection prevention and control in health care settings is critical in these efforts.

Viral genome sequencing is increasingly being used to assist with the identification of unknown sources and transmission chains, as viral genomes contain valuable information that complements contact tracing efforts. In the case of Ebola, Arias *et al.*[5] demonstrated that rapid outbreak sequencing in locally-established sequencing facilities can identify transmission chains linked to sporadic cases. Consequently, it is unsurprising that there have been many calls for making sequence data openly available in outbreak situations[3,6–8]. In addition to identifying specific transmission pathways, viral genome analyses can also shed light on the origins, evolution and transmission dynamics of a virus during an epidemic[9]. Early in the EBOV epidemic, analyses such as those by Gire and colleagues[10] demonstrated that the virus entered the human population in late 2013 and crossed from Guinea to Sierra Leone in May 2014 through sustained human-to-human transmission. The EBOV genome data that was generated also stimulated phylodynamic efforts to characterise transmission dynamics early in the epidemic (e.g. superspreading[11]) and to estimate critical epidemiological parameters such as the basic reproductive number[12]. Various molecular epidemiological studies subsequently attempted to trace Ebola spread[5,13–18] (see Holmes *et al.*[9] for a detailed overview), marking the beginning of large-scale real-time molecular epidemiology[9]. All these efforts culminated in an impressive collection of over 1,600 EBOV genome sequences corresponding to more than 5% of known cases[19]. These data represent a unique opportunity to learn important lessons about the evolutionary and epidemiological dynamics of an Ebola outbreak. Here, we aim to contribute to this progress by investigating how viral genome analyses can assess the impact that specific intervention strategies made, had or could have made if implemented.

Although Ebola viral genomes were reported across numerous studies, the collected genetic data cover the epidemic exceptionally well, and sampling intensity correlates well with the infection burden in different locations throughout the course of the outbreak[19]. This data set motivated a detailed phylogeographic study that identified the patterns and drivers of spatial spread[19]. Specifically, a generalised linear model (GLM) of transitions between discrete locations in a Bayesian statistical framework was used to test which causal factors might have

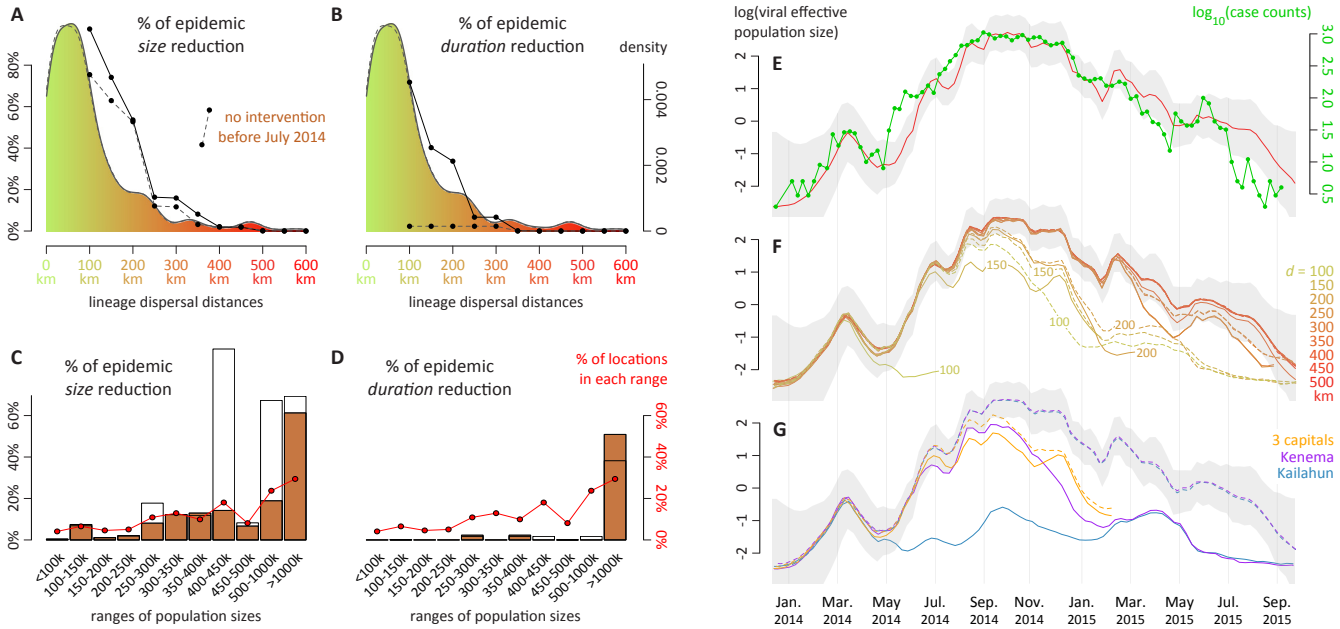

Figure 1. Hypothetical impact of intervention strategies. (**A**) Impact of halting transmission following long-distance dispersal events on the *epidemic size*. The plot depicts the epidemic size reduction when removing transmission following dispersal events over various distances, i.e. dispersal events between administrative areas separated by a least 100, 150, 200, 250, 300, 350, 400, 450 and 500 km. The effect on epidemic size is quantified by comparing (in %) the resulting tree length to the original transmission tree. In addition, we also report corresponding epidemic size reductions when transmission is prevented only after June 2014 (dashed line). These epidemic size reduction curves are superimposed on the distribution of lineage dispersal distances within the transmission tree (as inferred by discrete phylogeographic analysis of the outbreak in West Africa). (**B**) Impact of removing transmission following long-distance dispersal events on the *epidemic duration*. This plot corresponds to Figure 1A, but focuses on the impact on epidemic duration. In this case, the effect on epidemic duration is quantified by comparing (in %) the resulting tree height to the original transmission tree. (**C**) Impact of preventing dispersal events to specific administrative areas on the *epidemic size*. The plot reports the epidemic size reduction when removing transmission following dispersal events in administrative areas belonging to one of the following population sizes ranges: <100k, 100-150k, 150-200k, 200-250k, 250-300k, 300-350k, 350-400k, 400-450k, 450-500k, 500-1000k and >1000k people. As for Figure 1A, the effect on epidemic size is quantified by comparing (in %) the resulting tree length to the original transmission tree, and we also report corresponding epidemic size reductions when transmission is prevented only after June 2014 (dashed line). (**D**) Impact of removing transmission following dispersal events in specific administrative areas on the *epidemic duration*. This plot corresponds to 1C but focuses on the impact on epidemic duration. As for Figure 1B, the effect on epidemic duration is quantified by comparing (in %) the resulting tree height to the original transmission tree. (**E**) Estimation of viral population size (in red; 95% HPD in grey) and the evolution of the case counts (in green). (**F**) Impact of removing transmission following long-distance dispersal events on the viral effective population size through time. As in Figure 1E, 95% HPD of viral population size based on the entire dataset (no intervention strategy) is displayed in grey. On this graph, dashed lines correspond to viral population size evolution when transmission is prevented only after June 2014 (dashed line). (**G**) This plot corresponds to Figure 1F but focuses on the impact of removing transmission following dispersal events to specific locations on the viral effective population size through time.

influenced the spread of the virus at sub-national administrative levels (termed districts in Sierra Leone, prefectures in Guinea, and counties in Liberia). By considering a range of geographic, administrative, economic, climatic, infrastructural and demographic predictors, this GLM approach identified a gravity model of transmission, albeit one that was attenuated by international borders[19]. The gravity model emphasises the impact of population size on viral dispersal and implies that large urban populations acted as sources, re-seeding smaller limited epidemics in more outlying locations. Further, the epidemic was generally less likely to spread across international borders, but did so specifically both early and late in the epidemic, between locations that share such an international border. More detailed temporal analyses suggested that border attenuation may have resulted from border closures between Guinea, Sierra Leone and Liberia, although their containment effects were limited. Within the three affected countries, viral spread was not always maintained by continuous transmission in each location, but often by repeated introductions into a location, generating small, well-connected, clusters of cases. This dynamical pattern of connectivity characterises a metapopulation, highlighting the need for responsive, mobile and measured interventions.

Here, we extend the phylogeographic analyses of the West African Ebola epidemic in two different ways. First, we examine the implications of EBOV metapopulation dynamics on particular intervention strategies. Specifically, we assess to what extent limiting long-distance spread, or preventing spread to highly populated locations, might have impacted the epidemic. Second, we introduce continuous diffusion models as an alternative phylogeographic framework, which can characterise aspects of the process of Ebola spread that were not captured by the discrete approach employed by Dudas *et al*[19]. We quantify important parameters of spatial spread and demonstrate how a posterior predictive simulation procedure can be used to evaluate potential barriers to transmission, specifically, the impact of border closures. These new evolutionary approaches deepen our understanding of the public health implications of EBOV epidemic dynamics and the extent to which viral spread could be curbed by particular intervention strategies.

RESULTS

To understand the implications of EBOV metapopulation dynamics during the 2013-2016 epidemic, we first investigate the impact of hypothetical intervention strategies on epidemic size and duration. To model the effects of reducing long-distance dispersal, we prune the EBOV transmission trees when a long-distance lineage translocation between administrative areas was inferred, by effectively removing transmission following such dispersal events. We then quantify the effect of this intervention on the size (tree length) and duration (tree height) of the transmission trees (Fig. 1 A). This procedure therefore assesses how critical such long-distance events were to the expansion and maintenance of EBOV transmission. For comparison, we also perform this analysis by preventing transmission only after a particular point in time, i.e. by only removing viral lineage movements that occurred after June 2014. The latter analysis reflects a scenario in which hypothetical intervention strategies are implemented only some time after the onset of the outbreak, in this case after 6 months, at which time all three countries had already been seeded. This also avoids the effect of pruning early transmission bottlenecks that are specific to this epidemic, which could affect the generality of our conclusions. Figures 1A and 1B depict the impact of preventing long-distance viral lineage movements on relative epidemic size and duration, respectively, under both the immediate and delayed intervention strategies. Transitions between administrative areas involving distances greater than 300 km are rare and the transmission chains they generate do

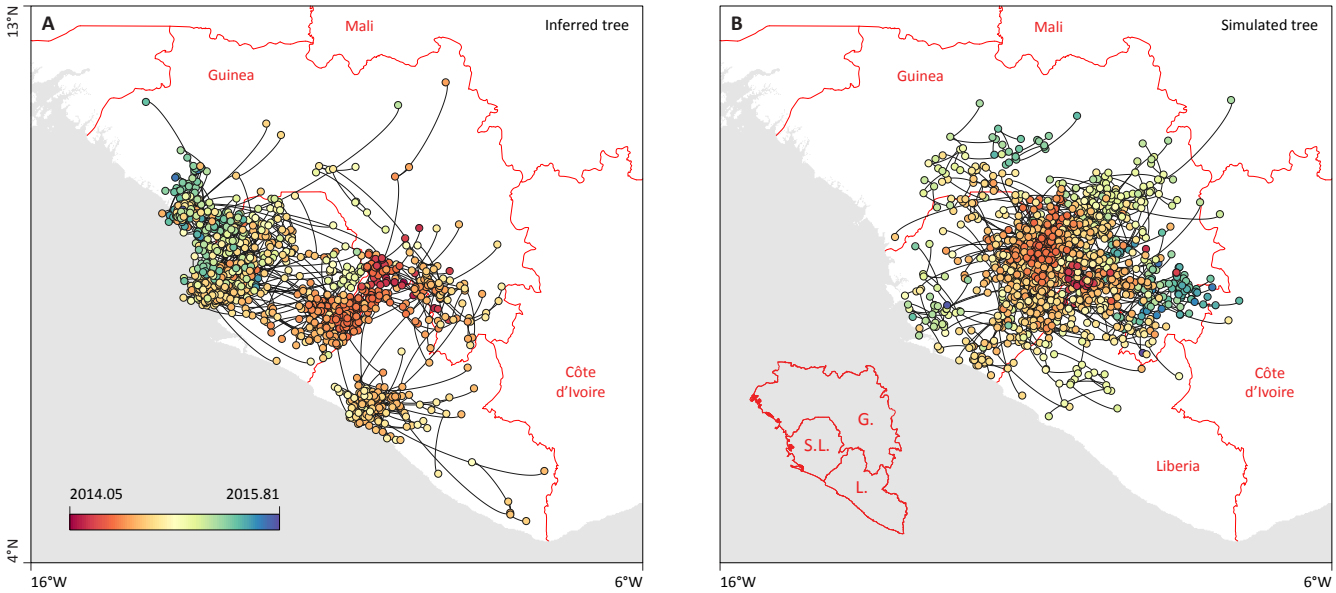

Figure 2. Example of a continuous phylogeographic estimate (A) and the corresponding posterior predictive simulation unaware of international borders (B). In both cases, phylogenetic nodes are coloured according to a colour scale ranging from red (most recent common ancestor) to blue (most recent sampling time).

not contribute substantially to the total epidemic size (Fig. 1A). Only if viral lineage movement is impeded over shorter distances (<250 km) do we start to observe a stronger impact on relative epidemic size. This is the case if lineage movements are prevented throughout the entire duration of the epidemic. If viral lineage movement is restricted only after June 2014, then the epidemic size effect diverges (97%/75% of reduction) at smaller distances (e.g. 100 km). Epidemic duration is similarly affected if short-distance transitions are prevented from the start of the epidemic, but not if the restriction is implemented only after June 2014 (Fig. 1B). This implies that specific transitions between administrative areas with distances larger than >100 km early in the epidemic were critical for generating long-term transmission chains, and that after June 2014, epidemic duration could have been largely maintained by viral lineage movements over shorter distances, albeit at smaller epidemic size.

To further investigate the effects of potential interventions on reducing epidemic size through time, we also undertook coalescent estimation of viral effective population size through time, which appears to be remarkably proportional to case counts (Fig. 1E). This relationship can be statistically tested by using an GLM-based extension of the coalescent approach that considers case counts as a potential covariate of viral effective population size[19]. The GLM coefficient for the association between case counts and estimated effective population size is high (0.55) and its credible intervals exclude zero (95% highest posterior density interval [HPD]: 0.18-0.90), indicating a significant association. The effective population size estimates indicate that the impact of pruning dispersal events between administrative areas ≥200 km apart from the transmission trees has a notable effect on epidemic size only after the time of the epidemic peak (Fig. 1F). These estimates also illustrate the pronounced effect on epidemic size and duration of applying the delay on preventing viral lineage movement over smaller distances. If viral lineage movement had been prevented over distances >100 km between administrative areas at the onset of the outbreak, then the epidemic would have been restricted to an initial small peak in epidemic size (Fig. 1C), which represents the emergence in the Guéckédou prefecture and neighbouring areas[19].

We next apply a similar procedure, but this time restricting viral lineage movement according to the population sizes of the 'destination' (administrative area) of viral lineage movement. Specifically, we bin areas according to their population size and remove all descendent transmission (subtrees) that follow lineage dispersal events to corresponding areas. We then examine the effect of this restriction on epidemic size and duration. Figures 1C and 1D summarise the relative reduction in epidemic size and duration, with and without a delay on transmission prevention (filled and open bars respectively), as well as the sample sizes from all areas within each population size range. For two population size ranges (350-400k and 500-1000k), the presence or absence of a delay on transmission prevention has a large effect on the observed reduction in epidemic size. By examining the impact of each administrative area separately (Supplementary Figure S1), we can attribute this difference to the impact of preventing viral lineage movement to the Kailahun and Kenema districts in Sierra Leone prior to June 2014. These administrative areas represent early transmission bottlenecks that are specific to this EBOV epidemic. The virus spread extremely rapidly from Kailahun district to several counties of Liberia[15] and Guinea[13,17]. However, preventing early viral lineage movement to Kailahun district would not have noticeably reduced the duration of the epidemic (Fig. 1G), because a basal phylogenetic lineage specific to Guinea would have remained unaffected by this restriction and would have continued to circulate, albeit with a limited epidemic burden. We acknowledge that in the interpretation of intervention scenarios, we assume that all other efforts would have remained unchanged when an introduction to a specific location is hypothetically prevented. However, a localised lineage causing limited cases as a consequence of such intervention would have been more easily contained. Preventing early spread to Kenema district would have also halted much of the westward spread from Kailahun district, but its impact on epidemic size is smaller (Fig. 1G), because Kenema district was much less important than Kailahun district for viral spread to Liberia. From these two districts, EBOV spread over relatively long distances to other administrative areas, which is why preventing early lineage movements over such distances generates a strong reduction in epidemic size in the distance-based examination.

When applying the delay to the intervention strategies, the effect of preventing viral lineage movement to sets of administrative areas with different population sizes generally had limited impact on predicted epidemic size (Fig. 1C) and virtually no impact on epidemic duration, except for the case where viral lineage movement was prevented to areas with population sizes >1,000,000. This category corresponds precisely to the areas encompassed by the three capitals, i.e. Montserrado, Freetown (and suburbs) and Conakry. Although about 27% of the genome samples were from these administrative areas, removal of viral lineages that moved into these areas lead to a disproportionate reduction in epidemic size of about 60%, which starts to take effect

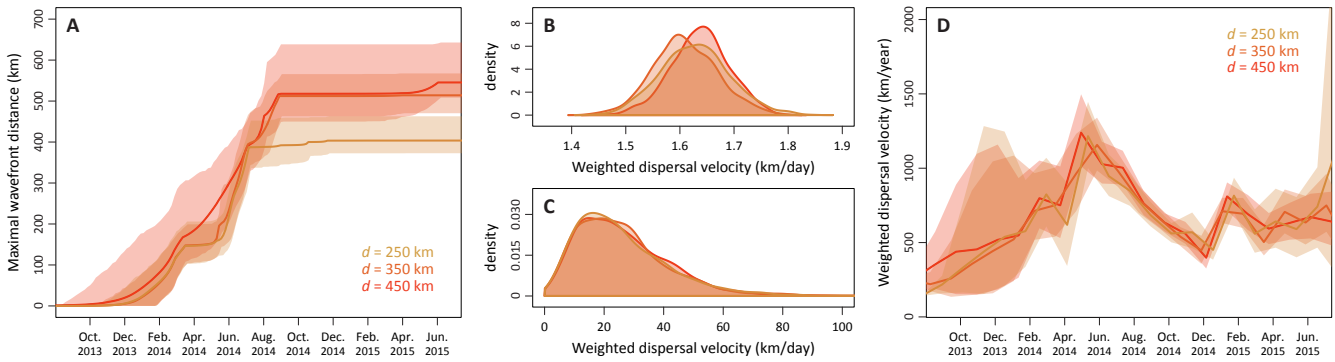

Figure 3. Evolution of maximal wavefront distance (A), mean dispersal velocity (B-C) and dispersal velocity through time (D) for each data set (d = 250, 350 and 450 km). These statistics were estimated from 1,000 trees sampled sampled from the posterior distribution of trees. Weighted dispersal velocities are both reported in km per day (**B**) and in km per infection (**C**). Weighted dispersal velocities in km per infection are obtained by multiplying mean branch velocities (in km/day) by serial interval values (in days between successive infections) randomly drawn from the generation time distribution estimated by the WHO Ebola Response Team (2014).

before the epidemic peak (Fig. 1G). This reflects the important role of highly populated locations in maintaining gravity-model transmission, as previously identified[19]. We note that the epidemic size reduction for the three capitals together is less than the sum of the reduction obtained by removing viral lineage movement to each capital individually (Supplementary Figure S1). This result shows that preventing viral lineage movement to these locations halts the dynamic spread of lineages, which in turn continue to generate numerous clusters of cases in other locations, even in different capitals. This stands in contrast to a model of separate, independent and local chains of transmission in each capital city. We also observe a substantial reduction in epidemic duration only when preventing the viral lineage movement to all the three capitals (i.e. administrative areas with >1,000k people; Fig. 1D), but not preventing the viral lineage movement to a single capital (Supplementary Figure S1). This result indicates that they were jointly important for the spread and maintenance of a handful of lineages that persisted over some time in 2015.

In the second part of this study, we use a new continuous phylogeographic approach to address further key aspects of EBOV spatial spread. Although viral spread driven by human mobility may not always be adequately modelled by a Brownian diffusion process, the strongly distance-dependent diffusion of the Ebola virus (Fig. 1A) justifies its use on this restricted geographic scale. Here we employ a relaxed random walk (RRW) model that accommodates diffusion rate heterogeneity among lineages. We assess the sensitivity of our continuous phylogeographic reconstructions to long-distance dispersal by analysing the data sets that were pruned based on discrete transitions larger than 450 km, 350 km and 250 km (cfr above). Figure 2A illustrates an EBOV transmission tree estimated under this model of continuous phylogeographic inference, which is spatially mapped onto the study area.

Our phylogeographic estimates of the epidemic wavefront through time indicate that EBOV spread up to ∼500 km from its location of origin in about 8 to 9 months (Fig. 3A). The same extent of spatial spread is not achieved for the data set restricted to dispersal events <250 km, indicating that relatively long-distance dispersal events contributed to the maximum epidemic wavefront distance. The mean velocity of the epidemic wavefront from the beginning of 2014 to early September 2014 is ∼1.9 km/day. Our continuous phylogeographic approach also estimates a dispersal velocity of 1.64 km/day (95% HPD [1.52, 1.74]), with little variability across data sets (Fig. 3B). Based on a mean serial interval of transmission of 15.3 days and its uncertainty (SD = 9.3 days[20]), this translates to a mean dispersal distance of 25.4 km per infection (95% HPD [5.33, 61.21]; Fig. 3C) confirming that viral infections may spread because of relatively high mobility of infected individuals in this region of Africa[21]. The mean dispersal velocity shows a remarkable variability during the course of the epidemic (Fig. 3D). It steeply increases from January 2014 to a peak around May/June, coinciding with the time of spread across the region from Kailahun. The peak in dispersal velocity is followed by a marked drop until the end of the year 2014. This drop appears to begin before the announced border closures (Sierra Leone on 11 June 2014, Liberia on 27 July 2014, and Guinea on 9 August 2014). In addition, we observe the same pattern even if we summarise dispersal velocity only for branches that do not cross borders (as determined by their inferred node locations), suggesting that the decrease in dispersal velocity may be attributed to the impact of a more general awareness of the outbreak, and of the emerging response against it.

Although the mean dispersal velocity may be affected by factors other than border closures, phylogeographic testing based on the discrete phylogeographic approach showed that the frequency at which national borders were crossed by viral lineages significantly decreased following the announced border closures between the countries[19]. In order to evaluate this using our continuous phylogeographic approach, we developed a test procedure that compares the estimated posterior frequency of border crossing events to the same frequency in posterior predictive simulations that are unaware of borders (cfr. Methods; see also Figure 2B for an example of such simulation). We quantify the deviation in estimated border frequency crossing from the expected frequency obtained by posterior simulation as a predictive odds ratio, and calculate posterior predictive p-values to assess statistical significance. We also perform this analysis using the data sets for d = 250, 350 and 450 km and include the frequency of within-country administrative border crossing as a negative control (because we do not expect any significant changes in this frequency as a result of border closure). In Figure 3, we plot the predictive odds ratio estimates per month (see also Supplementary Figure S2 for a more detailed representation of the monthly differences in crossing border events among inferred and simulated diffusion processes). These estimates provide strong evidence for a significantly reduced frequency in international border crossing from September 2014 (p-value <0.05), i.e. starting shortly after the announced border closures (Fig. 2), to about the end of the year. As expected, the frequency of within-country administrative border crossing does not depart significantly from the border-unaware expectation.

DISCUSSION

West Africa has experienced the largest outbreak of Ebola virus in history, with more cases and fatalities than all previous outbreaks combined since 1976. Although the region has been declared Ebola free since 2016, it remains critically important to learn as much as possible from this devastating epidemic. Our study builds on a recent viral genome analysis of EBOV in West Africa by investigating further the implications of the metapopulation dynamics revealed using phylogeographic inference[19]. Specifically, we investigate the two key elements of viral spread in a gravity model, distance and population size, by measuring the predicted reduction in epidemic size and duration that results from restricting viral lineage movement, either by preventing

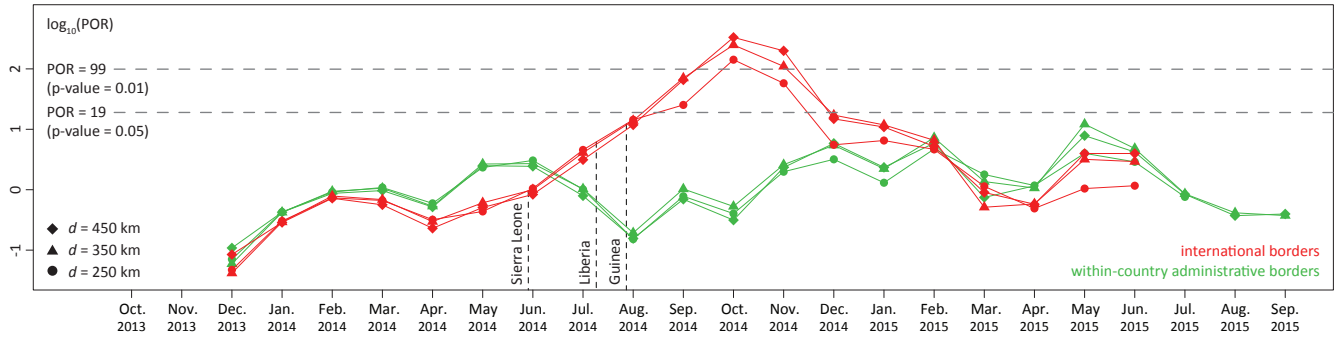

Figure 4. Analysis of the monthly impact of borders on the EBOV dispersal frequency. The plot depicts the predictive odds ratio (POR) estimates per month and per data set (d = 250, 350 and 450 km). PORs estimates for these borders are also included as a negative control as we do not expect any significant impact of within-country administrative borders on dispersal frequency. PORs >3 and >20 can be considered as 'positive' and 'strong' evidence for the impact of borders on the dispersal frequency. Vertical hashed lines indicate the time at which Sierra Leone, Liberia and Guinea announced their border closure. We refer to Supplementary Figure S2 for a more detailed representation of the monthly differences in terms of crossing border events among inferred and simulated diffusion processes.

viral lineage movements of different distances, or by preventing movement events to areas of different population size. The latter can be used to evaluate the impact of hypothetical intervention strategies. We found that long-distance dispersal events were not critical for epidemic expansion, and only when dispersal events are restricted to 200 km or less did we observe significant reductions in epidemic size. While this result does not immediately translate into practical intervention strategies, it suggests that frequent short-distance dispersal realised by human mobility may be more important than rarer long-distance dispersal events in driving the EBOV epidemic spread.

We demonstrate that the contribution of population size to the previously identified gravity model of EBOV spread[19] is almost entirely driven by viral lineage movement to and from the areas encompassing the three capital cities, which are the most highly populated areas in the outbreak region. The fact that the West African EBOV epidemic also affected urban areas, in addition to rural areas, represents a critical difference between it and previous EBOV outbreaks. If viral lineage movement to each of the capitals was prevented, beginning from the onset of the epidemic, then epidemic size could have been reduced by 15% to 37%. This emphasises the importance of these cities, but at the same time, it indicates that no single capital was critical for the maintenance of all co-circulating lineages. Only if viral lineage movement to all three capitals was prevented, would the epidemic size have been reduced to about one-third. Such an intervention would also have a considerable impact on epidemic duration. We therefore conclude that, collectively, urban transmission was a major contributor to the metapopulation dynamics underlying this epidemic. The inability to strongly reduce epidemic size by preventing viral lineage movement to other collections of locations underscores the highly pervasive and distributed nature of the metapopulation dynamic underlying the epidemic. The picture that emerges from our phylogeographic analyses is one of multiple moving targets: potential intervention strategies that are piecemeal, reactive and geographically restricted are predicted to have a limited impact on epidemic size and duration. This result argues for coordinated intervention strategies across the whole outbreak region. By applying a delay to the hypothetical intervention strategies, which will be realistic for most outbreaks, we avoided the impact of early transmission bottlenecks that were identified in Kailahun district and Kenema district in Sierra Leone. The importance of these transmission bottlenecks to the 2014-2016 EBOV outbreak, and their role in spreading the virus to other areas – in particular for Kailahun district – has been highlighted before[13,15,17].

Phylogeographic reconstructions in continuous space have been used primarily for animal viruses because relaxed random walk diffusion may be a reasonable approximation for the dispersal of their host[22–26]. However, the distance-dependent transmission dynamics of EBOV justify the use of a continuous diffusion process for phylogeographic reconstruction. We present this as an alternative and complementary approach for the study of spatial epidemiological dynamics in human populations, at least for relatively restricted geographic scales. Continuous phylogeographic reconstructions enable us to quantify several aspects of dispersal dynamics, such as mean dispersal velocity per infection (25.4 km per infection [5.33, 61.21]). We observe a strong heterogeneity in this dispersal velocity over time, with a significant decrease from May/June 2014 until the end of that year. This likely reflects the general impact of control strategies and awareness on human behaviour.

In our study we use posterior predictive simulation to identify a significant decline in international border crossing. Our findings confirm a significant decline following the announcements of border closures between Guinea, Sierra Leone and Liberia, which was also observed using a discrete phylogeographic approach[19]. Based on case occurrence data, Backer *et al.*[27] previously estimated that only 4-10% of newly infected EBOV cases migrated to another district and that, among these migrants, only 0-23% left their country of origin. Similarly, Kramer *et al.*[28] have used a spatial network approach to demonstrate that the probability of viral lineage movement between locations depended on international border closures. The procedure we used here has a number of advantages relative to the discrete phylogeographic approach that uses epoch modelling to incorporate time-varying predictors[19,29]. A continuous phylogeographic reconstruction does not require a prior specification of the number of change-points, which, in the previous EBOV analysis, was restricted to a single change-point[19]. Instead, we flexibly identify the relevant time period by using statistic that deviates from the null expectation. In addition, we avoid large state spaces in the discrete approach that arise from large numbers of locations and which are computationally expensive, despite the ability to employ multicore GPU architecture[30].

Our analysis expands on the findings from discrete phylogeographic reconstructions and underlines how border closure decisions may spatially structure an epidemic without necessarily having a strong containment effect. The methodology presented in this study could be used to study the impact of potential barriers on the epidemic spread of other important pathogens. Assessing the impact of both hypothesised intervention strategies, as well as the impact of border closures is of interest for public health agencies and policy makers and may provide a better general understanding of outbreak dynamics.

METHODS

Assessing the impact of preventing long-distance or location-specific viral lineage movements. We build on the phylogeographic reconstruction performed by Dudas *et al.*[19], who used a GLM-parameterization of discrete phylogenetic diffusion[31]. Based on a data set of 1,610 viral genomes sampled between 17 March 2014 and 24 October 2015 (available at https://github/ebov/space-time/data/), Dudas *et al.*[19] used this approach to reconstruct a history of lineage movements between 56 administrative regions in Guinea (prefectures),

Sierra Leone (districts) and Liberia (counties). Their Bayesian inference resulted in a posterior distribution of time-measured trees, each annotated with inferred ancestral locations, which was summarised as a maximum clade credibility (MCC) tree. In order to assess the impact of preventing viral lineage movement over specific geographic distances, or to specific locations, we condition on the full transition history in the MCC tree recorded using Markov jumps. Markov jump estimation provides a stochastic mapping of the realisations of the continuous-time Markov process throughout evolutionary history[32,33]. Because each transition between a pair of locations is associated with a geographic distance (the great-circle distance between the locations' population centroids), we are able to assess the impact of preventing viral lineage movement over distances $>d$ by pruning from the complete tree all subtrees that represent the transmission history following branches that accommodate such transition. As possible values for d, we test a series of decreasing distances: 500, 450, 400, 350, 300, 250, 200, 150 and 100 km; this yields pruned trees with 1607, 1567, 1567, 1498, 1368, 1242, 875, 383 and 53 sequences, respectively. We prune each of the posterior trees by removing the same extant taxa as identified in the MCC tree by the pruning selection process. From the resulting set of pruned posterior trees, we record the tree length (the sum of all branch lengths) and tree height (the time to the most recent common ancestor), which we interpret as measures of relative epidemic size and epidemic duration. In order to also obtain an estimate of the effect on relative epidemic size through time, we make use of coalescent estimates under a flexible Bayesian skygrid model[34]. For the complete data set, the effective population size estimates are highly correlated with case counts through time (Fig. 1A). Conditioning on a subset of 1,000 pruned empirical genealogies from the posterior tree distribution, we re-estimate effective population sizes through time using the Bayesian skygrid model and compare these estimates to the original coalescent estimates.

We follow a similar procedure to assess the impact of preventing viral lineage movement to a specific category of administrative areas or to individual administrative areas. Different categories of administrative areas were defined on the basis of their population size: $<$100k, 100-150k, 150-200k, 200-250k, 250-300k, 300-350k, 350-400k, 400-450k, 450-500k, 500-1000k and $>$1,000k people. In these analyses, we pruned subtrees from the complete tree that were the result of transitions from any location to the location(s) within the category under consideration. As before, we obtain the estimates of epidemic duration, relative size, and relative size through time.

Bayesian skygrid estimation with covariates. To assess the strength of association between case counts and effective population size through time, we use a recent extension of the non-parameteric Bayesian skygrid model that incorporates potential covariates[35]. This approach allows us to include external time series as covariates in a GLM framework while accounting for demographic uncertainty. By applying this GLM framework to the complete genome data set[19], we model the Ebola outbreak effective population size as a log-linear function of case counts and estimate the effect sizes for the latter as a GLM coefficient.

Continuous phylogeographic inference. As an alternative to discrete phylogeographic inference, we estimate the spatiotemporal dynamics of Ebola virus by inferring viral lineage movements in continuous space with a multivariate diffusion approach implemented in BEAST[36]. This approach infers the ancestral locations (in geographic coordinates) of internal nodes using a (relaxed) random walk diffusion process. In order to assess the impact of long-distance dispersal on the continuous diffusion estimates, we analyse the pruned data sets obtained using the procedure described in the previous section. Specifically, we remove the same taxa as in the subtrees that need to be pruned based on transitions with distances $> d$ on particular branches. For computational convenience, we restrict ourselves to fewer values of d, specifically d = 450 km, 350 km and 250 km. Additionally, we remove sequences without a known administrative region of sampling (e.g. when only the country of sampling is known), and we remove sequences such that monophyletic clusters of sequences sampled from the same administrative region are only represented by a single sequence. Such clusters would largely represent dispersal within administrative regions, which will be determined by the 'noise' assigned to their location within an administrative region (see below). The final data sets include 722 (d = 450 km), 676 (d = 350 km) and 527 (d = 250 km) sequences, respectively.

Inference under the multivariate diffusion models is problematic when different sequences are associated with identical trait values[36], or in this case, with the same geographic coordinates. When unique sampling coordinates are not available for every sequence, a common practice is to add a restricted amount of noise to duplicated traits. In BEAST[37], this is facilitated by a jitter option, that uniformly draws such noise from a user-defined window size for each dimension of the trait (i.e. each coordinate). Using a jitter may be problematic for two reasons in our case. Many administrative areas are along the coast, so the added noise may lead to sampling coordinates in the sea. Due to the different sizes and shapes of the administrative areas, the noise may also move coordinates to areas neighbouring their actual sampling area. To avoid these issues, we associate a random coordinate within the administrative area of sampling to each sequence.

We analyse each data set using a relaxed random walk (RRW) model with an underlying Cauchy distribution to represent among-branch heterogeneity in branch velocity[36], and with a multivariate Wishart distribution as a prior on the precision matrix[38]. We follow Dudas *et al.*[19] in choosing substitution, molecular clock and coalescent models, and their prior specifications, and reiterate those choices here. We model molecular evolution according to a HKY$+\Gamma_4$[39,40] substitution model independently across four partitions (codon positions 1, 2, 3 and non-coding intergenic regions), allowing for partition-specific relative rates. We use a non-parametric coalescent Bayesian skygrid model as a prior density over the tree[34,35] and model branch-specific evolutionary rates according to a relaxed molecular clock with an underlying log-normal distribution[41]. We specify a continuous-time Markov chain (CTMC) reference prior[42] for the overall evolutionary rate, while the rate multipliers for each partition were given an uninformative uniform prior over their bounds. All other priors in the phylogenetic inference are left at their default values. For each of the continuous phylogeographic inferences, we ran an MCMC chain using BEAST 1.8.4[37] for 130 (d = 250 km), 470 (d = 350 km) and 740 (d = 450 km) million iterations, removing the first 5% samples in each chain as burn-in. Based on these continuous phylogeographic inferences, we estimate, for each data set, the weighted dispersal velocity v_{weighted} defined as follows:

$$v_{\text{weighted}} = \frac{\sum_{i=1}^{n} d_i}{\sum_{i=1}^{n} t_i}$$

where n is the number of branches in the phylogeny, d_i the geographic distance travelled and t_i the time elapsed on each phylogeny branch. We estimate this statistic for 1,000 samples of the posterior distribution of trees and report a posterior distribution of estimated values. In addition, we also reconstruct, for each data set, the evolution of the mean dispersal velocity and of the maximal wavefront distance through time. We obtain all these statistics using the R package 'seraphim'[43,44].

Impact of borders on dispersal frequency. To test if administrative/international borders act as barriers to dispersal frequency, we adopt a Bayesian posterior predictive simulation procedure. Such a procedure allows the calculation of a Bayesian counterpart of the classical p-value, using posterior predictive replications of the data and employing a test statistic that depends on both data and unknown (nuisance) parameters[45,46]. In our setup, we record the number of times borders are crossed by tree branches in the posterior set of trees, as determined by the location at the parent and child node of the branches, and compare this to posterior predictive values for the same statis-

tic. We obtain the posterior predictive values by simulating a RRW process along each posterior tree using the sampled precision matrix parameters and location at the root node. We also condition on the branch-specific posterior rate scalars to generate the mixture of normals that characterises the RRW. In addition, we constrain the RRW simulations such that the simulated node locations are in continental African and cannot fall in the ocean. As illustrated in Figure 3, posterior and posterior predictive diffusion histories are roughly similar except for the position and orientation of branches in the West African study area. We integrate over all possible realisations, weighted by their posterior probabilities based on 1000 samples, to generate a test based on the border crossing frequency. In the absence of a border impact on Ebola movement in the West African epidemic, we expect a similar number of border crossing events between the posterior and posterior predictive diffusion histories, as the latter is deliberately unaware of borders. For each data set, we again use a sample of 1,000 trees from the post burn-in posterior distribution of phylogenies in order to accommodate phylogenetic uncertainty. Upon simulation of RRW diffusion along these sampled trees, we count and compare the number of crossing border events for each pair of inferred and simulated diffusion processes associated with a particular tree. Each 'inferred' N value (N_{inferred}) is thus compared to its corresponding 'simulated' value ($N_{\text{simulated}}$) to compute a predictive odds ratio (POR) as follows:

$$\text{POR} = \frac{p_e}{1-p_e} \Big/ \frac{0.5}{1-0.5}$$

where p_e is the posterior probability that $N_{\text{inferred}} < N_{\text{simulated}}$, i.e. the frequency at which $N_{\text{inferred}} < N_{\text{simulated}}$ in the sampled posterior distribution. In interpreting POR estimates, we adopt the Bayes factor scale defined by Kass & Raftery[47]: values higher than 3, 20 and 150 are considered as 'positive', 'strong' and 'very strong' evidence respectively for the impact of administrative/international borders on dispersal frequency. Both types of within-country administrative borders were obtained from the Global Administrative Areas database (GADM, www.gadm.org). The human population density and inaccessibility rasters were obtained from the Global Rural-Urban Mapping Project (GRUMP, sedac.ciesin.columbia.edu) and the Joint Research Centre of the European Commission (forobs.jrc.ec.europa.eu/products/gam) respectively. We implement the posterior predictive simulation procedure, along with a related tutorial, in the R package 'seraphim'[43,44].

REFERENCES

1. Baize, S. *et al.* Emergence of Zaire Ebola virus disease in Guinea. *N. Engl. J. Med.* **371**, 1418–1425 (2014).
2. World Health Organization. Ebola situation report - 10 june 2016 (2016).
3. Woolhouse, M. E. J., Rambaut, A. & Kellam, P. Lessons from Ebola: improving infectious disease surveillance to inform outbreak management. *Sci. Transl. Med.* **7**, 307rv5 (2015).
4. Bell, B. P. *et al.* Overview, control strategies, and lessons learned in the CDC response to the 2014-2016 Ebola epidemic. *MMWR Suppl.* **65**, 4–11 (2016).
5. Arias, A. *et al.* Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases. *Virus Evol.* **2**, vew016 (2016).
6. Gardy, J., Loman, N. J. & Rambaut, A. Real-time digital pathogen surveillance — the time is now. *Genome Biol.* **16**, 155 (2015).
7. Modjarrad, K. *et al.* Developing global norms for sharing data and results during public health emergencies. *PLoS Med.* **13**, e1001935 (2016).
8. Nature Microbiology editorial. Don't wait to share data on Zika. *Nat. Microbiol.* **1**, 16028 (2016).
9. Holmes, E. C., Dudas, G., Rambaut, A. & Andersen, K. G. The evolution of Ebola virus: insights from the 2013-2016 epidemic. *Nature* **538**, 193–200 (2016).
10. Gire, S. K. *et al.* Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369–72 (2014).
11. Volz, E. & Pond, S. Phylodynamic analysis of Ebola virus in the 2014 Sierra Leone epidemic. *PLoS Curr.* **6** (2014).
12. Stadler, T., Kühnert, D., Rasmussen, D. A. & du Plessis, L. Insights into the early epidemic spread of Ebola in Sierra Leone provided by viral sequence data. *Plos Curr.* **6** (2014).
13. Carroll, M. W. *et al.* Temporal and spatial analysis of the 2014-2015 Ebola virus outbreak in West Africa. *Nature* **524**, 97–101 (2015).
14. Kugelman, J. R. *et al.* Monitoring of Ebola virus Makona evolution through establishment of advanced genomic capability in Liberia. *Emerg. Infect. Dis.* **21**, 1135–1143 (2015).
15. Ladner, J. T. *et al.* Evolution and spread of Ebola virus in Liberia, 2014-2015. *Cell Host Microbe* **18**, 659–669 (2015).
16. Park, D. J. *et al.* Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. *Cell* **161**, 1516–1526 (2015).
17. Simon-Loriere, E. *et al.* Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic. *Nature* **524**, 102–104 (2015).
18. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232 (2016).
19. Dudas, G. *et al.* Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* **544**, 309–315 (2017).
20. WHO Ebola Response Team. Ebola virus disease in West Africa — The first 9 months of the epidemic and forward projections. *N. Engl. J. Med.* **371**, 1481–1495 (2014).
21. Wesolowski, A. *et al.* Commentary: containing the Ebola outbreak - the potential and challenge of mobile network data. *PLoS Curr.* **6** (2014).
22. Pybus, O. G. *et al.* Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 15066–15071 (2012).
23. Trovo, N. S., Suchard, M. A., Baele, G., Gilbert, M. & Lemey, P. Bayesian inference reveals host-specific contributions to the epidemic expansion of influenza a H5N1. *Mol. Biol. Evol.* **32**, 3264–3275 (2015).
24. Streicker, D. G. *et al.* Hostpathogen evolutionary signatures reveal dynamics and future invasions of vampire bat rabies. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 10926–10931 (2016).
25. The Global Consortium for H5N8 and Related Influenza Viruses. Role for migratory wild birds in the global spread of avian influenza H5N8. *Science* **354**, 213–217 (2016).
26. Dellicour, S. *et al.* Using viral gene sequences to compare and explain the heterogeneous spatial dynamics of virus epidemics. *Mol. Biol. Evol.* (in press).
27. Backer, J. A. & Wallinga, J. Spatiotemporal analysis of the 2014 Ebola epidemic in West Africa. *PLoS Comput. Biol.* **12**, e1005210 (2016).
28. Kramer, A. M. *et al.* Spatial spread of the West Africa Ebola epidemic. *R. Soc. Open Sci.* **3**, 160294 (2016).
29. Bielejec, F., Lemey, P., Baele, G., Rambaut, A. & Suchard, M. A. Inferring heterogeneous evolutionary processes through time: from sequence substitution to phylogeography. *Syst. Biol.* **63**, 493–504 (2014).
30. Ayres, D. L. *et al.* BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Syst. Biol.* **61**, 170–173 (2012).
31. Lemey, P. *et al.* Unifying viral genetics and humant transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS Path.* **10**, e1003932 (2014).
32. Minin, V. M. & Suchard, M. A. Fast, accurate and simulation-free stochastic mapping. *Phil. Trans. R. Soc. Lond. B Biol. Sci.* **363**, 3985–3995 (2008).
33. Minin, V. M. & Suchard, M. A. Counting labeled transitions in continuous-time Markov models of evolution. *J. Math. Biol.* **56**, 391–412 (2008).
34. Gill, M. *et al.* Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol.* **30**, 713–724 (2013).
35. Gill, M. S., Lemey, P., Bennett, S. N., Biek, R. & Suchard, M. A. Understanding past population dynamics: Bayesian coalescent-based modeling with covariates. *Systematic biology* **65**, 1041–1056 (2016).
36. Lemey, P., Rambaut, A., Welch, J. J. & Suchard, M. A. Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* **27**, 1877–1885 (2010).
37. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).
38. Wishart, J. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika* **20A**, 32–52 (1928).
39. Hasegawa, M., Kishino, H. & Yano, T.-a. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22**, 160–174 (1985).
40. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution* **39**, 306–314 (1994).
41. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol* **4**, e88 (2006).
42. Ferreira, M. A. R. & Suchard, M. A. Bayesian analysis of elapsed times in continuous-time markov chains. *Canadian Journal of Statistics* **36**, 355–368 (2008).
43. Dellicour, S., Rose, R. & Pybus, O. G. Explaining the geographic spread of emerging epidemics: a framework for comparing viral phylogenies and environmental landscape data. *BMC Bioinfo.* **17**, 1–12 (2016).
44. Dellicour, S., Rose, R., Faria, N. R., Lemey, P. & Pybus, O. G. SERAPHIM: studying environmental rasters and phylogenetically informed movements. *Bioinformatics* **32**, 3204–3206 (2016).
45. Meng, X. L. Posterior predictive p-values. *Ann. Statist.* **22**, 1142–1160 (1994).
46. Gelman, A., Meng, X. & Stern, H. Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Stat. Sin.* **6**, 733–807 (1996).
47. Kass, R. E. & Raftery, A. E. Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).

Supplementary Information Supplementary Figures S1-2 are available in the online version of the paper.

Acknowledgments SD is a post-doctoral research fellow funded by the Fonds Wetenschappelijk Onderzoek (FWO, Flanders, Belgium). GB, PL and AR acknowledge funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement no. 278433-PREDEMICS. OP received funding from the ERC under grant agreement 614725-PATHPHYLODYN PL, AR and MAS acknowledge funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 725422-ReservoirDOCS) and from the Wellcome Trust Collaborative Award, 206298/Z/17/Z. PL acknowledges support by the Special Research Fund, KU Leuven ('Bijzonder Onderzoeksfonds', KU Leuven, OT/14/115), and the Research Foundation – Flanders ('Fonds voor Wetenschappelijk Onderzoek – Vlaanderen', G066215N, G0D5117N and G0B9317N). MAS acknowledges funding through National Science Foundation grant DMS1264153 and National Institutes of Health grant R01 LM012080. GB acknowledges support from the Interne Fondsen KU Leuven / Internal Funds KU Leuven. GD is supported by the Mahan Post-doctoral Fellowship at Fred Hutchinson Cancer Research Center. The VIROGENESIS project receives funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 634650. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government - department EWI.