Virus genomes reveal the factors that spread and sustained the West African Ebola epidemic.

Gytis Dudas[1,2,*], Luiz Max Carvalho[1], Trevor Bedford[2], Andrew J. Tatem[3,4], Guy Baele[5], Nuno Faria[6], Daniel J. Park[7], Jason Ladner[8], Armando Arias[9,10], Danny Asogun[11,12], Filip Bielejec[5], Sarah Caddy[9], Matt Cotten[13], Jonathan Dambrozio[8], Simon Dellicour[5], Antonino Di Caro[14,12], Joseph W. Diclaro II[15], Sophie Duraffour[16,12], Mike Elmore[17], Lawrence Fakoli[18], Merle Gilbert[8], Sahr M Gevao[19], Stephen Gire[7,20], Adrianne Gladden-Young[7], Andreas Gnirke[7], Augustine Goba[21,22], Donald S. Grant[21,22], Bart Haagmans[23], Julian A. Hiscox[24,25], Umaru Jah[26], Brima Kargbo[22], Jeffrey Kugelman[8], Di Liu[27], Jia Lu[9], Christine M. Malboeuf[7], Suzanne Mate[8], David A. Matthews[28], Christian B. Matranga[7], Luke Meredith[9,26], James Qu[7], Joshua Quick[29], Susan D. Pas[23], My VT Phan[13], Georgios Poliakis[24], Chantal Reusken[23], Mariano Sanchez-Lockhart[8,30], Stephen F. Schaffner[7], John S. Schieffelin[31], Rachel S. Sealfon[7], Etienne Simon-Loriere[32,33], Saskia L. Smits[23], Kilian Stoecker[34,12], Lucy Thorne[9], Ekaete A. Tobin[11,12], Mohamed A. Vandi[21,22], Simon J. Watson[13], Kendra West[7], Shannon Whitmer[35,†], Michael R. Wiley[8,30], Sarah M. Winnicki[7,20], Shirlee Wohl[7,20], Roman Wölfel[34,12], Nathan L. Yozwiak[7,20], Kristian G. Andersen[36,37,7], Sylvia Blyden[22], Fatorma Bolay[18], Miles Carroll[17,12], Bernice Dahn[38], Boubacar Diallo[39], Pierre Formenty[40], Christophe Fraser[41], George F. Gao[27,42], Robert F. Garry[43], Ian Goodfellow[9,26], Stephan Günther[16,12], Christian Happi[44], Edward C Holmes[45], Brima Kargbo[22], Paul Kellam[13,47], Marion P. G. Koopmans[23], Nicholas J. Loman[29], N'Faly Magassouba[48], Dhamari Naidoo[40], Stuart T. Nichol[35,†], Tolbert Nyenswah[38], Gustavo Palacios[8], Oliver G Pybus[6], Pardis Sabeti[7,20], Amadou Sall[32], Keïta Sakoba[46], Ute Ströeher[35,†], Isatta Wurie[22], Marc A Suchard[49,50,51], Philippe Lemey[5,*] & Andrew Rambaut[1,52,53,*]

[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, WA, 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, Leuven, Belgium, [6]Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3PS, UK, [7]Broad Institute of Harvard and MIT, Cambridge, MA 02138, USA, [8]Center for Genome Sciences, U.S. Army Medical Research Institute of Infectious Diseases, Fort Detrick, Frederick, MD 21702, USA, [9]Department of Pathology, University of Cambridge, Addenbrooke's Hospital, Cambridge, CB2 2QQ, UK, [10]Section for Virology, National Veterinary Institute, Technical University of Denmark, Blowsvej 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, United Kingdom, [14]National Institute for Infectious Diseases "L. Spallanzani" - IRCCS, Via Portuense 292, 00149 Rome, Italy, [15]Naval Medical Research Unit 3, 3A Imtidad Ramses Street, Cairo, 11517, Egypt, [16]Bernhard Nocht Institute for Tropical Medicine, 20359 Hamburg, Germany, [17]National Infections Service, Public Health England, Porton Down, Salisbury, Wilts SP4 0JG, UK, [18]Liberian Institute for Biomedical Research, Charlesville, Liberia., [19]University of Sierra Leone, Freetown, Sierra Leone , [20]Center for Systems Biology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA, [21]Viral Hemorrhagic Fever Program, Kenema Government Hospital, 1 Combema Road, Kenema, Sierra Leone, [22]Ministry of Health and Sanitation, 4th Floor Youyi Building, Freetown, Sierra Leone , [23]Department of Viroscience, Erasmus University Medical Centre, P.O. Box 20140, 300 CA Rotterdam, the Netherlands, [24]Institute of Infection and Global Health, University of Liverpool, Liverpool L69 2BE, United Kingdom, [25]NIHR Health Protection Research Unit in Emerging and Zoonotic Infections, University of Liverpool, UK., [26]University of Makeni, Makeni, Sierra Leone, [27]Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China, [28]University of Bristol, BS8 1TD, United Kingdom, [29]Institute of Microbiology and Infection, University of Birmingham, Birmingham B15 2TT, United Kingdom, [30]University of Nebraska Medical Center, Omaha, NE, USA, [31]Department of Pediatrics, Section of Infectious Diseases, New Orleans, LA 70112, USA, [32]Institut Pasteur, Functional Genetics of Infectious Diseases Unit, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France, [33]CNRS URA3012, Paris 75015, France, [34]Bundeswehr Institute of Microbiology, Neuherbergstrasse 11, 80937 Munich, Germany, [35]Viral Special Pathogens Branch, Centers for Disease Control and Prevention, 1600 Clifton Rd.

NE, Atlanta, Georgia, USA, [36]The Scripps Research Institute, Department of Immunology and Microbial Science, La Jolla, CA 92037, USA., [37]Scripps Translational Science Institute, La Jolla, CA 92037, USA, [38]Minstry of Health Liberia, Monrovia, Liberia, [39]World Health Organization, Conakry, Guinea, [40]WHO Ebola Response Team, [41]Department of Infectious Disease Epidemiology, MRC Centre for Outbreak Analyses and Modelling, School of Public Health, Imperial College London, UK, [42]Chinese Center for Disease Control and Prevention (China CDC), Beijing 102206, China, [43]Department of Microbiology & Immunology, New Orleans, LA 70112, USA, [44]Redeemer's University, Ede, Osun State, Nigeria , [45]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, NSW 2006, Australia, [46]Ministry of Health Guinea, Conakry, Guinea, [47]Division of Infectious Diseases, Imperial College Faculty of Medicine, London W2 1PG, UK, [48]Universit Gamal Abdel Nasser de Conakry, Laboratoire des Fivres Hmorragiques en Guine, Conakry, Guinea, [49]Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles, CA, USA, [50]Department of Biomathematics David Geffen School of Medicine at UCLA, University of California, Los Angeles, CA, USA, [51]Department of Human Genetics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, CA, USA, [52]Centre for Immunology, Infection and Evolution, University of Edinburgh, King's Buildings, Edinburgh, EH9 3FL, UK, [53]Fogarty International Center, National Institutes of Health, Bethesda, MD, USA

September 15, 2016

*Corresponding authors (a.rambaut@ed.ac.uk, gdudas@fredhutch.org, philippe.lemey@rega.kuleuven.be)

†The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Summary

The 2013-2016 epidemic of Ebola virus disease in West Africa was of unprecedented magnitude, duration and impact. Extensive collaborative sequencing projects have produced a large collection of over 1600 Ebola virus genomes, representing over 5% of known cases, unmatched for any single human epidemic. In a comprehensive analysis of this entire dataset, we reconstruct in detail the history of migration, proliferation and decline of Ebola virus throughout the region. We test the association of geography, climate, administrative boundaries, demography and culture with viral movement among 56 administrative regions. Our results show that during the outbreak viral lineages moved according to a classic 'gravity' model, with more intense migration between larger and more proximate population centers. Despite a strong attenuation of international dispersal after border closures, localized cross-border transmission beforehand had already set the seeds for an international epidemic, rendering these measures relatively ineffective in curbing the epidemic. We use this empirical evidence to address why the epidemic did not spread into neighboring countries, showing that although these regions were susceptible to developing significant outbreaks, they were also at lower risk of viral introductions. Finally, viral genome sequence data uniquely reveals this large epidemic to be a heterogeneous and spatially dissociated collection of transmission clusters of varying size, duration and connectivity. These insights will help inform approaches to intervention in such epidemics in the future.

Main text

At least 28,000 cases and 11,000 deaths (World Health Organization, 2016a) have been attributed to the Makona variant of Ebola virus (EBOV) (Kuhn et al., 2014) in the two and a half years that 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 (Baize et al., 2014). Initial efforts to control the outbreak in Guinea were considered to be succeeding (World Health Organization Regional Office for Africa, 2014), but in early 2014 the virus crossed international borders into neighbouring Liberia (first cases diagnosed in late March) and Sierra Leone (first documented case in late February (Goba et al., 2016; Sack et al., 2014), first diagnosed cases in May (Gire et al., 2014)). Viral genomes sequenced from three patients in Guinea early in the epidemic (Baize et al., 2014) helped to establish that the progenitor of the Makona variant originated in Central Africa and arrived in West Africa within the last 15 years (Dudas and Rambaut, 2014; Gire et al., 2014). Rapid sequencing of the first reported cases in Sierra Leone confirmed that EBOV had crossed the border from Guinea and they were not the result of an independent zoonotic introduction (Gire et al., 2014). Subsequent studies analysed the genetic makeup of the Makona variant but focused on infections in either Guinea (Carroll et al., 2015; Quick et al., 2016; Simon-Loriere et al., 2015), Sierra Leone (Arias et al., 2016; Park et al., 2015) or Liberia (Kugelman et al., 2015; Ladner et al., 2015), identifying local viral lineages and patterns of transmission within each country.

Although virus sequencing has covered considerable fractions of the epidemic in each country, individual studies focused on either limited geographical areas or periods of time, so that the regional level patterns and drivers of the epidemic across its entire duration have remained uncertain. Using 1610 genome sequences collected throughout the epidemic, which represent over 5% of known Ebola virus disease (EVD) cases (Figures 1 & S1), we apply phylogenetic approaches to reconstruct a detailed history of the movement of the virus within and among the three most affected countries. Using a recently developed approach for integrating covariates of spatial spread within a phylogeographic model (Lemey et al., 2014), we test which features of each region (administrative, economic, climatic, infrastructural and demographic) were important in shaping the spatial dynamics of EBOV. 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 in early December 2013 (95% highest posterior density interval: Oct 2013, Feb 2014) and phylogeographic estimation

assigns this ancestor to the Guéckédou préfecture with a high degree of confidence (96% posterior support) (Figure 2). In addition, we find that initial lineages deriving from this common ancestor circulated among Guéckédou and its neighbouring préfectures of Macenta and Kissidougou until late February 2014 (Figure 2). These results, based on a comprehensive sample of EBOV genomes, support the epidemiological evidence that the West African epidemic began in late 2013 in Guéckédou préfecture of Guinea (Baize et al., 2014).

The first introduction of EBOV from Guinea into another country that resulted in sustained transmission is estimated to have occurred in early April 2014 (Figure 2), when the virus spread to the Kailahun district of Sierra Leone (Goba et al., 2016; Sack et al., 2014). This lineage was first detected in Kailahun at the end of May 2014, from where it spread across the region (Figure 3 & S2). From Kailahun EBOV spread extremely rapidly in May 2014 into several counties of Liberia (Lofa, Montserrado and Margibi) (Ladner et al., 2015) and Guinea (Conakry, back into Guéckédou) (Carroll et al., 2015; Simon-Loriere et al., 2015). The virus continued to spread westwards through Sierra Leone, and by July 2014 it was present in the capital city, Freetown.

By mid-September 2014 Liberia was reporting >500 new EVD cases per week, mostly driven by a large outbreak in Montserrado county, which encompasses the capital city, Monrovia. Sierra Leone reported as many as 700 new cases per week by mid-November, driven by 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 dropping. By March 2015 the epidemic was largely under control in Liberia and eastern Guinea, although sustained transmission was still occurring in western Guinea and western Sierra Leone, near the border between the two countries. By the following month prevalence had declined such that only a handful of relatively distantly related lineages survived from the exponential growth phase of the epidemic (Arias et al., 2016; Quick et al., 2016) (Figure 3).

The last Ebola virus resulting from a conventionally-acquired infection was collected and sequenced in October 2015 in Forecariah préfecture (Guinea) (Quick et al., 2016). Following this, only sporadic cases of EVD were detected: in Margibi (Liberia) in June 2015, Montserrado (Liberia) in November 2015, Tonkolili (Sierra Leone) in January and February 2016, and Nzérékoré (Guinea) in March 2016. All these sporadic cases likely result from transmission from EVD survivors with established persistent infections (Blackley et al., 2016; Mate et al., 2015).

Factors associated with EBOV dispersal.

To determine the factors that influenced the spread of EBOV among administrative regions at the district (Sierra Leone), préfecture (Guinea) and county (Liberia) levels we employed a phylogeographic generalized linear model (GLM) (Lemey et al., 2014). Of the 25 factors assessed (see Table S2 for a full list and description) five were included in the model with categorical support (Table 1). In summary, EBOV migration events tend to occur between geographically close regions (great circle distance: Bayes factor (BF) support for inclusion BF>50). Half of all virus lineage movements occurred between locations <72 km apart and only 5% involved movement over 232 km (Figure 5a). Population sizes are very strongly (BF>50) positively correlated with viral dissemination, with a stronger effect for the population size of the origin location than that for the destination population size. The result, when combined with the inverse effect of geographic distance, implies that the epidemic's spread followed a classic gravity-model dynamic. Gravity models, widely used in economic and geographic studies, describe the movement of people between locations as a function of their population sizes and distance apart. They are a natural choice for modelling infectious disease transmission (Truscott and Ferguson, 2012; Viboud et al., 2006) and have been used in spatio-temporal modelling of EBOV transmission in Sierra Leone (Yang et al., 2015). Here we use viral genomes to provide empirical evidence that such a process drove viral dissemination during the epidemic.

In addition to geographical distance, we found a significant propensity for migration events to occur among administrative regions within each country, as opposed to international viral dispersal (National effect, BF>50), suggesting that country borders acted to curb the geographic spread of EBOV. Within-country viral migration is higher than international movement even after the direct effect of distance is accounted for. When international migrations do take place, they are more intense between administrative regions that meet on an international border (IntBoSh, BF>50).

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

Factors associated with local EBOV proliferation.

The analysis above identified factors that predict virus movement between administrative regions. These factors represent the degree of importation risk, i.e. the likelihood that a viral lineage initiates at least one infection in a new region, and are dominated by geographical and administrative factors. However, the epidemiological consequences of each introduction — the size and duration of resulting transmission chains — may be affected by different factors. To investigate this we explored which demographic, economic and climatic factors might predict cumulative case counts (World Health Organization, 2016a) for each region (Bayesian generalized linear model; see Supplementary Methods).

We find that cumulative case counts in each location 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 >50,000 inhabitants (tt50K, BF 32.4). These results confirm the common perception that, compared 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 make the Ebola virus more environmentally stable, especially in light of frequent post-mortem transmission of the virus (Fischer et al., 2015). Although we found no evidence of an association between EBOV migration and any aspects of local climate, we find that regions with less seasonal variation in temperature, and more rainfall, tended to have larger EVD outbreaks (TempSS, BF >50 and Precip, BF 4.4 respectively).

Did international travel restrictions have an effect?

It has been suggested that porous borders between Liberia, Sierra Leone and Guinea allowed unimpeded spread of EBOV during the 2013-2016 epidemic (Bausch and Schwarz, 2014; Chan, 2014; Wesolowski et al., 2014). Our results suggest that, on average, international borders were associated with a decreased rate of transmission events compared to national borders (Figure S3), but there were still frequent international cross-border transmission events. Specifically, these events were concentrated in Guéckédou (Guinea), Kailahun (Sierra Leone) and Lofa (Liberia) during the early phases of the epidemic (Figure S4b), and in the later stage of the epidemic (Figure S4b) between neighbouring Forécariah (Guinea) and Kambia (Sierra Leone). These later movements significantly hindered efforts to interrupt the final chains of transmission in late 2015, with a number of such chains moving back and forth across this border (Arias et al., 2016; Goodfellow et al., 2015; Quick et al., 2016). Sierra Leone announced border closures on 11 June 2014, followed by Liberia on 27 July 2014, and Guinea on 9 August 2014, although there is little information on what these border closures actually entailed. As a consequence, even though our results show that international viral spread was more intense before these borders were closed (mean change point: Aug-Sept 2014; 80.0% posterior support; (Figure 3b; see also Figure S5), it is difficult to ascertain whether it was the border closures themselves that were responsible for the apparent reduction in cross-border transmissions, as opposed to concomitant control efforts or public information campaigns. Overall, these results suggest that border closures may have reduced international traffic, particularly over longer distances and between larger population centres, but by the time Sierra Leone and later Liberia closed their borders the epidemic had become firmly established in both countries (Figure 3).

Why did the epidemic not spread further?

With the exception of a few documented exportations (Abdoulaye et al., 2015; Folarin et al., 2016; Hoenen et al., 2015), the West African Ebola virus epidemic did not spread into neighbouring regions of Guinea-Bissau, Senegal, Mali, or Côte d'Ivoire, and no cases were reported in seven préfectures of northern Guinea. By extending our GLM (i.e., the supported predictors and their estimated coefficients) to include these regions we can address whether these regions were spared EBOV cases through good fortune, or because they had 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 regions with cases (see supplementary methods). Overall, the contiguous regions in neighbouring countries were all predicted to low numbers of introductions (Figure 4a). They were not, however, predicted to have particularly low levels of transmission if an outbreak had been seeded (Figure 4b). Thus, it is likely that some of these surrounding regions and their countries overall were at risk of an EVD epidemic, but that their geographical distance from areas of active transmission and the attenuating effect of international borders prevented their epidemic potential from being realized. The Kati region in Mali and Tonkpi region in Côte d'Ivoire are to some extent exceptions to this general result, being more susceptible to viral introductions under the gravity model because of their large populations (Kati, 1 million; Tonkpi 950,000), (Figure 4a) and are predicted to have experienced many cases had EVD become established (Figure 4b).

The metapopulation structure and dynamics of the epidemic.

Figure 3 shows that after the initial establishment of transmission in Sierra Leone and Liberia, Guinea experienced repeated reintroductions of viral lineages from the escalating epidemics in these other two countries. From the 5% of cases that were sequenced, our analysis reveals that there were at least 21 (95% credible interval, CI: 18 - 24) re-introductions into Guinea from April 2014 to February 2015. Although an early epidemic lineage was established in the region around the Guinean capital, Conakry, and persisted for the duration of the epidemic (GN-1 in Figures 2 & 3), the continual 'seeding' of EBOV lineages into Guinea without a clear peak in transmission suggests that the virus may have been struggling to maintain transmission in that country. There were also numerous introductions into Sierra Leone over a similar time period (median: 9, 95% CI: 7 - 11) but the resulting transmission chains constituted a tiny proportion of the Sierra Leonean epidemic, with the bulk of transmission resulting from one early introduction (Figure 3a).

The importance of repeated seeding as a factor in the longevity of the epidemic is also suggested by the pattern of viral movement among administrative regions within each country (Figure S6). Regional epidemics were the result of multiple overlapping introduction events followed by within-region spread and occasional onward transmission to other regions. This observation suggests a metapopulation model in which viral persistence is driven by introduction into novel contact networks rather than by mass-action susceptible-infectious-recovered (SIR) dynamics (Ferrari et al., 2008). We find that, on average, EBOV migrates between administrative regions at a rate of 0.85 events per lineage per year (95% CI: 0.72, 0.97). If we assume a serial interval of 15.3 days (WHO Ebola Response Team, 2014), this translates to a 3.6% chance (95% CI: 3.0%, 4.1%) that a single step in the transmission chain migrates between regions. The detection and isolation of these mobile cases may have a disproportionate effect on the control of the epidemic.

Many regions experienced numerous independent introductions (Figure 5b) but the size of the clusters of cases that result from these introductions was generally small (with a mean cluster size of 4.3 and only 5% larger than 17 in our sample; Figure 5c) and their persistence of limited duration (a mean persistence time of 41.3 days with only 5% greater than 181 days; Figure 5d). Here, we define a 'cluster' as a group of sequenced cases that derive from a single introduction event into a region without including subsequent infections in other regions and persistence as the time between the introduction event and the last sampled case in the cluster. These definitions are conservative with regards to sampling intensity as we expect additional samples would split apart clusters rather than join them. Furthermore, introductions that were not detected will be disproportionately smaller, and so the cluster size estimate will be biased towards larger sizes. Thus, with 5.8% sampling, we arrive at a conservative estimate of approximately 75 regional cases per introduction event. Although larger population centres, in particular the capital cities, generally had more introductions (Figure S7a) the cluster sizes are less strongly associated with population size (Figure S7b). The frequent extinction of these clusters even though a small fraction of individuals were infected suggests that they were

constrained by the degree of connectedness among contact networks. Thus, it appears the West African epidemic was sustained by frequent seeding that resulted in numerous small local clusters of cases, some of which went on to seed further local clusters.

Viral genomics as a tool for outbreak response.

The 2013-2016 EVD epidemic in West Africa has unfortunately become a costly lesson in dealing with an infectious disease outbreak when both the exposed population and the international community are unprepared. It also demonstrates the value of pathogen genome sequencing in a public healthcare emergency situation and the value of timely pre-publication data sharing in order to identify the origins of imported lineages, to track viral transmission as the epidemic progresses, and to follow up on individual cases as the epidemic subsides. Real-time virus genome sequencing at the point of diagnosis can provide additional insights, especially when conventional epidemiological contact tracing is challenging. Other sources of human mobility data, mobile phone network data in particular, are promising but currently such data is difficult to obtain in a timely fashion (Wesolowski et al., 2014). It is inevitable that as sequencing becomes cheaper, more portable and accurate, real-time viral surveillance and molecular epidemiology will be routinely deployed on the frontlines of infectious disease outbreaks (Gardy et al., 2015; Yozwiak et al., 2015; Woolhouse et al., 2015; Quick et al., 2016). As viral genome sequencing is scaled up and gets closer to the time-scale of viral evolution, the pressure will increasingly fall on analysis techniques to provide the necessary temporal resolution to inform outbreak response. The analysis of the comprehensive EBOV genome set collected during the 2013-2016 epidemic, including the findings presented here and in other studies (Arias et al., 2016; Carroll et al., 2015; Gire et al., 2014; Kugelman et al., 2015; Ladner et al., 2015; Park et al., 2015; Simon-Loriere et al., 2015; Stadler et al., 2014; Tong et al., 2015) will provide a framework for predicting the behaviour of future outbreaks for EBOV, other filoviruses, and perhaps other human pathogens.

Many open 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 (Blackley et al., 2016; Mate et al., 2015; World Health Organization, 2016c,b) . Although, in hindsight, such sequelae were not entirely unexpected (Rowe et al., 1999), 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 hence of future outbreaks of this devastating disease.

Methods Summary

A total of 1610 nearly complete EBOV genome sequences were collated, aligned and annotated with date of sampling and likely location of infection (all data available from `https://github.com/ebov/space-time`). Geographical, demographic and climatic variables were collated for each of 63 regions in three focal countries, and for a further 18 regions in surrounding countries that reported no cases or no sustained transmission (see supplemental information for details). Time structured phylogenies were inferred using BEAST (Drummond et al., 2012; Ayres et al., 2012) and these formed the basis of a phylogenetic generalized linear model (Lemey et al., 2014) that infers the probability of inclusion, and degree of correlation, of each of the predictor variables for the spatial pattern of virus lineage migration. Along each branch of the tree we infer change among regions (Minin and Suchard, 2008). For those variables in the model with significant support, we extended the analysis to allow a single step-change in coefficient and inferred the time of this change-point. Furthermore, we used the inferred spatial model to estimate the expected number of migrations into regions which experience no known cases of EVD including in the surrounding countries. Finally, to assess which of the demographic and climatic variable were predictive of the magnitude of outbreak once introduced into a region, we employed generalized linear models and Bayesian model averaging, with cumulative case counts in each affected region as a response variable.

Supplementary Methods

Sequence data

We compiled a data set of 1610 publicly available full Ebola virus (EBOV) genomes sampled between 17 March 2014 and 24 October 2015 (see https://github/ebov/space-time/data/ for full list and metadata). The number of sequences and the proportion of cases sequenced varies with country; our data set contains 209 sequences from Liberia (3.8% of known and suspected cases), 982 from Sierra Leone (8.0%) and 368 from Guinea (9.2%) (Table S1). Most (1100) genomes are of high quality, with ambiguous sites and gaps comprising less than 1% of total alignment length, followed by sequences with between 1% and 2% of sites comprised of ambiguous bases or gaps (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 (Blackley et al., 2016; Mate et al., 2015). Sequences were aligned using MAFFT (Katoh et al., 2002) and edited manually. The alignment was partitioned into coding regions and non-coding intergenic regions with a final alignment length of 18992 nucleotides (available from https://github/ebov/space-time/data/).

Masking putative ADAR edited sites

As noticed by Tong et al. (2015), Park et al. (2015) and other studies, 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 (Bass and Weintraub, 1988). ADARs have been suggested to act on RNAs from numerous groups of viruses (Gélinas et al., 2011). 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 (Cattaneo et al., 1988; Rueda et al., 1994; Carpenter et al., 2009). Multiple T-to-C mutations are introduced simultaneously via ADAR-mediated RNA editing which would interfere with molecular clock estimates and, by extension, the tree topology. We thus designate 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, *i.e.* every T-to-C mutation in a stretch occurred on a single branch. We detect 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 survive hypermutation are transmitted (Smits et al., 2015). 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 (Hasegawa et al., 1985; Yang, 1994) substitution model 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 (Drummond et al., 2006). A non-parametric coalescent 'Skygrid' tree prior was employed for demographic inference (Gill et al., 2013). The overall evolutionary rate was given an uninformative continuous-time Markov chain (CTMC) reference prior (Ferreira and Suchard, 2008), 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/.

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: préfecture 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=222), we inferred the sequence location as a latent variable with equal prior probability over all available regions within that country. In the absence of any geographic information (N=2) we inferred both the country and the region of a sequence.

We deploy an asymmetric continuous-time Markov chain (CTMC) (Lemey et al., 2009; Edwards et al., 2011) matrix to infer instantaneous transitions between regions. For 56 regions with recorded EVD cases, a total of 3080 independent transition rates would be challenging to infer from one realisation of the process, even when reduced to a sparse migration matrix using stochastic search variable selection (SSVS) (Lemey et al., 2009).

Thus, to infer the spatial phylogenetic diffusion history between the $K = 56$ locations, we adopt a sparse generalized linear model (GLM) formulation of continuous-time Markov chain (CTMC) diffusion (Lemey et al., 2014). This model parameterizes the instantaneous movement rate Λ_{ij} from location i to location j as a log-linear function of P potential predictors $\mathbf{X}_{ij} = (x_{ij1}, \ldots, x_{ijP})'$ with unknown coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)'$ and diagonal matrix $\boldsymbol{\delta}$ with entries $(\delta_1, \ldots, \delta_P)$. These latter unknown indicators $\delta_p \in \{0, 1\}$ determine predictor p's inclusion in or exclusion from the model. 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

$$\begin{aligned} \log \Lambda_{ij} &= \mathbf{X}_{ij}' \boldsymbol{\delta}\boldsymbol{\beta} + \epsilon_i + \epsilon_j, \\ \epsilon_k &\sim \text{Normal}(0, \sigma^2) \text{ for } k = 1, \ldots, K, \text{ and} \\ \sigma^2 &\sim \text{Inverse-Gamma}(0.001, 0.001), \end{aligned} \tag{1}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_K)$ 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 Lemey et al. (2014) in specifying that *a priori* all β_p are independent and normally distributed with mean 0 and a relatively large variance of 4 and in 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 it is straightforward to see that $q = 1 - w^{1/P}$, 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 (Table S2). 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 (cfr. Fig. S1), but these predictors did not yield any noticeable support (data not shown).

To draw posterior inference, we follow Lemey et al. (2014) integrating $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$, and further employ a random-walk Metropolis transition kernel on $\boldsymbol{\epsilon}$ 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, two independent MCMC chains were run in BEAST 1.8.4 (Drummond et al., 2012) for 100 million states, sampling every 10 000 states. The first 1000 samples in each chain were removed as burnin, and the remaining 18 000 samples combined between the two runs. These 18 000 samples were used to estimate a maximum clade credibility tree and to estimate posterior densities for individual parameters.

To obtain realisations of the phylogenetic CTMC process, including both transitions (Markov jumps) between states and waiting times (Markov rewards) within states, we employ posterior inference of the complete

Markov jump history through time (Minin and Suchard, 2008; Lemey et al., 2014). 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:

$$\zeta_{ij} = \tau_i \mu \Lambda_{ij} \pi_i / c \tag{2}$$

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 normalising 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 parameterised as a log linear function of the predictors, we can obtain these rate through the coefficient estimates from the analysis and and 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 calculate the expected number of transitions from a particular phylogeographic location to any destination location, we sum over all destination locations (with and without cases, $n = 81$). 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 1000 trees from the posterior distribution of the main phylogenetic analysis (Lemey et al., 2014). We summarise mean posterior estimates for the transition expectations based on the samples obtained by our MCMC analysis; we note that also the value of c is sample-specific.

To consider time-inhomogeneity in the spatial diffusion process, we start by borrowing epoch modelling concepts from Bielejec et al. (2014). 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 $\mathbf{X}_{ijt} = (x_{ijt1}, \ldots, x_{ijtP})'$ with constant-through-time, unknown coefficients $\boldsymbol{\beta}$. We generalize this model to incorporate time-varying contribution of the predictors through time-varying coefficients $\boldsymbol{\beta}(t)$ using a series of change-point processes. Specifically, the time-varying epoch GLM models

$$\begin{aligned} \log \Lambda_{ijt} &= \mathbf{X}'_{ijt} \boldsymbol{\beta}(t) \\ \boldsymbol{\beta}(t) &= [\mathbf{I} - \boldsymbol{\phi}(t)]\, \boldsymbol{\beta}_{\mathrm{B}} + [\boldsymbol{\phi}(t)]\, \boldsymbol{\beta}_{\mathrm{A}}, \end{aligned} \tag{3}$$

where $\boldsymbol{\beta}_{\mathrm{B}} = (\beta_{\mathrm{B}1}, \ldots, \beta_{\mathrm{B}P})'$ are the unknown coefficients before the change-points, $\boldsymbol{\beta}_{\mathrm{A}} = (\beta_{\mathrm{A}1}, \ldots, \beta_{\mathrm{A}P})'$ are the unknown coefficients after the change-points, diagonal matrix $\boldsymbol{\phi}(t)$ has entries $(1_{t>t_1}(t), \ldots, 1_{t>t_P}(t))$, $1_{(\cdot)}(t)$ is the indicator function and $\mathbf{T} = (t_1, \ldots, t_P)$ are the unknown change-point times. In this general form, the contribution of predictor p before its change-point time t_p is $\beta_{\mathrm{B}p}$ and its contribution after is $\beta_{\mathrm{A}p}$ 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 that *a priori* all $\beta_{\mathrm{B}p}$ and $\beta_{\mathrm{A}p}$ are independent and normally distributed with mean 0 and a relatively large variance of 4. Under the prior, each t_p is equally likely to lie before any epoch.

We employ random-walk Metropolis transition kernels on $\boldsymbol{\beta}_{\mathrm{B}}$, $\boldsymbol{\beta}_{\mathrm{A}}$ 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 (t_{dis}), within country effect (t_{wco}), shared international border (t_{sib}) and origin and destination population size (t_{pop_o} and t_{pop_d}) change-points. To quantify the evidence in favour of each change-point, we calculate Bayes factor support based on 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 within country effect, we subsequently estimate the effect sizes before and after t_{wco}, keeping the remaining four predictors homogeneous through time.

Within-location generalized linear models

Case counts

Ebola virus disease (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 lab-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 sequence have 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 Table S2. We also employed SSVS as described above, in order to compute Bayes factors (BF) 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.

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

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 employed 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 Markov chain Monte Carlo (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 (ESS) greater than 200. We ran multiple chains to ensure 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.

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme [FP7/2007-2013] under Grant Agreement No 278433-PREDEMICS — Philippe Lemey and Andrew Rambaut, and ERC Grant Agreement No 260864 — Philippe Lemey, Andrew Rambaut and Marc A. Suchard. This work was supported by the European Unions Horizon 2020 research and innovation program (grant agreement no. 666100; EVIDENT) and the Directorate-General for International Cooperation and Development of the European Commission (service contract IFS/2011/272-372, EMLab) — Miles Carroll, David A. Matthews, Julian A. Hiscox, Antonino Di Caro, Roman Wlfel, Danny Asogun, Ekaete Alice Tobin, Joshua Quick, Nicholas J. Loman, Sophie Duraffour and Stephan Günther. European Unions Horizon 2020 research and innovation program (grant agreement No 643476.; COMPARE) — Marion Koopmans and Andrew Rambaut. National Institutes of Health (R01 AI107034, R01 HG006139 and R01 LM011827) and the National Science Foundation (IIS 1251151 and DMS 1264153) — Marc A. Suchard. National Health & Medical Research Council (Australia) — Edward C. Holmes. NIH AI081982, AI082119, AI082805 AI088843, AI104216, AI104621, AI115754, HSN272200900049C, and HHSN272201400048C — Robert F. Garry. The 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 U.S. Department of Defense — Gustavo Palacios. Bill and Melinda Gates Foundation (OPP1106427, 1032350, OPP1134076), Wellcome Trust Sustaining Health Grant (106866/Z/15/Z), Clinton Health Access Initiative — Andrew J. Tatem. This work was supported by the National Institute for Health Research Health Protection Research Unit in Emerging and Zoonotic Infections — Julian A. Hiscox. Key Research and Development Program (grant no. 2016YFC1200800) from the Ministry of Science and Technology of China — Di Lui. National Natural Science Foundation of China (NSFC, grant Nos. 81590760 and 81321063) — George F. Gao.

Colour-blind-friendly colour palettes by Cynthia Brewer, Pennsylvania State University (http://colorbrewer2.org). We gratefully acknowledge support from NVIDIA Corporation with the donation of parallel computing resources used for this research. Finally, we would like to recognize the contributions made by our colleagues who tragically died from Ebola virus disease whilst fighting the epidemic. In particular, we honor the memory of Dr. Sheik Humarr Khan and Nurse Mbalu Fonnie, whose careers were dedicated to viral hemorrhagic fever research.

References

Abdoulaye B, Moussa S, Daye K, et al. (11 co-authors). 2015. Experience on the management of the first imported ebola virus disease case in senegal. The Pan African medical journal. 22 Suppl 1:6.

Arias A, Armando A, Watson SJ, et al. (11 co-authors). 2016. Rapid outbreak sequencing of ebola virus in sierra leone identifies transmission chains linked to sporadic cases. Virus Evolution. 2:vew016.

Ayres DL, Darling A, Zwickl DJ, et al. (11 co-authors). 2012. Beagle: an application programming interface and high-performance computing library for statistical phylogenetics. Systematic biology. 61:170173.

Baize S, Pannetier D, Oestereich L, et al. (11 co-authors). 2014. Emergence of zaire ebola virus disease in guinea. The New England journal of medicine. 371:14181425.

Bass BL, Weintraub H. 1988. An unwinding activity that covalently modifies its double-stranded RNA substrate. Cell. 55:1089–1098.

Bausch DG, Schwarz L. 2014. Outbreak of ebola virus disease in guinea: Where ecology meets economy. PLoS neglected tropical diseases. 8:e3056.

Bielejec F, Lemey P, Baele G, Rambaut A, Suchard MA. 2014. Inferring heterogeneous evolutionary processes through time: from sequence substitution to phylogeography. Syst. Biol. 63:493–504.

Blackley DJ, Wiley MR, Ladner JT, et al. (11 co-authors). 2016. Reduced evolutionary rate in reemerged ebola virus transmission chains. Science advances. 2:e1600378.

Carpenter JA, Keegan LP, Wilfert L, O'Connell MA, Jiggins FM. 2009. Evidence for ADAR-induced hypermutation of the Drosophila sigma virus (Rhabdoviridae). BMC Genetics. 10:75.

Carroll MW, Matthews DA, Hiscox JA, et al. (11 co-authors). 2015. Temporal and spatial analysis of the 2014-2015 ebola virus outbreak in west africa. Nature. 524:97101.

Cattaneo R, Schmid A, Eschle D, Baczko K, ter Meulen V, Billeter MA. 1988. Biased hypermutation and other genetic changes in defective measles viruses in human brain infections. Cell. 55:255–265.

Chan M. 2014. Ebola virus disease in west africa–no early end to the outbreak. The New England journal of medicine. 371:11831185.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with beauti and the beast 1.7. Molecular biology and evolution. 29:19691973.

Dudas G, Rambaut A. 2014. Phylogenetic analysis of guinea 2014 ebov ebolavirus outbreak. PLoS Currents. 6.

Edwards CJ, Suchard MA, Lemey P, et al. (18 co-authors). 2011. Ancient Hybridization and an Irish Origin for the Modern Polar Bear Matriline. Current Biology. 21:1251–1258.

Ferrari MJ, Grais RF, Bharti N, Conlan AJK, Bjrnstad ON, Wolfson LJ, Guerin PJ, Djibo A, Grenfell BT. 2008. The dynamics of measles in sub-saharan africa. Nature. 451:679684.

Ferreira MAR, Suchard MA. 2008. Bayesian analysis of elapsed times in continuous-time markov chains. Canadian Journal of Statistics. 36:355–368.

Fischer R, Judson S, Miazgowicz K, Bushmaker T, Prescott J, Munster VJ. 2015. Ebola virus stability on surfaces and in fluids in simulated outbreak environments. Emerging infectious diseases. 21:12431246.

Folarin OA, Ehichioya D, Schaffner SF, et al. (11 co-authors). 2016. Ebola virus epidemiology and evolution in nigeria. The Journal of infectious diseases. .

Gardy J, Loman NJ, Rambaut A. 2015. Real-time digital pathogen surveillance — the time is now. Genome biology. 16:155.

Gélinas JF, Clerzius G, Shaw E, Gatignol A. 2011. Enhancement of Replication of RNA Viruses by ADAR1 via RNA Editing and Inhibition of RNA-Activated Protein Kinase. Journal of Virology. 85:8460–8466.

Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. 2013. Improving bayesian population dynamics inference: A coalescent-based model for multiple loci. Molecular Biology and Evolution. 30:713–724.

Gire SK, Goba A, Andersen KG, et al. (11 co-authors). 2014. Genomic surveillance elucidates ebola virus origin and transmission during the 2014 outbreak. Science. 345:13691372.

Goba A, Khan SH, Fonnie M, et al. (11 co-authors). 2016. An outbreak of ebola virus disease in the lassa fever zone. The Journal of infectious diseases. .

Goodfellow I, Reusken C, Koopmans M. 2015. Laboratory support during and after the ebola virus endgame: towards a sustained laboratory infrastructure. Euro surveillance: bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin. 20.

Hasegawa M, Kishino H, Yano Ta. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution. 22:160–174.

Hoenen T, Safronetz D, Groseth A, et al. (11 co-authors). 2015. Mutation rate and genotype variation of ebola virus from mali case sequences. Science. 348:117119.

Katoh K, Misawa K, Kuma Ki, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research. 30:3059–3066.

Kugelman JR, Wiley MR, Mate S, et al. (11 co-authors). 2015. Monitoring of ebola virus makona evolution through establishment of advanced genomic capability in liberia. Emerging infectious diseases. 21:11351143.

Kuhn JH, Andersen KG, Baize S, et al. (11 co-authors). 2014. 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. 6:47604799.

Ladner JT, Wiley MR, Mate S, et al. (11 co-authors). 2015. Evolution and spread of ebola virus in liberia, 2014–2015. Cell host and microbe. 18:659669.

Lemey P, Rambaut A, Bedford T, et al. (11 co-authors). 2014. Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3n2. PLOS Pathog. 10:e1003932.

Lemey P, Suchard M, Rambaut A. 2009. Reconstructing the initial global spread of a human influenza pandemic. PLoS Currents. 1.

Mate SE, Kugelman JR, Nyenswah TG, et al. (11 co-authors). 2015. Molecular evidence of sexual transmission of ebola virus. The New England journal of medicine. 373:24482454.

Minin VN, Suchard MA. 2008. Fast, accurate and simulation-free stochastic mapping. Philosophical Transactions of the Royal Society of London B: Biological Sciences. 363:3985–3995.

Park DJ, Dudas G, Wohl S, et al. (11 co-authors). 2015. Ebola virus epidemiology, transmission, and evolution during seven months in sierra leone. Cell. 161:15161526.

Quick J, Loman NJ, Duraffour S, et al. (11 co-authors). 2016. Real-time, portable genome sequencing for ebola surveillance. Nature. 530:228232.

Rowe AK, Bertolli J, Khan AS, et al. (11 co-authors). 1999. Clinical, virologic, and immunologic follow-up of convalescent ebola hemorrhagic fever patients and their household contacts, kikwit, democratic republic of the congo. commission de lutte contre les epidmies kikwit. The Journal of infectious diseases. 179 Suppl 1:S2835.

Rueda P, García-Barreno B, Melero JA. 1994. 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.

Sack K, Fink S, Belluck P, Nossiter A, Berehulak D. 2014. How ebola roared back.

Simon-Loriere E, Faye O, Faye O, et al. (11 co-authors). 2015. Distinct lineages of ebola virus in guinea during the 2014 west african epidemic. Nature. 524:102104.

Smits SL, Pas SD, Reusken CB, et al. (17 co-authors). 2015. Genotypic anomaly in Ebola virus strains circulating in Magazine Wharf area, Freetown, Sierra Leone, 2015. Euro Surveillance: Bulletin Européen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin. 20.

Stadler T, Khnert D, Rasmussen DA, du Plessis L. 2014. Insights into the early epidemic spread of ebola in sierra leone provided by viral sequence data. PLoS currents. 6.

Tong YG, Shi WF, Liu D, et al. (11 co-authors). 2015. Genetic diversity and evolutionary dynamics of ebola virus in sierra leone. Nature. 524:9396.

Truscott J, Ferguson NM. 2012. Evaluating the adequacy of gravity models as a description of human mobility for epidemic modelling. PLoS computational biology. 8:e1002699.

Viboud C, Bjrnstad ON, Smith DL, Simonsen L, Miller MA, Grenfell BT. 2006. Synchrony, waves, and spatial hierarchies in the spread of influenza. Science. 312:447451.

Wesolowski A, Buckee CO, Bengtsson L, Wetter E, Lu X, Tatem AJ. 2014. Commentary: containing the ebola outbreak - the potential and challenge of mobile network data. PLoS currents. 6.

WHO Ebola Response Team. 2014. Ebola virus disease in west africa–the first 9 months of the epidemic and forward projections. The New England journal of medicine. 371:14811495.

Woolhouse MEJ, Rambaut A, Kellam P. 2015. Lessons from ebola: Improving infectious disease surveillance to inform outbreak management. Science translational medicine. 7:307rv5.

World Health Organization. 2016a. Ebola situation report - 10 june 2016.

World Health Organization. 2016b. End of ebola transmission in guinea.

World Health Organization. 2016c. New ebola case in sierra leone. who continues to stress risk of more flare-ups.

World Health Organization Regional Office for Africa. 2014. Ebola virus disease, west africa (situation as of 25 april 2014).

Yang W, Zhang W, Kargbo D, et al. (11 co-authors). 2015. Transmission network of the 20142015 ebola epidemic in sierra leone. Journal of the Royal Society, Interface / the Royal Society. 12:20150536.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. Journal of Molecular Evolution. 39:306–314.

Yozwiak NL, Schaffner SF, Sabeti PC. 2015. Data sharing: Make outbreak research open access. Nature. 518:477.

Figures

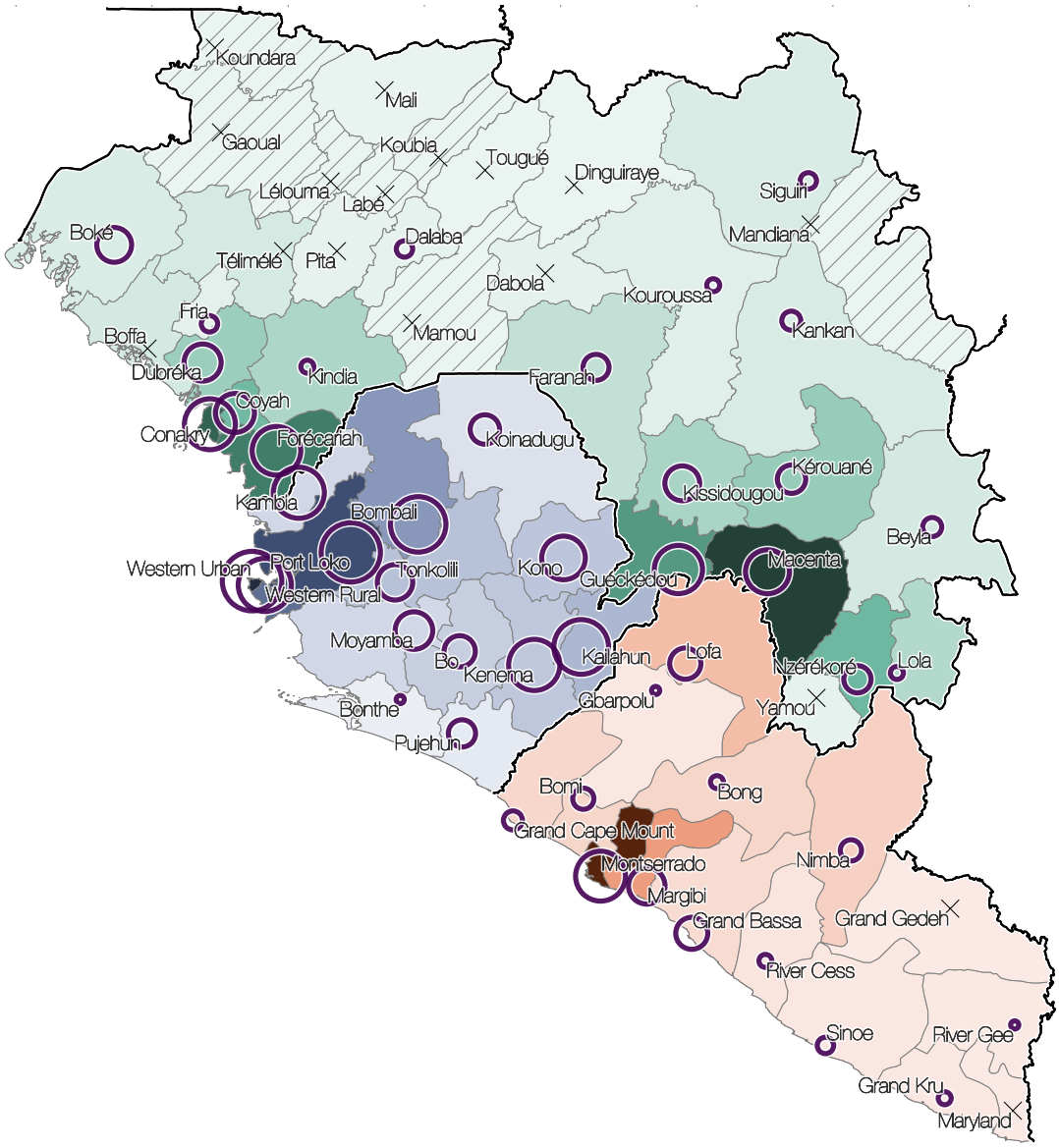


Figure 1. Distribution of EBV cases and virus sequences. Administrative regions within Guinea (green), Sierra Leone (blue) and Liberia (red); shading is proportional to the cumulative number of known and suspected EVD cases in each region. Darkest shades represent 784 cases for Guinea (Macenta), 3219 cases for Sierra Leone (Western Urban) and 2925 cases for Liberia (Montserrado); hatched areas indicate regions without any reported EVD cases. Circle diameters are proportional to the number of sequences available from that region over the entire 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. The number of sequences and number of cases for each region where cases were recorded are strongly correlated (Spearman rank correlation coefficient 0.93; Supplementary Figure S1).

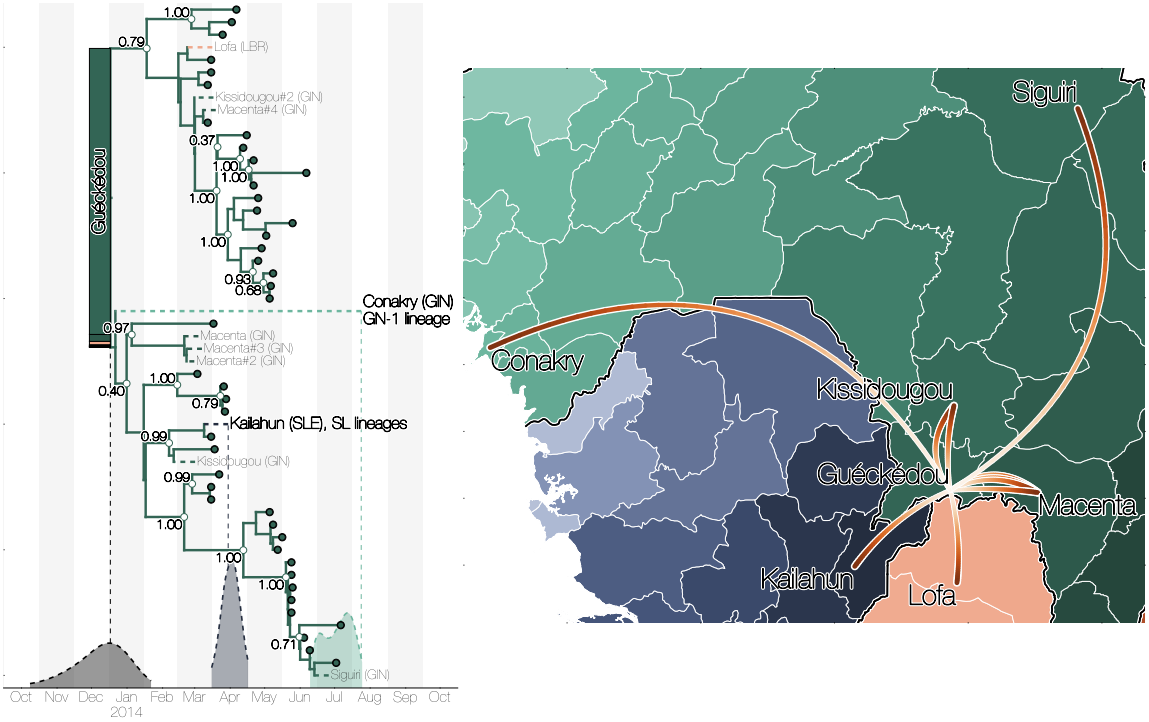


Figure 2. Summary of early epidemic events. a) The time-scaled phylogeny of the early sampled cases in Guéckédou, Guinea and their relationships to the initial dispersal events into other neighbouring and more distant regions. Stacked bars at the root of the tree indicate posterior probabilities for the origin of the epidemic (0.96 for Guéckédou, 0.02 for Macenta, 0.01 for Lofa and negligible probabilities for other locations). 95% posterior densities of the time of the common ancestor of all lineages (grey) and far-dispersing lineages into Kailahun district (blue, introduction gave rise to SL lineages) and to Conakry préfecture (green, introduction leads to lineage GN-1) are shown at the bottom of the tree. Nodes with three or more tips have posterior probabilities shown if > 0.3. b) These same dispersal events (marked by dashed lineages on the phylogeny) projected on a map with directionality indicated by colour intensity (from white to red). Lineages that migrated to Conakry and Kailahun have led to the vast majority of cases throughout the region.

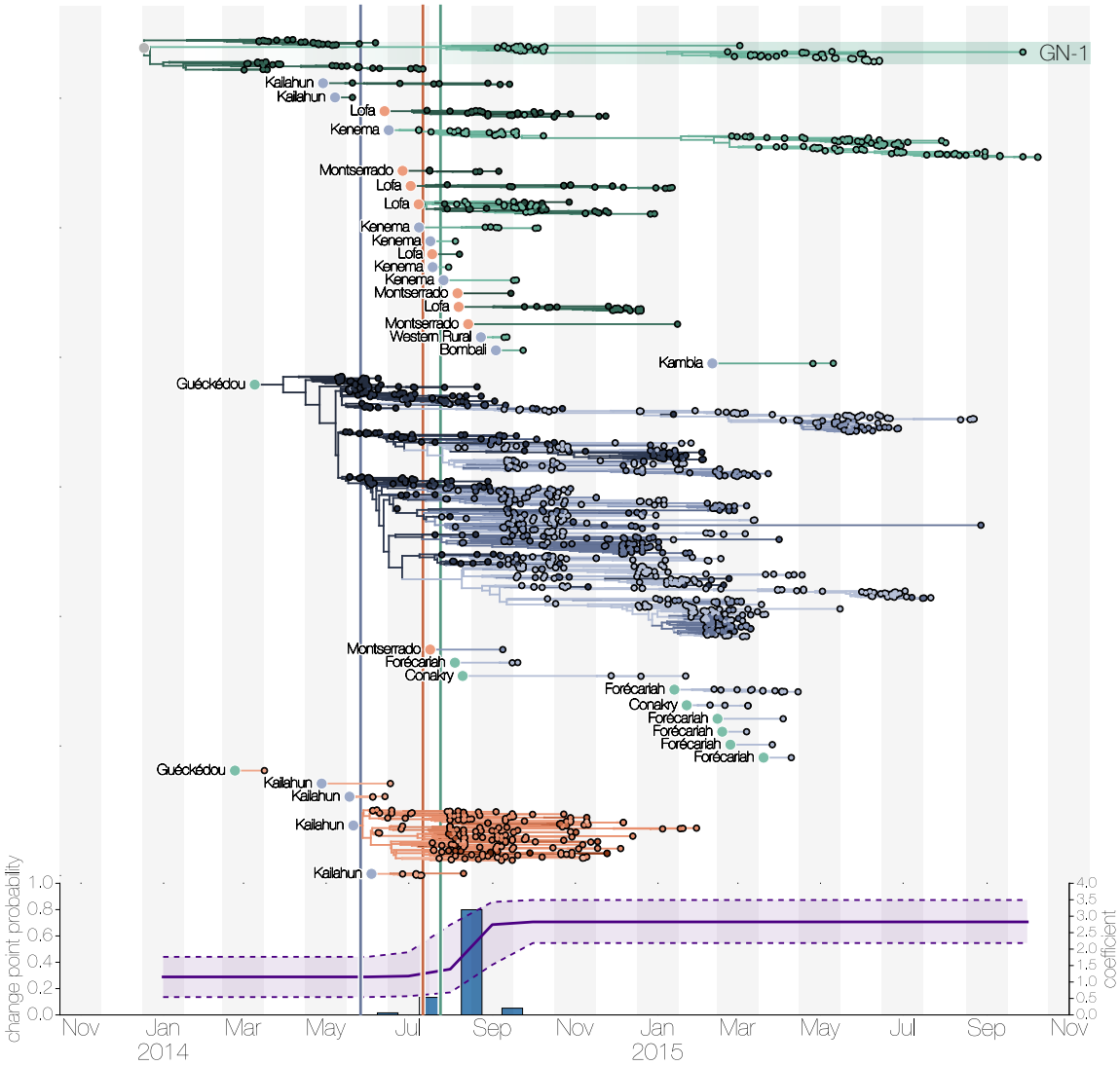
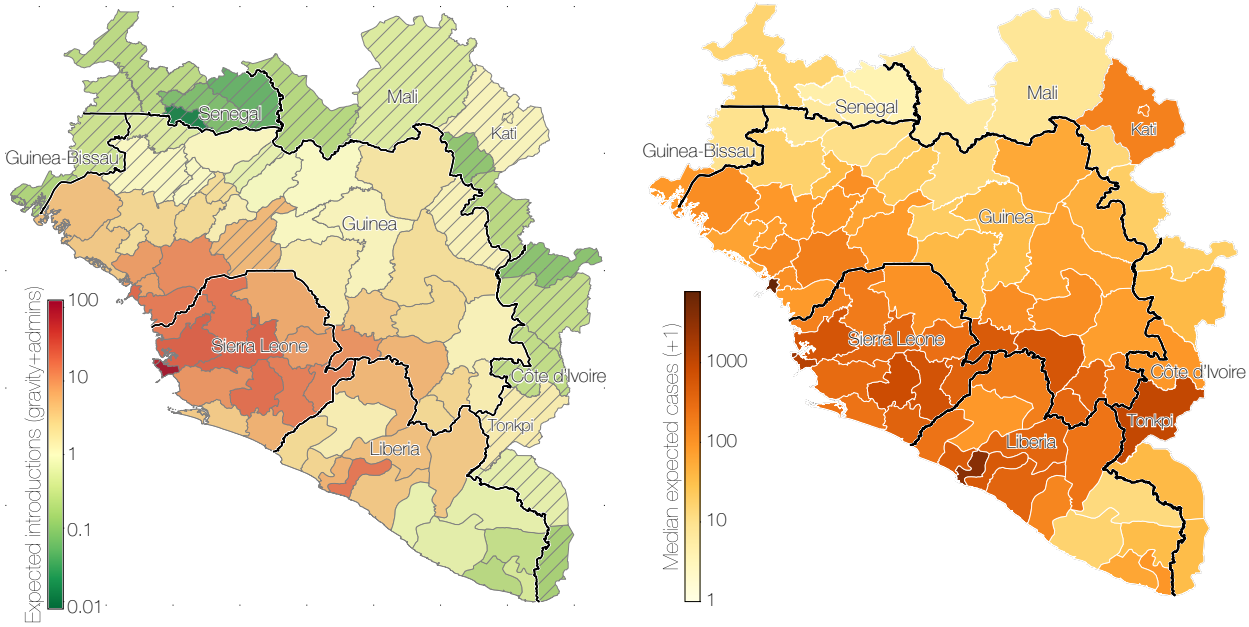


Figure 3. Time-scaled phylogeny deconstructed into country-specific transmission chains arising from independent international movements. a) EBOV lineages, tracked until the sampling date of their last known descendants, sorted by country (Guinea, green; Sierra Leone, blue; Liberia, red) and earliest possible introduction date. Tips are shared by longitude (lightest to the West, darkest to the East). Circles at root of each subtree denote the country of origin for the introduced lineage. The four introductions into Liberia in May-June 2014 are all inferred to have come from Kailahun, Sierra Leone. These may represent a few, or just one, movement events; however, the genetic similarity of these Liberian genomes to viruses from Kailahun makes further resolution impossible. In contrast, the multiple introductions into Guinea are very likely the result of multiple separate movement events over a 10 month period. b) Epoch estimates of the change point probability (primary Y-axis) and log coefficient (mean and credible interval; secondary Y-axis) for the within country-effect (the only effect with support for epoch dynamics; see Supplementary Information). The highest change point probability and an associated doubling of log effect size for within country transmission is estimated between August and September 2014 (blue columns). Vertical lines represent dates of border closures by the respective countries (Sierra Leone, blue; Liberia, red; Guinea, green).



a) b)

Figure 4. Predicted destinations and consequences of viral migrations. a) Predicted number of imports into each of 63 regions in Guinea, Sierra Leone and Liberia (including 7 with no recorded cases in Guinea) and the surrounding 18 regions from the neighbouring countries of Guinea-Bissau, Senegal, Mali and Côte d'Ivoire. The expected number of exports from locations in the phylogeographic tree and imports to any location are calculated based on the phylogeographic GLM model estimates and associated predictors that were extended to apparently EVD-free locations (see Supplementary Methods). b) Predicted outbreak sizes from the generalized linear model fitted to case data.

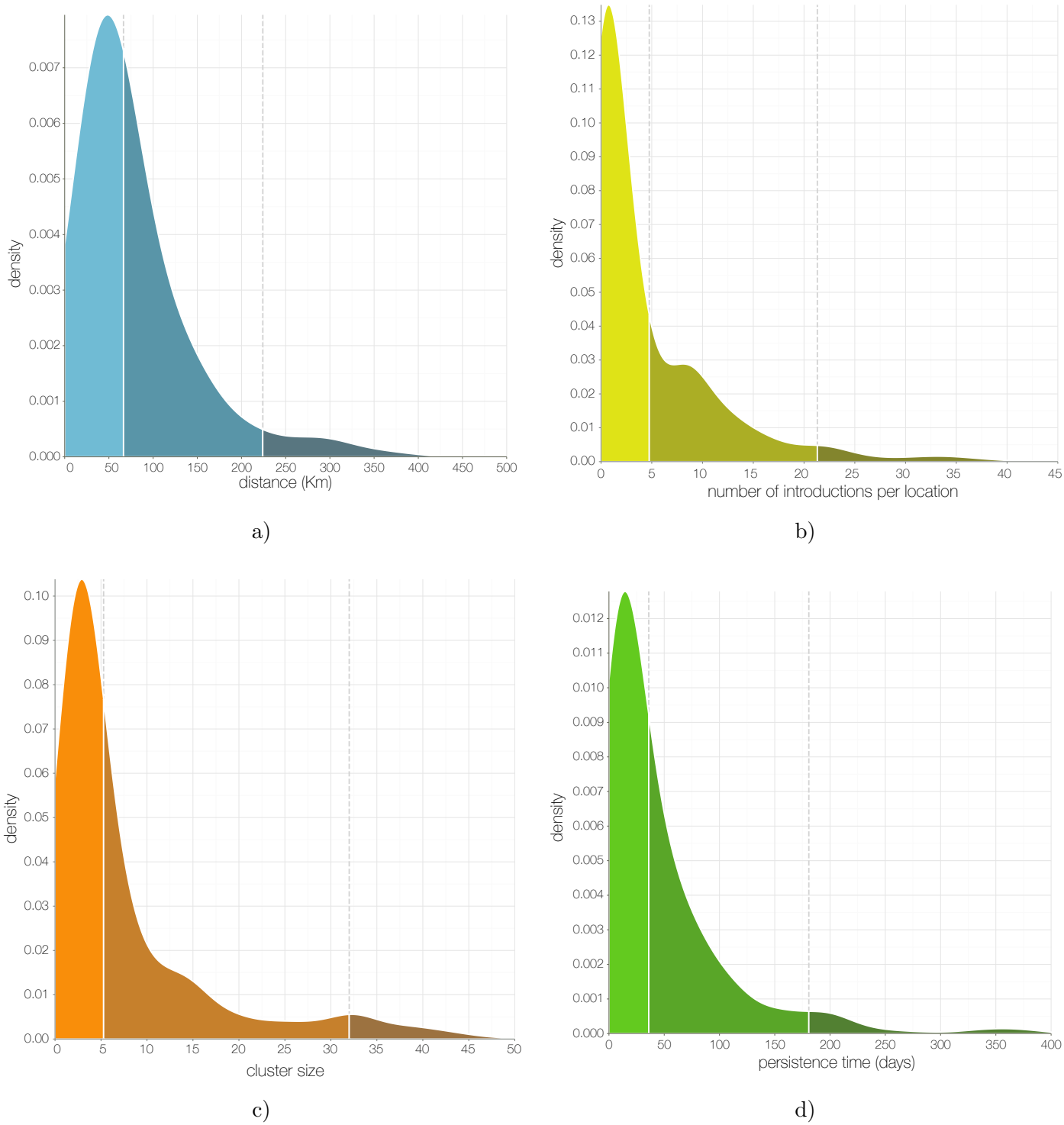


Figure 5. The metapopulation structure of the epidemic. a) Kernel density estimate (KDE) of distance for all inferred migrations: 50% occur over distances <72 km and <5% occur over distances >232 km. b) KDE of the number of independent introductions into each administrative region: 50% have fewer than 4.8 and <5% greater than 21.3. c) KDE of the mean size of sampled cases resulting from each introduction with at least 2 sampled cases: 50% < 5.3, 95% <32. d) KDE of the persistence of clusters in days (from time of introduction to time of the last sampled case): 50% < 36 days, 95% < 181 days.

Tables

Predictor	Description	Coefficient (-1 0 1 2 3 4)	mean [95% C.I.]	Inclusion probability (0.0 – 1.0)	B.F.
IntBoSh	Two locations share an international border		3.46 [2.46, 4.50]		>50
National	Two locations are in the same country		3.13 [2.46, 3.85]		>50
OrPop	Population size in the origin location		1.42 [0.95, 2.01]		>50
Distances	Great circle distances between the locations' population centroids		-0.76 [-0.91, -0.63]		>50
DestPop	Population size in the destination location		0.57 [-0.00, 0.97]		>50
DestPopDens	Population density in the destination location		0.15 [0.00, 0.72]		12.6
LibGinAsym	Asymmetry between Liberia and Guinea		0.17 [0.00, 1.37]		5.6

B.F. Threshold: 3, 15, 50

Table 1. Summary of the phylogenetic generalized linear model results. The estimated coefficients and model inclusion probabilities for spatial movement predictors supported with a Bayes factor (BF) > 3. Positive coefficients are shown in green, negative in red. The remainder are not supported and are not shown (see supplementary document for a full list).

Predictor	Description	Coefficient (-1 0 1)	mean [95% C.I.]	Inclusion probability (0.0 – 1.0)	B.F.
TempSS	Temperature seasonality		-1.1 [-1.6, -0.5]		>50
tt50K	Time to travel to a population centre of 50,000 people		-0.9 [-1.4, -0.4]		32.4
PopSize	Population size		0.9 [0.3, 1.6]		29.6
Precip	Precipitation		0.8 [0.2, 1.3]		4.4
tt100K	Time to travel to a population centre of 0.1 million people		-0.8 [-1.7, -0.1]		3.8

B.F. Threshold: 3, 15, 50

Table 2. Summary of generalized linear model results with case counts as the response variable. The estimated coefficients and model inclusion probabilities for per-region predictors supported with a Bayes factor (BF) > 3. Positive coefficients are shown in green, negative in red. The remainder are not supported and are not shown (see supplementary document for a full list).

Supplementary Information

Tables

Table S1. Number of cases and sampled sequences per region and country, where 'Admin' is the name of the administrative level used ('préfect..' being préfecture and 'départ.' being département) and 'Sampling' is sequences/cases × 100.

Country	Name	Admin.	Population	Sequences	Cases	Sampling
GIN	Beyla	préfect.	218,698	4	52	7.69
GIN	Boffa	préfect.	195,019	0	52	0
GIN	Boké	préfect.	465,824	18	36	50
GIN	Conakry	préfect.	1,513,554	73	630	11.61
GIN	Coyah	préfect.	108,104	26	258	10.12
GIN	Dabola	préfect.	156,599	0	15	0
GIN	Dalaba	préfect.	178,343	3	10	30
GIN	Dinguiraye	préfect.	178,550	0	1	0
GIN	Dubréka	préfect.	273,945	22	167	13.17
GIN	Faranah	préfect.	187,820	8	88	9.09
GIN	Forécariah	préfect.	389,052	60	503	11.95
GIN	Fria	préfect.	115,696	3	16	18.75
GIN	Gaoual	préfect.	171,616	0	0	NA
GIN	Guéckédou	préfect.	720,289	58	390	14.87
GIN	Kankan	préfect.	374,445	4	38	10.53
GIN	Kérouané	préfect.	259,648	10	176	5.68
GIN	Kindia	préfect.	466,987	2	132	1.53
GIN	Kissidougou	préfect.	275,182	18	138	13.04
GIN	Koubia	préfect.	111,176	0	0	NA
GIN	Koundara	préfect.	108,639	0	0	NA
GIN	Kouroussa	préfect.	197,242	2	22	9.09
GIN	Labé	préfect.	313,715	0	0	NA
GIN	Lélouma	préfect.	144,433	0	0	NA
GIN	Lola	préfect.	214,082	2	118	1.69
GIN	Macenta	préfect.	495,845	40	787	5.1
GIN	Mali	préfect.	208,339	0	5	0
GIN	Mamou	préfect.	371,426	0	0	NA
GIN	Mandiana	préfect.	252,272	0	0	NA
GIN	Nzérékoré	préfect.	372,266	9	269	3.35
GIN	Pita	préfect.	263,471	0	8	0
GIN	Siguiri	préfect.	427,947	3	38	7.89
GIN	Télimélé	préfect.	258,398	0	43	0
GIN	Tougué	préfect.	152,448	0	2	0
GIN	Yamou	préfect.	300,674	0	12	0
LBR	Bomi	county	124,080	5	220	2.27
LBR	Bong	county	334,921	2	219	0.91
LBR	Gbapolu	county	91,366	1	28	3.57
LBR	GrandBassa	county	219,024	14	164	8.54
LBR	Grand Cape Mount	county	132,777	4	207	1.93
LBR	GrandGedeh	county	127,295	0	4	0
LBR	GrandKru	county	64,797	2	25	8
LBR	Lofa	county	287,555	13	511	2.54
LBR	Margibi	county	335,736	21	878	2.39
LBR	Maryland	county	132,024	0	7	0
LBR	Montserrado	county	1,016,221	67	2925	2.29
LBR	Nimba	county	483,036	5	282	1.77
LBR	River Cess	county	73,960	2	48	4.17

LBR	River Gee	county	83,020	1	13	7.69
LBR	Sinoe	county	103,789	3	33	9.09
SLE	Bo	district	552,742	13	450	2.89
SLE	Bombali	district	443,868	108	1212	9.82
SLE	Bonthe	district	148,892	1	5	20
SLE	Kailahun	district	409,182	101	756	13.36
SLE	Kambia	district	302,083	74	326	19.33
SLE	Kenema	district	563,021	75	553	13.56
SLE	Koinadugu	district	299,798	11	185	5.95
SLE	Kono	district	324,103	39	568	6.87
SLE	Moyamba	district	263,788	23	317	7.26
SLE	Port Loko	district	540,439	150	2208	6.75
SLE	Pujehun	district	278,897	9	68	13.24
SLE	Tonkolili	district	386,112	19	630	3.01
SLE	Western Rural	district	435,323	88	1736	4.84
SLE	Western Urban	district	528,224	152	3219	4.04
SEN	Kédougou	départ.	70,072	0	0	NA
SEN	Salémata	départ.	19,887	0	0	NA
SEN	Saraya	départ.	55,879	0	0	NA
SEN	Vélingara	départ.	254,319	0	0	NA
SEN	Tambacounda	départ.	259,657	0	0	NA
MLI	Kéniéba	cercle	195,927	0	0	NA
MLI	Kita	cercle	451,019	0	0	NA
MLI	Kangaba	cercle	103,684	0	0	NA
MLI	Kati	cercle	1,076,713	0	0	NA
MLI	Yanfolila	cercle	217,429	0	0	NA
GNB	Gabu	region	220,218	0	0	NA
GNB	Tombali	region	102,893	0	0	NA
CIV	San-Pédro	région	185,465	0	0	NA
CIV	Folon	région	88,493	0	0	NA
CIV	Kabadougou	région	212,545	0	0	NA
CIV	Cavally	région	421,559	0	0	NA
CIV	Tonkpi	région	946,740	0	0	NA
CIV	Bafing	région	188,328	0	0	NA

Table S2. 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	National	Two locations are in the same country
Administrative	IntBoSh	Location pairs that are in different countries and share a border
Administrative	NatBoSh	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	destinationTT100k	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	Location pairs that are in different countries and share at least one of 17 vernacular languages
Cultural	NatLangShared	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

Figures

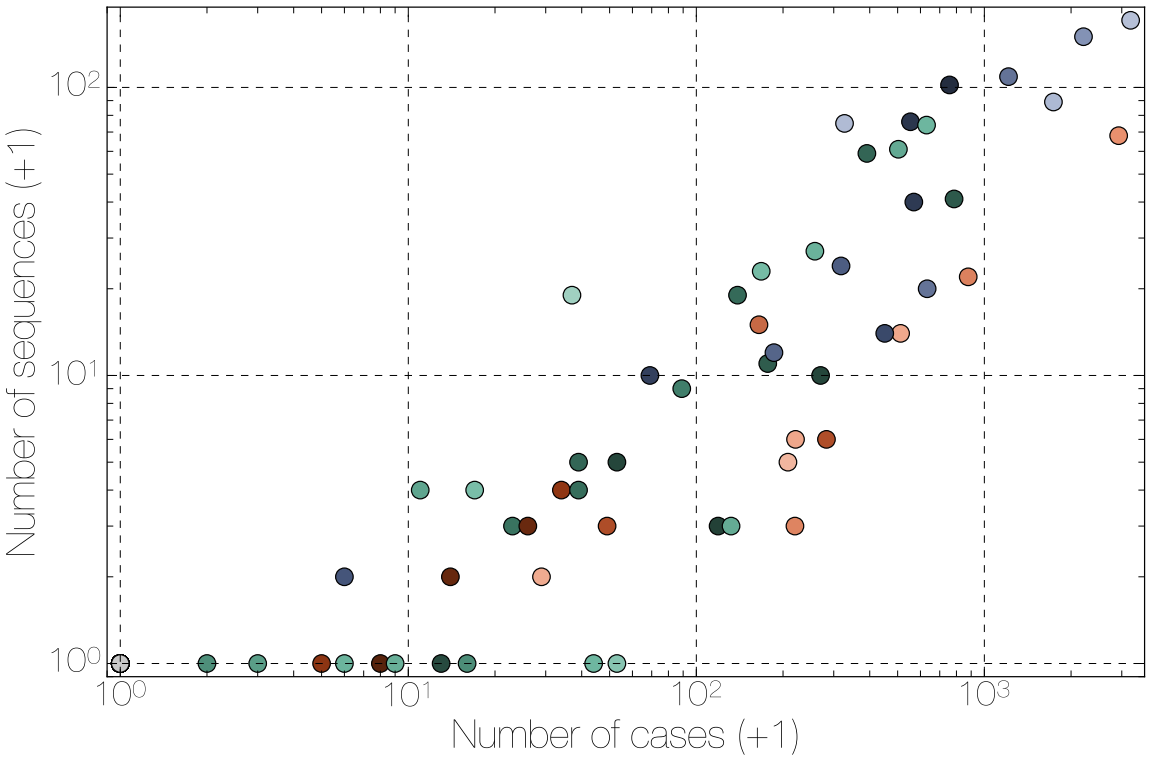


Figure S1. Correlation between number of cases and number of sequences for each location. A plot of number of EBOV genomes sampled against the known and suspected cumulative EVD case numbers. Regions in Guinea are denoted in green, Sierra Leone in blue and Liberia in red. Spearman correlation coefficient: 0.93.

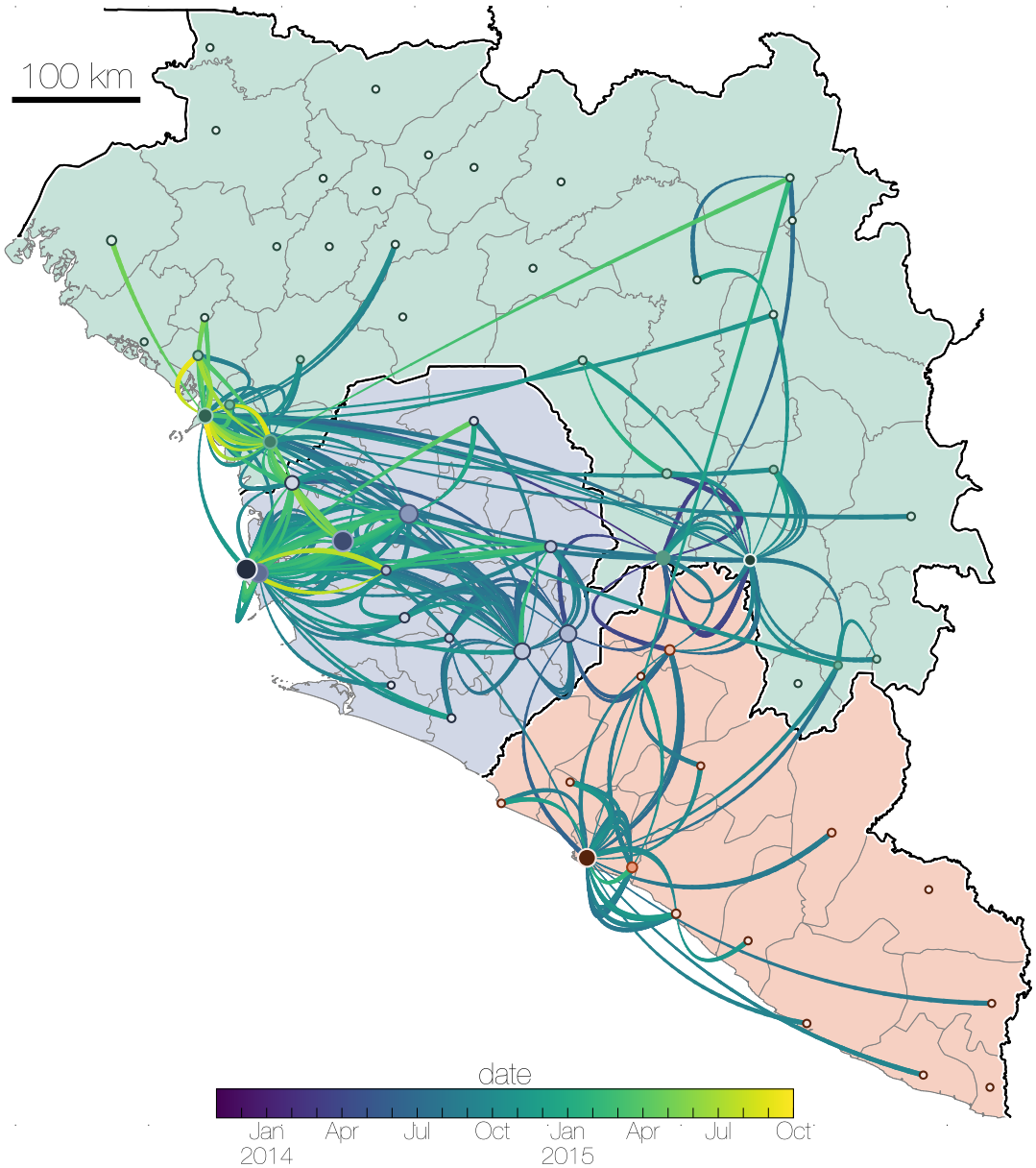


Figure S2. Dispersal of virus lineages over time. Virus dispersal between administrative regions estimated under the GLM phylogeography model (see Supplementary Methods). The arcs are between population centroids of each region, show directionality from thin end to thick end and are coloured in a scale 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.

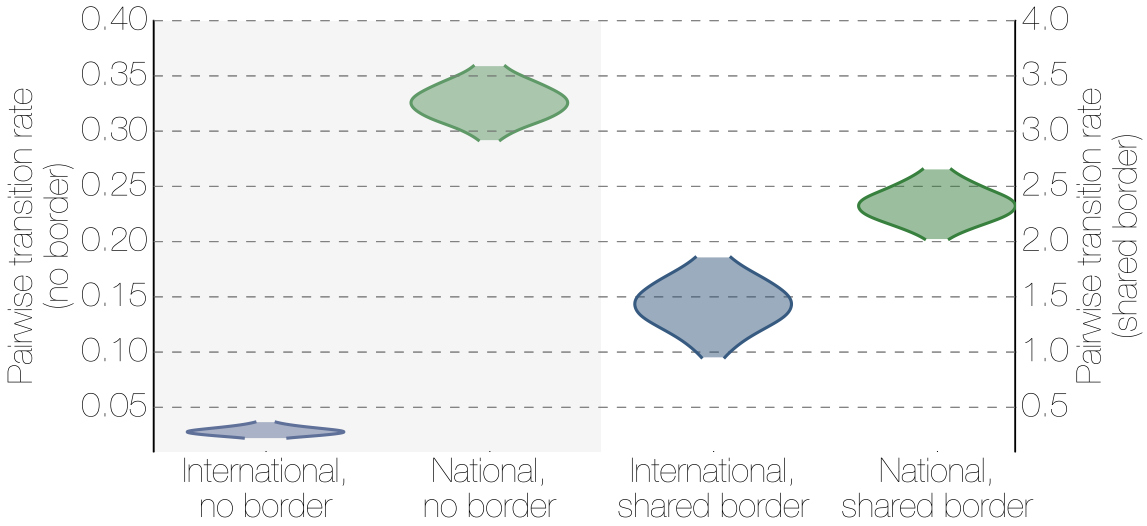


Figure S3. The effect of borders on EBOV migration rates between regions. Posterior densities of the migration rates between locations that share a geographical border (left) and those that do not (right) for international migrations and national migrations. Where two regions share a border, national 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, international migrations are much rarer than national migrations.

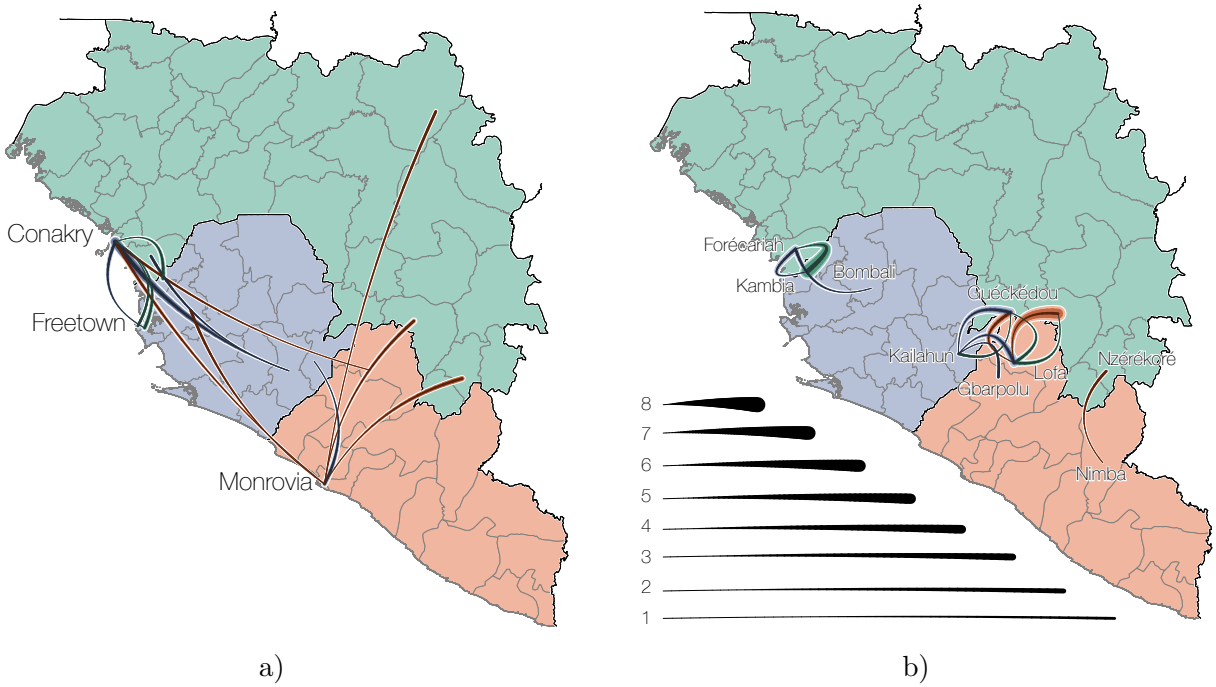


Figure S4. Summarized epidemic international migration history. All viral movement events between counties (Guinea, green; Sierra Leone, blue; Liberia, red) are shown split by whether they are between a) geographically distant regions or b) regions that share the international border. Curved lines indicate 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.

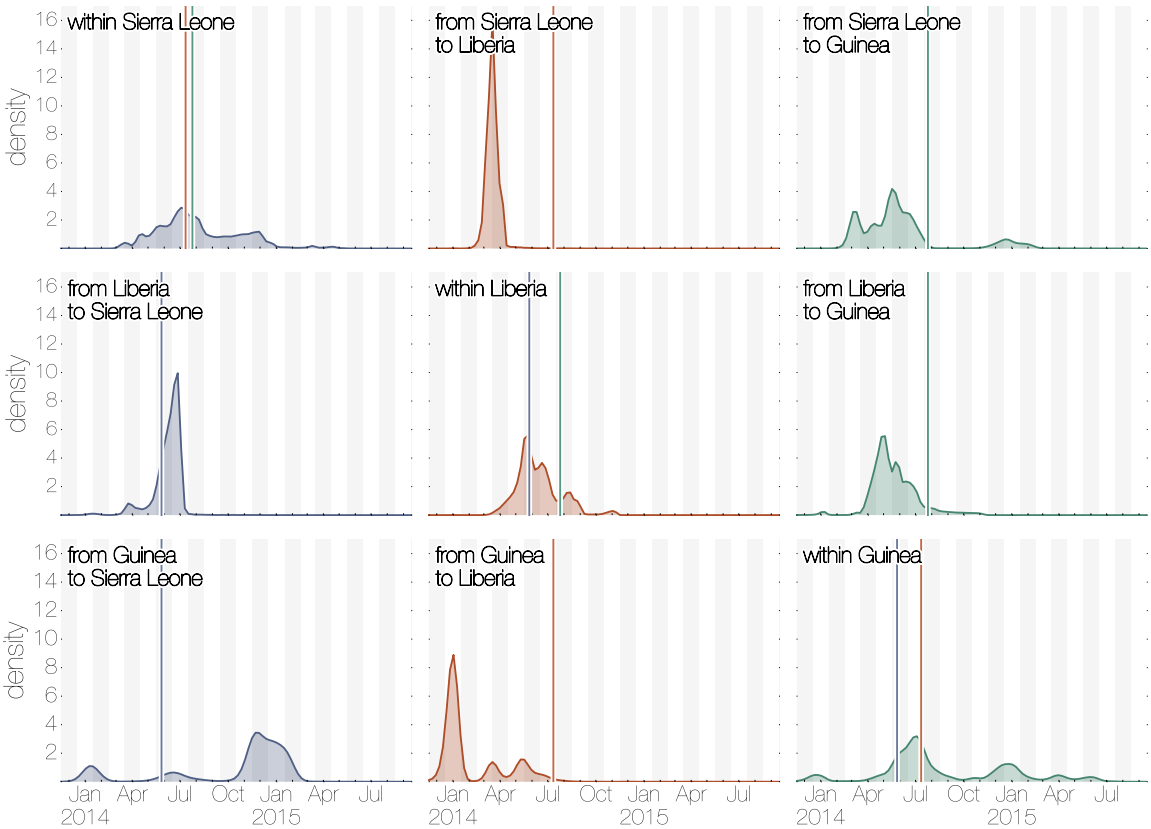


Figure S5. Summary of migration intensity over time in the region. Each cell shows the posterior probability density of temporal migration intensity. Vertical lines within each cell indicate the dates of declared border closures by each of the three countries: 11 June 2014 in Sierra Leone (blue), 27 July 2014 in Liberia (red), and 09 August 2014 in Guinea (green). Densities are rescaled and directly comparable across cells.

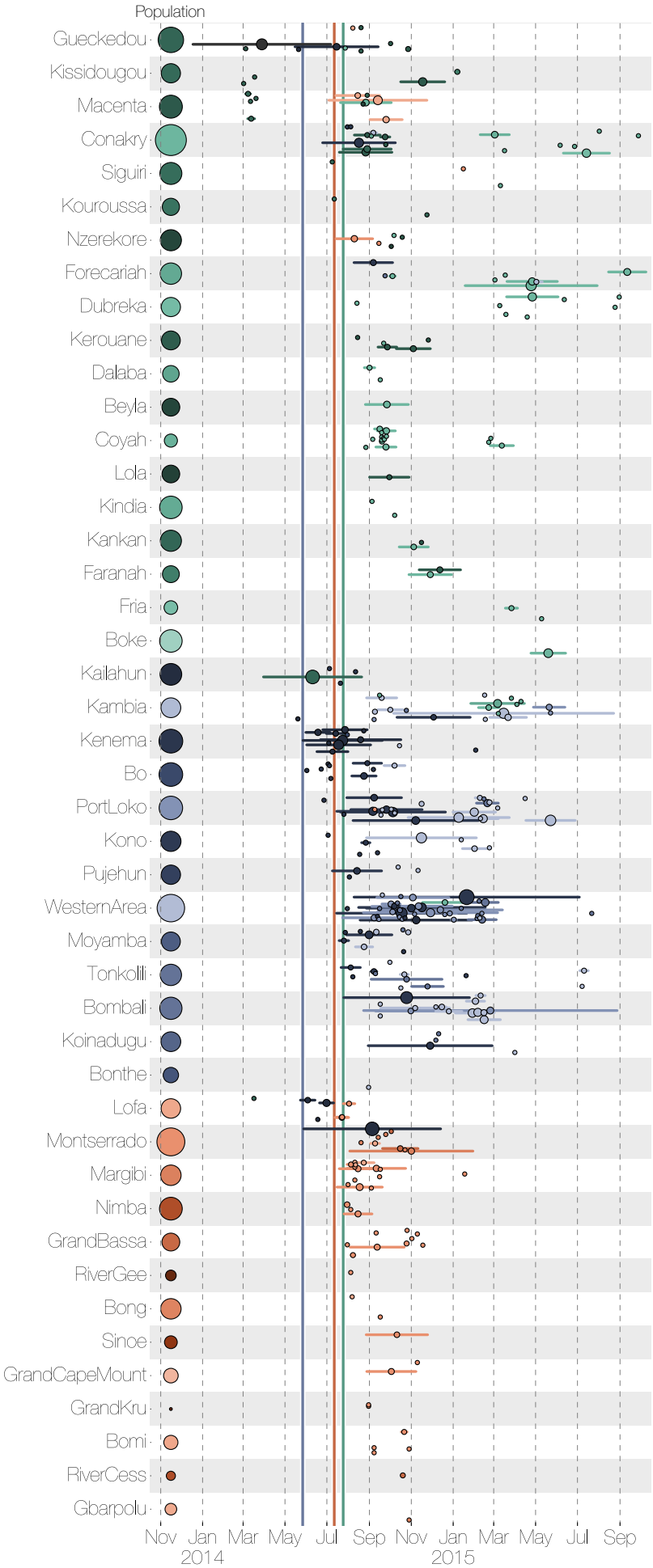
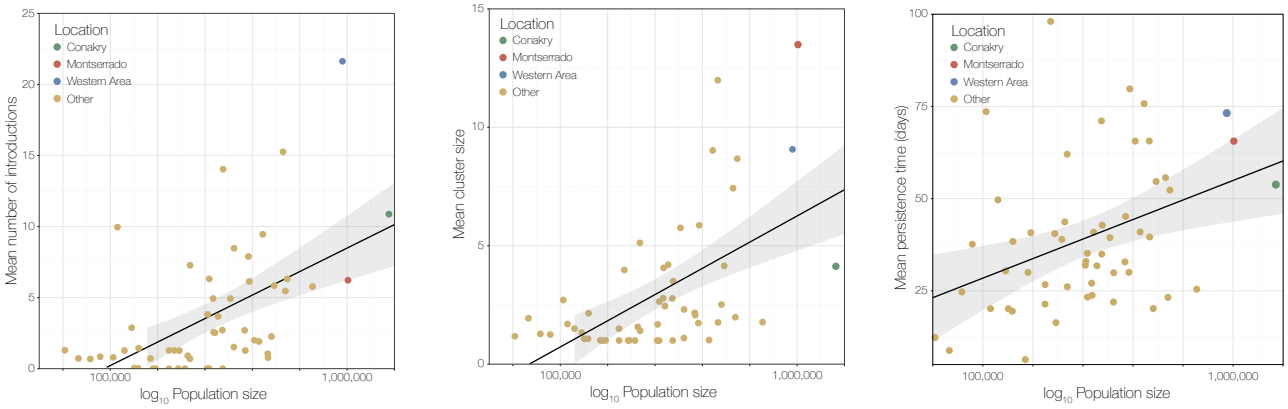


Figure S6. Region specific introductions, cluster sizes and persistence. Independent introductions into each administrative region and the size of each resulting cluster . The horizontal lines represent the persistence of each cluster from the time of introduction to the last sampled case. The areas of the circles in the middle of the lines are proportional to the number of sampled cases in the cluster. The areas of the circles next to the labels 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).



a) b) c)

Figure S7. Relationship of cluster size, introductions and persistence to population size. a) The mean number of introductions into each location against (log) population sizes. The WesternArea (in Sierra Leone) received the most introductions, whilst Conakry and Montserrado were closer to the average. The association between population sizes and 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 here is weaker ($R^2 = 0.11$, 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 ($R^2 = 0.12$, pearson correlation = 0.37, Spearman correlation = 0.36). All computations based on a sample of 10, 000 trees from the posterior distribution.