INTERFACE

rsif.royalsocietypublishing.org

Research

Cite this article: Gustafson KB, Proctor JL. 2017 Identifying spatiotemporal dynamics of Ebola in Sierra Leone using virus genomes. *J. R. Soc. Interface* **14**: 20170583. http://dx.doi.org/10.1098/rsif.2017.0583

Received: 7 August 2017
Accepted: 2 November 2017

Subject Category:
Life Sciences–Physics interface

Subject Areas:
computational biology, systems biology

Keywords:
Ebola, phylodynamics, disease modelling, spatial epidemiology

Author for correspondence:
Kyle B. Gustafson
e-mail: kgustafson@idmod.org

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3931891.

THE ROYAL SOCIETY PUBLISHING

Identifying spatiotemporal dynamics of Ebola in Sierra Leone using virus genomes

Kyle B. Gustafson and Joshua L. Proctor

Institute for Disease Modeling, Bellevue, WA 98005, USA

KBG, 0000-0002-1903-9015

Containing the recent West African outbreak of Ebola virus (EBOV) required the deployment of substantial global resources. Despite recent progress in analysing and modelling EBOV epidemiological data, a complete characterization of the spatiotemporal spread of Ebola cases remains a challenge. In this work, we offer a novel perspective on the EBOV epidemic in Sierra Leone that uses individual virus genome sequences to inform population-level, spatial models. Calibrated to phylogenetic linkages of virus genomes, these spatial models provide unique insight into the *disease mobility* of EBOV in Sierra Leone without the need for human mobility data. Consistent with other investigations, our results show that the spread of EBOV during the beginning and middle portions of the epidemic strongly depended on the size of and distance between populations. Our phylodynamic analysis also revealed a change in model preference towards a spatial model with power-law characteristics in the latter portion of the epidemic, correlated with the timing of major intervention campaigns. More generally, we believe this framework, pairing molecular diagnostics with a dynamic model selection procedure, has the potential to be a powerful forecasting tool along with offering operationally relevant guidance for surveillance and sampling strategies during an epidemic.

1. Introduction

Arresting the West African Ebola virus (EBOV) epidemic of 2013–2016 required a significant international intervention and exposed a global vulnerability to emerging epidemics. Advances in genetic-sequencing technologies have enabled the near real-time analysis of infectious pathogen genomes [1–3], thereby improving forecasts for emerging epidemics [4,5], enhancing surveillance of endemic diseases [6] and identifying strategies for eradication [7,8]. During the West African EBOV crisis, publicly available data facilitated a series of prominent analyses aimed at identifying basic epidemiological parameters, i.e. the reproductive number [9,10]. Further, the release of EBOV genome data, coupled with phylogenetic methods, provided fundamental insight into the origin and spatial properties of the epidemic [11,12]. Despite the prominent role mathematical and statistical modelling played during the epidemic, there has been a significant delay in characterizing the spatiotemporal spread of EBOV. For future epidemics, the design of operationally relevant, spatially distributed interventions requires the identification of predictive models that are able to assimilate case and genetic data. In this article, we describe how EBOV molecular data, specifically virus genomes, can be used to directly model the spatiotemporal dynamics of the Sierra Leone epidemic.

Recent investigations of disease propagation on modern transportation networks have pointed to the importance of characterizing the spatial behaviour of vectors and pathogens due to human movements [13,14]. An influential development in the study of human mobility and disease spread is the adoption of the *gravity model* from the field of economics [15]. Analogous to the attracting force between physical masses, the gravity model describes human movements as dependent on the size of and distance between human populations [16,17].

Other spatial models, such as the well-known, scale-free Lévy flights, depend solely on travelling distance and have been used to describe a wide-ranging set of phenomena from epidemiology [18], ecology [19] and plasma physics [20].

Mathematically, the gravity and Lévy flight models are closely related. However, when the model parameters are fit to country-specific data, their dynamic behaviour can be qualitatively different. The gravity model parameters are often fit using proxy data such as cell phone, transportation and individual survey records [21]. This spatial model can then be coupled to a disease transmission model [22]. Alternatively, molecular data offer direct insight into disease mobility [4,23]. Genomic data have been used to construct distance-dependent spatial models for West Nile virus in North America [2] as well as dengue virus in Thailand [24] and Vietnam [25]. Other phylodynamic approaches, focused on mapping transmission trees [26], have been widely applied to infectious disease data including outbreaks of foot-and-mouth disease [27], severe acute respiratory syndrome (SARS) [28] and tuberculosis [29].

Previous spatial analyses and modelling efforts for the EBOV epidemic have identified population as an influential factor using a generalized gravity model parametrized to case-reporting data [30–32]. A comprehensive phylodynamic analysis of all available West African EBOV genomes also concluded that population distribution and distance between cases are important explanatory factors [33]. Other phylodynamic analyses of EBOV incorporated multiple countries and revealed the importance of social clustering to transmission risk [34,35]. Despite these recent investigations, which identify potential drivers of the EBOV epidemic, a fully characterized understanding of the spread of the West African epidemic remains a challenge.

We present a novel investigation of the EBOV genome data, allowing for a more resolved characterization of the spatiotemporal dynamics during the epidemic. Transmission of EBOV within Sierra Leone was almost completely within its borders [36], which provided a constrained and representative dataset to investigate the utility of virus genomes to construct population-level spatial models. Paired with advances in phylogenetics that identify linkages between cases [8], EBOV genome data offer powerful insight into spatiotemporal, transmission events. These genomic linkages, in combination with geographical and demographical characteristics included in our framework, help infer the parameters of gravity and Lévy flight models. We focus on identifying data-driven, spatial models that are interpretable and consistent with established patterns of human population movement. Adaptive model selection during the course of the epidemic reveals a significant change in virus mobility in Sierra Leone: dependence on population size decreases towards the end of the epidemic. For future epidemics, we believe that this framework could be implemented to improve forecasting efforts and help design efficient intervention campaigns that adapt to real-time phylodynamics.

2. Study data and methods

2.1. Genomic data

Genetic sequences from 1031 human infections of EBOV in Sierra Leone were obtained from a openly accessible compilation [33] of previously published sequencing data [36–39]. In figure 1*a* and electronic supplementary material, figure S4, we show the time course of all confirmed EBOV cases (black trace) in Sierra Leone [40] compared with the number of sequenced virus genomes [33] (red trace). The FASTA file with the genomes and metadata was downloaded from http://github.com/ebov/space-time/tree/master/Data/Makona_1610_genomes_2016-06-23.fasta on 9 August 2016. We then used BEAUTI 1.8.3 [41] with default options to generate an XML file with the metadata of spatial and temporal coordinates for each sequence.

2.2. Partially observed transmission network

We used a recently developed phylogenetic method [8], known as the partially observed transmission network (POTN) algorithm, to determine genetic linkages between EBOV infections in Sierra Leone. The POTN algorithm computes a likelihood ratio based on a Poisson model of the mutation rate to identify genomes that are most likely to be direct relatives. This contrasts with widely used phylogenetic analyses that infer common ancestors, such as Bayesian Evolutionary Analysis Sampling Trees (BEAST) [41]. The POTN algorithm produces a pairwise, time-directed network of ancestor and descendant genomes, linked by the relative change in their sequences between collection dates. For EBOV, we used an average nucleotide substitution rate of 2×10^{-3} bp/site/yr, a value measured during the 2013–2016 epidemic; see Fig. 4F of [11]. A false discovery rate for each linkage is computed with a single degree-of-freedom χ^2 test, with a cut-off at $p = 0.05$. Figure 1*b* shows a visualization of the EBOV POTN for Sierra Leone pruned to the shortest generation time for each ancestor. The blue arrow highlights a single linkage between virus genomes collected in the districts of Western Urban and Kenema.

2.3. Population distribution and driving distances

Population distribution maps from the 2010 and 2014 Worldpop models were downloaded from http://worldpop.org.uk/data/ in June 2016. These maps were segmented into 153 Admin-3 units (chiefdoms) using the Sierra Leone shapefiles from Global Administrative Areas http://gadm.org/download, illustrated in figure 1*c*. Driving distances were used as the distance measure between chiefdoms. The shortest time driving distances between all chiefdom pairs were collected from the Google Maps API. Figure 1*c* shows the major roads in Sierra Leone. The blue arrow indicates the path along the roads between chiefdoms of median population in the districts of Western Urban and Kenema.

2.4. Distance statistics for genetic linkages

We examined the statistics of EBOV transmissions using the distribution of driving distances between POTN-linked cases. These are denoted as transmission distances, d_{ij}. We plot in figure 1*d* the probability of observing a d_{ij} above a certain magnitude, which is defined as a complementary cumulative distribution function (cCDF), and is useful for identifying a power-law distribution from empirical data [42]. We computed d_{ij} for each genetic linkage by assigning each sequence to a chiefdom, either known from the metadata or approximated by population size in its annotated district. We omitted 119 genomes without a district (Admin-2) localization from the analysis. For the first part of the epidemic, chiefdom localizations are available for 187 genomes [36]. When the chiefdom localization is unknown for a virus genome, we selected a chiefdom based on assumptions of population size within the known district, such as maximum, mean, median or minimum.

2.5. Probabilistic spatial models

Previous analyses have pointed to the importance of size and distance between populations as factors that influenced the spread

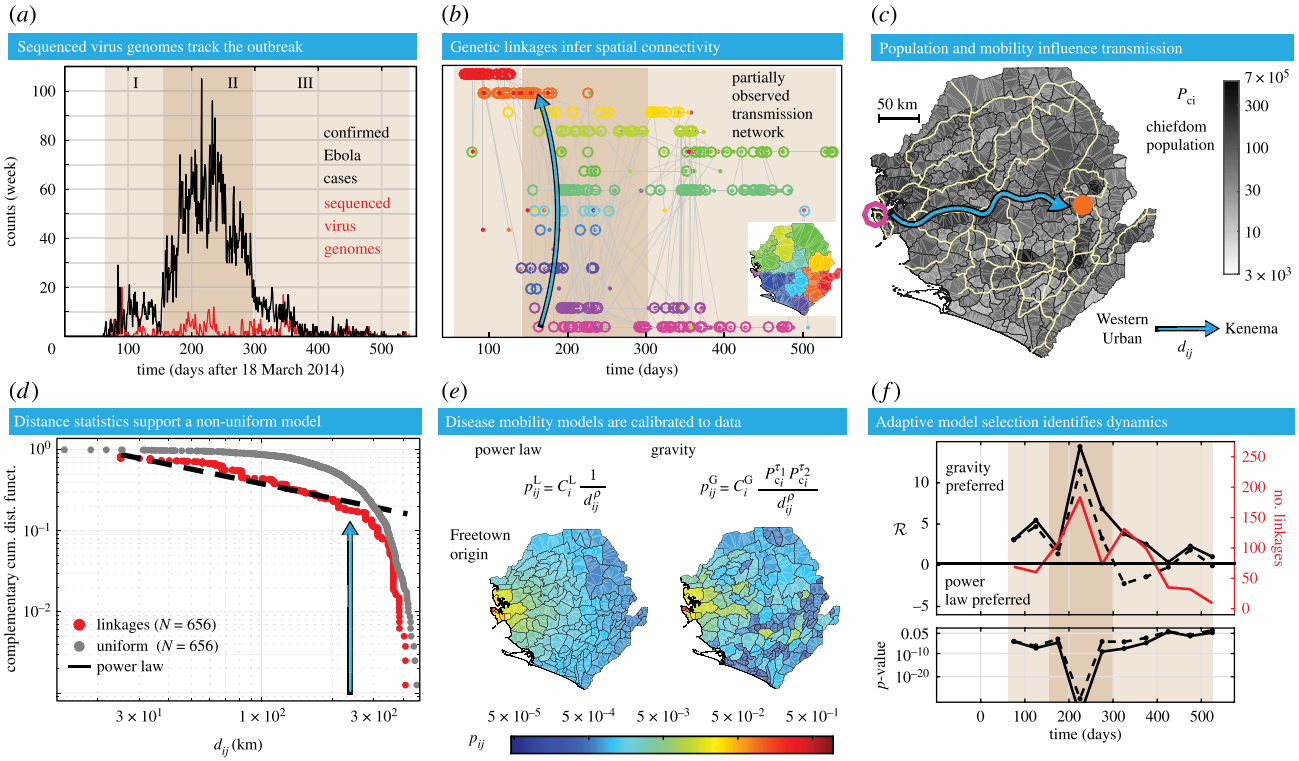

Figure 1. Virus genome data from EBOV cases in Sierra Leone characterizes the spatial spread of the epidemic. (*a*) The time course is shown for the number of confirmed cases [40] and sequenced EBOV genomes [33]. Three stages of the epidemic are highlighted. (*b*) Genetic linkages are illustrated with ancestors (open circles) and descendants (closed dots), both coloured by the origin district shown in the map key. The blue arrow highlights a linkage from the Western Urban to Kenema districts. (*c*) Chiefdom populations (greyscale) and major roads (yellow traces) are illustrated on the map of Sierra Leone. The blue arrow highlights the fastest driving route between the Western Urban to Kenema district. (*d*) All transmission distances are shown in a cCDF. The distribution of transmission distances are fit by a power law with $\rho = 1.66$. The blue arrow follows the linkage from (*b*) and (*c*). (*e*) Two spatial models are plotted as maps representing the probability of observing a new case linked to the Western Urban district, using ($\rho^\star = 1$, $\tau_2^\star = 1$) for the gravity model and $\rho = 1.66$ for the power law. (*f*) The log-likelihood ratio, $\mathcal{R}$, comparing the gravity and power-law models, is plotted for 50-day windows. The dashed black line represents ($\rho = 1.66$, $\tau_2 = 1$) fixed in time; the solid black line of $\mathcal{R}$ uses the MLE ($\rho^\star(t)$, $\tau_2^\star(t)$), computed for each window. The solid red trace describes the number of linkages.

of EBOV in West Africa [31,33]. Here, we specify a gravity model for a discrete spatial network of populations that describes the probability of a virus being transmitted from chiefdom i to chiefdom j: $p_{ij}^{\mathrm{G}} = C_i^{\mathrm{G}} P^{\tau_1}{}_i P^{\tau_2}{}_j / d_{ij}^{\rho}$, where the origin population is P_i, the destination population is P_j and C_i^{G} normalizes the probability distribution for each origin. The exponents τ_1, τ_2 and ρ are parameters that determine the influence of population and distance for the gravity model. The normalization for each origin is computed by the following: $C_i^{\mathrm{G}} = 1/(\sum_j P_i^{\tau_1} P_j^{\tau_2} / d_{ij}^{\rho})$. Note that the normalization depends solely on the destination population and distance. This formulation of the gravity model predicts where a future linked case will appear.

A closely related probabilistic model is the Lévy flight model, which has a rich mathematical basis in the framework of fractional diffusion equations and scale-free non-diffusive random processes [43]. We write the discrete space power-law model as $p_{ij}^{\mathrm{L}} = C_i^{\mathrm{L}} / d_{ij}^{\rho}$, where C_i^{L} normalizes the probability for each origin. Again, we are interested in characterizing the probability of viral transmission to chiefdom j. The resting probability for both models is uniformly approximated to $p_{ii} = 0.5$; see electronic supplementary material, figure S1 for a district-level analysis of stationary linkages. This approach can be extended to include a wide variety of spatial models with context-appropriate parameters for the underlying stochastic process.

2.6. Maximum-likelihood estimates for gravity model parameters

For the gravity model, the parameters ρ and τ_2 that best fit the data can be determined through a maximum-likelihood estimate (MLE). The joint likelihood for the parametric gravity model, $\mathcal{L}^{\mathrm{G}}$, is defined as the product of model evaluations over the set of virus genome linkages $\mathcal{S}$: $\mathcal{L}^{\mathrm{G}} = \prod_{\mathcal{S}} p_{ij}^{\mathrm{G}}(\rho, \tau_2)$. We define ($\rho^\star$, $\tau_2^\star$) as the MLE of the parameters for the gravity model determined by evaluating the likelihood for a range of (ρ, τ_2) values. We establish a 95% CI for ($\rho^\star$, $\tau_2^\star$) via the well-known Fisher information criterion [44].

2.7. Adaptive model selection

We computed a time-dependent likelihood ratio that quantifies the relative preference between models over the course of the epidemic. Note that the power-law model is considered nested within the gravity model if $\tau_2 \to 0$. The likelihood ratio, $\mathcal{R}$, is computed for a set of virus genome linkages $\mathcal{S}$. The normalized log-likelihood ratio of a gravity model to a Lévy flight model is $\mathcal{R}(\rho, \tau_2) = \sum_{\mathcal{S}} [\ln(p_{ij}^{\mathrm{G}}) - \ln(p_{ij}^{\mathrm{L}})] / \sqrt{N}$, where N is the number of linkages in $\mathcal{S}$. If $\mathcal{R} > 0$, the gravity model is preferred, but if $\mathcal{R} < 0$, the power-law model is preferred. The significance of this preference is computed by a χ^2 test according to Wilks' theorem [45]. We made $\mathcal{R}$ time-dependent by partitioning $\mathcal{S}$ into subsets of linkages, $\mathcal{S}_t$. In figure 1*f*, each subset includes all linkages with the descendant genomes collected in each 50 day interval centred around t. This model selection approach can be extended to include non-nested models by using an information criterion such as the Akaike information criterion [46].

3. Results

3.1. A transmission network links most virus genomes

We constructed a POTN using 880 virus genomes from Sierra Leone that revealed 798 transmission events. Of these, 355

have a unique descendant and 670 have fewer than three likely descendants. The POTN algorithm is designed to return several possible descendants associated with the same ancestor when these are supported by the data. See electronic supplementary material, figure S2A for the empirical distribution of all linkage durations. For illustrative purposes, figure 1*b* shows the POTN pruned to include only the shortest linkage duration from each ancestor. The median linkage time from the POTN is 31 days. Based on an EBOV serial interval of 15 days [47], this implies approximately one unobserved transmission event per linkage. The median linkage duration is reduced to 11 days if the POTN is pruned to the shortest generation times, illustrated in electronic supplementary material, figure S2B.

The likelihood ratio and MLE calculations include all likely descendants associated with a single ancestor. However, for the adaptive model selection procedure, descendants outside of each time interval are excluded. For example, figure 1*f* illustrates the likelihood ratio for 50 day intervals. Despite the challenges associated with partially observed transmission chains, we find that our subsequent analyses of spatial model calibration and model selection are robust to a wide variety of linkage exclusion criteria, such as restricting the maximum allowable linkage duration. See electronic supplementary material, figure S3 for more details. Our results are also robust when considering longer time intervals. We define three sequential stages for the epidemic: Stage I (0–150 days), Stage II (150–300 days) and Stage III (300–550 days). The geographical distribution of linked cases across districts for the three stages shows that the number of genomes sequenced is proportional to the number of confirmed cases [40], except when the number of confirmed cases is larger than 1000 (electronic supplementary material, figure S4).

3.2. Transmission distances follow a power-law model

Several analytic techniques were used to test for a power law in the distribution of d_{ij} for all linkages. Cumulatively, for 656 linkages with $d_{ij} > 0$ km, we computed a power-law scaling exponent of $\rho = 1.66 \pm 0.02$ for the discrete distribution of d_{ij}, as shown in figure 1*d*. We also found that ρ is consistent across different stages of the epidemic, as shown in figure 2 and electronic supplementary material, figure S5. This estimate for ρ was computed using a well-known maximum-likelihood method for power-law distributions [42]. As a note of caution, the methodology in [42] provides a lower bound on the distance to define a power-law tail, whereas we have explicitly included all transmission distances here to remain unbiased. We verified that the power law is preferred by the likelihood ratio over a Weibull or exponential probability distribution.

As a separate investigation, a two-sample Kolmogorov–Smirnov test showed that the distribution of d_{ij} is not likely drawn from the same distribution as all the possible driving distances between chiefdoms. Therefore, the transmission events do not match a uniform random process on the driving network. We also examined the sensitivity of the model fit by sampling from the inferred model and driving distance distribution. By randomly drawing a similar number of samples from the inferred model and driving network, we found that the model fit is robust to the number of samples collected during the epidemic; see electronic supplementary material, figure S6 and accompanying text for more details.

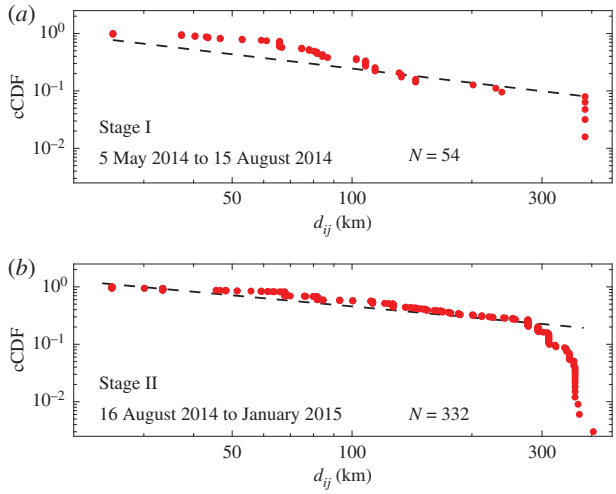

Figure 2. The empirical power law for the transmission distances. (*a*) The complementary cumulative distribution function (cCDF) for Stage I, ($50 \leq t < 200$ days), is plotted along with the power-law model using the MLE value of $\rho^* = 1.8 \pm 0.1$ for $N = 54$ linkages. (*b*) The cCDF for Stage II, ($200 \leq t < 350$ days), is plotted with the power-law model using the MLE value of $\rho = 1.6 \pm 0.1$ for $N = 332$ linkages.

The observed distribution of d_{ij} is closely related to a power law for a significant portion of the data, shown in electronic supplementary material, figure S6.

However, there are clear differences between the simulation of the idealized power-law model and the distribution of d_{ij} that suggest the influence of other factors. For example, the geographical structure of Sierra Leone constrains the number of possible trips over 300 km. This is consistent with the observed drop-off in transmission distances, indicating a limitation of the standard power-law model for the complex effects of national borders and local administrative divisions. Further, including other factors, such as the chiefdom-level population distribution, will allow for more flexibility in characterizing the observed distribution of transmission distances.

3.3. Gravity model at epidemic peak was driven by Freetown

Inferring the parameters of the gravity model with the genetic linkage data, population was found to be an important variable in characterizing spatial transmission events, especially in Stage II of the epidemic when the Western Area is involved in 244 of the 363 genetic linkages. In Stage II, the MLE of the gravity model parameters found $\tau_2^\star = 1.2 \pm 0.3$ and $\rho^\star = 0.9 \pm 0.5$ with a 95% CI; see electronic supplementary material, figure S7 for more details on the MLE calculation. The likelihood landscapes, with varying (ρ, τ_2), are shown in figure 3*d* for Stages I–III. Values of the log-likelihood for each stage and both models are shown in electronic supplementary material, figure S9. The likelihood ratio, comparing the gravity and power-law models during Stage II, indicated a strong preference for gravity shown in figure 1*f*. Further, figure 3*b* illustrates that the POTN for Stage II contains a significant number of transmission events in the Western Area of Sierra Leone supporting the population-dependent model. When setting the population parameter τ_2 to the canonical value of 1, the gravity model was still preferred over the power law for this portion of the epidemic. Figure 1*e* illustrates gravity and power-law

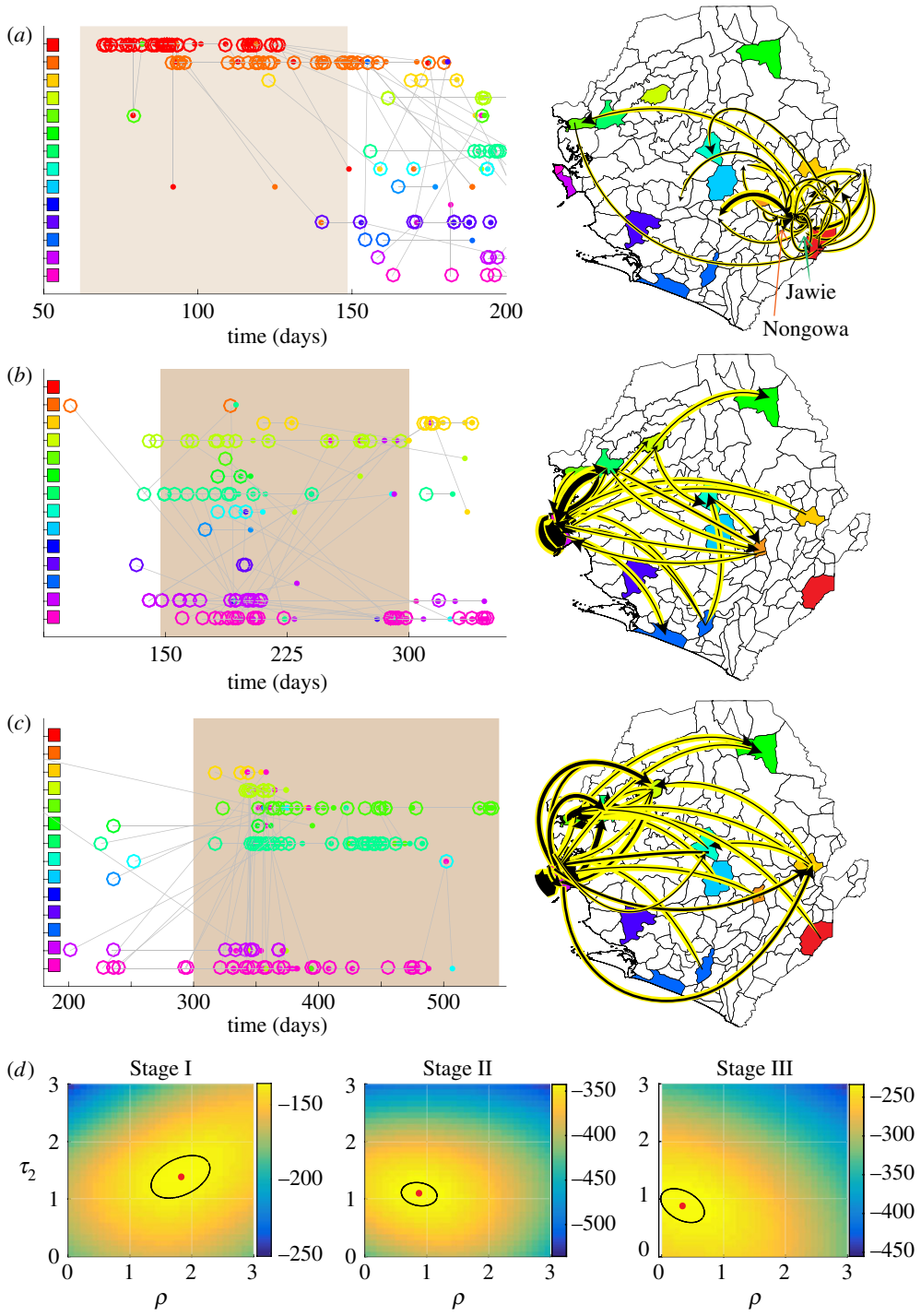

Figure 3. The partially observed transmission network (POTN) and estimation of parameters for the gravity model. (*a*)–(*c*) The left column illustrates the POTN for Stages I–III. The open circles and closed dots represent ancestors and descendants, respectively. Each are coloured by the ancestor district. Linkages of shortest duration are shown. The right column illustrates the POTN linkages on a map with black arrows highlighted in yellow. The black arrow width is proportional to the number of linkages. (*d*) The likelihood evaluation is illustrated for each stage along a grid of values (ρ, τ_2). The MLE and 95% CI are illustrated by a red dot and black ellipse, respectively.

models as chiefdom-level maps of Sierra Leone with Freetown as the origin of a virus genome. The Stage II virus sequence data were consistent with a destination-population gravity model dominated by Freetown linkages.

3.4. Adaptive model selection identifies a change in dynamics

The likelihood ratio helped identify a changing preference of the gravity model over the course of the epidemic. Figure 1*f* illustrates this preference change with 50-day windows. Stage I exhibited a weaker preference for the gravity model than Stage II. The 119 sequences of Stage I came from the work of a single team [36] and included chiefdom localization linking 70% of cases in Stage I to either Jawie chiefdom in Kailahun district or Nongowa chiefdom in Kenema district, shown in figure 3*a*. Most of the linkages occurred in these larger population chiefdoms in the eastern province of Sierra Leone, illustrated on the map of figure 3*a*. The MLE estimate for the population parameter of the gravity model in Stage I found $\tau_2^\star = 1.5 \pm 0.5$; see electronic supplementary material, figure S7 for each stage. The likelihood landscape

also shows a distinct shift to a stronger population dependence and smaller expected transmission distances than found in Stage II, shown in figure 3*d*.

The preference for the gravity model decreased substantially after 300 days. In Stage III, for 261 linkages with descendants in the final 250 days of recorded genomes, the MLE for the gravity model parameters found $\tau_2^\star = 0.8 \pm 0.3$ and $\rho^\star = 0.5 \pm 0.5$. However, the likelihood ratio revealed that the Stage III gravity model was not significantly preferred over the power law whether using the MLE of ($\rho^\star$, $\tau_2^\star$) or the canonical gravity model. Further, a power law model was *weakly preferred* after February 2015 when considering shorter 50-day windows and setting the population parameter $\tau_2 = 1$, as shown in figure 1*f* . From the Stage III map in figure 3*c*, we note that a more diverse scattering of linkages across Sierra Leone supports a qualitative change in the dynamic behaviour of the epidemic.

3.5. Sensitivity analysis of missing chiefdom data

The results in this article are largely consistent regardless of the chiefdoms assigned to genomes with only district-level localization. For sequences with unknown chiefdoms, we selected the median population chiefdom from the known district. Electronic supplementary material, figure S8 illustrates how the likelihood ratio trajectory over the course of the epidemic depended on this assumption. A similar qualitative trend is identified whether choosing the maximum, minimum, median or mean population chiefdom. However, the identified statistical change in model preference from gravity to power law after February 2015 was sensitive to this assumption, especially when using the maximum population chiefdom for each district. In Fang *et al.* [40], a majority of confirmed cases have chiefdom annotations except in the Western Area. Both the confirmed cases and virus genomes recorded during Stage I indicate that most cases in the Kenema district are from highly populated chiefdoms. However, in Stage III, most confirmed cases are in chiefdoms closer to the median population; see electronic supplementary material, table S2 for more details.

4. Discussion

Understanding the changing spatiotemporal dynamics of an emerging epidemic is fundamental to designing real-time disease interventions. Data, gathered from case-contact tracing and molecular diagnostics, can identify individual transmission events that inform population-level models of disease spread. For example, analyses of recent epidemics, including EBOV outbreak in West Africa [31,33,48], the SARS outbreak in 2003 [49] and Middle East Respiratory Syndrome (MERS) outbreaks in 2012 [50], each used detailed individual-level data to infer epidemic parameters and factors influencing large-scale dynamics. Despite the encouraging progress of mathematical modelling and statistical analyses for the 2013–2016 EBOV epidemic [10,33,51,52], the characterization and spatial modelling of the outbreak is incomplete. The ability to rapidly quantify spatiotemporal spread *during* an epidemic would allow for near real-time forecasts and the design of operationally relevant, spatially targeted interventions. The primary contribution of this work is the development of an adaptive framework for analysing epidemics that incorporates detailed transmission information from linked virus genomes to characterize interpretable, population-level spatial models.

Recent advances in phylogenetic reconstruction of transmission networks promise accurate and actionable models of epidemic dynamics [4,27–29,53,54]. Here, we chose the POTN method [8] as an efficient and direct likelihood-based tool to link EBOV cases in space and time. These high-fidelity, space–time couplings between individual cases allowed the parametrization of spatial models describing *disease mobility*, without the need for proxy human mobility data. This framework offers a principled and extensible methodology for investigating the relevant factors for disease mobility.

Our results are largely consistent with other investigations of the spatiotemporal spread of the EBOV epidemic. Previous spatial modelling, with or without virus genome sequences, has concluded that distance, population density and international border closures are covariates that help predict the probability of transmission [31,33]. Other modelling studies have indicated that large population centres, such as Kenema and Port Loko in Sierra Leone, are responsible for initiating self-sustaining local outbreaks [55]. In our investigation, we confirmed that a population-dependent model is preferred when aggregating all transmission events during the epidemic [33].

We have broadened the scope of previous analyses by identifying how the influences of population and distance on the spread of EBOV *change* over the course of an epidemic. A wide variety of probabilistic models can be proposed to describe the stochastic spatial process underlying disease transmission during an epidemic. For this study, we posited two parsimonious models, well known in the epidemiology and ecology literature, to investigate the influence of population and distance on the spatial spread of cases. We discovered that the stochastic propagation of cases is best described by a probabilistic gravity model where dependence on the population and distance varies over the course of the epidemic. The gravity model was preferred in the early part of the epidemic when EBOV was circulating near cities in the east of Sierra Leone. Once the virus migrated to more densely populated areas of the capital area, such as Freetown, Kenema and Port Loko, the gravity model preference became much stronger. During this portion of the epidemic, the transmission events were highly local with a large proportion of linkages staying between large population centres. This observation is also consistent with recent studies of the superspreader phenomenon in the Western Area [52,56].

The probabilistic gravity model can be considered a generalization of a random walk process, weighted by country-specific population distributions. This population influence changed over the course of the EBOV epidemic. In fact, after March 2015, the population dependence diminished significantly. This suggests that EBOV mobility in the last stage of the epidemic can be accurately modelled as a spatial process dependent solely on distance. The MLE of the parameters for the gravity model showed a large uncertainty in the population exponent τ_2. Further, the distance exponent was $\rho < 1$, indicating a higher probability of larger distances between linked cases. In the pure power-law model, the disease mobility during this period has a Lévy flight exponent of order $\alpha = \rho - 1 = 0.6$, suggesting a space-fractional diffusion process. This result is consistent with the observation that confirmed EBOV cases decreased in the large cities of the Western Area

and appeared sporadically in less populated areas far from the Western Area after March 2015. Further, this shift away from a strong preference for a gravity model coincided with an intervention campaign by the government of Sierra Leone called *Operation Western Area Surge* (OWAS) that occurred on 17 December 2014 [57]. Sociological observations, after the OWAS, described an increase in health centre avoidance, return trips to home villages and transmission away from population centres [58]. These results highlight the importance of continual collection of genomic data for characterizing the change in dynamic behaviour along with evaluating the effectiveness of interventions.

Surveillance difficulties during an epidemic pose constraints on our framework being used as a forecasting tool. Despite the EBOV data spanning the entire country and nearly the full time course of the epidemic, the collection of virus genomes was not part of a unified programme. Moreover, the metadata for each sequence, i.e. the global positioning system location and demographical information, are not completely resolved. Uncertainty in reporting due to collection and laboratory processing introduces delays that could impact the utility of predictive spatial models. Our model selection also currently consists of two classes of spatial models: gravity and Lévy flight. We expect to expand our framework to a wider variety of models, but are aware of the challenge in finding parsimonious descriptions of human mobility [59]. As a retrospective study, we have analysed the robustness of our methodology to uncertainties, but inherent difficulties in data collection and modelling will challenge real-time deployment of this tool.

Notwithstanding these limitations, our study can provide operational guidance into the number of collected virus genomes and acceptable time frames required to inform spatial models for prediction. Our model selection technique showed that virus genomes can potentially help characterize the impact of intervention campaigns during an epidemic. Looking towards the next emergence of a dangerous pathogen, molecular diagnostics paired with dynamic models are poised to become a new benchmark for uncovering epidemiological patterns [6], forecasting disease propagation [5] and informing interventions [60] for a wide variety of infectious diseases.

Data accessibility. We have made the source code and data files for our analysis available at https://github.com/kgustafIDM/disease-mobility.

Author's contributions. K.B.G. processed the data, conducted the analyses, made the figures and drafted the manuscript. J.L.P. conceived the study, supervised the analyses, designed the figures and drafted the manuscript. Both the authors gave their final approval for publication.

Competing interests. We declare we have no competing interests.

Funding. The authors thank Bill and Melinda Gates for their active support of this work and their sponsorship through the Global Good Fund.

Acknowledgments. We wish to honour the memory of the healthcare workers, researchers and all people who lost their lives in the 2013–2016 Ebola epidemic. We appreciate helpful conversations with Mike Famulare, Philip Welkhoff, Edward Wenger, Hao Hu, Ben Althouse, Laurent Hébert-Dufresne, Niall Mangan, Gytis Dudas, Jon Wakefield, Dennis Harding and many others.

References

1. Rasmussen DA, Ratmann O, Koelle K. 2011 Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Comput. Biol.* **7**, 1–11. (doi:10.1371/journal.pcbi.1002136)
2. 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)
3. Rasmussen AL, Katze MG. 2016 Genomic signatures of emerging viruses: a new era of systems epidemiology. *Cell Host Microbe* **19**, 611–618. (doi:10.1016/j.chom.2016.04.016)
4. Kühnert D, Stadler T, Vaughan TG, Drummond AJ. 2014 Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth–death sir model. *J. R. Soc. Interface* **11**, 20131106. (doi:10.1098/rsif.2013.1106)
5. Gandon S, Metcalf CJE, Grenfell BT. 2016 Forecasting epidemiological and evolutionary dynamics of infectious diseases. *Trends Ecol. Evol.* **31**, 776–88. (doi:10.1016/j.tree.2016.07.010)
6. Daniels RF *et al.* 2015 Modeling malaria genomics reveals transmission decline and rebound in senegal. *Proc. Natl Acad. Sci. USA* **112**, 7067–7072. (doi:10.1073/pnas.1505691112)
7. Upfill-Brown AM, Lyons HM, Pate MA, Shuaib F, Baig S, Hu H, Eckhoff PA, Chabot-Couture G. 2014 Predictive spatial risk model of poliovirus to aid prioritization and hasten eradication in nigeria. *BMC Med.* **12**, 92. (doi:10.1186/1741-7015-12-92)
8. Famulare M, Hu H. 2015 Extracting transmission networks from phylogeographic data for epidemic and endemic diseases: Ebola virus in Sierra Leone, 2009 H1N1 pandemic influenza and polio in Nigeria. *Int. Health* **7**, 130–138. (doi:10.1093/inthealth/ihv012)
9. Althaus CL. 2014 Estimating the reproduction number of Ebola virus (ebov) during the 2014 outbreak in West Africa. *PLOS Curr. Outbreaks* **6**. (doi:10.1371/currents.outbreaks.91afb5e0f279e7f29e7056095255b288)
10. Camacho A *et al.* 2015 Temporal changes in Ebola transmission in Sierra Leone and implications for control requirements: a real-time modelling study. *PLoS Curr. Outbreaks* **7**. (doi:10.1371/currents.outbreaks.406ae55e83ec0b5193e30856b9235ed2)
11. Gire SK *et al.* 2014 Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369–1372. (doi:10.1126/science.1259657)
12. Carroll MW *et al.* 2015 Temporal and spatial analysis of the 2014–2015 Ebola virus outbreak in West Africa. *Nature* **524**, 97–101. (doi:10.1038/nature14594)
13. Hufnagel L, Brockmann D, Geisel T. 2004 Forecast and control of epidemics in a globalized world. *Proc. Natl Acad. Sci. USA* **101**, 15 124–15 129. (doi:10.1073/pnas.0308344101)
14. Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco JJ, Vespignani A. 2009 Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl Acad. Sci. USA* **106**, 21 484–21 489. (doi:10.1073/pnas.0906910106)
15. Tinbergen J. 1962 *Shaping the world economy; suggestions for an international economic policy*. New York, NY: Twentieth Century Fund.
16. Yingcun X, Bjornstad ON, Grenfell BT. 2004 Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. *Am. Nat.* **164**, 267–281. (doi:10.1086/422341)
17. Viboud C, Bjørnstad ON, Smith DL, Simonsen L, Miller MA, Grenfell BT. 2006 Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science* **312**, 447–451. (doi:10.1126/science.1125237)
18. Meyer S, Held L. 2014 Power-law models for infectious disease spread. *Ann. Appl. Stat.* **8**, 1612–1639. (doi:10.1214/14-AOAS743)
19. Humphries NE, Weimerskirch H, Queiroz N, Southall EJ, Sims DW. 2012 Foraging success of biological lévy flights recorded *in situ*. *Proc. Natl Acad. Sci. USA* **109**, 7169–7174. (doi:10.1073/pnas.1121201109)

20. Carreras BA, Lynch VE, Zaslavsky GM. 2001 Anomalous diffusion and exit time distribution of particle tracers in plasma turbulence model. *Phys. Plasmas* **8**, 5096–5103. (doi:10.1063/1.1416180)
21. Truscott J, Ferguson NM. 2012 Evaluating the adequacy of gravity models as a description of human mobility for epidemic modelling. *PLoS Comput. Biol.* **8**, e1002699. (doi:10.1371/journal.pcbi.1002699)
22. Wesolowski A, Qureshi T, Boni MF, Sundsøy PR, Johansson MA, Rasheed SB, Eng-Monsen K, Buckee CO. 2015 Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proc. Natl Acad. Sci. USA* **112**, 11 887–11 892. (doi:10.1073/pnas.1504964112)
23. Ratmann O, Donker G, Meijer A, Fraser C, Koelle K. 2012 Phylodynamic inference and model assessment with approximate Bayesian computation: influenza as a case study. *PLoS. Comput. Biol.* **8**, 1–14. (doi:10.1371/journal.pcbi.1002835)
24. Salje H *et al.* 2017 Dengue diversity across spatial and temporal scales: local structure and the effect of host population size. *Science* **355**, 1302–1306. (doi:10.1126/science.aaj9384)
25. Rasmussen DA, Boni MF, Koelle K. 2013 Reconciling phylodynamics with epidemiology: the case of dengue virus in southern vietnam. *Mol. Biol. Evol.* **31**, 258–271. (doi:10.1093/molbev/mst203)
26. Holmes EC, Nee S, Rambaut A, Garnett GP, Harvey PH. 1995 Revealing the history of infectious disease epidemics through phylogenetic trees. *Phil. Trans. R. Soc. Lond. B* **349**, 33–40. (doi:10.1098/rstb.1995.0088)
27. 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)
28. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. 2014 Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput. Biol.* **10**, 1–14. (doi:10.1371/journal.pcbi.1003457)
29. Didelot X, Fraser C, Gardy J, Colijn C. 2017 Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Mol. Biol. Evol.* **34**, 997–1007. (doi:10.1093/molbev/msw275)
30. Rainisch G, Shankar MB, Wellman M, Merlin T, Meltzer MI. 2015 Regional spread of Ebola virus, West Africa, 2014. *Emerg. Infect. Dis.* **21**, 444–447. (doi:10.3201/eid2103.141845)
31. Kramer AM, Tomlin Pulliam J, Alexander LW, Park AW, Rohani P, Drake JM. 2016 Spatial spread of the West Africa Ebola epidemic. *R. Soc. open. sci.* **3**, 160294. (doi:10.1098/rsos.160294)
32. Backer JA, Wallinga J. 2016 Spatiotemporal analysis of the 2014 Ebola epidemic in West Africa. *PLoS Comput. Biol.* **12**, 1–17. (doi:10.1371/journal.pcbi.1005210)
33. Dudas G *et al.* 2017 Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* **544**, 309–315. (doi:10.1038/nature22040).
34. Scarpino SV *et al.* 2015 Epidemiological and viral genomic sequence analysis of the 2014 Ebola outbreak reveals clustered transmission. *Clin. Infect. Dis.* **60**, 1079–1082. (doi:10.1093/cid/ciu1131)
35. Lau MSY, Dalziel BD, Funk S, McClelland A, Tiffany A, Riley S, Metcalf CJE, Grenfell BT. 2017 Spatial and temporal dynamics of superspreading events in the 2014-2015 West Africa Ebola epidemic. *Proc. Natl Acad. Sci. USA* **114**, 2337–2342. (doi:10.1073/pnas.1614595114)
36. Park DJ *et al.* 2015 Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. *Cell* **161**, 1516–1526. (doi:10.1016/j.cell.2015.06.007)
37. Tong Y-G *et al.* 2015 Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. *Nature* **524**, 93–96. (doi:10.1038/nature14490)
38. Smits SL *et al.* 2015 Genotypic anomaly in Ebola virus strains circulating in Magazine Wharf area, Freetown, Sierra Leone, 2015. *Eurosurveillance* **20**, 30035. (doi:10.2807/1560-7917.ES.2015.20.40.30035)
39. Arias A *et al.* 2016 Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases. *Virus Evol.* **2**, vew016. (doi:10.1093/ve/vew016)
40. Fang L-Q *et al.* 2016 Transmission dynamics of Ebola virus disease and intervention effectiveness in Sierra Leone. *Proc. Natl Acad. Sci. USA* **113**, 4488–4493. (doi:10.1073/pnas.1518587113)
41. Pybus OG *et al.* 2012 Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973. (doi:10.1093/molbev/mss075)
42. Clauset A, Shalizi CR, Newman MEJ. 2009 Power-law distributions in empirical data. *SIAM Rev.* **51**, 661–703. (doi:10.1137/070710111)
43. Mainardi F, Luchko Y, Gianni P. 2001 The fundamental solution of the space-time fractional diffusion equation. *Fractional Calc. Appl. Anal.* **4**, 153–192.
44. Fisher RA. 1959 *Statistical methods and scientific inference*. London, UK: Oliver and Boyd.
45. Wilks SS. 1938 The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* **9**, 60–62. (doi:10.1214/aoms/1177732360)
46. Akaike H. 1974 A new look at the statistical model identification. *IEEE. Trans. Automat. Contr.* **19**, 716–723. (doi:10.1109/TAC.1974.1100705)
47. Van Kerkhove MD, Bento AI, Mills HL, Ferguson NM, Donnelly CA. 2015 A review of epidemiological parameters from Ebola outbreaks to inform early public health decision-making. *Sci. Data.* **2**, 150019. (doi:10.1038/sdata.2015.19)
48. Team, WHO Ebola Response. 2016 After Ebola in West Africa: unpredictable risks, preventable epidemics. *N. Engl. J. Med.* **2016**, 587–596. (doi:10.1056/NEJMsr1513109)
49. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. 2005 Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359. (doi:10.1038/nature04153)
50. Cauchemez S, Fraser C, Van Kerkhove MD, Donnelly CA, Riley S, Rambaut A, Enouf V, van der Werf S, Ferguson NM. 2014 Middle east respiratory syndrome coronavirus: quantification of the extent of the epidemic. *Lancet Infect. Dis.* **14**, 50–56. (doi:10.1016/S1473-3099(13)70304-9)
51. Fisman D, Khoo E, Tuite A. 2014 Early epidemic dynamics of the West African 2014 Ebola outbreak: estimates derived with a simple two-parameter model. *PLoS Curr. Outbreaks* **6**. (doi:10.1371/currents.outbreaks.89c0d3783f36958d96ebbae97348d571)
52. Lau MSY, Dalziel BD, Funk S, McClelland A, Tiffany A, Riley S, Metcalf CJE, Grenfell BT. 2017 Spatial and temporal dynamics of superspreading events in the 2014–2015 West Africa Ebola epidemic. *Proc. Natl Acad. Sci. USA* **114**, 2337–2342. (doi:10.1073/pnas.1614595114)
53. Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. 2013 Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and Hepatitis C. Virus (hcv). *Proc. Natl Acad. Sci. USA* **110**, 228–233. (doi:10.1073/pnas.1207965110)
54. Rasmussen DA, Volz EM, Koelle K. 2014 Phylodynamic inference for structured epidemiological models. *PLoS Comput. Biol.* **10**, 1–16. (doi:10.1371/journal.pcbi.1003570)
55. Yang W *et al.* 2015 Transmission network of the 2014–2015 Ebola epidemic in Sierra Leone. *J. R. Soc. Interface* **12**, 20150536. (doi:10.1098/rsif.2015.0536)
56. Althaus CL. 2015 Ebola superspreading. *Lancet Infect. Dis.* **15**, 507–508. (doi:10.1016/S1473-3099(15)70135-0)
57. USAID/CDC. 2014 West Africa Ebola outbreak Fact Sheet.
58. Richards P, Amara J, Ferme MC, Kamara P, Mokuwa E, Sheriff AI, Suluku R, Voors M. 2015 Social pathways for Ebola virus disease in rural Sierra Leone, and some implications for containment. *PLoS Negl. Trop. Dis.* **9**, e0003567. (doi:10.1371/journal.pntd.0003567)
59. Wesolowski A, O'Meara WP, Eagle N, Tatem AJ, Buckee CO. 2015 Evaluating spatial interaction models for regional mobility in sub-saharan africa. *PLoS Comput. Biol.* **11**, 1–16. (doi:10.1371/journal.pcbi.1004267)
60. Cori A *et al.* 2017 Key data for outbreak evaluation: building on the Ebola experience. *Phil. Trans. R. Soc. Lond. B* **372**, 20160371. (doi:10.1098/rstb.2016.0371)