ARTICLE

doi:10.1038/nature22040

Virus genomes reveal factors that spread and sustained the Ebola epidemic

A list of authors and their affiliations appears at the end of the paper

The 2013–2016 West African epidemic caused by the Ebola virus was of unprecedented magnitude, duration and impact. Here we reconstruct the dispersal, proliferation and decline of Ebola virus throughout the region by analysing 1,610 Ebola virus genomes, which represent over 5% of the known cases. We test the association of geography, climate and demography with viral movement among administrative regions, inferring a classic 'gravity' model, with intense dispersal between larger and closer populations. Despite attenuation of international dispersal after border closures, cross-border transmission had already sown the seeds for an international epidemic, rendering these measures ineffective at curbing the epidemic. We address why the epidemic did not spread into neighbouring countries, showing that these countries were susceptible to substantial outbreaks but at lower risk of introductions. Finally, we reveal that this large epidemic was a heterogeneous and spatially dissociated collection of transmission clusters of varying size, duration and connectivity. These insights will help to inform interventions in future epidemics.

At least 28,646 cases and 11,323 deaths[1] have been attributed to the Makona variant of Ebola virus (EBOV)[2] in the two and a half years it circulated in West Africa. The epidemic is thought to have begun in December 2013 in Guinea, but was not detected and reported until March 2014 (ref. 3). Initial efforts to control the outbreak in Guinea were considered to be succeeding[4], but in early 2014 the virus crossed international borders into the neighbouring countries Liberia (where the first cases were diagnosed in late March) and Sierra Leone (first documented case in late February[5,6], first diagnosed cases in May[7]). EBOV genomes sequenced from three patients in Guinea early in the epidemic[3] demonstrated that the progenitor of the Makona variant originated in Middle Africa and arrived in West Africa within the last 15 years[7,8]. Rapid sequencing from the first reported cases in Sierra Leone confirmed that EBOV had crossed the border from Guinea and that these cases were not the result of an independent zoonotic introduction[7]. Subsequent studies have analysed the genetic makeup of the Makona variant, focusing on Guinea[9,10,13], Sierra Leone[14,15] or Liberia[16,17], and have identified local viral lineages and transmission patterns within each country.

Although virus sequencing data have covered considerable fractions of the epidemic in each affected country, individual studies focused on either limited geographical areas or time periods, so that the regional level patterns and drivers of the epidemic across its entire duration have remained uncertain. Using 1,610 genome sequences collected throughout the epidemic, representing over 5% of recorded Ebola virus disease (EVD) cases (Extended Data Fig. 1), we reconstruct a detailed phylogenetic history of the movement of EBOV within and between the three most affected countries. Using a recently developed phylogeographic approach that integrates covariates of spatial spread[18], we test which features of each region (administrative, economic, climatic, infrastructural or demographic factors) were important in shaping the spatial dynamics of EVD. We also examine the effectiveness of international border closures on controlling virus dissemination. Finally, we investigate why regions that immediately border the most affected countries did not develop protracted outbreaks similar to those that ravaged Sierra Leone, Guinea and Liberia.

Origin, ignition and trajectory of the epidemic

Molecular clock dating indicates that the most recent common ancestor of the epidemic existed between December 2013 and February 2014 (mean, 22 Jan 2014; 95% credible interval (CI), 16 Dec 2013–20 Feb 2014) and phylogeographic estimation assigns this ancestor to the Guéckédou prefecture, Nzérékoré region, Guinea, with high credibility (Fig. 1). In addition, we find that initial EBOV lineages that were derived from this common ancestor circulated among the Guéckédou prefecture and its neighbouring prefectures of Macenta and Kissidougou until late February 2014 (Fig. 1). These results support the epidemiological evidence that the West African epidemic began in late 2013 in Guéckédou prefecture[3].

The first EBOV introduction from Guinea into another country that resulted in sustained transmission is estimated to have occurred in early April 2014 (Fig. 1), when the virus spread to the Kailahun district of Sierra Leone[5,6]. This lineage was first detected in Kailahun at the end of May 2014, from where it spread across the region (Figs 1, 2 and Extended Data Fig. 2). From Kailahun, EBOV spread very rapidly in May 2014 into several counties of Liberia (Lofa, Montserrado and Margibi)[17] and Guinea (Conakry, back into Guéckédou)[9,13]. The virus continued to spread westwards through Sierra Leone, and by July 2014 EBOV was present in the capital city, Freetown.

By mid-September 2014, Liberia was reporting more than 500 new EVD cases per week, mostly driven by a large outbreak in Montserrado county, which encompasses the capital city, Monrovia. Sierra Leone reported more than 700 new cases per week by mid-November, with large outbreaks in Port Loko, Western Urban (Freetown) and Western Rural districts (Freetown suburbs). December 2014 brought the first signs that efforts to control the epidemic in Sierra Leone were effective, as EVD incidence began to drop. By March 2015, the epidemic was largely under control in Liberia and eastern Guinea, although sustained transmission continued in the border area of western Guinea and western Sierra Leone. By the following month, prevalence had declined such that only a handful of lineages persisted[10,14] (Fig. 2).

The last EBOV genome obtained from a conventionally acquired infection was collected and sequenced in October 2015 in Forécariah prefecture (Guinea)[10]. After this, only sporadic cases of EVD were detected: in Montserrado (Liberia) in November 2015, Tonkolili (Sierra Leone) in January and February 2016, and Nzérékoré (Guinea) in March 2016. All these sporadic cases probably resulted from transmission from EVD survivors with established, persistent infections[11,12,14].

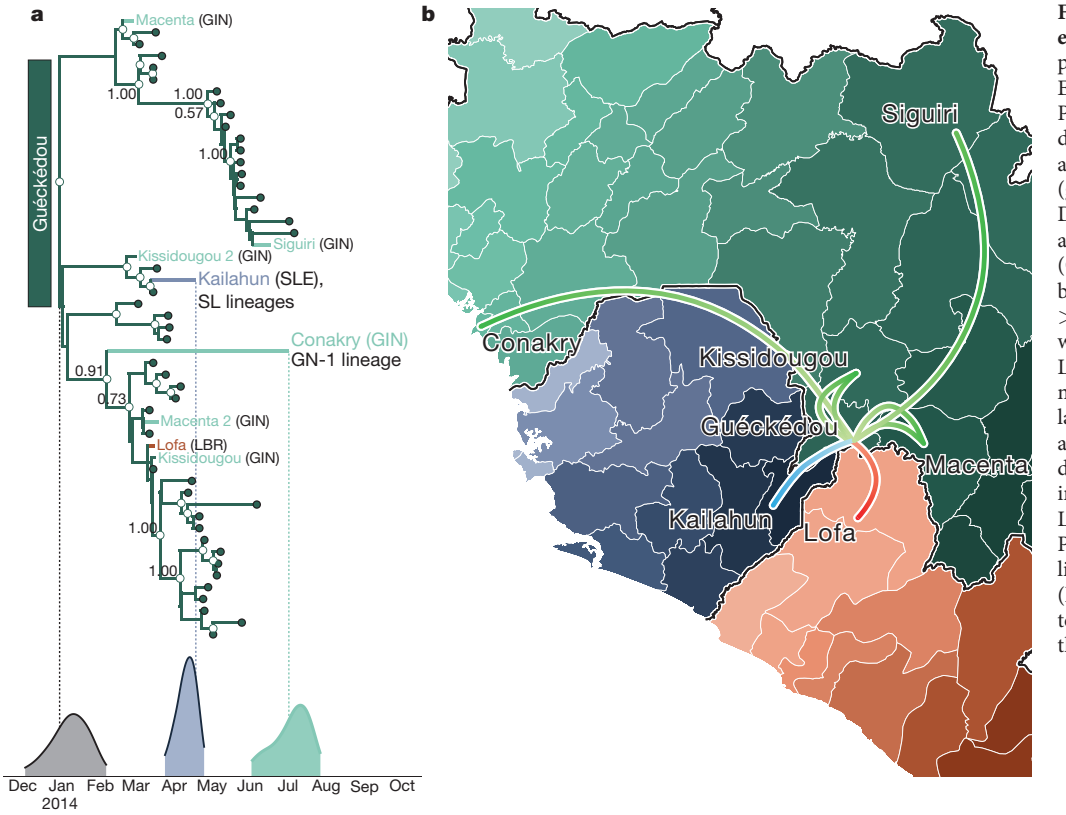


Figure 1 | Summary of early epidemic events. a, Temporal phylogeny of earliest sampled EBOV lineages in Guéckédou Prefecture, Guinea. 95% posterior densities of most recent common ancestor estimates for all lineages (grey) and lineages into Kailahun District, Sierra Leone (SLE; blue) and to Conakry Prefecture, Guinea (GIN; green) are shown at the bottom. Posterior probabilities >0.5 are shown for lineages with >5 descendent sequences. LBR, Liberia. **b**, Dispersal events marked by coloured lineages and labelled by name on the phylogeny are projected on a map with directionality indicated by colour intensity (from light to dark). Lineages that migrated to Conakry Prefecture (labelled as GN-1 lineage) and Kailahun District (labelled as SL lineages) have led to the vast majority of EVD cases throughout the region.

Factors associated with EBOV dispersal

To determine the factors that influenced the spread of EBOV among administrative regions at the district (Sierra Leone), prefecture (Guinea) and county (Liberia) levels, we used a phylogeographic generalized linear model (GLM)[18]. Of the 25 factors assessed (see Extended Data Table 1 for a full list and description), 5 were included in the model with categorical support (Table 1). In summary, EBOV tends to disperse between geographically close regions (great circle distance, Bayes factor (BF) support for inclusion: BF > 50). Half of all virus dispersals occurred between locations less than 72 km apart and only 5% involved movement over 232 km (Fig. 3a). Both origin and destination population sizes are very strongly (BF > 50) positively correlated with viral dissemination, with a stronger effect for origin population size. The positive effect of population sizes combined with the inverse effect of the geographic distance implies that the spread of the epidemic followed a classic gravity-model dynamic. Gravity models, widely used in economic and geographic studies and a natural choice for modelling infectious disease transmission[19–21], describe the movement of people between locations as a function of their population sizes and the distance that separates them. Here we use viral genomes to provide empirical evidence that such a process drove viral dissemination during the EVD epidemic.

In addition to geographical distance, we found a significant propensity for virus dispersal to occur within each country, relative to international dispersal (nat./int. effect, BF > 50), suggesting that country borders acted to curb the geographic spread of EBOV. When international dispersals do take place, they are more intense between administrative regions that are adjacent at an international border (IntBoSh, BF > 50).

We tested whether sharing of any of 17 vernacular languages explains virus spread, as common languages might reflect cultural links, including between non-contiguous or international regions, but we found no evidence that such linguistic links were correlated with EBOV spread. A variety of other possible predictors of EBOV transmission, such as aspects of urbanization (economic output, population density, travelling times to large settlements) as well as climatic effects, were not significantly associated with virus dispersal. However, these factors may have contributed to the size and longevity of transmission chains after introduction to a region (see below).

Finally, to investigate the potential of 'real-time' viral genome sequencing, we considered the degree to which the findings could have been obtained at the height of the epidemic, had sequences been available shortly after the samples were taken (see Methods for details). For the factors associated with EBOV dispersal, the results were highly comparable to those for the full dataset whereby the same five factors were strongly supported and these had similar effect sizes (Extended Data Fig. 3).

Factors associated with local EBOV proliferation

The analysis above identified predominantly geographical and administrative factors that predict the degree of importation risk, that is, the likelihood that a viral lineage initiates at least one infection in a new region. However, the epidemiological consequences of each introduction—the size and duration of resulting transmission chains—may be affected by different factors. Therefore, we investigated which demographic, economic and climatic factors might predict cumulative case counts[1] for each region (Bayesian GLM; see Methods) and found that these were associated with factors related to urbanization (Table 2): primarily population sizes (PopSize, BF = 29.6) and a significant inverse association with travel times to the nearest settlement with more than 50,000 inhabitants (TT50K, BF = 32.4). These results confirm the common perception that, in contrast to previous EVD outbreaks, widespread transmission within urban regions in West Africa was a major contributing factor to the scale of the epidemic of the Makona variant.

As the epidemic in West Africa progressed, there were fears that increased rainfall and humidity might prolong the environmental persistence of EBOV particles, increasing the likelihood of transmission[22]. Although we found no evidence of an association between EBOV dispersal and any aspects of local climate, we find that regions with less

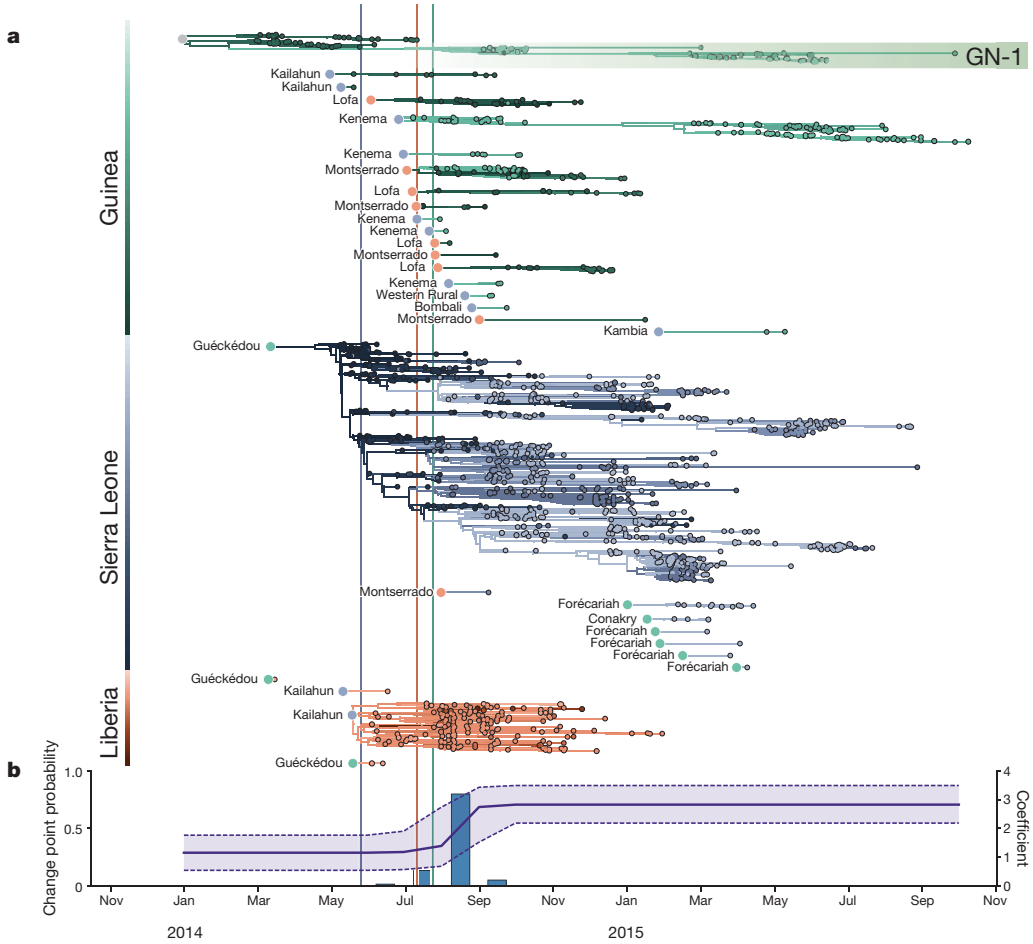


Figure 2 | Transmission chains arising from independent international movements. **a**, EBOV lineages by country (Guinea, green; Sierra Leone, blue; Liberia, red), tracked until the sampling date of their last known descendants. Circles at the roots of each subtree denote the country of origin for the introduced lineage. **b**, Estimates of the change point probability (left *y* axis) and log coefficient (mean and credible interval; right *y* axis) for the nat./int. factor. Vertical lines represent dates that border closures were announced by the respective countries.

seasonal variation in temperature, and with more rainfall, tended to have larger EVD outbreaks (TempSS, BF > 50 and Precip, BF = 4.4, respectively).

The impact of international travel restrictions

Porous borders between Liberia, Sierra Leone and Guinea may have allowed the unimpeded EBOV spread during the 2013–2016 epidemic[23–25]. Our results indicate that international borders were associated with a decreased rate of transmission events compared to national borders (Extended Data Fig. 4), but that frequent international cross-border transmission events still occurred. These events were concentrated in the Guéckédou prefecture (Guinea), Kailahun district (Sierra Leone) and Lofa county (Liberia) during the early stages of the epidemic (Extended Data Fig. 5a), and between the Forécariah prefecture (Guinea) and Kambia district (Sierra Leone) at the later stage (Extended Data Fig. 5b). These later EBOV movements hindered efforts to interrupt the final chains of transmission in late 2015, with EBOV from these chains moving back and forth across this border[10,14,26]. Sierra Leone announced border closures on 11 June 2014, followed by Liberia on 27 July 2014, and Guinea on 9 August 2014, but little information is available about what these border closures actually entailed. Although we show that the relative contribution of international spread to overall viral migration was lower after country borders were closed (mean nat./int. coefficient increasing from 1.15 to 2.83 between August and September 2014; 80.0% posterior support; (Fig. 2b)), it is difficult to ascertain whether the border closures themselves were responsible for the apparent reduction in cross-border transmissions, as opposed to concomitant control efforts or public information campaigns. However, even if border closures reduced international traffic, particularly over longer distances and between larger population centres, by the time that Sierra Leone and Liberia had closed their borders, the epidemic had become firmly established in both countries.

Table 1 | Summary of phylogenetic generalized linear model results

Predictor*	Description	Coefficient†	95% CI‡	Inclusion§	BF‖
Nat./int.	National dispersal relative to international	3.07	2.36, 3.77	1.0	>50
Distances	Great circle distances between the locations' population centroids¶	−0.77	−0.91, −0.63	1.0	>50
OrPop	Population size at the location of origin	1.36	0.86, 1.84	1.0	>50
DestPop	Population size at the destination location	0.74	0.43, 1.06	1.0	>50
IntBoSh	Two locations share an international border	3.39	2.42, 4.33	1.0	>50
OrTempSS	Index of temperature seasonality at origin	−0.47	−0.88, −0.11	0.1	3.79

*Predictors included in the model with Bayes factor >3.
†Mean coefficient.
‡95% highest posterior density credible interval (CI).
§Probability that the predictor was included in the model.
‖BF, Bayes factor.
¶Population centroids indicate the centre of a location weighted by population.

Why did the epidemic not spread further?

A few EBOV exportations were documented from Guinea by road transport into Mali and Senegal[27,28] and by air from Liberia to Nigeria and the USA[29,30]. However, apart from these limited exceptions, the West African Ebola virus epidemic did not spread into the neighbouring regions of Côte d'Ivoire, Guinea-Bissau, Mali and Senegal. By extending our GLM (the supported predictors and their estimated coefficients) to include these regions we were able to address whether

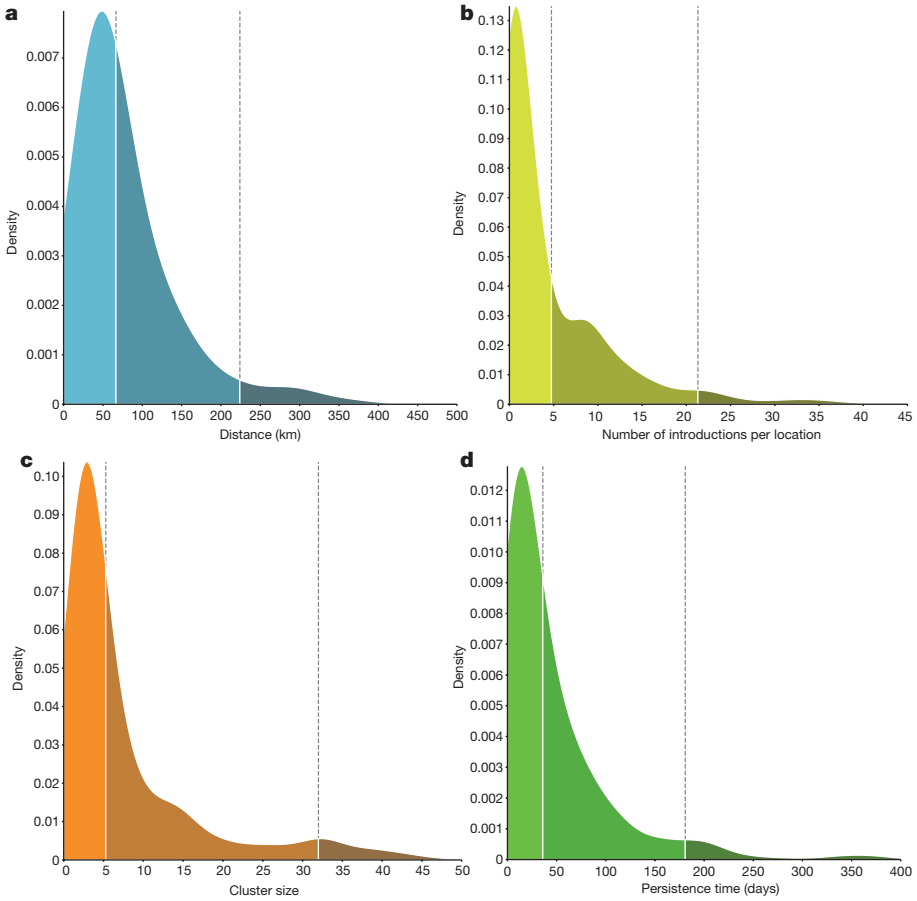


Figure 3 | The metapopulation structure of the epidemic. a, Kernel density estimate of distances associated with inferred EBOV dispersal events: 50% occur over distances <72 km and <5% occur over distances >232 km. **b**, Kernel density estimate of the number of independent EBOV introductions into each administrative region: 50% have fewer than 4.8 and <5% greater than 21.3. **c**, Kernel density estimate of the mean size of sampled cases resulting from each introduction with at least 2 sampled cases: 50% <5.3 cases, 95% <32 cases. **d**, Kernel density estimate of the persistence of clusters in days (from time of introduction to time of the last sampled case): 50% <36 days, 95% <181 days. **a**–**d**, 50% and 95% are indicated by the dashed lines.

these regions were spared EVD cases through good fortune, or because they were associated with an inherently lower risk of EBOV spread and transmission. We estimated the degree to which these, apparently EVD-free, regions had the potential to be exposed to viral introductions from affected regions (see Methods).

Overall, the contiguous regions in unaffected neighbouring countries were all predicted to have low numbers of EBOV introductions (Fig. 4a and Extended Data Fig. 6a) based on the phylogeographic history of the sampled cases. They were not, however, predicted to have particularly low levels of transmission if an outbreak had started (Fig. 4b and Extended Data Fig. 6b). Therefore, it is likely that some of these regions were at risk of becoming part of the EVD epidemic, but that their geographical distance from areas of active transmission and the attenuating effect of international borders prevented this from occurring. The Kati cercle in Mali and Tonkpi region in Côte d'Ivoire are to some extent exceptions to this general result, as these were more susceptible to EBOV introductions under the gravity model because of their large populations (1 million and 950,000, respectively) (Fig. 4a), and are predicted to have experienced many cases had EVD become established (Fig. 4b).

Table 2 | Summary of generalized linear model results with case counts as the response variable

Predictor*	Description	Coefficient†	95% CI‡	Inclusion§	BF‖
TempSS	Temperature seasonality	−1.1	−1.6, −0.5	0.83	>50
TT50K	Time to travel to a population centre of 50,000 people	−0.9	−1.4, −0.4	0.62	32.4
PopSize	Population size	0.9	0.3, 1.6	0.60	29.6
Precip	Precipitation	0.8	0.2, 1.3	0.18	4.4
TT100K	Time to travel to a population centre of 0.1 million people	−0.8	−1.7, −0.1	0.16	3.8

*Predictors included in the model with Bayes factor >3.
†Mean coefficient.
‡95% highest posterior density credible interval (CI).
§Probability that the predictor was included in the model.
‖BF, Bayes factor.

Metapopulation structure and dynamics of the epidemic

After the initial establishment of transmission in Sierra Leone and Liberia, Guinea experienced repeated reintroductions of viral lineages from disparate transmission chains from both countries (Fig. 2). Our analysis reveals that there were at least 21 (95% CI, 16–25) reintroductions into Guinea from April 2014 to February 2015. An early epidemic lineage was established around the Guinean capital, Conakry, and persisted for the duration of the epidemic (GN-1 in Figs 1, 2). However, the continual reintroduction of EBOV into Guinea without a clear peak in transmission suggests that the virus may have been failing to maintain transmission elsewhere. There were also numerous introductions into Sierra Leone over a similar time period (median, 9; 95% CI, 6–12), but the resulting transmission chains constituted a very small proportion of the country's EVD cases, with the bulk of transmission resulting from one early introduction (Fig. 2a).

In all three countries, repeated introductions into administrative regions seems to have been a large factor in the longevity of the EVD epidemic (Extended Data Fig. 7). As such, regional case numbers were generally the result of multiple overlapping introduction events followed by within-region spread and occasional onward transmission to other regions. This suggests a metapopulation model in which the persistence of the epidemic was driven by introduction into novel contact networks rather than by mass-action transmission, such as susceptible-infectious-removed dynamics[31,32]. We found that, on average, EBOV migrates between administrative regions at a rate of

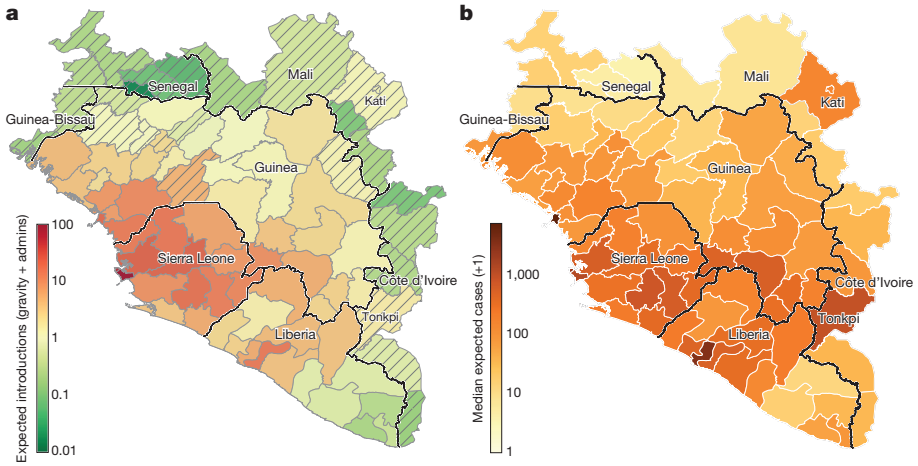


Figure 4 | Predicted destinations and consequences of viral dispersal. a, Predicted number of EBOV imports into each of the 63 regions in Guinea, Sierra Leone and Liberia (including 7 without recorded cases in Guinea) and the surrounding 18 regions of the neighbouring countries of Guinea-Bissau, Senegal, Mali and Côte d'Ivoire. The expected number of EBOV exports from locations in the phylogeographic tree and imports to any location were calculated on the basis of the phylogeographic GLM model estimates and associated predictors that were extended to apparently EVD-free locations (see Supplementary Methods). **b**, Predicted EVD cluster sizes from the Bayesian GLM fitted to case data.

0.85 events per lineage per year (95% CI, 0.72–0.97). Assuming a serial interval of 15.3 days[33], this rate translates to a 3.6% chance (95% CI, 3.0–4.1%) that over the course of a single infection, the transmission chain moved between regions. Given the key role that virus dispersal played in sustaining the epidemic, the detection and isolation of these relatively low proportions of mobile cases may have a disproportionate effect on the control of an EVD epidemic.

From our spatial phylogenetic model we conclude that many regions experienced numerous independent EBOV introductions (Fig. 3b). However, these introductions gave rise to clusters of cases that were generally small (a mean cluster size of 4.3 and only 5% larger than 17 in our sample; Fig. 3c) and of limited duration (a mean persistence time of 41.3 days with only 5% greater than 181 days; Fig. 3d). Here, we define a cluster as a group of sequenced cases in a region that derive from a single introduction event and define persistence as the time between the introduction event and the last sampled case in the cluster. These definitions are conservative regarding sampling intensity, as we expect additional samples would have split clusters apart rather than join them. Furthermore, introductions that were not detected will be disproportionately smaller, and so the cluster size estimate will be biased upwards. Segregating these observations by country (Extended Data Fig. 8 (left)) shows that districts of Sierra Leone had more introductions and that Guinea generally had smaller clusters, but that persistence was similar between the three countries. A comparison between introductions that occurred before October 2014 and those that occurred after this date shows that the number of introductions per location was comparable, whereas those that occurred early generally resulted in larger and more persistent clusters (Extended Data Fig. 8 (right)).

Therefore, with 5.8% sampling, we arrive at a conservative estimate of approximately 75 regional cases per introduction event. Although larger population centres, in particular capital cities, generally experienced more introductions (Extended Data Fig. 9a), the cluster sizes are less strongly associated with population size (Extended Data Fig. 9b), further highlighting the role of virus movement into urban areas as major factor for the high caseloads in large population centres. Frequent cluster extinction, despite a small fraction of individuals being infected, suggests that individual outbreaks were constrained by the degree of connectedness among contact networks. Thus, it appears that the West African EVD epidemic was sustained by frequent introductions that resulted in numerous small local clusters of cases, some of which went on to further seed clusters in other locations.

Viral genomics as a tool for outbreak response

The 2013–2016 EVD epidemic in West Africa has unfortunately become a costly lesson in addressing an infectious disease outbreak in the absence of preparedness of both the exposed population and the international community. Our work demonstrates the value of pathogen genome sequencing in a public healthcare emergency and the value of timely pre-publication data sharing to identify the origins of imported disease case clusters, to track pathogen transmission as the epidemic progresses, and to follow up on individual cases as the epidemic subsides.

It is inevitable that as sequencing costs decrease, accuracy increases and sequencing instruments become more portable, real-time viral surveillance and molecular epidemiology will be routinely deployed on the front lines of infectious disease outbreaks[10,14,16,34–36]. Although we have shown here that the broad pattern of EBOV spatial movement was discernible from virus genomes derived from samples collected up until October 2014 only, there was a notable hiatus in sequencing at this time[35] and the genomes in the present dataset from that time were sequenced retrospectively from archived material. The West African EVD epidemic has demonstrated that a steady sequencing pace[34–36], local sequencing capacity[10,14,16] and rapid dissemination of data[7] are key requirements in generating actionable sequence data from an infectious disease outbreak. However, as viral genome sequencing is scaled up and approaches the timescale of viral evolution, the analysis techniques will increasingly represent the bottleneck for timely communication of information for an outbreak response.

The analysis of the comprehensive EBOV genome set that was collected during the 2013–2016 EVD epidemic, including the findings presented here and in other studies[7,9,13–17,37,38], provides a framework for predicting the behaviour of future disease outbreaks caused by EBOV, other filoviruses and perhaps other human pathogens. However, many questions remain about the biology of EBOV. As sustained human-to-human transmission waned, West Africa experienced several instances of recrudescent transmission, often in regions that had not seen cases for many months as a result of persistent sub-clinical infections[11,12,39]. Although, in hindsight, such sequelae were not entirely unexpected[40], the magnitude of the 2013–2016 epidemic has put the region at ongoing risk of sporadic EVD re-emergence. Similarly, the nature of the reservoir of EBOV, and its geographic distribution, remain as fundamental gaps in our knowledge. Resolving these questions is critical to predicting the risk of zoonotic transmission and therefore of future EVD outbreaks.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 31 August 2016; accepted 2 March 2017.
Published online 12 April 2017.

1. World Health Organization. *Ebola Situation Report—10 June 2016* http://apps.who.int/iris/bitstream/10665/208883/1/ebolasitrep_10Jun2016_eng.pdf (2016).
2. Kuhn, J. H. *et al.* Nomenclature- and database-compatible names for the two Ebola virus variants that emerged in Guinea and the Democratic Republic of the Congo in 2014. *Viruses* **6,** 4760–4799 (2014).

3. Baize, S. *et al.* Emergence of Zaire Ebola virus disease in Guinea. *N. Engl. J. Med.* **371,** 1418–1425 (2014).
4. World Health Organization Regional Office for Africa. *Ebola Virus Disease, West Africa (situation as of 25 April 2014)* http://www.afro.who.int/en/clusters-a-programmes/dpc/epidemic-a-pandemic-alert-and-response/4121-ebola-virus-disease-west-africa-25-april-2014.html (2014).
5. Goba, A. *et al.* An outbreak of Ebola virus disease in the Lassa fever zone. *J. Infect. Dis.* **214,** S110–S121 (2016).
6. Sack, K., Fink, S., Belluck, P., Nossiter, A. & Berehulak, D. *How Ebola roared back* http://nyti.ms/1wwG5VX (2014).
7. Gire, S. K. *et al.* Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345,** 1369–1372 (2014).
8. Dudas, G. & Rambaut, A. Phylogenetic analysis of Guinea 2014 EBOV Ebolavirus outbreak. *PLoS Curr.* **6,** http://dx.doi.org/10.1371/currents.outbreaks.84eefe5ce43ec9dc0bf0670f7b8b417d (2014).
9. Carroll, M. W. *et al.* Temporal and spatial analysis of the 2014–2015 Ebola virus outbreak in West Africa. *Nature* **524,** 97–101 (2015).
10. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530,** 228–232 (2016).
11. Blackley, D. J. *et al.* Reduced evolutionary rate in reemerged Ebola virus transmission chains. *Sci. Adv.* **2,** e1600378 (2016).
12. Mate, S. E. *et al.* Molecular evidence of sexual transmission of Ebola virus. *N. Engl. J. Med.* **373,** 2448–2454 (2015).
13. Simon-Loriere, E. *et al.* Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic. *Nature* **524,** 102–104 (2015).
14. 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).
15. Park, D. J. *et al.* Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. *Cell* **161,** 1516–1526 (2015).
16. 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).
17. Ladner, J. T. *et al.* Evolution and spread of Ebola virus in Liberia, 2014–2015. *Cell Host Microbe* **18,** 659–669 (2015).
18. Lemey, P. *et al.* Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS Pathog.* **10,** e1003932 (2014).
19. Viboud, C. *et al.* Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science* **312,** 447–451 (2006).
20. Truscott, J. & Ferguson, N. M. Evaluating the adequacy of gravity models as a description of human mobility for epidemic modelling. *PLOS Comput. Biol.* **8,** e1002699 (2012).
21. Yang, W. *et al.* Transmission network of the 2014–2015 Ebola epidemic in Sierra Leone. *J. R. Soc. Interface* **12,** 20150536 (2015).
22. Fischer, R. *et al.* Ebola virus stability on surfaces and in fluids in simulated outbreak environments. *Emerg. Infect. Dis.* **21,** 1243–1246 (2015).
23. Bausch, D. G. & Schwarz, L. Outbreak of Ebola virus disease in Guinea: where ecology meets economy. *PLoS Negl. Trop. Dis.* **8,** e3056 (2014).
24. Chan, M. Ebola virus disease in West Africa—no early end to the outbreak. *N. Engl. J. Med.* **371,** 1183–1185 (2014).
25. Wesolowski, A. *et al.* Commentary: containing the Ebola outbreak—the potential and challenge of mobile network data. *PLoS Curr.* **6,** http://dx.doi.org/10.1371/currents.outbreaks.0177e7fcf52217b8b634376e2f3efc5e (2014).
26. Goodfellow, I., Reusken, C. & Koopmans, M. Laboratory support during and after the Ebola virus endgame: towards a sustained laboratory infrastructure. *Euro Surveill.* **20,** 21074 (2015).
27. World Health Organization. *Ebola Response Roadmap Situation Report Update—12 November 2014* http://apps.who.int/iris/bitstream/10665/141468/1/roadmapsitrep_12Nov2014_eng.pdf (2014).
28. Folarin, O. A. *et al.* Ebola virus epidemiology and evolution in Nigeria. *J. Infect. Dis.* **214,** S102–S109 (2016).
29. Abdoulaye, B. *et al.* Experience on the management of the first imported Ebola virus disease case in Senegal. *Pan Afr. Med. J.* **22,** 6 (2015).
30. Whitmer, S. L. M. *et al.* Preliminary evaluation of the effect of investigational Ebola virus disease treatments on viral genome sequences. *J. Infect. Dis.* **214,** S333–S341 (2016).
31. Xia, Y., Bjørnstad, O. N. & Grenfell, B. T. Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. *Am. Nat.* **164,** 267–281 (2004).
32. Ferrari, M. J. *et al.* The dynamics of measles in sub-Saharan Africa. *Nature* **451,** 679–684 (2008).
33. 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).
34. Gardy, J., Loman, N. J. & Rambaut, A. Real-time digital pathogen surveillance — the time is now. *Genome Biol.* **16,** 155 (2015).
35. Yozwiak, N. L., Schaffner, S. F. & Sabeti, P. C. Data sharing: make outbreak research open access. *Nature* **518,** 477–479 (2015).
36. 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).
37. 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,** http://dx.doi.org/10.1371/currents.outbreaks.02bc6d927ecee7bbd33532ec8ba6a25f (2014).
38. Tong, Y.-G. *et al.* Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. *Nature* **524,** 93–96 (2015).
39. Diallo, B. *et al.* Resurgence of Ebola virus disease in Guinea linked to a survivor with virus persistence in seminal fluid for more than 500 days. *Clin. Infect. Dis.* **63,** 1353–1356 (2016).
40. Rowe, A. K. *et al.* Clinical, virologic, and immunologic follow-up of convalescent Ebola hemorrhagic fever patients and their household contacts, Kikwit, Democratic Republic of the Congo. *J. Infect. Dis.* **179,** S28–S35 (1999).

Supplementary Information is available in the online version of the paper.

Acknowledgements The authors acknowledge support from: European Union Seventh Framework 278433-PREDEMICS (P.L., A.R.) and ERC 260864 (P.L., A.R., M.A.S.) European Union Horizon 2020 643476-COMPARE (M.P.G.K., A.R.), 634650-VIROGENESIS (P.L., M.P.G.K.), 666100-EVIDENT and European Commission IFS/2011/272-372, EMLab (S.G.), National Institutes of Health R01 AI107034, R01 AI117011 and R01 HG006139 and National Science Foundation IIS 1251151 and DMS 1264153 (M.A.S.), NIH AI081982, AI082119, AI082805 AI088843, AI104216, AI104621, AI115754, HSN272200900049C, HHSN272201400048C (R.F.G.), NIH R35 GM119774-01 (T.B.) National Health & Medical Research Council (Australia) (E.C.H.). The Research Foundation - Flanders G0D5117N (G.B., P.L.), Work in Liberia was funded by the Defense Threat Reduction Agency, the Global Emerging Infections System and the Targeted Acquisition of Reference Materials Augmenting Capabilities (TARMAC) Initiative agencies from the US Department of Defense (G.Pa.), Bill and Melinda Gates Foundation OPP1106427, 1032350, OPP1134076, Wellcome Trust 106866/Z/15/Z, Clinton Health Access Initiative (A.J.T.), National Institute for Health Research Health Protection Research Unit in Emerging and Zoonotic Infections (J.A.H.), Key Research and Development Program from the Ministry of Science and Technology of China 2016YFC1200800 (D.L.), National Natural Science Foundation of China 81590760 and 81321063 (G.F.G.), Mahan Post-doctoral fellowship Fred Hutchinson Cancer Research Center (G.D.), National Institute of Allergy and Infectious Disease U19AI110818, 5R01AI114855-03, United States Agency for International Development OAA-G-15-00001 and the Bill and Melinda Gates Foundation OPP1123407 (P.C.S.), NIH 1U01HG007480-01 and the World Bank ACE019 (C.T.H.), PEW Biomedical Scholarship, NIH UL1TR001114, and NIAID contract HHSN272201400048C (K.G.A.). J.H.K., an employee of Tunnell Government Services, Inc., is a subcontractor under Battelle Memorial Institute's prime contract with the NIAID (contract HHSN272200700016I). Colour-blind-friendly colour palettes were designed by C. Brewer, Pennsylvania State University (http://colorbrewer2.org). Matplotlib (http://matplotlib.org) was used extensively throughout this article for data visualisation. We acknowledge support from NVIDIA Corporation with the donation of parallel computing resources used for this research. Finally, we recognize the contributions made by our colleagues who died from Ebola virus disease whilst fighting the epidemic.

Author Contributions G.D., L.M.C., T.B., C.F., M.A.S., P.L. and A.R. designed the study. G.D., L.M.C., T.B., A.J.T., G.B., P.L. and A.R. performed the analysis. G.D., T.B., M.A.S, P.L. and A.R. wrote the manuscript. L.M.C., A.J.T., G.B., N.R.F., J.T.L., M.C., S.F.S., K.G.A., M.W.C., R.F.G., I.G., E.C.H., P.K., M.P.G.K., J.H.K., S.T.N., G.Pa., O.G.P., P.C.S. and U.S. edited the manuscript. The other authors were critical for the coordination, collection, processing of virus samples or the sequencing and bioinformatics of virus genomes. All authors read and approved the contents of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to A.R. (a.rambaut@ed.ac.uk), G.D. (gdudas@fredhutch.org) or P.L. (philippe.lemey@kuleuven.be).

Reviewer Information *Nature* thanks R. Biek, C. Viboud, M. Worobey and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Gytis Dudas[1,2], Luiz Max Carvalho[1], Trevor Bedford[2], Andrew J. Tatem[3,4], Guy Baele[5], Nuno R. Faria[6], Daniel J. Park[7], Jason T. Ladner[8], Armando Arias[9,10], Danny Asogun[11,12], Filip Bielejec[5], Sarah L. Caddy[9], Matthew Cotten[13,14], Jonathan D'Ambrozio[8], Simon Dellicour[5], Antonino Di Caro[12,15], Joseph W. Diclaro II[16], Sophie Duraffour[12,17], Michael J. Elmore[18], Lawrence S. Fakoli III[19], Ousmane Faye[20], Merle L. Gilbert[8], Sahr M. Gevao[21], Stephen Gire[7,22], Adrianne Gladden-Young[7], Andreas Gnirke[7], Augustine Goba[23,24], Donald S. Grant[23,24], Bart L. Haagmans[14], Julian A. Hiscox[25,26], Umaru Jah[27], Jeffrey R. Kugelman[8], Di Liu[28], Jia Lu[9], Christine M. Malboeuf[7], Suzanne Mate[8], David A. Matthews[29], Christian B. Matranga[7], Luke W. Meredith[9,27], James Qu[7], Joshua Quick[30], Suzan D. Pas[14], My V. T. Phan[13,14], Georgios Pollakis[25], Chantal B. Reusken[14], Mariano Sanchez-Lockhart[8,31], Stephen F. Schaffner[7], John S. Schieffelin[32], Rachel S. Sealfon[7,33,34], Etienne Simon-Loriere[35,36], Saskia L. Smits[14], Kilian Stoecker[12,37], Lucy Thorne[9], Ekaete Alice Tobin[11,12], Mohamed A. Vandi[23,24], Simon J. Watson[13], Kendra West[7], Shannon Whitmer[38], Michael R. Wiley[8,31], Sarah M. Winnicki[7,32], Shirlee Wohl[7,22], Roman Wölfel[12,37],

Nathan L. Yozwiak[7,22], Kristian G. Andersen[39,40], Sylvia O. Blyden[41], Fatorma Bolay[19], Miles W. Carroll[12,18,26,42], Bernice Dahn[43], Boubacar Diallo[44], Pierre Formenty[45], Christophe Fraser[46], George F. Gao[28,47], Robert F. Garry[48], Ian Goodfellow[9,27], Stephan Günther[12,17], Christian T. Happi[49,50], Edward C. Holmes[51], Brima Kargbo[24], Sakoba Keïta[52], Paul Kellam[13,53], Marion P. G. Koopmans[14], Jens H. Kuhn[54], Nicholas J. Loman[30], N'Faly Magassouba[55], Dhamari Naidoo[45], Stuart T. Nichol[38], Tolbert Nyenswah[43], Gustavo Palacios[8], Oliver G. Pybus[6], Pardis C. Sabeti[7,22], Amadou Sall[20], Ute Ströher[38], Isatta Wurie[21], Marc A. Suchard[56,57,58], Philippe Lemey[5] & Andrew Rambaut[1,59,60]

[1]Institute of Evolutionary Biology, University of Edinburgh, King's Buildings, Edinburgh EH9 3FL, UK. [2]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA. [3]WorldPop, Department of Geography and Environment, University of Southampton, Highfield, Southampton SO17 1BJ, UK. [4]Flowminder Foundation, Stockholm, Sweden. [5]Department of Microbiology and Immunology, Rega Institute, KU Leuven – University of Leuven, 3000 Leuven, Belgium. [6]Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK. [7]Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. [8]Center for Genome Sciences, US Army Medical Research Institute of Infectious Diseases, Fort Detrick, Frederick, Maryland 21702, USA. [9]Department of Pathology, University of Cambridge, Addenbrooke's Hospital, Cambridge CB2 2QQ, UK. [10]National Veterinary Institute, Technical University of Denmark, Bülowsvej 27, 1870, Frederiksberg C, Denmark. [11]Institute of Lassa Fever Research and Control, Irrua Specialist Teaching Hospital, Irrua, Nigeria. [12]The European Mobile Laboratory Consortium, 20359 Hamburg, Germany. [13]Virus Genomics, Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. [14]Department of Viroscience, Erasmus University Medical Centre, PO Box 2040, 300 CA Rotterdam, the Netherlands. [15]National Institute for Infectious Diseases 'L. Spallanzani'—IRCCS, Via Portuense 292, 00149 Rome, Italy. [16]Naval Medical Research Unit 3, 3A Imtidad Ramses Street, Cairo 11517, Egypt. [17]Bernhard Nocht Institute for Tropical Medicine, 20359 Hamburg, Germany. [18]National Infections Service, Public Health England, Porton Down, Salisbury, Wilts SP4 0JG, UK. [19]Liberian Institute for Biomedical Research, Charlesville, Liberia. [20]Institut Pasteur de Dakar, Arbovirus and Viral Hemorrhagic Fever Unit, 36 Avenue Pasteur, BP 220, Dakar, Sénégal. [21]University of Sierra Leone, Freetown, Sierra Leone. [22]Center for Systems Biology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [23]Viral Hemorrhagic Fever Program, Kenema Government Hospital, 1 Combema Road, Kenema, Sierra Leone. [24]Ministry of Health and Sanitation, 4th Floor Youyi Building, Freetown, Sierra Leone. [25]Institute of Infection and Global Health, University of Liverpool, Liverpool L69 2BE, UK. [26]NIHR Health Protection Research Unit in Emerging and Zoonotic Infections, University of Liverpool, Liverpool L69 3GL, UK. [27]University of Makeni, Makeni, Sierra Leone. [28]Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China. [29]University of Bristol, Bristol BS8 1TD, UK. [30]Institute of Microbiology and Infection, University of Birmingham, Birmingham B15 2TT, UK. [31]University of Nebraska Medical Center, Omaha, Nebraska 68198, USA. [32]Department of Pediatrics, Section of Infectious Diseases, New Orleans, Louisiana 70112, USA. [33]Center for Computational Biology, Flatiron Institute, New York, New York 10010, USA. [34]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey 08544, USA. [35]Institut Pasteur, Functional Genetics of Infectious Diseases Unit, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France. [36]Génétique Fonctionelle des Maladies Infectieuses, CNRS URA3012, Paris 75015, France. [37]Bundeswehr Institute of Microbiology, Neuherbergstrasse 11, 80937 Munich, Germany. [38]Viral Special Pathogens Branch, Centers for Disease Control and Prevention, 1600 Clifton Road NE, Atlanta, Georgia 30333, USA. [39]The Scripps Research Institute, Department of Immunology and Microbial Science, La Jolla, California 92037, USA. [40]Scripps Translational Science Institute, La Jolla, California 92037, USA. [41]Ministry of Social Welfare, Gender and Children's Affairs, New Englandville, Freetown, Sierra Leone. [42]University of Southampton, South General Hospital, Southampton SO16 6YD, UK. [43]Minstry of Health Liberia, Monrovia, Liberia. [44]World Health Organization, Conakry, Guinea. [45]World Health Organization, Geneva, Switzerland. [46]Oxford Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, Oxford OX3 7FZ, UK. [47]Chinese Center for Disease Control and Prevention (China CDC), Beijing 102206, China. [48]Department of Microbiology and Immunology, New Orleans, Louisiana 70112, USA. [49]Department of Biological Sciences, Redeemer's University, Ede, Osun State, Nigeria. [50]African Center of Excellence for Genomics of Infectious Diseases (ACEGID), Redeemer's University, Ede, Osun State, Nigeria. [51]Marie Bashir Institute for Infectious Diseases and Biosecurity, Charles Perkins Centre, School of Life and Environmental Sciences and Sydney Medical School, the University of Sydney, Sydney, New South Wales 2006, Australia. [52]Ministry of Health Guinea, Conakry, Guinea. [53]Division of Infectious Diseases, Faculty of Medicine, Imperial College London, London W2 1PG, UK. [54]Integrated Research Facility at Fort Detrick, National Institute of Allergy and Infectious Diseases, National Institutes of Health, B-8200 Research Plaza, Fort Detrick, Frederick, Maryland 21702, USA. [55]Université Gamal Abdel Nasser de Conakry, Laboratoire des Fièvres Hémorragiques en Guinée, Conakry, Guinea. [56]Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles, California 90095, USA. [57]Department of Biomathematics David Geffen School of Medicine at UCLA, University of California, Los Angeles, California 90095, USA. [58]Department of Human Genetics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, California 90095, USA. [59]Centre for Immunology, Infection and Evolution, University of Edinburgh, King's Buildings, Edinburgh, EH9 3FL, UK. [60]Fogarty International Center, National Institutes of Health, Bethesda, Maryland 20892, USA.

METHODS

Sequence data. We compiled a dataset of 1,610 publicly available full EBOV genomes sampled between 17 March 2014 and 24 October 2015 (see https://github/ebov/space-time/data/ for the full list and metadata). The number of sequences and the proportion of cases sequenced varies between countries; our dataset contains 209 sequences from Liberia (3.8% of known and suspected cases), 982 from Sierra Leone (8.0%) and 368 from Guinea (9.2%) (Supplementary Table 1). Most ($n = 1{,}100$) genomes are of high quality, with ambiguous sites and gaps comprising less than 1% of the total alignment length, followed by sequences with between 1% and 2% of sites that comprised ambiguous bases or gaps ($n = 266$), 98 sequences with 2–5%, 120 sequences with 5–10% and 26 sequences with more than 10% of sites that are ambiguous or are gaps. Sequences known to be associated with sexual transmission or latent infections were excluded, as these viruses often exhibit anomalous molecular clock signals[11,12]. Sequences were aligned using MAFFT[41] and edited manually. The alignment was partitioned into coding regions and non-coding intergenic regions with a final alignment length of 18,992 nucleotides (available from https://github/ebov/space-time/data/).

Masking putative ADAR-edited sites. As noticed in previous studies[15,38], some EBOV isolates contain clusters of T-to-C mutations within relatively short stretches of the genome. Interferon-inducible adenosine deaminases acting on RNA (ADAR) are known to induce adenosine to inosine hypermutations in double-stranded RNA[43]. ADARs have been suggested to act on RNAs from numerous groups of viruses[42]. When negative-sense single-stranded RNA virus genomes are edited by ADARs, A-to-G hypermutations seem to preferentially occur on the negative strand, which results in U/T-to-C mutations on the positive strand[44–46]. Multiple T-to-C mutations are introduced simultaneously by ADAR-mediated RNA editing which would interfere with molecular clock estimates and, by extension, the tree topology. We therefore designated that four or more T-to-C mutations within 300 nucleotides of each other as a putative hypermutation tract, whenever there is evidence that all T-to-C mutations within such stretches were introduced at the same time, that is, every T-to-C mutation in a stretch occurred on a single branch. We detected a total of 15 hypermutation patterns with up to 13 T-to-C mutations within 35 to 145 nucleotides. Of these patterns, 11 are unique to a single genome and 4 are shared across multiple isolates, suggesting that occasionally viruses that survive hypermutation are transmitted[47]. Putative tracts of T-to-C hypermutation almost exclusively occur within non-coding intergenic regions, where their effects on viral fitness are presumably minimal. In each case, we mask out these sites as ambiguous nucleotides, but leave the first T-to-C mutation unmasked to provide phylogenetic information on the relatedness of these sequences.

Phylogenetic inference. Molecular evolution was modelled according to a HKY+Γ_4 substitution model (refs 48, 49) independently across four partitions (codon positions 1, 2, 3 and non-coding intergenic regions). Site-specific rates were scaled by relative rates in the four partitions. Evolutionary rates were allowed to vary across the tree according to a relaxed molecular clock that draws branch-specific rates from a log-normal distribution[50]. A non-parametric coalescent 'Skygrid' model was used to act as a prior density on trees[51]. The overall evolutionary rate was given an uninformative continuous-time Markov chain (CTMC) reference prior[52], while the rate multipliers for each partition were given an uninformative uniform prior over their bounds. All other priors used to infer the phylogenetic tree were left at their default values. BEAST XML files are available from https://github/ebov/space-time/data/. We ran an additional analysis with a subset of data (787 sequences collected up to November 2014—the peak of case numbers in Sierra Leone) to test the robustness of inference if they had been performed mid-epidemic.

Geographic history reconstruction. The level of administrative regions within each country was chosen so that population sizes between regions are comparable. For each country the appropriate administrative regions were: prefecture for Guinea (administrative subdivision level 2), county for Liberia (level 1) and district for Sierra Leone (level 2). We refer to them as regions (63 in total, but only 56 are recorded to have had EVD cases) and each sequence, where available, was assigned the region where the patient was recorded to have been infected as a discrete trait. When the region within a country was unknown ($n = 223$), we inferred the sequence location as a latent variable with equal prior probability over all available regions within that country. Most of the sequences with unknown regional origins were from Sierra Leone ($n = 151$), followed by Liberia ($n = 69$) and Guinea ($n = 3$). In the absence of any geographic information ($n = 2$) we inferred both the country and the region of a sequence.

We used an asymmetric CTMC[53,55] matrix to infer instantaneous transitions between regions. For 56 regions with recorded EVD cases, a total of 3,080 independent transition rates would be challenging to infer from one realization of the process, even when reduced to a sparse migration matrix using stochastic search variable selection[53].

Therefore, to infer the spatial phylogenetic diffusion history between the $K = 56$ locations, we adopt a sparse GLM formulation of CTMC diffusion[18]. This model parameterizes the instantaneous movement rate Λ_{ij} from location i to location j as a log-linear function of P potential predictors $\boldsymbol{X}_{ij} = (x_{ij1}, \ldots, x_{ijP})'$ with unknown coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)'$ and diagonal matrix δ with entries $(\delta_1, \ldots, \delta_p)$. These latter unknown indicators $\delta_p \in \{0,1\}$ determine the inclusion in or exclusion from the model of a single predictor. We generalize this formulation here to include two-way random effects that allow for location origin- and destination-specific variability. Our two-way random effects GLM becomes

$$\log(\Lambda_{ij}) = \boldsymbol{X}'_{ij}\boldsymbol{\delta}\boldsymbol{\beta} + \varepsilon_i + \varepsilon_j \quad (1)$$

where ε_k is distributed as normal$(0, \sigma^2)$ for $k = 1, ..., K$, and σ^2 is distributed as inverse-$\Gamma(0.001, 0.001)$, and where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_K)$ are the location-specific effects. These random effects account for unexplained variability in the diffusion process that may otherwise lead to spurious inclusion of predictors.

We follow ref. 18 by specifying that *a priori* all β_p are independent and normally distributed with mean 0 and a relatively large variance of 4 and by assigning independent Bernoulli prior probability distributions on δ_p.

Let q be the inclusion probability and w be the probability of no predictors being included. Then, using the distribution function of a binomial random variable $q = 1 - w^{1/P}$, where P is the number of predictors, as before. We use a small success probability on each predictor's inclusion that reflects a 50% prior probability (w) on no predictors being included.

In our main analysis, we consider 25 individual predictors that can be classified as geographic, administrative, demographic, cultural and climatic covariates of spatial spread (Extended Data Table 1). Where measures are region-specific (rather than pairwise region measures), we specify both an origin and destination predictor. We also tested for sampling bias by including an additional origin and destination predictor based on the residuals for the regression of sample size against case count (Extended Data Fig. 1b), but these predictors did not receive any support (data not shown).

To draw posterior inference, we follow ref. 18 by integrating $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$, and further employ a random-walk Metropolis transition kernel on ε and sample σ^2 directly from its full conditional distribution using Gibbs sampling.

To obtain a joint posterior estimate from this joint genetic and phylogeographic model, an MCMC chain was run in BEAST 1.8.4 (ref. 54) for 100 million states, sampling every 10,000 states. The first 1,000 samples in each chain were removed as burnin, and the remaining 9,000 samples used to estimate a maximum clade credibility tree and to estimate posterior densities for individual parameters. A second independent run of 100 million states was performed to check convergence of the first.

To consider the feasibility of 'real-time' inference from virus genome data from the height of the EVD epidemic we took only those sequences derived from samples taken up until the end of October 2014 ($n = 787$). We undertook the same joint phylogenetic and spatial GLM analysis as for the full dataset including the same set of 25 predictors. We ran this analysis for 200 million states, sampling every 20,000 states and removing the first 10% of samples.

To obtain realizations of the phylogenetic CTMC process, including both transitions (Markov jumps) between states and waiting times (Markov rewards) within states, we used posterior inference of the complete Markov jump history through time[18,56]. In addition to transitions 'within' the phylogeny, we also estimate the expected number of transitions 'from' origin location i in the phylogeographic tree to arbitrary 'destination' location j as follows:

$$\zeta_{ij} = \tau_i \mu \Lambda_{ij} \pi_i / c \quad (2)$$

where τ_i is the waiting time (or Markov reward) in 'origin' state i throughout the phylogeny, μ is the overall rate scalar of the location transition process, π_i is the equilibrium frequency of 'origin' state i, and c is the normalizing constant applied to the CTMC rate matrices in BEAST. To obtain the expected number of transitions to a particular destination location from any phylogeographic location (integrating over all possible locations across the phylogeny), we sum over all 56 origin locations included in the analysis. We note that the destination location can also be a location that was not included in the analysis because we only need to consider destination j in the instantaneous movement rates Λ_{ij}; since the log of these rates are parameterized as a log-linear function of the predictors, we can obtain these rates through the coefficient estimates from the analysis and the predictors extended to include these additional locations. Specifically, we use this to predict introductions in regions in Guinea, for which no cases were reported ($n = 7$) and for regions in neighbouring countries along the borders with Guinea or Liberia that remained disease free ($n = 18$). To obtain such estimates under different predictors or predictor combinations, we perform a specific analysis under the GLM

model including only the relevant predictors or predictor combinations without the two-way random effects. For computational expedience, we performed these analyses, as well as the time-inhomogeneous analyses below, by conditioning on a set of 1,000 trees from the posterior distribution of the main phylogenetic analysis[18]. We summarize mean posterior estimates for the transition expectations based on the samples obtained by our MCMC analysis; we also note that the value of c is sample-specific.

Time-dependent spatial diffusion. To consider time-inhomogeneity in the spatial diffusion process, we start by borrowing epoch modelling concepts from ref. 57. The epoch GLM parameterizes the instantaneous movement rate Λ_{ijt} from state i to state j within epoch t as a log-linear function of P epoch-specific predictors $\boldsymbol{X}_{ijt} = (x_{ijt1}, \ldots, x_{ijtP})'$ with constant-through-time, unknown coefficients β. We generalize this model to incorporate a time-varying contribution of the predictors through time-varying coefficients $\beta(t)$ using a series of change-point processes. Specifically, the time-varying epoch GLM models

$$\begin{aligned} \log \Lambda_{ijt} &= \boldsymbol{X}'_{ijt}\beta(t) \\ \beta(t) &= (\boldsymbol{I} - \phi(t))\beta_B + (\phi(t))\beta_A \end{aligned} \tag{3}$$

where $\boldsymbol{\beta}_B = (\beta_{B1}, \ldots, \beta_{BP})'$ are the unknown coefficients before the change-points, $\beta_A = (\beta_{A1}, \ldots, \beta_{AP})'$ are the unknown coefficients after the change-points, diagonal matrix $\varphi(t)$ has entries $(1_{t>t1}(t), \ldots, 1_{t>tP}(t))$, $1_{(\cdot)}(t)$ is the indicator function and $\boldsymbol{T} = (t_1, \ldots, t_P)$ are the unknown change-point times. In this general form, the contribution of predictor p before its change-point time t_p is β_{Bp} and its contribution after is β_{Ap} for $P = 1, \ldots, P$. Fixing t_p to be less than the time of the first epoch or greater than the time of the last epoch results in a time-invariant coefficient for that predictor.

Similar to the constant-through-time GLM, we specify *a priori* that all β_{Bp} and β_{Ap} are independent and normally distributed with mean 0 and a relatively large variance of 4. Under the prior, each t_p is equally probable to lie before any epoch.

We used random-walk Metropolis transition kernels on $\boldsymbol{\beta}_B$, $\boldsymbol{\beta}_A$ and T.

In a first epoch GLM analysis, we keep the five predictors that are convincingly supported by the time-homogeneous analysis included in the model and estimate an independent change-point t_p for their associated effect sizes: distance, nat./int. effect, shared international border and origin and destination population size change-points. To quantify the evidence in favour of each change-point, we calculate Bayes factor support on the basis of the prior and posterior odds that t_p is less than the time of the first epoch or greater than the time of the last epoch. Because we find only very strong support for a change-point in the nat./int. effect, we subsequently estimate the effect sizes before and after its associated change-point, keeping the remaining four predictors homogeneous through time.

Within-location generalized linear models. EVD case numbers are reported by the WHO for every country division (region) at the appropriate administrative level, split by epidemiological week. For every region and for each epidemiological week four numbers are reported: new cases in the patient and situation report databases as well as whether the new cases are confirmed or probable. At the height of the epidemic many cases went unconfirmed, even though they were likely to have been genuine EVD. As such, we treat probable EVD cases in WHO reports as confirmed and combine them with laboratory-confirmed EVD case numbers. Following this we take the higher combined case number of situation report and patient databases. The latest situation report in our data goes up to the epidemiological week spanning 8 to 14 February 2016, with all case numbers being downloaded on 22 February 2016. There are apparent discrepancies between cumulative case numbers reported for each country over the entire epidemic and case numbers reported per administrative division over time, such that our estimate for the final size of the epidemic, based on case numbers over time reported by the WHO, is on the order of 22,000 confirmed and suspected cases of EVD compared to the official estimate of around 28,000 cases across the entire epidemic. This likely arose because case numbers are easier to track at the country level, but become more difficult to narrow down to administrative subdivision level, especially over time (only 86% of the genome sequences had a known location of infection).

We studied the association between disease case counts using generalized linear models in a very similar fashion to the framework presented above. A list of the location-level predictors we used for these analyses can be found in Extended Data Table 1. We also employed stochastic search variable selection as described above, in order to compute Bayes factors (BFs) for each predictor. In keeping with the genetic GLM analyses, we also set the prior inclusion probabilities such that there was a 50% probability of no predictors being included.

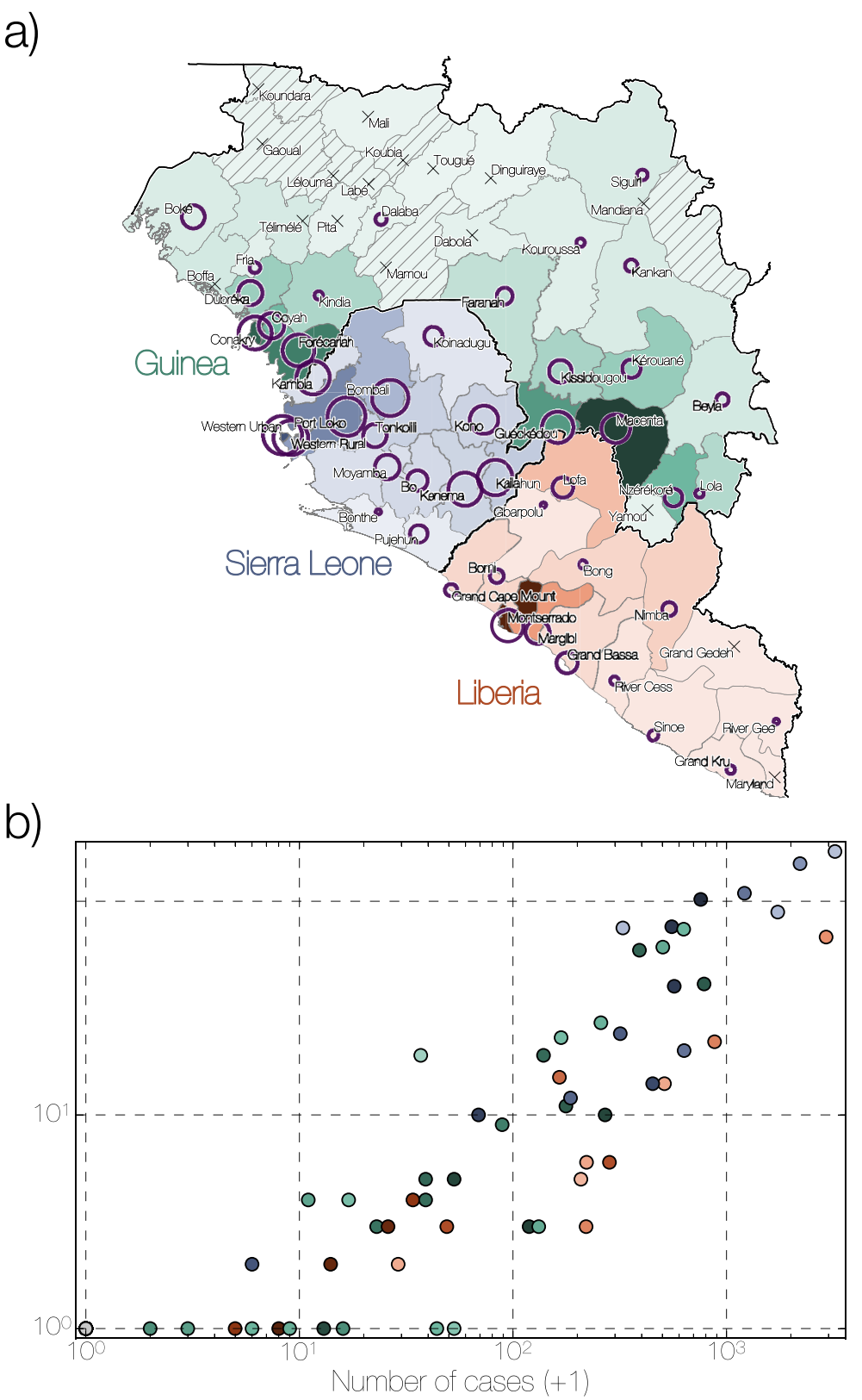
$$\begin{aligned} Y_i &\sim \text{negative}-\text{binomial}(p_i, r) \\ p_i &= \frac{r}{(r+\lambda_i)} \\ \log(\lambda_i) &= \alpha + \beta_1\delta_1 x_{i1} + \ldots + \beta_P\delta_P x_{iP} \end{aligned}$$

where r is the over-dispersion parameter, δ_i are the indicators as before. Prior distributions on model parameters for these analyses were the same as those used for the genetic analyses whenever possible. We then use this model to predict how many cases the locations which reported zero EVD cases would have gathered, that is, the potential size of the epidemic in each location.

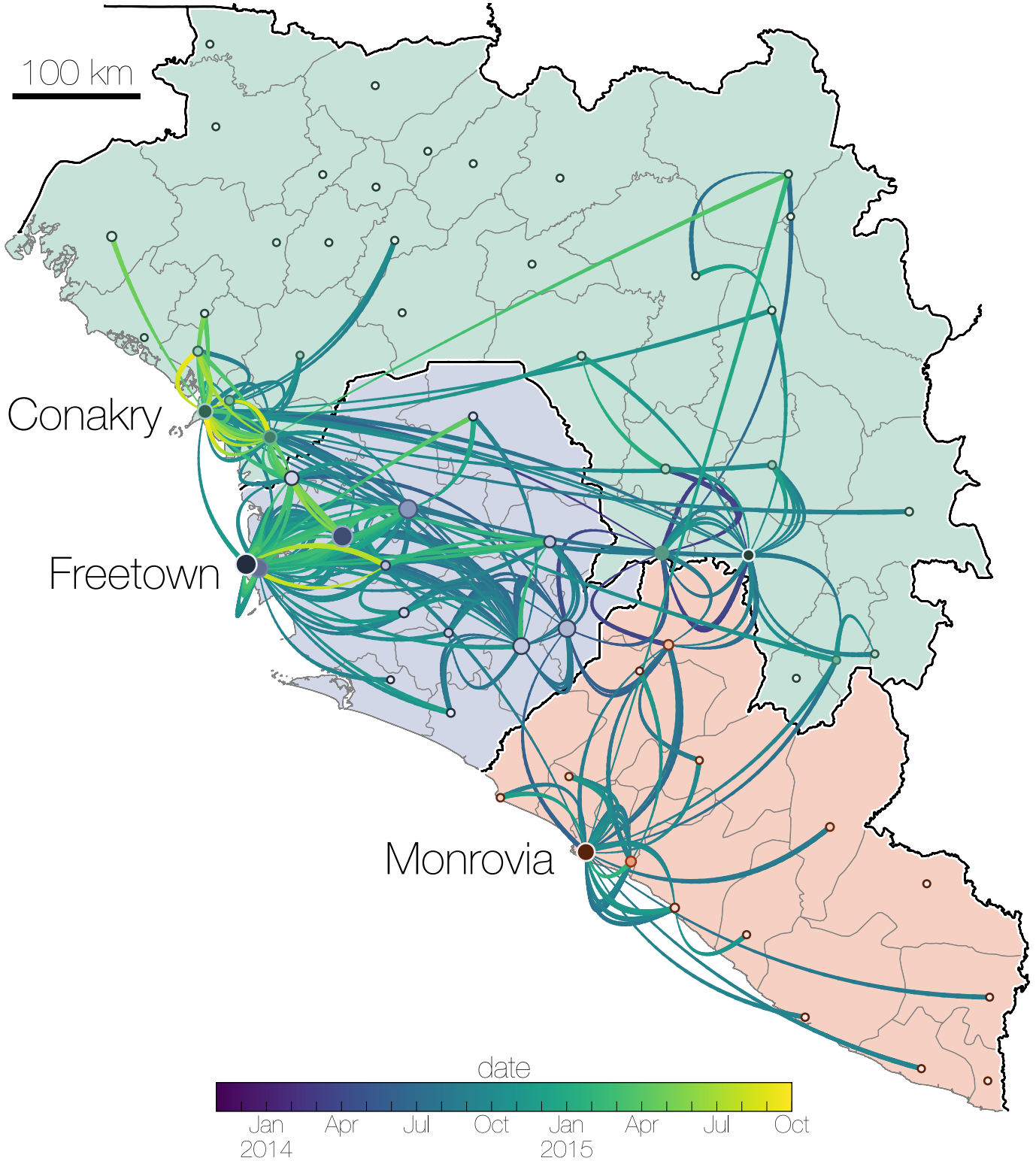
Computational details. To fit the models described above we took advantage of the routines already built in BEAST (https://github.com/beast-dev/beast-mcmc) but in a non-phylogenetic setting. Once again, posterior distributions for the parameters were explored using MCMC. We ran each chain for 50 million iterations and discarded at least 10% of the samples as burn-in. Convergence was checked by visual inspection of the chains and checking that all parameters had effective sample sizes greater than 200. We ran multiple chains to ensure that results were consistent. To make predictions, we used 50,000 Monte Carlo samples from the posterior distribution of coefficients and the overdispersion parameter (r) to simulate case counts for all locations with zero recorded EVD cases.

Data availability. All collated data, genetic sequence alignments, phylogenetic trees, analysis scripts and analysis output are available at https://github.com/ebov/space-time and http://dx.doi.org/10.7488/ds/1711. Individual virus genetic sequences are published in earlier works and are available from NCBI GenBank (see https://github.com/ebov/space-time for a list of accession numbers and references).

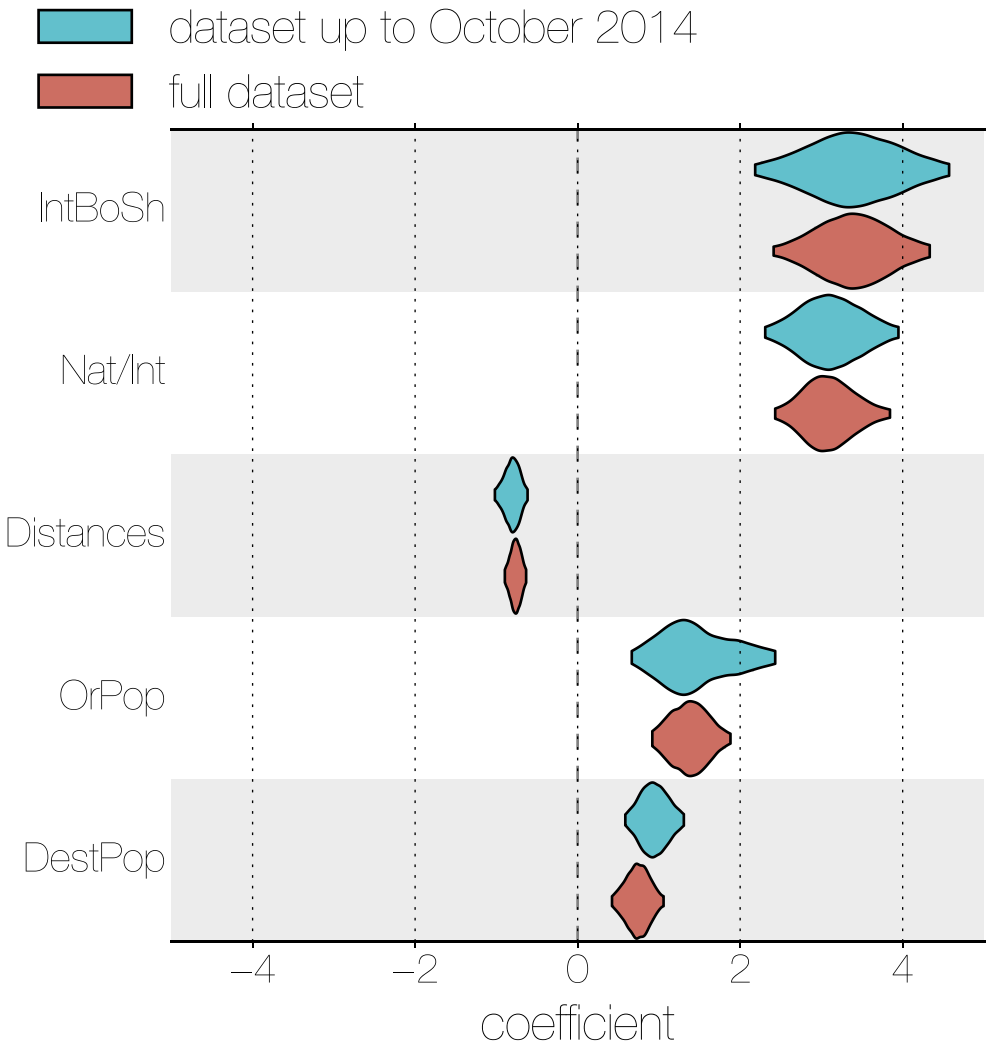
41. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30,** 3059–3066 (2002).
42. Gélinas, J.-F., Clerzius, G., Shaw, E. & Gatignol, A. Enhancement of replication of RNA viruses by ADAR1 via RNA editing and inhibition of RNA-activated protein kinase. *J. Virol.* **85,** 8460–8466 (2011).
43. Bass, B. L. & Weintraub, H. An unwinding activity that covalently modifies its double-stranded RNA substrate. *Cell* **55,** 1089–1098 (1988).
44. Cattaneo, R. *et al.* Biased hypermutation and other genetic changes in defective measles viruses in human brain infections. *Cell* **55,** 255–265 (1988).
45. Rueda, P., García-Barreno, B. & Melero, J. A. Loss of conserved cysteine residues in the attachment (G) glycoprotein of two human respiratory syncytial virus escape mutants that contain multiple A–G substitutions (hypermutations). *Virology* **198,** 653–662 (1994).
46. Carpenter, J. A., Keegan, L. P., Wilfert, L., O'Connell, M. A. & Jiggins, F. M. Evidence for ADAR-induced hypermutation of the Drosophila sigma virus (Rhabdoviridae). *BMC Genet.* **10,** 75 (2009).
47. Smits, S. L. *et al.* Genotypic anomaly in Ebola virus strains circulating in Magazine Wharf area, Freetown, Sierra Leone, 2015. *Euro Surveill.* **20,** 30035 (2015).
48. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22,** 160–174 (1985).
49. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39,** 306–314 (1994).
50. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4,** e88 (2006).
51. Gill, M. S. *et al.* Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol.* **30,** 713–724 (2013).
52. Ferreira, M. A. R. & Suchard, M. A. Bayesian analysis of elapsed times in continuous-time Markov chains. *Can. J. Stat.* **36,** 355–368 (2008).
53. Lemey, P., Suchard, M. & Rambaut, A. Reconstructing the initial global spread of a human influenza pandemic: a Bayesian spatial-temporal model for the global spread of H1N1pdm. *PLoS Curr.* **1,** RRN1031 (2009).
54. 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).
55. Edwards, C. J. *et al.* Ancient hybridization and an Irish origin for the modern polar bear matriline. *Curr. Biol.* **21,** 1251–1258 (2011).
56. Minin, V. N. & Suchard, M. A. Fast, accurate and simulation-free stochastic mapping. *Phil. Trans. R. Soc. B* **363,** 3985–3995 (2008).
57. 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).



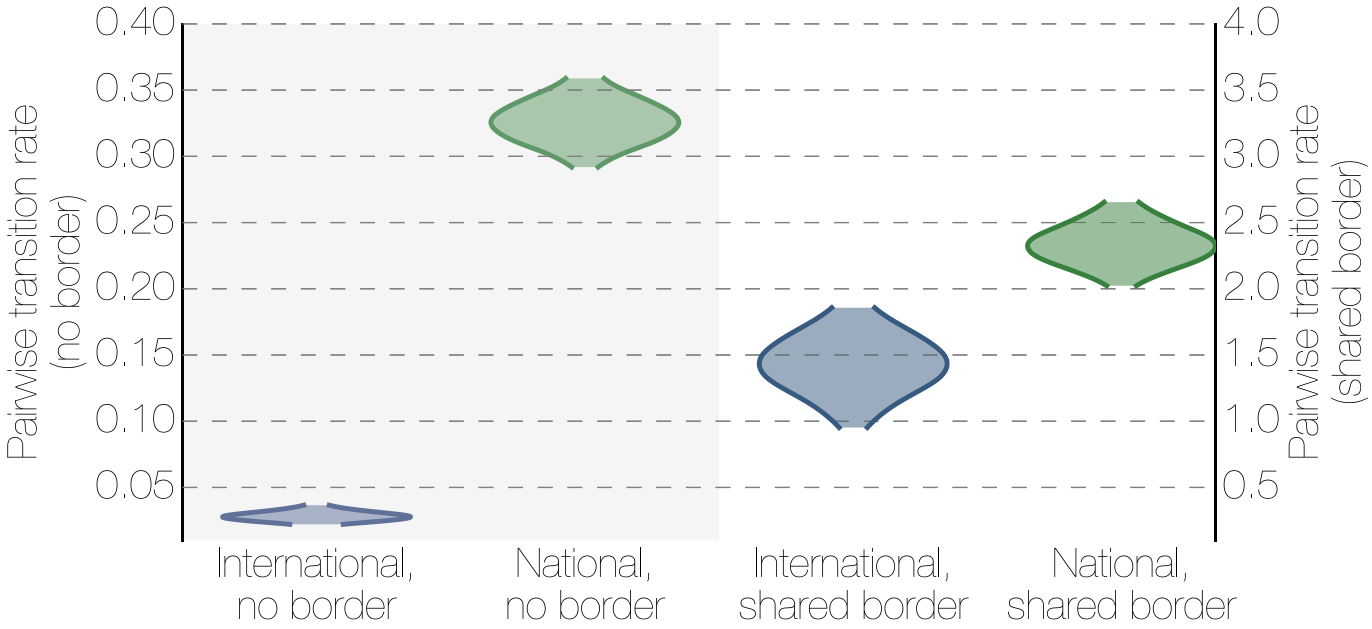
Extended Data Figure 1 | Distribution and correlation of EVD cases and EBOV sequences. a, Administrative regions within Guinea (green), Sierra Leone (blue) and Liberia (red); shading is proportional to the cumulative number of known and suspected EVD cases in each region. Darkest shades represent 784 cases for Guinea (Macenta prefecture); 3,219 cases for Sierra Leone (Western Area urban district); and 2,925 cases for Liberia (Montserrado county); hatching indicate regions without reported EVD cases. Circle diameters are proportional to the number of EBOV genomes available from that region over the entire EVD epidemic with the largest circle representing 152 sequences. Crosses mark regions for which no sequences are available. Circles and crosses are positioned at population centroids within each region. **b**, A plot of number of EBOV genomes sampled against the known and suspected cumulative EVD case numbers. Regions in Guinea are denoted in green, Sierra Leone in blue and Liberia in red. Spearman correlation coefficient: 0.93.



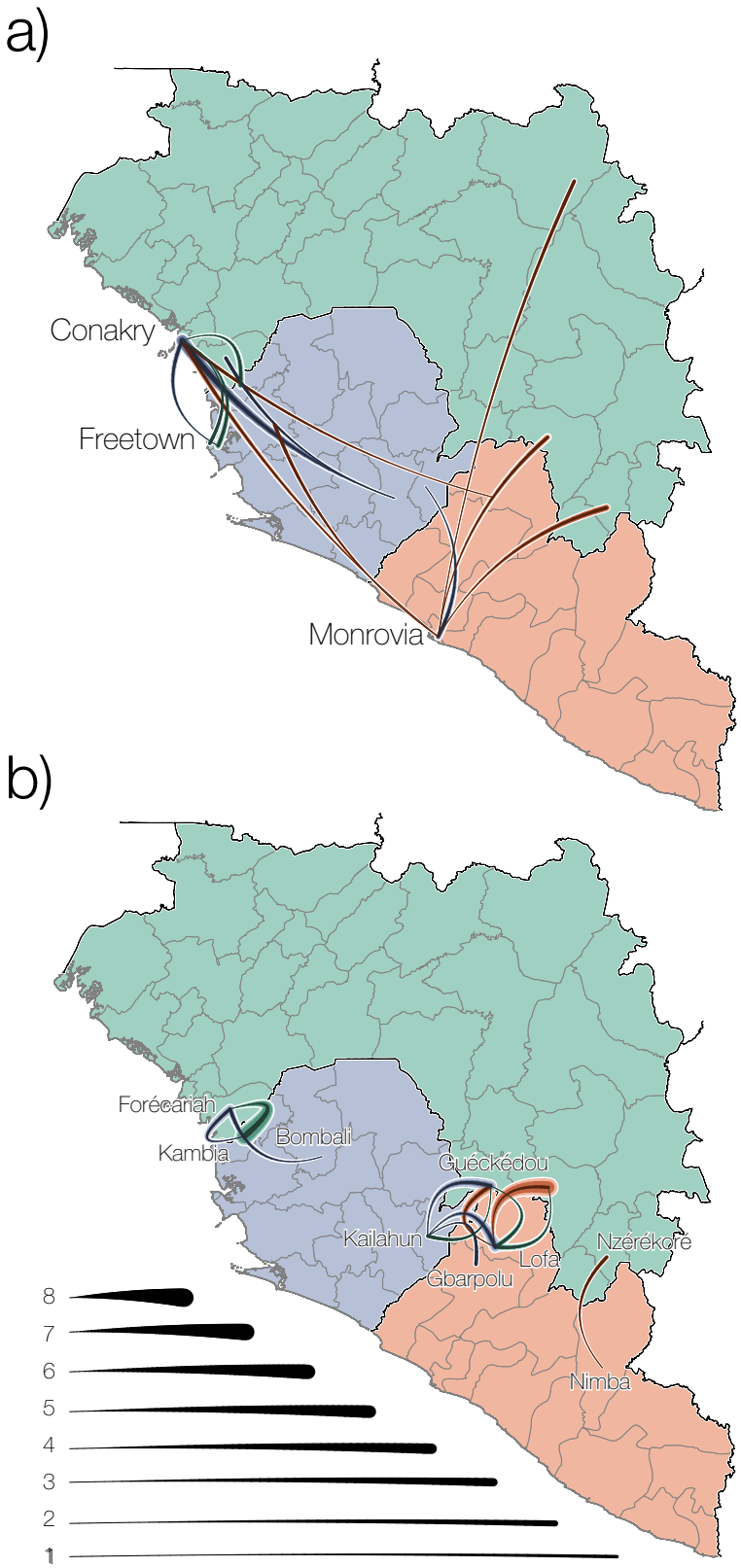
Extended Data Figure 2 | Dispersal of virus lineages over time. Virus dispersal between administrative regions estimated using the GLM phylogeography model (see Methods). The arcs are between population centroids of each region, show directionality from the thin end to the thick end and are coloured in a scale denoting time from December 2013 in blue to October 2015 in yellow. Countries are coloured with Liberia in red, Guinea in green and Sierra Leone in blue.



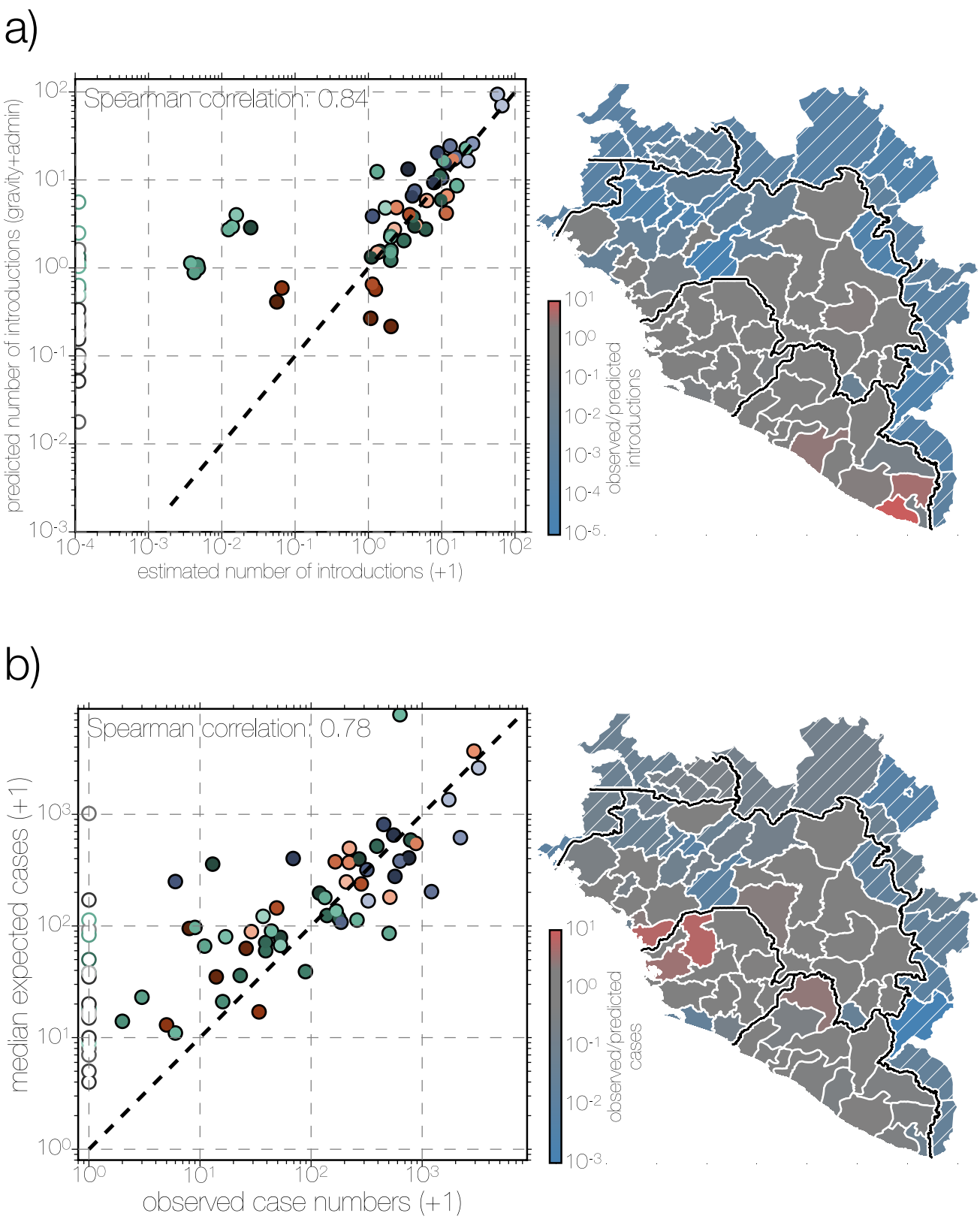
Extended Data Figure 3 | Inference of GLM predictors in a 'real-time' context. For the dataset constructed from EBOV genome sequences derived from samples taken up until October 2014 (blue), the same 5 spatial EBOV movement predictors were given categorical support (inclusion probabilities = 1.0) as for the full dataset (red). Likewise, the coefficients for these predictors are consistent in their sign and magnitude.



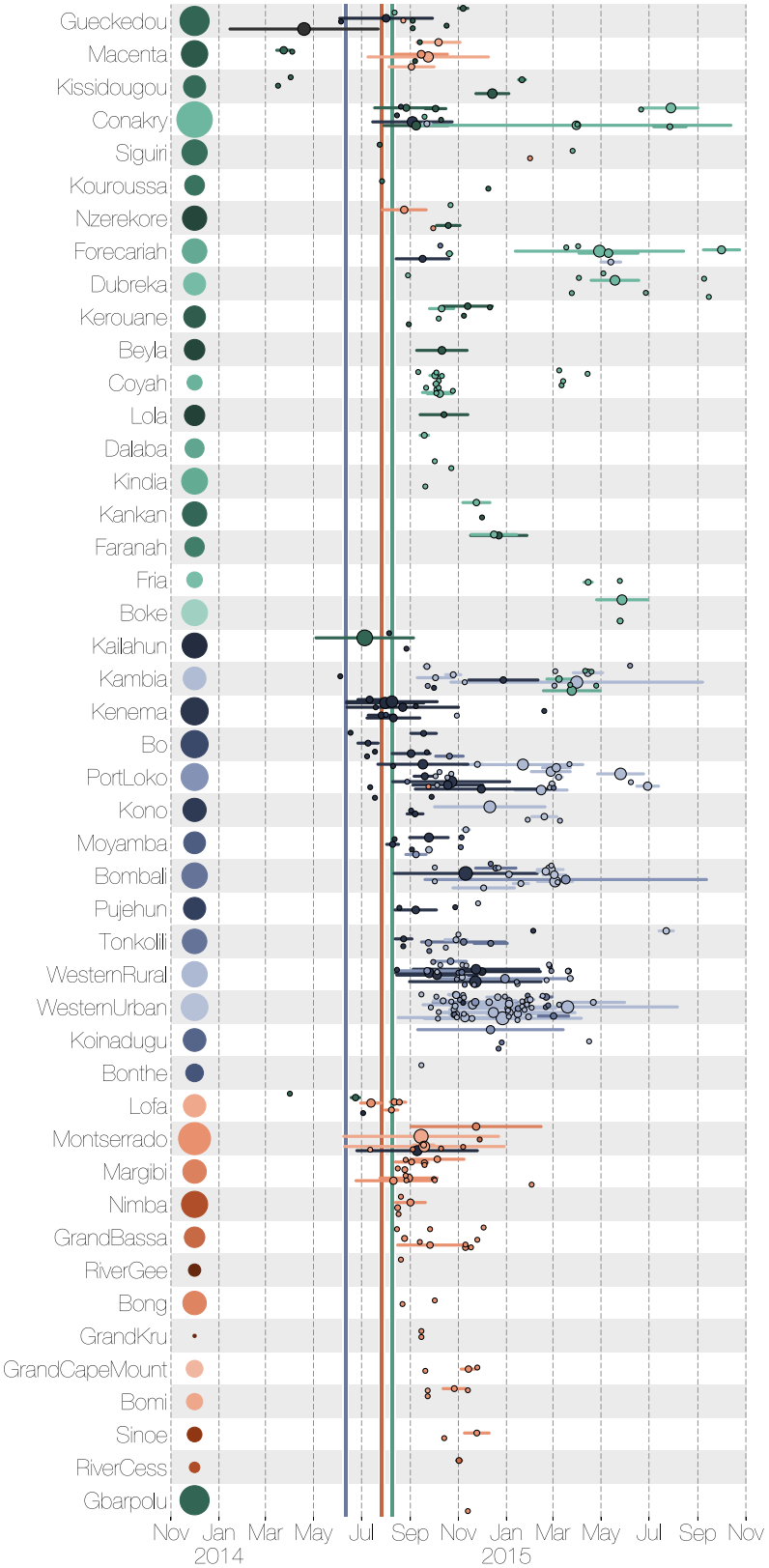
Extended Data Figure 4 | The effect of borders on EBOV migration rates between regions. Posterior densities for the migration rates between locations that share a geographical border and those that do not share borders for international migrations and national migrations. Where two regions share a border (right *y* axis), national migrations are only marginally more frequent than international migrations showing that both types of borders are porous to short local movement. Where the two regions are not adjacent (left *y* axis), international migrations are much rarer than national migrations.



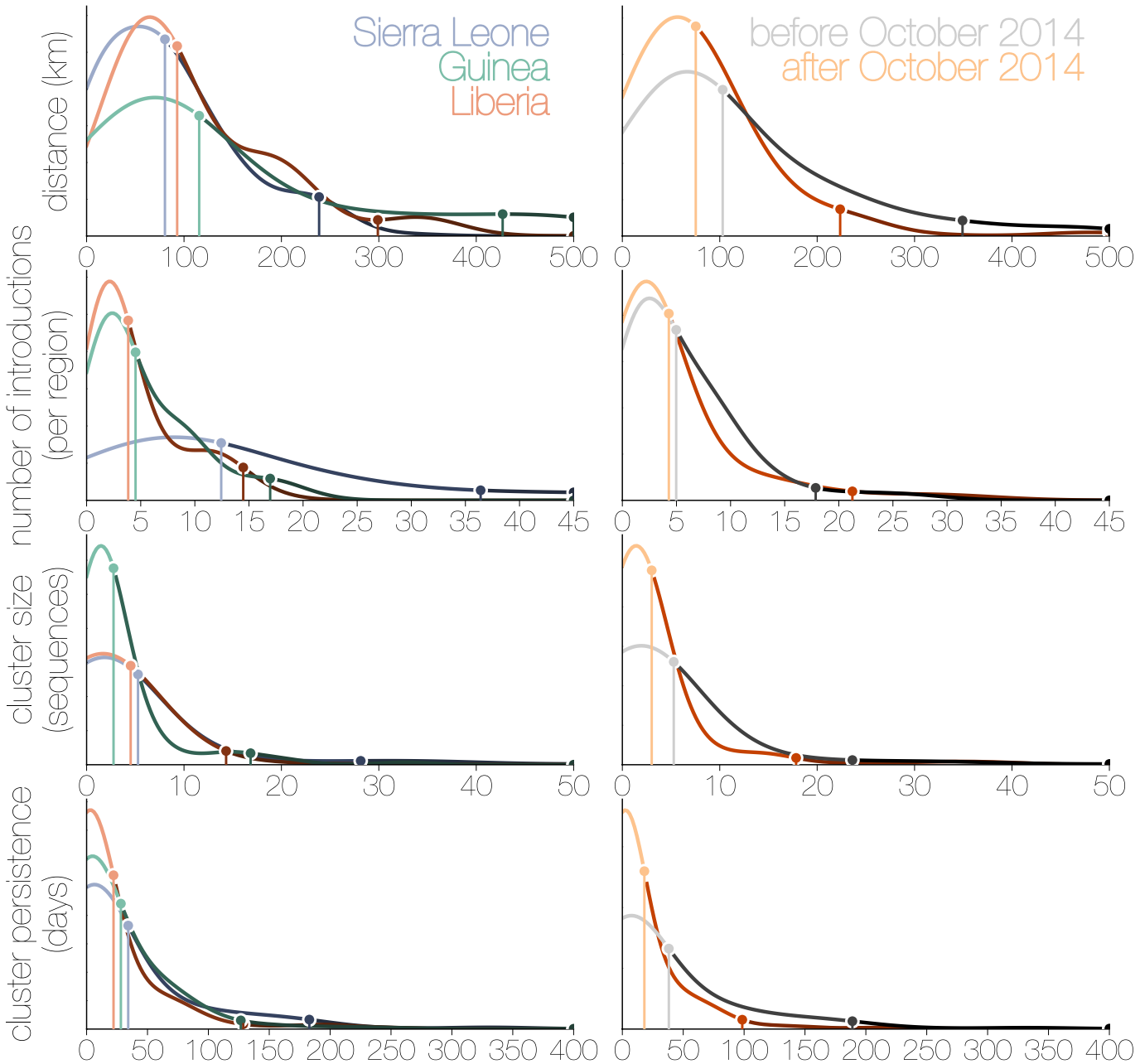
Extended Data Figure 5 | Summarized international migration history of the epidemic. a, **b**, All viral movement events between countries (Guinea, green; Sierra Leone, blue; Liberia, red) are shown split by whether they are between regions that are geographically distant (**a**) or regions that share the international border (**b**). Curved lines indicate median (intermediate colour intensity), and 95% highest posterior density intervals (lightest and darkest colour intensities) for the number of migrations that are inferred to have taken place between countries.



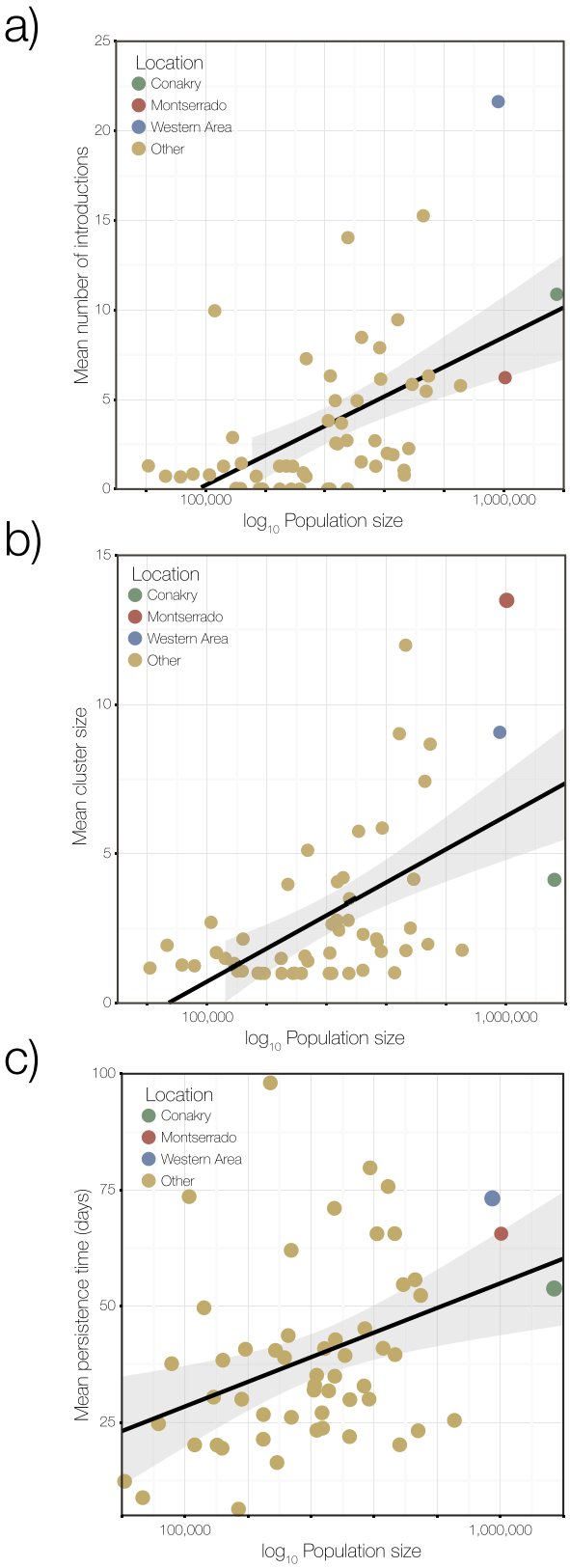
Extended Data Figure 6 | Comparison of predicted and observed numbers of introductions and case numbers. a, **b**, Left, scatter plots show inferred introduction numbers (**a**) or observed case numbers (**b**), coloured by region as in Extended Data Fig. 1. Administrative regions that did not report any cases are indicated with empty circles on the scatter plot. Right, administrative regions on the map are coloured by the residuals (as observed/predicted) of the scatter plot. Regions are coloured grey where $0.5 <$ observed/predicted < 2.0 and transition into red or blue colours for overestimation or underestimation, respectively.



Extended Data Figure 7 | Region-specific introductions, cluster sizes and persistence. Each row summarizes independent introductions and the sizes (as numbers of sequences) of resulting outbreak clusters. Clusters are coloured by their inferred region of origin (colours are the same as in Extended Data Fig. 1). The horizontal lines represent the persistence of each cluster from the time of introduction to the last sampled case (individual tips have persistence 0). The areas of the circles in the middle of the lines are proportional to the number of sequenced cases in the cluster. The areas of the circles next to the labels on the left represent the population sizes of each administrative region. Vertical lines within each cell indicate the dates of declared border closures by each of the three countries: 11 June 2014 in Sierra Leone (blue), 27 July 2014 in Liberia (red) and 09 August 2014 in Guinea (green).



Extended Data Figure 8 | Kernel density estimates for inferred epidemiological statistics. From top to bottom, distance travelled (distance between population centroids, in kilometres); number of introductions that each location experienced; cluster size (number of sequences collected in a location as a result of a single introduction); cluster persistence (days from the common ancestor of a cluster to its last descendent, single tips have persistence of 0. Left, analysis for Sierra Leone (blue), Liberia (red) and Guinea (green). Right, analysis for before October 2014 (grey) and after October 2014 (orange). Points with vertical lines connected to the *x* axis indicate the 50% and 95% quantiles of the parameter density estimates. Within Sierra Leone, Liberia and Guinea, 50% of all migrations occurred over distances of around 100 km and persisted for around 25 days. Exceptions were for Sierra Leone, which experienced more introductions per location (around 12) than Guinea and Liberia (around 4); and Guinea, where migrations tended to occur over larger distances owing to the size of the country and whose cluster sizes following introductions tended to be lower (3 sequences versus Liberia and Sierra Leone, which had 5 sequences each). Between the first (grey) and second (orange) years of the epidemic there were considerable reductions in cluster persistence, cluster sizes and distances travelled by viruses, whereas dispersal intensity remained largely the same.



Extended Data Figure 9 | Relationship between cluster size, introductions or persistence and population size. a, The mean number of introductions into each location against (log) population sizes. The Western Area (in Sierra Leone) received the most introductions, whereas Conakry and Montserrado were closer to the average. The association between population size and the number of introductions was not very strong ($R^2 = 0.28$, Pearson correlation = 0.54, Spearman correlation = 0.57). **b**, The mean cluster size for each location plotted against (log) population sizes. The association is weaker than for **a** ($R^2 = 0.11$, Pearson correlation = 0.35, Spearman correlation = 0.57). **c**, The mean persistence times (per cluster, in days) against population sizes. A similarly weak association is observed as in **b** ($R^2 = 0.12$, Pearson correlation = 0.37, Spearman correlation = 0.36). All computations were based on a sample of 10,000 trees from the posterior distribution.

Extended Data Table 1 | Predictors included in the time-homogenous GLM

Predictor type	Abbreviation	Predictor description
Geographic	Distances	Great circle distances between the locations' population centroids, log-transformed, standardized
Administrative	Nat/Int	Two locations are in the same country verses in different countries
Administrative	Nat/Int	The relative preference of transitioning between locations in the same country over transitioning between locations in two different countries
Administrative	IntBoSh	The relative preference of transitioning between location pairs that are in different countries and share a border
Administrative	NatBoSh	The relative preference of transitioning between location pairs that are in the same country and share a border
Administrative	LibGinAsym	Between Liberia-Guinea asymmetry
Administrative	LibSLeAsym	Between Liberia-Sierra Leone asymmetry
Administrative	GinSLeAsym	Between Guinea-Sierra Leone asymmetry
Demographic	OrPop	Origin population size, log-transformed, standardized
Demographic	DestPop	Destination population size, log-transformed, standardized
Demographic	OrPopDens	Origin population density, log-transformed, standardized
Demographic	DestPopDens	Destination population density, log-transformed, standardized
Demographic	OrTT100k	Estimated mean travel time in minutes to reach the nearest major settlement of at least 100,000 people at origin, log-transformed, standardized
Demographic	DestTT100k	estimated mean travel time in minutes to reach the nearest major settlement of at least 100,000 people at destination, log-transformed, standardized
Demographic	OrGrEcon	Origin Gridded economic output, log-transformed, standardized
Demographic	DestGrEcon	Destination Gridded economic output, log-transformed, standardized
Cultural	IntLangShared	The relative preference of transitioning between location pairs that are in different countries and share at least one of 17 vernacular languages
Cultural	NatLangShared	The relative preference of transitioning between location pairs that are in the same country and share at least one of 17 vernacular languages
Climatic	OrTemp	Temperature annual mean at origin, log-transformed, standardized
Climatic	DestTemp	Temperature annual mean at destination, log-transformed, standardized
Climatic	OrTempSS	Index of temperature seasonality at origin, log-transformed, standardized
Climatic	DestTempSS	Index of temperature seasonality at destination, log-transformed, standardized
Climatic	OrPrecip	Precipitation annual mean at origin, log-transformed, standardized
Climatic	DestPrecip	Precipitation annual mean at destination, log-transformed, standardized
Climatic	OrPrecipSS	Index of precipitation seasonality at origin, log-transformed, standardized
Climatic	DestPrecipSS	Index of precipitation seasonality at destination, log-transformed, standardized