

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

Gytis Dudas^{1,2}, Luiz Max Carvalho¹, Trevor Bedford², Andrew J. Tatem^{3,4}, Guy Baele⁵, Nuno Faria⁶, Daniel J. Park⁷, Jason Ladner⁸, Armando Arias^{9,10}, Danny Asogun^{11,12}, Filip Bielejec⁵, Sarah Caddy⁹, Matt Cotten¹³, MAJ Jonathan Dambrozio⁸, Simon Dellicour⁵, Antonino Di Caro^{14,12}, Joseph W. Diclaro II¹⁵, Sophie Duraffour^{16,12}, Mike Elmore¹⁷, Lawrence Fakoli¹⁸, CPT Merle Gilbert⁸, Sahr M Gevao¹⁹, Stephen Gire^{7,20}, Adrianne Gladden-Young⁷, Andreas Gnirke⁷, Augustine Goba^{21,22}, Donald S. Grant^{21,22}, Bart Haagmans²³, Julian A. Hiscox^{24,25}, Umaru Jah²⁶, Brima Kargbo²², CPT Jeffrey Kugelman⁸, Di Liu²⁷, Jia Lu⁹, Christine M. Malboeuf⁷, CPT Suzanne Mate⁸, David A. Matthews²⁸, Christian B. Matranga⁷, Luke Meredith⁹, James Qu⁷, Joshua Quick²⁹, Susan Pas²³, My VT Phan¹³, Georgios Poliakis²⁴, Chantal Reusken²³, Mariano Sanchez-Lockhart^{8,30}, Stephen F. Schaffner⁷, John S. Schieffelin³¹, Rachel S. Sealfon⁷, Etienne Simon-Loriere^{32,33}, Saskia Smits²³, Kilian Stoecker^{34,12}, Lucy Thorne⁹, Ekaete Alice Tobin^{11,12}, Mohamed A. Vandi^{21,22}, Simon J. Watson¹³, Kendra West⁷, Shannon Whitmer³⁵, Michael R Wiley^{8,30}, Sarah M. Winnicki^{7,20}, Shirlee Wohl^{7,20}, Roman Wlfel^{34,12}, Nathan L. Yozwiak^{7,20}, Kristian G. Andersen^{36,37,7}, Sylvia Blyden²², Fatorma Bolay¹⁸, Miles Carroll^{17,12}, Boubacar Diallo³⁸, Pierre Formenty³⁹, Christophe Fraser⁴⁰, George F. Gao^{27,41}, Robert F. Garry⁴², Ian Goodfellow⁹, Stephan Gnther^{16,12}, Christian Happi⁴³, Edward C Holmes⁴⁴, Brima Kargbo²², Sakoba Keta⁴⁵, Paul Kellam^{13,46}, Marion P. G. Koopmans²³, Nicholas J. Loman²⁹, N?Faly Magassouba⁴⁷, Dhamari Naidoo³⁹, Stuart T. Nichol³⁵, Gustavo Palacios⁸, Oliver G Pybus⁶, Pardis Sabeti^{7,20}, Amadou Sall³², Ute Stroehrer³⁵, Isatta Wury⁴⁵, Marc A Suchard^{48,49,50}, Philippe Lemey⁵ & Andrew Rambaut^{1,51,52}

¹Institute of Evolutionary Biology, University of Edinburgh, King's Buildings, Edinburgh, EH9 3FL, UK, ²Vaccine and Infectious Disease Division, Fred

Hutchinson Cancer Research Center, Seattle, WA, USA, ³WorldPop, Department of Geography and Environment, University of Southampton, Highfield, Southampton SO17 1BJ, UK, ⁴Flowminder Foundation, Stockholm, Sweden, ⁵Department of Microbiology and Immunology, Rega Institute, KU Leuven ? University of Leuven, Leuven, Belgium, ⁶Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3PS, UK, ⁷Broad Institute of Harvard and MIT, Cambridge, MA 02138, USA, ⁸Center for Genome Sciences, U.S. Army Medical Research Institute of Infectious Diseases, Fort Detrick, Frederick, MD 21702, USA, ⁹Department of Pathology, University of Cambridge, Addenbrooke's Hospital, Cambridge, CB2 2QQ, UK, ¹⁰Section for Virology, National Veterinary Institute, Technical University of Denmark, Artillerivej 27, 1870, Frederiksberg C, Denmark, ¹¹Institute of Lassa Fever Research and Control, Irrua Specialist Teaching Hospital, Irrua, Nigeria, ¹²The European Mobile Laboratory Consortium, 20359 Hamburg, Germany, ¹³Virus Genomics, Wellcome Trust Sanger Institute, Hinxton, United Kingdom, ¹⁴National Institute for Infectious Diseases "L. Spallanzani" - IRCCS, Via Portuense 292, 00149 Rome, Italy, ¹⁵Naval Medical Research Unit 3, 3A Intidad Ramses Street, Cairo, 11517, Egypt, ¹⁶Bernhard Nocht Institute for Tropical Medicine, 20359 Hamburg, Germany, ¹⁷National Infections Service, Public Health England, Porton Down, Salisbury, Wilts SP4 0JG, UK, ¹⁸Liberian Institute for Biomedical Research, Charlesville, Liberia., ¹⁹University of Sierra Leone, Freetown, Sierra Leone, ²⁰Center for Systems Biology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA, ²¹Viral Hemorrhagic Fever Program, Kenema Government Hospital, 1 Combema Road, Kenema, Sierra Leone, ²²Ministry of Health and Sanitation, 4th Floor Youyi Building, Freetown, Sierra Leone, ²³Department of Viroscience, Erasmus University Medical Centre, P.O. Box 20140, 300 CA Rotterdam, the Netherlands, ²⁴Institute of Infection and Global Health, University of Liverpool, Liverpool L69 2BE, United Kingdom, ²⁵NIHR Health Protection Research Unit in Emerging and Zoonotic Infections, University of Liverpool, UK., ²⁶University of Makeni, Makeni, Sierra Leone, ²⁷Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China, ²⁸University of Bristol, BS8 1TD, United Kingdom, ²⁹Institute of Microbiology and Infection, University of Birmingham, Birmingham B15 2TT, United Kingdom, ³⁰University of Nebraska Medical Center, Omaha, NE, USA, ³¹Department of Pediatrics, Section of Infectious Diseases, New Orleans, LA 70112, USA, ³²Institut Pasteur, Functional Genetics of Infectious Diseases Unit, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France, ³³CNRS URA3012, Paris 75015, France, ³⁴Bundeswehr Institute of Microbiology, Neuherbergstrasse 11, 80937 Munich, Germany, ³⁵Viral Special Pathogens Branch, Centers for Disease Control and Prevention, 1600 Clifton Rd. NE, Atlanta, Georgia, USA, ³⁶The Scripps Research Institute, Department of Immunology and Microbial Science, La Jolla, CA 92037, USA., ³⁷Scripps Translational Science Institute, La Jolla, CA 92037, USA, ³⁸World Health Organization, Conakry, Guinea, ³⁹WHO Ebola Response Team, ⁴⁰Department of Infectious Disease Epidemiology, MRC Centre for Outbreak Analyses and Modelling, School of Public Health, Imperial College London, UK, ⁴¹Chinese Center for Disease Control and Prevention (China CDC), Beijing 102206, China, ⁴²Department of Microbiology & Immunology, New Orleans, LA 70112, USA, ⁴³Redeemer's University, Ede, Osun State, Nigeria, ⁴⁴Marie 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, ⁴⁵Ministry of Health Guinea, Conakry, Guinea, ⁴⁶Division of Infectious Diseases, Imperial College Faculty of Medicine, London W2 1PG, UK, ⁴⁷Universit Gamal Abdel Nasser de Conakry, Laboratoire des Fivres Hmorragiques en Guine, Conakry, Guinea, ⁴⁸Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles, CA, USA, ⁴⁹Department of Biomathematics David Geffen School of Medicine at UCLA, University of California, Los Angeles, CA, USA, ⁵⁰Department of Human Genetics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, CA, USA, ⁵¹Centre for Immunology, Infection and Evolution, University of Edinburgh, King's Buildings, Edinburgh, EH9 3FL, UK, ⁵²Fogarty International Center, National Institutes of Health, Bethesda, MD, USA

August 9, 2016

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 comprehensive collection of Ebola virus genomes, representing over 5% of known cases, unprecedented for a single epidemic. In the first comprehensive analysis of this entire collection, we reconstruct a detailed history of migration, proliferation and decline of the virus throughout the region. We test the association of geographical, climatic, administrative, demographic and cultural factors with viral movement between administrative regions. We identify a classic 'gravity' model as the core dynamic, with more intense migration between larger population centers particularly when geographically close. Notably, we show that despite a strong attenuating effect of border closures on international dispersal, localized cross-border transmission had already set the seeds for an international epidemic, rendering these measures relatively ineffective in curbing the epidemic. Finally, 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.

Main text

Over the two and a half years of Ebola virus (EBOV) circulation in West Africa in 2013-2016, at least 28,000 cases and 11,000 deaths ¹ have been attributed to the Makona variant of EBOV ². The epidemic is thought to have begun in December 2013 in Guinea, but was not detected and reported until March 2014 ³. Initial efforts to control the outbreak were considered to be succeeding in Guinea ⁴, but the virus crossed international borders into neighbouring Liberia (first cases diagnosed in late March) and Sierra Leone (first cases diagnosed in May) in early 2014 ⁵. Viral genomes sequenced from three patients in Guinea early in the epidemic ³, helped to establish that the progenitor of the Makona variant originated in Central Africa and arrived to West Africa within the last 15 years ^{5,6}. Rapid sequencing of the first reported cases in Sierra Leone confirmed that EBOV had crossed the border from Guinea and were not the result of an independent zoonotic introduction ⁵. Subsequent studies analyzing the genetic makeup of the Makona variant focused on Guinea ⁷⁻⁹, Sierra Leone ¹⁰⁻¹² and Liberia ^{13,14} in relative isolation, identifying local lineages of the virus and the origins and outbreak patterns within each country.

Although virus sequencing has covered considerable fractions of the epidemic in each country, individual studies have focused on either limited geographical areas or periods of time, so that the patterns and drivers of the epidemic in the region as a whole and through its entire duration are uncertain. Using 1610 genome sequences collected throughout the epidemic, representing over 5% of known EVD cases (Figures 1 & S1), we reconstruct a detailed history of the movement of the virus within and between the three most affected countries. Using a recently developed approach for integrating covariates of spatial spread in a phylogeographic model ¹⁵, we test which administrative, economic, climatic, infrastructure and demographic features of subregions were the crucial factors in the spatial dynamics of EBOV. We examine the effectiveness of international border closures between the three

countries. Finally, we investigate why regions immediately bordering 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 all sampled lineages existed in early December 2013 (95% highest posterior density interval: Oct 2013 – Feb 2014) and phylogeographic estimation confidently assigns it to the Guéckédou préfecture (**96%** posterior support) (Figure 2). In addition, we find that initial lineages deriving from this common ancestor circulated between Guéckédou and its neighbouring préfectures of Macenta and Kissidougou until late February 2014 (Figure 2). These results, based on a large sample of EBOV genomes, support the epidemiological evidence that the West African epidemic began in late December 2013 in Guéckédou préfecture of Guinea ³.

The first introduction of EBOV from Guinea into another country that resulted in sustained transmission is estimated to have occurred in early April 2014, when the virus spread to the Kailahun district of Sierra Leone ^{16,17}. This lineage was first detected in Kailahun at the end of May 2014, from where it spread across the region (Figure 3). From Kailahun EBOV spread extremely rapidly into several counties of Liberia (Lofa, Montserrado and Margibi) ¹⁴ and Guinea (Conakry, back into Guéckédou) in May 2014 ^{7,8}. The Makona variant continued spreading west through Sierra Leone, and by July 2014 it was present in the capital city, Freetown.

Liberia was reporting over 500 new EVD cases per week by mid-September, 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, similarly driven by large scale 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 Ebola virus 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. The following month, prevalence had declined such that only a handful of relatively distantly related lineages survived from the exponential phase of the epidemic ^{9,12} (Figure 3).

The last Ebola virus resulting from a conventionally acquired infection was collected and sequenced in October 2015 in Forecariah prefecture, Guinea ⁹. Following this, only sporadic cases of EVD were detected: in Montserrado, Liberia and Kono, Sierra Leone in November 2015, in Tonkolili, Sierra Leone in January and February 2016, and in Nzérékoré, Guinea in March 2016, all likely the result of transmission of Ebola virus from survivors with established persistent infections ^{18,19}.

Factors associated with EBOV dispersal

To determine the factors influencing the spread of EBOV between administrative regions at the *district* (Sierra Leone), *préfecture* (Guinea) and *county* (Liberia) levels we employed a

phylogeographic generalized linear model¹⁵. Of the 25 factors assessed (see Table S2 for a full list and descriptions) 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 support for inclusion $BF > 50$). Half of all virus migrations between regions were less than 100 km and only 5% were greater than 340 km (Figure S3a). Population sizes are very strongly ($BF > 50$) positively correlated with viral movement, with a stronger effect for origin compared to destination population size, which when combined with the inverse effect of distance, implies the existence of a classic gravity migration dynamic. Gravity models, widely used in economic and geographic studies, describe the movement of people between locations based on their population sizes and distance apart. They are a natural choice for modelling infectious disease transmission^{20,21} and have been used in spatio-temporal modelling of EBOV transmission in Sierra Leone²²; here we provide strong empirical evidence of such a process driving viral dissemination in an epidemic.

In addition to geographical distance, we find a significant propensity for migration events to have occurred between administrative regions within each of the three countries, 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 even accounting for the dispersal distance effect. When international migrations do take place, they are more intense between administrative divisions that meet on the international border (IntBoSh, $BF > 50$).

The only other factor included in the model with any notable support was the seasonality in temperature for the origin region (TempSS, $BF = 8.3$) which was negatively correlated with EBOV spread. We also considered the sharing of any of 17 vernacular languages as a covariate that reflects local cultural links including between non-contiguous or international regions, but no evidence that such links were correlated with EBOV spread was found. A variety of other variables that might intuitively contribute to EBOV transmission, such as aspects of urbanisation (economic output, population density, travel times to large settlements) and other 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 once seeded in a region (see below).

Factors associated with local EBOV proliferation

By considering the factors that predict virus movement between administrative regions we built a model of the degree of importation risk; that is, the 'sparks' that can ignite outbreaks within each region. This model is dominated by geographical and administrative factors. However, the result of these sparks – the size and duration of resulting outbreaks – may be affected by different factors. To investigate this we considered which of our demographic, economic and climatic factors were predictive of cumulative case counts (WHO patient database; ¹) and EBOV endurance (measured as mean survival time of a genetic lineage after introduction) for each region (Bayesian generalized linear model; see Supplementary Methods).

We find that cumulative case counts in each location are associated with factors related to urbanisation: primarily population sizes (PopSize, $BF = 29.6$) and a significant inverse

association with traveling times to settlements of at least 50,000 inhabitants (tt50K, BF 32.4). These results are consistent with the perception that widespread transmission within urban regions was a major contributing factor to the scale of the West African epidemic compared to previous EVD outbreaks.

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 virus²³. Although we found no evidence of associations 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^{24–26}. Our results support this view, in that most migrations were between geographically close regions and, furthermore, international dispersals were more likely to be between regions sharing a geographical border. Specifically, repeated movement occurred between Guéckédou (Guinea), Kailahun (Sierra Leone) and Lofa (Liberia) during the early phases of the epidemic (Figure 4)

In the later stage of the epidemic there was also cross-border movements between neighbouring Forécariah (Guinea) and Kambia (Sierra Leone) on the coast (Figure 4). These were a significant hindrance to efforts to interrupt the final chains of transmission in late 2015 with a number of such chains moving back and forth across this border^{9,12,27}.

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 such border closures actually entailed. As such, even though our results show that international spread of Ebola virus was more intense prior to border closures than afterwards (mean change point: Aug-Sept 2014; 80.0% posterior support of a greater coefficient after this time; Figure 2b), it is difficult to ascertain whether border closures themselves, rather than renewed control efforts or intensified public information campaigns, were responsible for the apparent reduction in cross-border transmissions. However, we did not observe in our analysis an effect on dispersal patterns at the intra-country level, which would also be expected to be affected by these latter control efforts. Overall, these results suggest that border closures may have reduced international traffic, in particular over longer distances and between larger population centres (Figure S3b), but by the time Sierra Leone and later Liberia closed their borders the epidemic had become firmly established in both countries (Figure 2a).

Why did the epidemic not spread further?

With the exception of a few documented exportations^{28–30} the EVD epidemic did not spread into the contiguous neighbouring districts of Guinea-Bissau, Senegal, Mali, or Côte d'Ivoire no cases were reported in seven préfectures of Guinea. By extending our GLM model (supported predictors and estimated coefficients) to include these regions we can address whether these were inherently at lower risk of EBOV spread and transmission. We estimated

the expected number of times that viral lineages (see supplementary methods), migrating from regions with cases, might have jumped into these apparently EVD-free regions. Overall, the contiguous regions in the neighbouring countries were all predicted to have relatively low numbers of introductions (Figure 5b). They were not, however, predicted to have particularly low levels of transmission if an outbreak had been seeded (Figure 5c). 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 peak transmission and attenuating effect of international borders prevented too many sparks from landing. The exceptions to this are Kati region in Mali and Tonkpi region in Côte d'Ivoire, which, because of their large populations (Kati, 1 million; Tonkpi 950,000), were susceptible to introductions and are predicted to have considerable numbers of cases had EVD become established.

It is evident from Figure 3 that after the initial seeding of Sierra Leone and Liberia, Guinea experienced repeated reintroductions from the escalating epidemics in these other two countries. Our analysis reveals that there were, given the 5% sample of cases, at least 21 (95% credible interval, CI: 17 - 26) re-introductions into Guinea from April 2014 to February 2015, more than Liberia and Sierra Leone combined (Figure S6). Although there were numerous introductions into Sierra Leone over a similar time period (median: 9, 95% CI: 7 - 12), the resulting outbreaks constituted a tiny proportion of the Sierra Leonean epidemic. Although an early lineage established around the Guinean capital, Conakry, and persisted for the duration of the epidemic, the continual 'seeding' into Guinea without a clear peak in transmission suggests that EBOV was struggling to maintain transmission in the country. The necessity of repeated seeding to maintain transmission in certain regions is also suggested by the fact that regions with more cases than expected are also those with more introductions (Figure S7).

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 utility of pathogen sequencing in a public healthcare emergency situation at all scales and the value of rapid data sharing prior to publication to identify origins of imported lineages initially, 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 a unique insight into spatial movement of infectious disease especially when epidemiological tracing is challenging. Other sources of human mobility data, in particular, mobile phones are promising but currently such data is difficult to obtain in a timely fashion²⁶. 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^{9,31,32}. As 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.

EBOV, like most RNA viruses, accumulates genetic changes over very short timescales³³⁻³⁵. As a result, continuous and comprehensive sequencing of viruses from the earliest opportunity in an epidemic can provide valuable information about the geographic dispersal of the virus. Patterns such as repeated seeding of local outbreaks from neighbouring regions

may inform the need for, and degree of, movement restrictions to attenuate the epidemic. For EBOV and closely related filoviruses, the analysis of the comprehensive genome set collected during the 2013-2016 epidemic, including the findings presented here and in other studies^{5,7,8,10–14,36,37} will provide a framework for predicting the behaviour of future outbreaks.

Finally, many open questions remain about the biology of EBOV. As sustained human-to-human transmission waned, the region experienced a number of ‘flare-ups’ often in regions that hadn’t seen cases for many months^{18,19,38,39} as a result of persistent sub-clinical infections. Although sequelae like these were not entirely unexpected⁴⁰, the magnitude of the 2013-2016 epidemic has put the entire region at ongoing risk of sporadic EVD re-emergence. Similarly, the nature of the reservoir of EBOV, or its geographic distribution, remain as fundamental gaps in our knowledge that are 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 genome sequences were collated, aligned and annotated with available information about date of sampling and likely location of infection (all data available from <https://github.com/ebov/space-time>). Geographical, demographic and climatic variables were collected for each of the 63 regions in the 3 focal countries and a further 18 in surrounding countries that saw no cases or no sustained transmission (see supplemental information for details and sources). Time structured phylogenetic trees were constructed using BEAST⁴¹ and these formed the basis of a phylogenetic generalized linear model¹⁵ that infers the probability of inclusion, and degree of correlation, of each of the predictor variables with the pattern of migration of virus lineages across the region. Along each branch of the tree we infer transitions between regions⁴². For the variables included 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 are predictive of the magnitude of outbreak once introduced into a region, we employed generalised linear models and Bayesian model averaging with cumulative case counts and persistence time in each affected region as response variables.

Acknowledgments

The work at USAMRIID was funded by JSTO/DTRA, project CB10246. Opinions, interpretations, conclusions, and recommendations are those of the authors and do not necessarily reflect the official policy or position of the US Army, US Department of Defense, nor the US Government.

Colour-blind friendly colour palettes by Cynthia Brewer, Pennsylvania State University (<http://colorbrewer2.org>).

References

1. World Health Organization. *Ebola Situation Report -- 10 June 2016*. (Available at 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. Ebola virus disease, West Africa (Situation as of 25 April 2014) - WHO | Regional Office for Africa. Available at: <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>. (Accessed: 1st August 2016)
5. Gire, S. K. *et al.* Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369–1372 (2014).
6. Dudas, G. & Rambaut, A. Phylogenetic Analysis of Guinea 2014 EBOV Ebolavirus Outbreak. *PLoS Curr.* **6**, (2014).
7. Carroll, M. W. *et al.* Temporal and spatial analysis of the 2014-2015 Ebola virus outbreak in West Africa. *Nature* **524**, 97–101 (2015).
8. Simon-Loriere, E. *et al.* Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic. *Nature* **524**, 102–104 (2015).
9. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232 (2016).
10. Tong, Y.-G. *et al.* Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. *Nature* **524**, 93–96 (2015).
11. Park, D. J. *et al.* Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. *Cell* **161**, 1516–1526 (2015).
12. Arias, A. *et al.* Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases. *Virus Evolution* **2**, vew016 (2016).
13. 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).
14. Ladner, J. T. *et al.* Evolution and Spread of Ebola Virus in Liberia, 2014--2015. *Cell Host Microbe* **18**, 659–669 (2015).
15. 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).
16. Goba, A. *et al.* An Outbreak of Ebola Virus Disease in the Lassa Fever Zone. *J. Infect. Dis.* (2016). doi:10.1093/infdis/jiw239
17. Sack, K., Fink, S., Belluck, P., Nossiter, A. & Berhulak, D. How Ebola roared back. *NY Times* (2014).
18. Blackley, D. J. *et al.* Reduced evolutionary rate in reemerged Ebola virus transmission chains. *Sci Adv* **2**, e1600378 (2016).
19. Mate, S. E. *et al.* Molecular Evidence of Sexual Transmission of Ebola Virus. *N. Engl. J. Med.* **373**, 2448–2454 (2015).
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. Viboud, C. *et al.* Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science* **312**, 447–451 (2006).
22. Yang, W. *et al.* Transmission network of the 2014–2015 Ebola epidemic in Sierra Leone. *J. R. Soc. Interface* **12**, 20150536 (2015).
23. Fischer, R. *et al.* Ebola Virus Stability on Surfaces and in Fluids in Simulated Outbreak Environments. *Emerg. Infect. Dis.* **21**, 1243–1246 (2015).
24. Chan, M. Ebola virus disease in West Africa--no early end to the outbreak. *N. Engl. J. Med.* **371**, 1183–1185 (2014).
25. Bausch, D. G. & Schwarz, L. Outbreak of Ebola Virus Disease in Guinea: Where Ecology Meets Economy. *PLoS Negl. Trop. Dis.* **8**, e3056 (2014).
26. Wesolowski, A. *et al.* Commentary: containing the ebola outbreak - the potential and challenge of mobile network data. *PLoS Curr.* **6**, (2014).

27. Goodfellow, I., Reusken, C. & Koopmans, M. Laboratory support during and after the Ebola virus endgame: towards a sustained laboratory infrastructure. *Euro Surveill.* **20**, (2015).
28. Hoenen, T. *et al.* Mutation rate and genotype variation of Ebola virus from Mali case sequences. *Science* **348**, 117–119 (2015).
29. Folarin, O. A. *et al.* Ebola Virus Epidemiology and Evolution in Nigeria. *J. Infect. Dis.* (2016). doi:10.1093/infdis/jiw190
30. Abdoulaye, B. *et al.* Experience on the management of the first imported Ebola virus disease case in Senegal. *Pan Afr. Med. J.* **22 Suppl 1**, 6 (2015).
31. 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).
32. Gardy, J., Loman, N. J. & Rambaut, A. Real-time digital pathogen surveillance – the time is now. *Genome Biol.* **16**, 155 (2015).
33. Jenkins, G. M., Rambaut, A., Pybus, O. G. & Holmes, E. C. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J. Mol. Evol.* **54**, 156–165 (2002).
34. Hedge, J., Lycett, S. J. & Rambaut, A. Real-time characterization of the molecular epidemiology of an influenza pandemic. *Biol. Lett.* **9**, 20130331 (2013).
35. Duffy, S., Shackelton, L. A. & Holmes, E. C. Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* **9**, 267–276 (2008).
36. Volz, E. & Pond, S. Phylodynamic Analysis of Ebola Virus in the 2014 Sierra Leone Epidemic. *PLoS Curr.* **6**, (2014).
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**, (2014).
38. WHO | New Ebola case in Sierra Leone. WHO continues to stress risk of more flare-ups. (2016).
39. End of Ebola transmission in Guinea. Available at: <http://www.afro.who.int/media-centre/pressreleases/item/8676-end-of-ebola-transmission-in-guinea.html>. (Accessed: 1st August 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. Commission de Lutte contre les Epidémies à Kikwit. *J. Infect. Dis.* **179 Suppl 1**, S28–35 (1999).
41. 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).
42. Minin, V. N. & Suchard, M. A. Fast, accurate and simulation-free stochastic mapping. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **363**, 3985–3995 (2008).

Figures

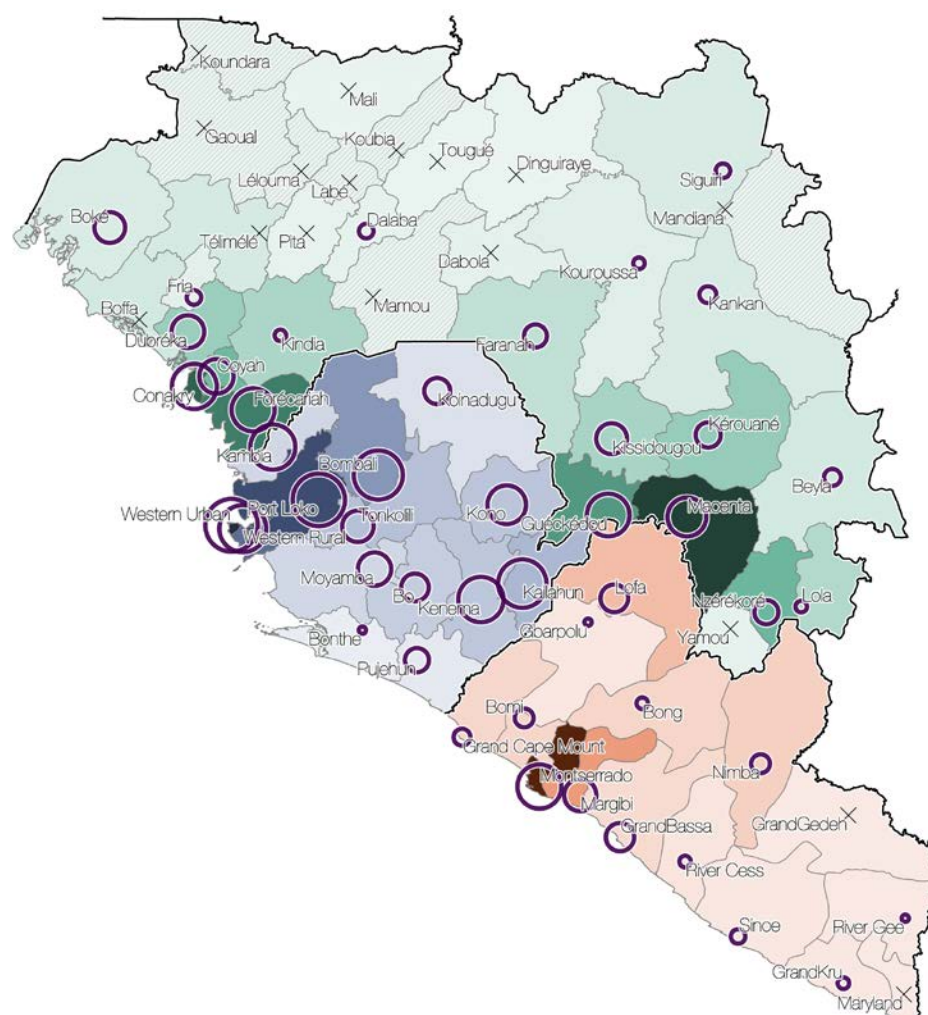


Figure 1 | Distribution of EBOV case numbers and sequences across West Africa.

Administrative areas within Guinea (green), Sierra Leone (blue) and Liberia (red) are coloured according to known and suspected cumulative EVD case numbers¹. Darkest colours represent 784 cases for Guinea (Macenta), 3219 cases for Sierra Leone (Western Urban) and 2925 cases for Liberia (Montserrado). Circles are positioned at population centroids within each region and circle diameters are proportional to the number of sequences available from that region over the entire epidemic with the largest circle representing 152 sequences. Regions marked with X represent those for which sequences are not available; hatched areas indicate regions without any reported EVD cases. The number of sequences and number of cases for each region where cases were recorded are highly correlated (Spearman rank correlation coefficient 0.93; Supplementary Figure S1).

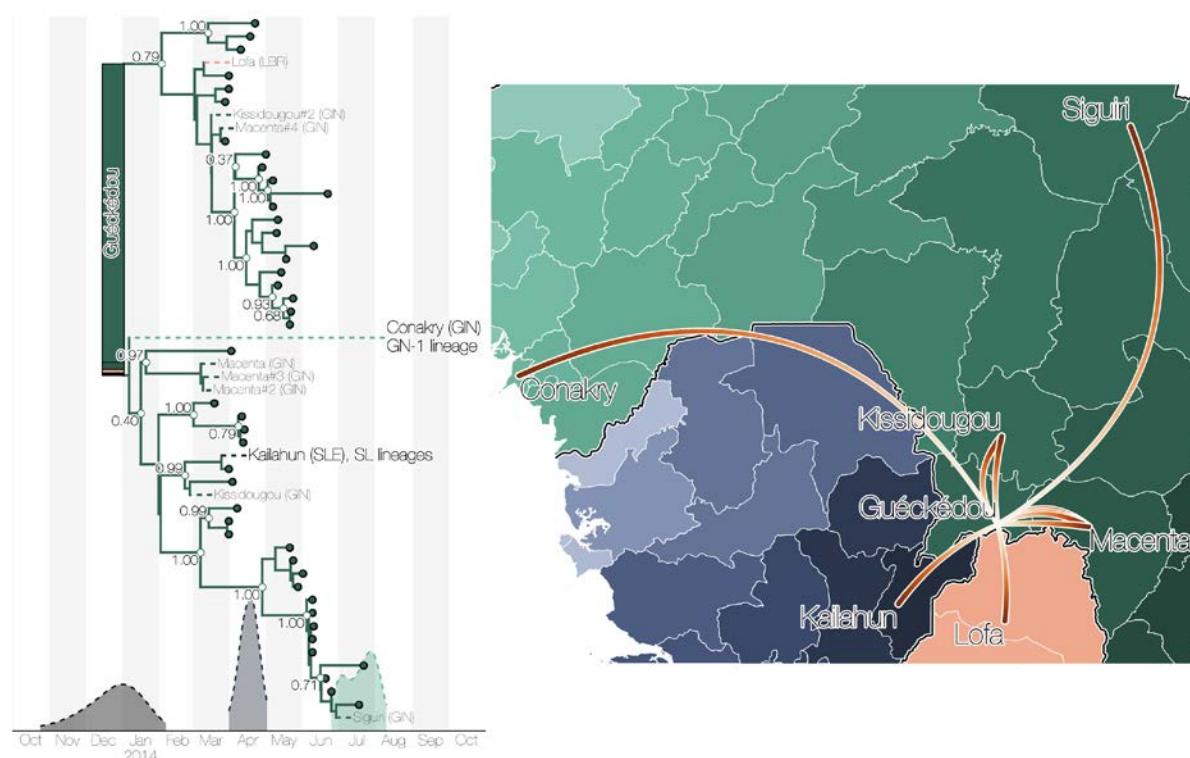


Figure 2 | Summary of initial events of the epidemic. a) The phylogeny of the initially sampled cases in Guéckédou, Guinea and their relationship to the initial dispersal events into neighbouring (and more distant) regions. Stacked bars at the root of the tree indicate posterior probabilities for the 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 heading to Kailahun district (blue, introduction gave rise to SL lineages) and to Coyah prefecture (green, introduction leads to lineage A) are shown at the bottom of the tree. Nodes with three or more tips have posterior probabilities shown if they are >0.3 . **b)** These same dispersal events on a map with directionality 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.

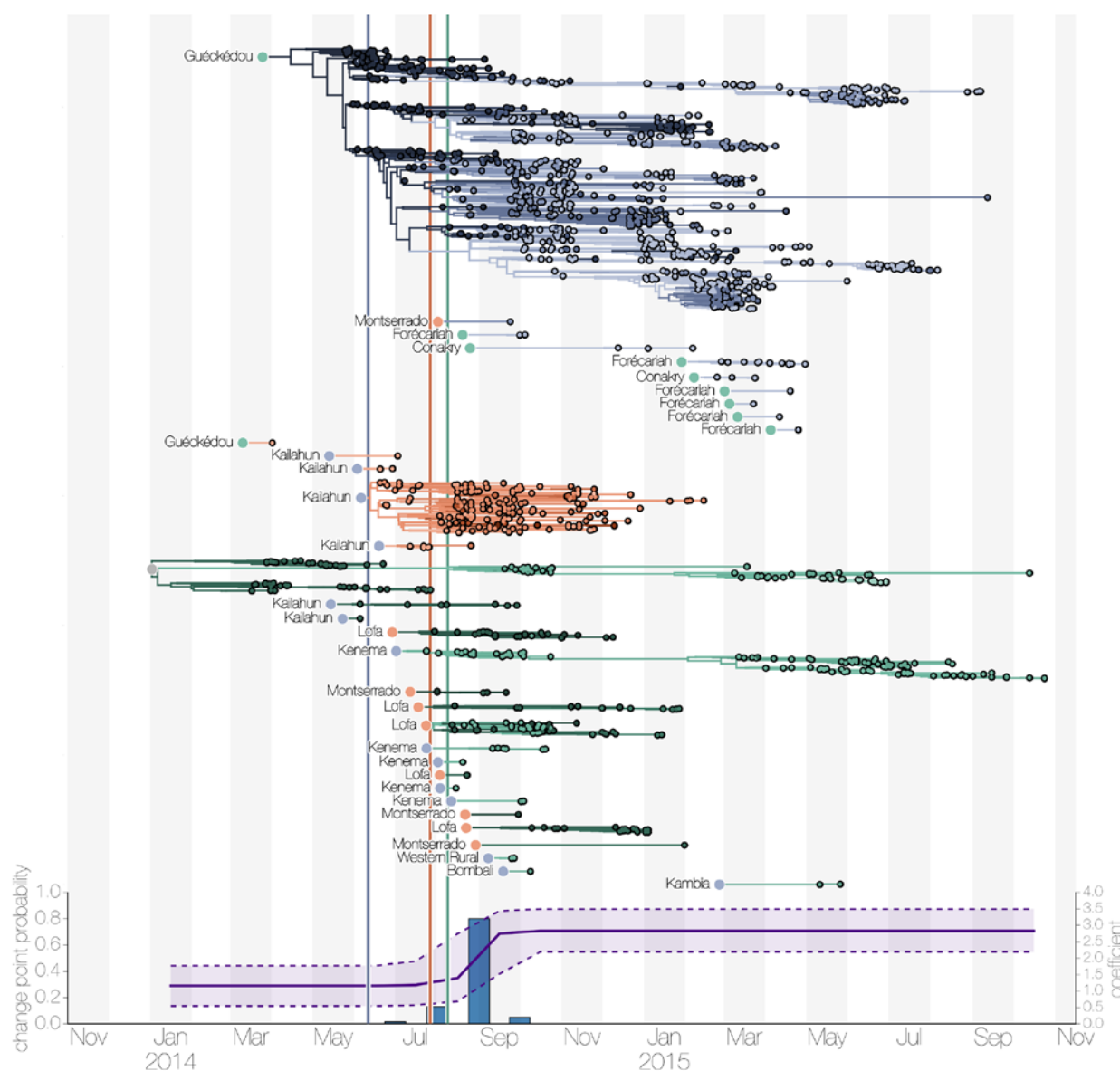


Figure 3 | Time tree deconstructed into international introductions. a) EBOV lineages tracked until their last known descendants are sampled, sorted by country (Sierra Leone, blue; Liberia, red; Guinea, green) and earliest possible introduction point. Colour intensity of tips are by longitude (lightest to the West, darkest to the East). Circles at root of each subtree denote country of origin for the introduced lineage. The 4 introductions into Liberia in May-June 2014 are all reconstructed to have come from Kailahun, Sierra Leone, and plausibly represent a few or just one event but genetic similarity to viruses from Kailahun makes resolution impossible. In contrast, the many introductions into Guinea are reconstructed as being from multiple origins over a 10 month period. **b)** Epoch estimates of the change point probability (primary Y-axis) and log coefficient (mean and HPD, secondary Y-axis) for the within country effect (the only effect with support for epoch dynamics, cfr. 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. .

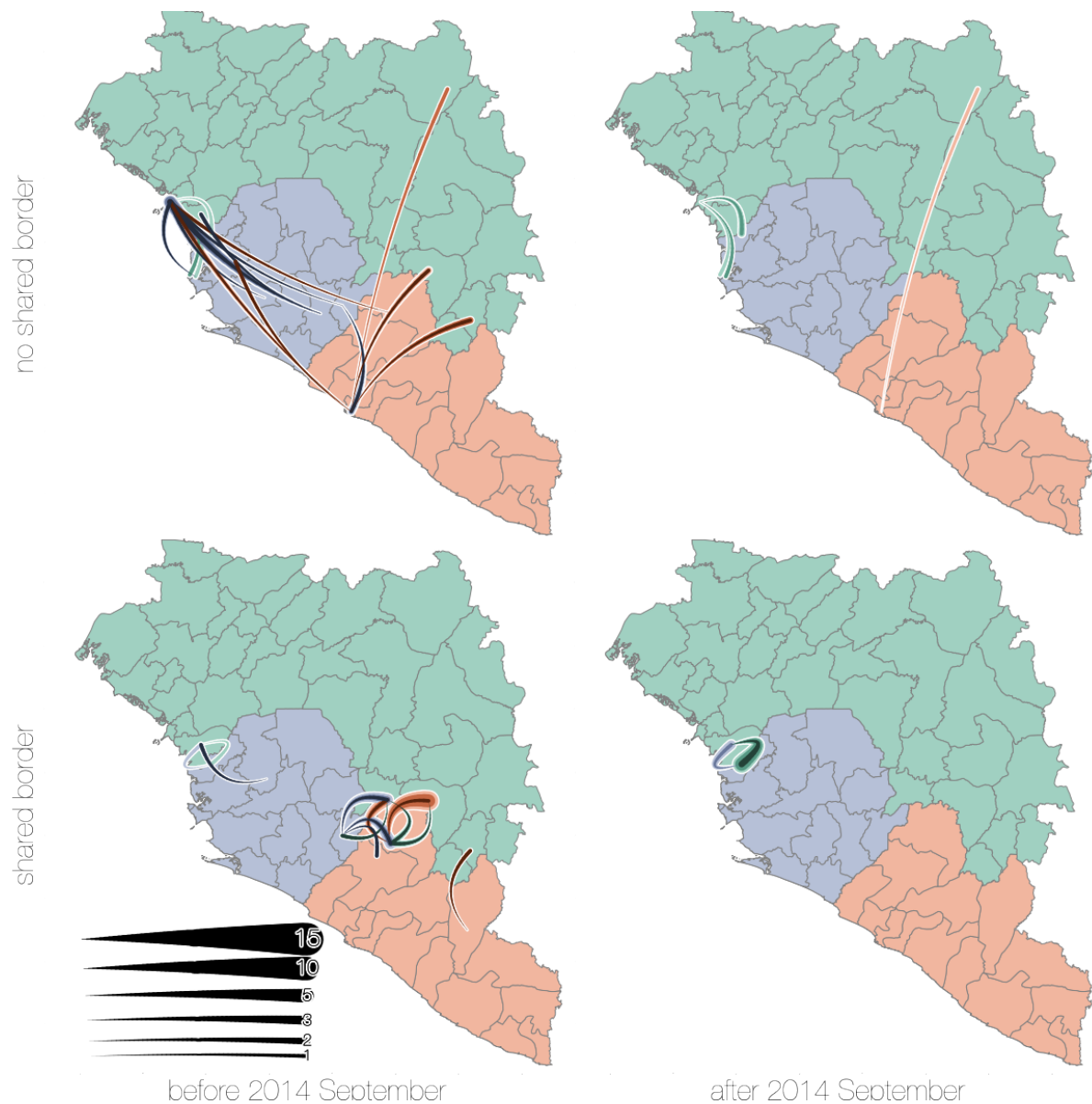


Figure 4 | Summarised migration history during the epidemic split by timing and border sharing. Curved lines indicate median (intermediate colour intensity), and 95% highest posterior density intervals (lightest and darkest colour intensities) for the number of migrations that are inferred to have taken place between locations, split by whether they occurred before or after September 2014 (inferred epoch change point, Fig 3) and whether migrating lineages migrated to a neighbouring location or further.

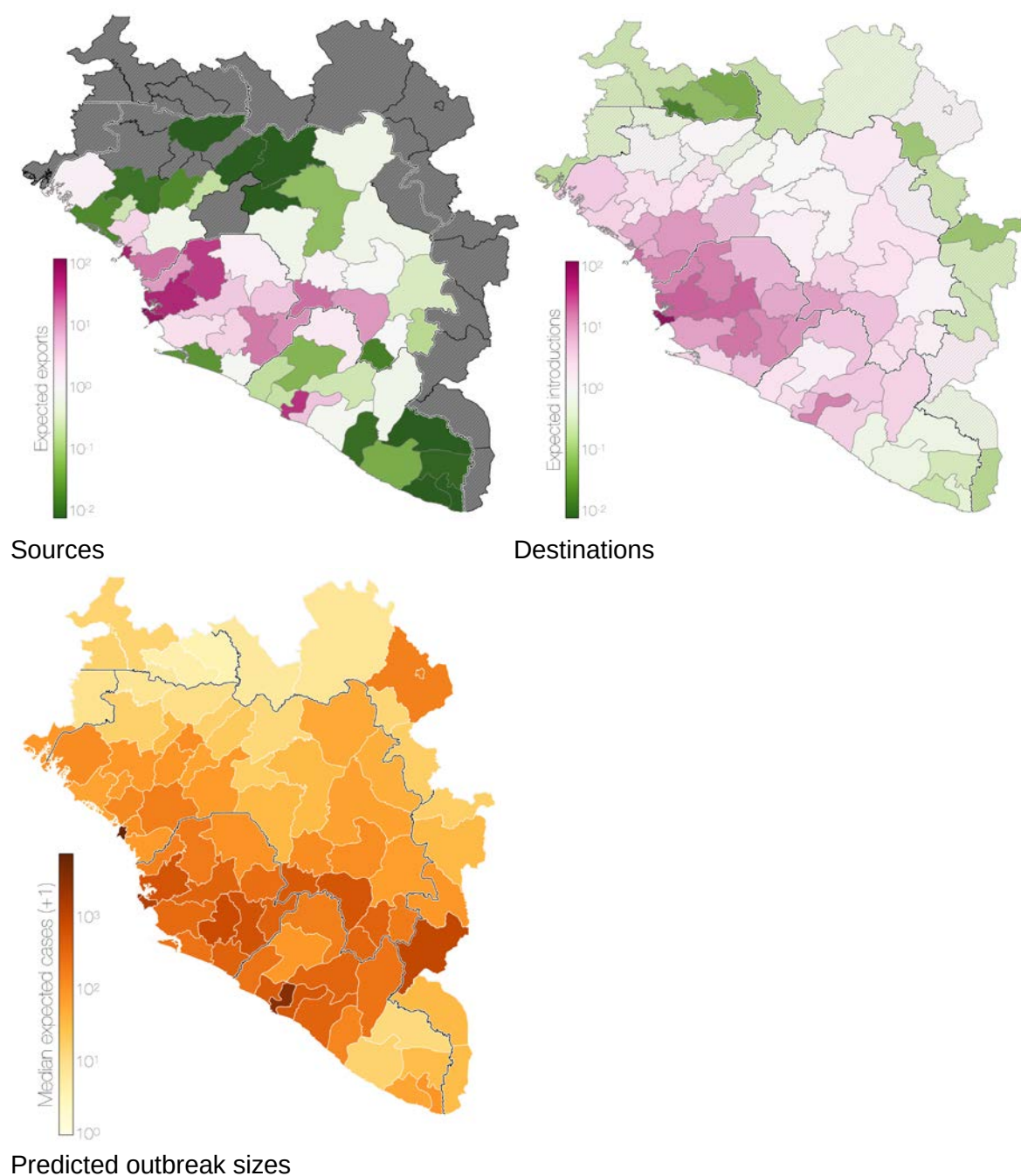


Figure 5 | Predicted sources, destinations and consequences of viral migrations. a) Expected number of exports from each of 56 regions in Guinea, Sierra Leone and Liberia with recorded cases and the surrounding 18 regions from the neighbouring countries of Guinea-Bissau, Senegal, Mali and Côte d'Ivoire. Grey regions had no recorded cases. **b)** Predicted number of imports into each region. **c)** Predicted outbreak sizes from the generalised linear model fitted to case data.

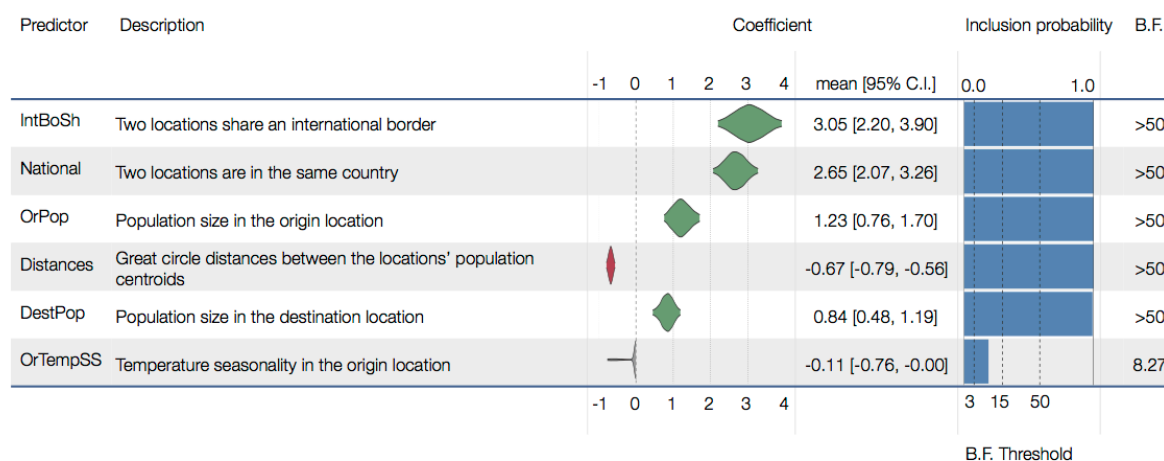


Table 1 | Summary of the phylogenetic generalized linear model results. The estimated coefficients and model inclusion probabilities for spatial movement predictors supported with a Bayes factor (BF) > 3. Positive coefficients are shown in green, negative in red. The remainder are not supported and are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

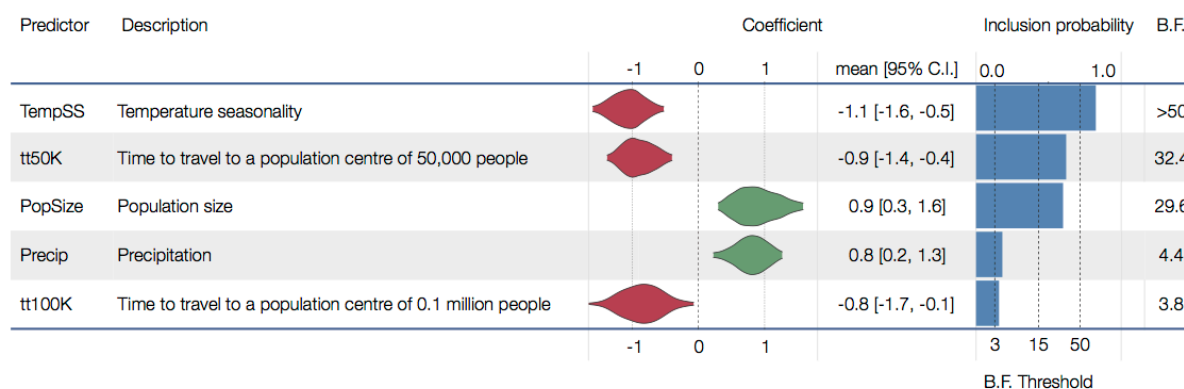


Table 2 | Summary of generalized linear model results with case counts as the response variable. The estimated coefficients and model inclusion probabilities for per-region predictors supported with a Bayes factor (BF) > 3. Positive coefficients are shown in green, negative in red. The remainder are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

Supplementary Methods

Sequence data

The data set consists of 1610 full Ebola virus (EBOV) genomes sampled between 17 March 2014 and 24 October 2015. 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, although it is difficult to ascertain whether sequences from such cases were not included in the final data set if they were collected at the height of the epidemic. Sequences were aligned using MAFFT [Katoh et al., 2002] and edited manually. The alignment was partitioned into coding regions and non-coding intergenic regions and concatenated such that coding sequences are connected end to end, followed by intergenic regions separated from the coding sequences by a spacer of three N's. The final alignment length was 18992 nucleotides.

Masking putative ADAR edited sites

As noticed by Tong et al. [2015], Park et al. [2015] and other studies, some EBOV isolates contain excessive numbers 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 the 15 hypermutation patterns we detect 11 that are found in single genomes and four that are shared across 67 strains. Putative tracts of T-to-C hypermutation almost exclusively occur within non-coding intergenic regions, where their effects to viral fitness are presumably minimal. Hypermutation tracts are usually found in individual sequences, further supporting the hypothesis that they are deleterious to EBOV fitness. Nevertheless, some lineages that survive hypermutation are transmitted and are subsequently found in

more than one isolate [Smits et al., 2015]. In these cases we 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). Evolutionary rates were modelled with branch-specific rates drawn from a relaxed molecular clock following a log-normal distribution [Drummond et al., 2006] and site-specific rates scaled by relative rates in the four partitions. A non-parametric coalescent ‘Skygrid’ tree prior was employed for demographic inference [Gill et al., 2013]. The overall evolutionary rate was given an uninformative 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.

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.

Geographic history reconstruction

The level of administrative division within each country was chosen so that population sizes between divisions are comparable. For each country the appropriate administrative divisions 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 locations (63 in total) and each sequence, where available, was assigned the location where the patient was recorded to have been infected as a discrete trait. When exact location was unknown, but other information, such as country, was available we inferred the sequence location as a latent variable with equal prior probability over all available locations within that country. In the absence of all geographic information we inferred both the country and the location of a sequence.

An asymmetric continuous-time Markov chain (CTMC) [Lemey et al., 2009, Edwards et al., 2011] matrix was used to infer instantaneous transitions between locations. We restricted the analysis to the 56 locations with recorded EVD cases but even then, a total of 3080 independent transition rates would be challenging to infer from one realisation of the process.

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 $\mathbf{X}_{ij} = (x_{ij1}, \dots, x_{ijP})'$ with unknown coefficients $\boldsymbol{\beta} = (\beta_1, \dots, \beta_P)'$ and diagonal

matrix δ with entries $(\delta_1, \dots, \delta_P)$. 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

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

where $\epsilon = (\epsilon_1, \dots, \epsilon_K)$ 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 easy 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.

To draw posterior inference, we enjoy the success of Lemey et al. [2014] on integrating β and δ , further employ a random-walk Metropolis transition kernel on ϵ and sample σ^2 directly from its full conditional distribution using Gibbs sampling.

We estimate the expected number of transitions (ζ_j) to any location j not included in the analysis – specifically for districts in Guinea, Sierra Leone and Liberia for which no cases were reported ($n = 7$) and for districts in neighbouring countries along the borders with Guinea or Liberia that remained disease free ($n = 18$) – as follows:

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

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 normalising constant applied to the CTMC rate matrices in BEAST. We sum over all possible origin states (states included in the analysis) to integrate over all possible positions in the phylogeny. To obtain estimates (ζ_j) 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. We summarise mean posterior estimates for ζ_j based on the samples obtained by our MCMC analysis; we note that also the value of c is sample-specific. To contextualise these expectations, we also calculate these quantities for the 56 sampled locations.

For analyses that consider time-inhomogeneity in the 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 $\mathbf{X}_{ijt} = (x_{ijt1}, \dots, x_{ijtP})'$ with constant-through-time,

unknown coefficients β . We generalize this model to incorporate 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

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

where $\beta_B = (\beta_{B1}, \dots, \beta_{BP})'$ are the unknown coefficients before the change-points, $\beta_A = (\beta_{A1}, \dots, \beta_{AP})'$ are the unknown coefficients after the change-points, diagonal matrix $\phi(t)$ has entries $(1_{t > t_1}(t), \dots, 1_{t > t_P}(t))$, $1_{(\cdot)}(t)$ is the indicator function and $\mathbf{T} = (t_1, \dots, t_P)$ 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 $p = 1, \dots, P$. 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 β_{Bp} and β_{Ap} are independent and normally distributed with mean 0 and a relatively large variance of 4. When random, each t_p is equally likely to lie before any epoch. We employ random-walk Metropolis transition kernels on β_B , β_A and T .

In a first epoch GLM analysis, we keep the five predictors that are 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 (Table ??). 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} (Figure ??), keeping the remaining four predictors homogeneous through time.

Ebola virus disease (EVD) case numbers are reported by the WHO for every country division (district) at the appropriate administrative level, split by epidemiological week. For every district 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 epi week lasting from 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. This probably 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.

Within-location generalised linear models

We studied the association between both disease case counts and lineage persistence using generalised linear models and stochastic search variable selection (SSVS), 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 ???. We also employed SSVS as described above, in order to compute Bayes factors 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.

Case counts

Using the disease case counts collected as detailed above, we then fit a negative binomial generalised linear model using case data from all locations which reported cases.

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

where r is the overdispersion parameter.

To complete our model we need to specify a prior distribution for the model's parameters:

$$\begin{aligned} \sigma &\sim \text{InverseGamma}(0.001, 0.001) \\ r &\sim \text{Uniform}(0, 50) \\ \alpha &\sim \text{Normal}(0, \tau) \\ \tau &\sim \text{InverseGamma}(0.01, 0.01) \\ Pr(\delta_k = 1) &= q = 1 - \left(\frac{1}{2}\right)^{1/P}, \quad k = 1, 2, \dots, P \end{aligned}$$

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.

Viral persistence

PARAGRAPH DESCRIBING HOW WE GOT THE PERSISTENCE DATA

For the locations which never experience an introduction we set the persistence time to zero.

As this is a continuous response variable, we log-transformed the persistence times and employed a Gaussian GLM:

$$\begin{aligned} T_i &\sim \text{Normal}(u_i, \sigma_{\text{persistence}}^2) \\ \mu_i &= \alpha + \beta_1 \delta_1 x_{i1} + \dots + \beta_P \delta_P x_{iP} \end{aligned}$$

with $\sigma_{\text{persistence}}^2 \sim \text{InverseGamma}(0.01, 0.01)$.

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>) and slightly modified them to a non-phylogenetic setting. We also wrote a suite of accompanying functions in the R package RBeast () to make it easier for users to create XML files with their models and data.

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 counts for all locations with zero counts.

References

- Brenda L. Bass and Harold Weintraub. An unwinding activity that covalently modifies its double-stranded RNA substrate. *Cell*, 55(6):1089–1098, December 1988. ISSN 0092-8674. doi: 10.1016/0092-8674(88)90253-X. URL <http://www.sciencedirect.com/science/article/pii/009286748890253X>.
- F. Bielejec, P. Lemey, G. Baele, A. Rambaut, and M. A. Suchard. Inferring heterogeneous evolutionary processes through time: from sequence substitution to phylogeography. *Syst. Biol.*, 63(4):493–504, Jul 2014.
- Jennifer A. Carpenter, Liam P. Keegan, Lena Wilfert, Mary A. O’Connell, and Francis M. Jiggins. Evidence for ADAR-induced hypermutation of the *Drosophila* sigma virus (Rhabdoviridae). *BMC Genetics*, 10(1):75, November 2009. ISSN 1471-2156. doi: 10.1186/1471-2156-10-75. URL <http://www.biomedcentral.com/1471-2156/10/75/abstract>.
- Roberto Cattaneo, Anita Schmid, Daniel Eschle, Knut Baczko, Volker ter Meulen, and Martin A. Billeter. Biased hypermutation and other genetic changes in defective measles viruses in human brain infections. *Cell*, 55(2):255–265, October 1988. ISSN 0092-8674. doi: 10.1016/0092-8674(88)90048-7. URL <http://www.sciencedirect.com/science/article/pii/0092867488900487>.
- Alexei J Drummond, Simon Y. W Ho, Matthew J Phillips, and Andrew Rambaut. Relaxed phylogenetics and dating with confidence. *PLoS Biol*, 4(5):e88, March 2006. doi: 10.1371/journal.pbio.0040088. URL <http://dx.doi.org/10.1371/journal.pbio.0040088>.

- Alexei J. Drummond, Marc A. Suchard, Dong Xie, and Andrew Rambaut. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29 (8):1969–1973, August 2012. ISSN 0737-4038, 1537-1719. doi: 10.1093/molbev/mss075. URL <http://mbe.oxfordjournals.org/content/29/8/1969>.
- Ceiridwen J. Edwards, Marc A. Suchard, Philippe Lemey, John J. Welch, Ian Barnes, Tara L. Fulton, Ross Barnett, Tamsin C. O’Connell, Peter Coxon, Nigel Monaghan, Cristina E. Valdiosera, Eline D. Lorenzen, Eske Willerslev, Gennady F. Baryshnikov, Andrew Rambaut, Mark G. Thomas, Daniel G. Bradley, and Beth Shapiro. Ancient Hybridization and an Irish Origin for the Modern Polar Bear Matriline. *Current Biology*, 21(15):1251–1258, August 2011. ISSN 0960-9822. doi: 10.1016/j.cub.2011.05.058. URL <http://www.sciencedirect.com/science/article/pii/S0960982211006452>.
- Marco A. R. Ferreira and Marc A. Suchard. Bayesian analysis of elapsed times in continuous-time markov chains. *Canadian Journal of Statistics*, 36(3):355–368, 2008. ISSN 1708-945X. doi: 10.1002/cjs.5550360302. URL <http://onlinelibrary.wiley.com/doi/10.1002/cjs.5550360302/abstract>.
- Jean-François G  linas, Guerline Clerzius, Eileen Shaw, and Anne Gatignol. Enhancement of Replication of RNA Viruses by ADAR1 via RNA Editing and Inhibition of RNA-Activated Protein Kinase. *Journal of Virology*, 85(17):8460–8466, September 2011. ISSN 0022-538X, 1098-5514. doi: 10.1128/JVI.00240-11. URL <http://jvi.asm.org/content/85/17/8460>.
- Mandev S. Gill, Philippe Lemey, Nuno R. Faria, Andrew Rambaut, Beth Shapiro, and Marc A. Suchard. Improving bayesian population dynamics inference: A coalescent-based model for multiple loci. *Molecular Biology and Evolution*, 30(3):713–724, March 2013. ISSN 0737-4038, 1537-1719. doi: 10.1093/molbev/mss265. URL <http://mbe.oxfordjournals.org/content/30/3/713>.
- Masami Hasegawa, Hirohisa Kishino, and Taka-aki Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2): 160–174, October 1985. ISSN 0022-2844, 1432-1432. doi: 10.1007/BF02101694. URL <http://link.springer.com/article/10.1007/BF02101694>.
- Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, July 2002. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkf436. URL <http://nar.oxfordjournals.org/content/30/14/3059>.
- Philippe Lemey, Marc Suchard, and Andrew Rambaut. Reconstructing the initial global spread of a human influenza pandemic. *PLoS Currents*, 1, September 2009. ISSN 2157-3999. doi: 10.1371/currents.RRN1031. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2762761/>.
- Philippe Lemey, Andrew Rambaut, Trevor Bedford, Nuno Faria, Filip Bielejec, Guy Baele, Colin A. Russell, Derek J. Smith, Oliver G. Pybus, Dirk Brockmann, and Marc A. Suchard. Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3n2. *PLOS Pathog*, 10(2):e1003932,

February 2014. ISSN 1553-7374. doi: 10.1371/journal.ppat.1003932. URL <http://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1003932>.

Daniel J. Park, Gytis Dudas, Shirlee Wohl, Augustine Goba, Shannon L. M. Whitmer, Kristian G. Andersen, Rachel S. Sealfon, Jason T. Ladner, Jeffrey R. Kugelman, Christian B. Matranga, Sarah M. Winnicki, James Qu, Stephen K. Gire, Adrienne Gladden-Young, Simbirie Jalloh, Dolo Nosamiefan, Nathan L. Yozwiak, Lina M. Moses, Pan-Pan Jiang, Aaron E. Lin, Stephen F. Schaffner, Brian Bird, Jonathan Towner, Mambu Mamoh, Michael Gbakie, Lansana Kanneh, David Kargbo, James L. B. Massally, Fatima K. Kamara, Edwin Konuwa, Josephine Sellu, Abdul A. Jalloh, Ibrahim Mustapha, Momoh Foday, Mohamed Yillah, Bobbie R. Erickson, Tara Sealy, Dianna Blau, Christopher Paddock, Aaron Brault, Brian Amman, Jane Basile, Scott Bear-den, Jessica Belser, Eric Bergeron, Shelley Campbell, Ayan Chakrabarti, Kimberly Dodd, Mike Flint, Aridh Gibbons, Christin Goodman, John Klena, Laura McMullan, Laura Morgan, Brandy Russell, Johanna Salzer, Angela Sanchez, David Wang, Irwin Jungreis, Christopher Tomkins-Tinch, Andrey Kislyuk, Michael F. Lin, Sinead Chapman, Bronwyn MacInnis, Ashley Matthews, James Bochicchio, Lisa E. Hensley, Jens H. Kuhn, Chad Nusbaum, John S. Schieffelin, Bruce W. Birren, Marc Forget, Stuart T. Nichol, Gustavo F. Palacios, Daouda Ndiaye, Christian Happi, Sahr M. Gevaio, Mohamed A. Vandi, Brima Kargbo, Edward C. Holmes, Trevor Bedford, Andreas Gnirke, Ute Ströher, Andrew Rambaut, Robert F. Garry, and Pardis C. Sabeti. Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. *Cell*, 161(7):1516–1526, June 2015. ISSN 0092-8674. doi: 10.1016/j.cell.2015.06.007. URL <http://www.sciencedirect.com/science/article/pii/S009286741500690X>.

Paloma Rueda, Blanca García-Barreno, and José A. Melero. 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(2):653–662, February 1994. ISSN 0042-6822. doi: 10.1006/viro.1994.1077. URL <http://www.sciencedirect.com/science/article/pii/S0042682284710774>.

Saskia L. Smits, Suzan D. Pas, Chantal B. Reusken, Bart L. Haagmans, Peirro Pertile, Corrado Cancedda, Kerry Dierberg, Isata Wurie, Abdul Kamara, David Kargbo, Sarah L. Caddy, Armando Arias, Lucy Thorne, Jia Lu, Umaru Jah, Ian Goodfellow, and Marion P. Koopmans. 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(40), October 2015. ISSN 1560-7917. doi: 10.2807/1560-7917.ES.2015.20.40.30035.

Yi-Gang Tong, Wei-Feng Shi, Di Liu, Jun Qian, Long Liang, Xiao-Chen Bo, Jun Liu, Hong-Guang Ren, Hang Fan, Ming Ni, Yang Sun, Yuan Jin, Yue Teng, Zhen Li, David Kargbo, Foday Dfae, Alex Kanu, Cheng-Chao Chen, Zhi-Heng Lan, Hui Jiang, Yang Luo, Hui-Jun Lu, Xiao-Guang Zhang, Fan Yang, Yi Hu, Yu-Xi Cao, Yong-Qiang Deng, Hao-Xiang Su, Yu Sun, Wen-Sen Liu, Zhuang Wang, Cheng-Yu Wang, Zhao-Yang Bu, Zhen-Dong Guo, Liu-Bo Zhang, Wei-Min Nie, Chang-Qing Bai, Chun-Hua Sun, Xiao-Ping An, Pei-Song Xu, Xiang-Li-Lan Zhang, Yong Huang, Zhi-Qiang Mi, Dong Yu, Hong-Wu Yao, Yong Feng, Zhi-Ping Xia, Xue-Xing Zheng, Song-Tao Yang, Bing Lu, Jia-Fu Jiang, Brima Kargbo, Fu-Chu He, George F. Gao, Wu-Chun Cao, and The China

Mobile Laboratory Testing Team in Sierra Leone. Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. *Nature*, 524(7563):93–96, August 2015. ISSN 0028-0836. doi: 10.1038/nature14490. URL <http://www.nature.com/nature/journal/v524/n7563/full/nature14490.html>.

Ziheng Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, 39(3): 306–314, September 1994. ISSN 0022-2844, 1432-1432. doi: 10.1007/BF00160154. URL <http://link.springer.com/article/10.1007/BF00160154>.

Supplementary Information

Table S1. Number of cases and sampled sequences per location and country, where ‘Location’ is standardised location name and ‘Sampling frequency’ is sequences/cases $\times$ 100.

Country	Location	Sequences	Cases	Sampling freq.
GIN	Macenta	40	784	5.10
GIN	Conakry	73	629	11.61
GIN	Forecariah	60	502	11.95
GIN	Gueckedou	58	390	14.87
GIN	Nzerekore	9	269	3.35
GIN	Coyah	26	257	10.12
GIN	Kerouane	10	176	5.68
GIN	Dubreka	22	167	13.17
GIN	Kissidougou	18	138	13.04
GIN	Kindia	2	131	1.53
GIN	Lola	2	118	1.69
GIN	Faranah	8	88	9.09
GIN	Boffa	0	52	0.00
GIN	Beyla	4	52	7.69
GIN	Telimele	0	43	0.00
GIN	Siguiri	3	38	7.89
GIN	Kankan	4	38	10.53
GIN	Boke	18	36	50.00
GIN	Kouroussa	2	22	9.09
GIN	Fria	3	16	18.75
GIN	Dabola	0	15	0.00
GIN	Yamou	0	12	0.00
GIN	Dalaba	3	10	30.00
GIN	Pita	0	8	0.00
GIN	Mali	0	5	0.00
GIN	Tougue	0	2	0.00
GIN	Dinguiraye	0	1	0.00
GIN	Labe	0	0	NA
GIN	Koundara	0	0	NA
GIN	Mandiana	0	0	NA
GIN	Gaoual	0	0	NA
GIN	Lelouma	0	0	NA
GIN	Koubia	0	0	NA
GIN	Mamou	0	0	NA
LBR	Montserrado	67	2925	2.29
LBR	Margibi	21	878	2.39
LBR	Lofa	13	511	2.54
LBR	Nimba	5	282	1.77
LBR	Bomi	5	220	2.27
LBR	Bong	2	219	0.91

LBR	GrandCapeMount	4	207	1.93
LBR	GrandBassa	14	164	8.54
LBR	RiverCess	2	48	4.17
LBR	Sinoe	3	33	9.09
LBR	Gbarpolu	1	28	3.57
LBR	GrandKru	2	25	8.00
LBR	RiverGee	1	13	7.69
LBR	Maryland	0	7	0.00
LBR	GrandGedeh	0	4	0.00
SLE	WesternUrban	130	3219	4.04
SLE	PortLoko	149	2208	6.75
SLE	WesternRural	84	1736	4.84
SLE	Bombali	119	1212	9.82
SLE	Kailahun	101	756	13.36
SLE	Tonkolili	19	632	3.01
SLE	Kono	39	568	6.87
SLE	Kenema	75	553	13.56
SLE	Bo	13	450	2.89
SLE	Kambia	63	326	19.33
SLE	Moyamba	23	317	7.26
SLE	Koinadugu	11	185	5.95
SLE	Pujehun	9	68	13.24
SLE	Bonthe	1	5	20.00

Table S2. Predictors included in the time-homogenous GLM.

Predictor type	Internal predictor code	Predictor description
Geographic	greatCircleDistances	great-circle distances standardized, ln-transformed, standardized
Administrative	withinCountry	within country effect
Administrative	internationalBorderShared	location pairs that are in different countries and share a border
Administrative	nationalBorderShared	location pairs that are in the same country and share a border
Administrative	betweenLBR_GIN_Asymmetry	between Liberia-Guinea asymmetry
Administrative	betweenLBR_SLE_Asymmetry	between Liberia-Sierra Leone asymmetry
Administrative	betweenGIN_SLE_Asymmetry	between Guinea-Sierra Leone asymmetry
Demographic	originPopSize	origin population size, ln-transformed, standardized
Demographic	destinationPopSize	destination population size, ln-transformed, standardized
Demographic	originPopDens	origin population density, ln-transformed, standardized

Demographic	destinationPopDens	destination population density, ln-transformed, standardized
Demographic	originTT100k	estimated mean travel time in minutes to reach the nearest major settlement of at least 100,000 people at origin, ln-transformed, standardized
Demographic	destinationTT100k	estimated mean travel time in minutes to reach the nearest major settlement of at least 100,000 people at destination, ln-transformed, standardized
Demographic	originGEcon	origin Gridded economic output, ln-transformed, standardized
Demographic	destinationGEcon	destination Gridded economic output, ln-transformed, standardized
Cultural	internationalLanguageShared	location pairs that are in different countries and share at least one of 17 vernacular languages
Cultural	nationalLanguageShared	location pairs that are in the same country and share at least one of 17 vernacular languages
Climatic	originTemp	Temperature annual mean at origin, ln-transformed, standardized
Climatic	destinationTemp	Temperature annual mean at destination, ln-transformed, standardized
Climatic	originTempSeason	index of temperature seasonality at origin, ln-transformed, standardized
Climatic	destinationTempSeason	index of temperature seasonality at destination, ln-transformed, standardized
Climatic	originPrecip	Precipitation annual mean at origin, ln-transformed, standardized
Climatic	destinationPrecip	Precipitation annual mean at destination, ln-transformed, standardized
Climatic	originPrecipSeason	index of precipitation seasonality at origin, ln-transformed, standardized
Climatic	destinationPrecipSeason	index of precipitation seasonality at destination, ln-transformed, standardized

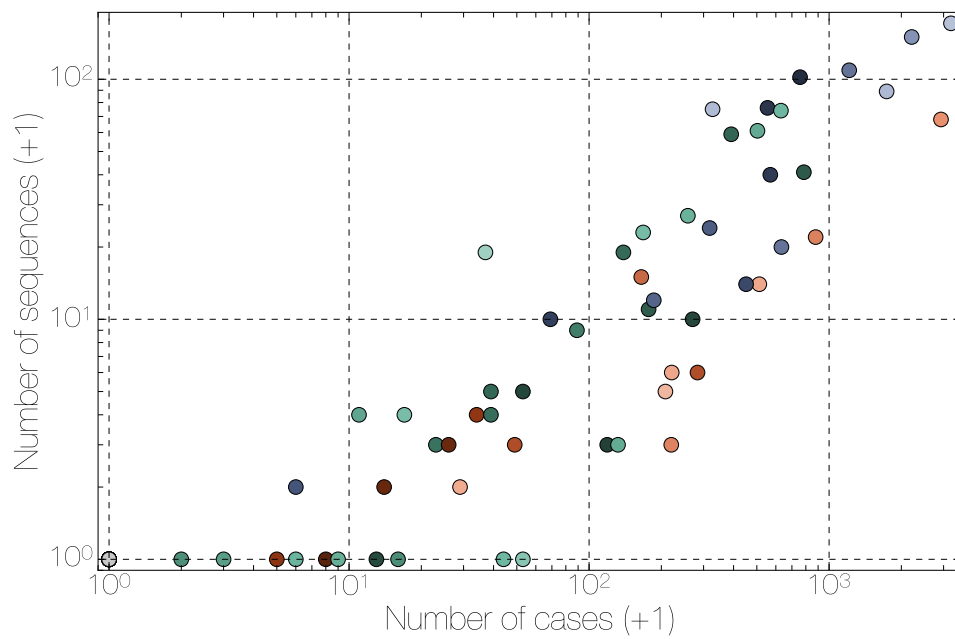


Figure S1. Correlation between number of cases and number of sequences for each location. A plot of number of EBOV genomes sampled against the known and suspected cumulative EVD case numbers. Regions in Guinea are denoted in green, Sierra Leone in blue and Liberia in red. Spearman correlation coefficient: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

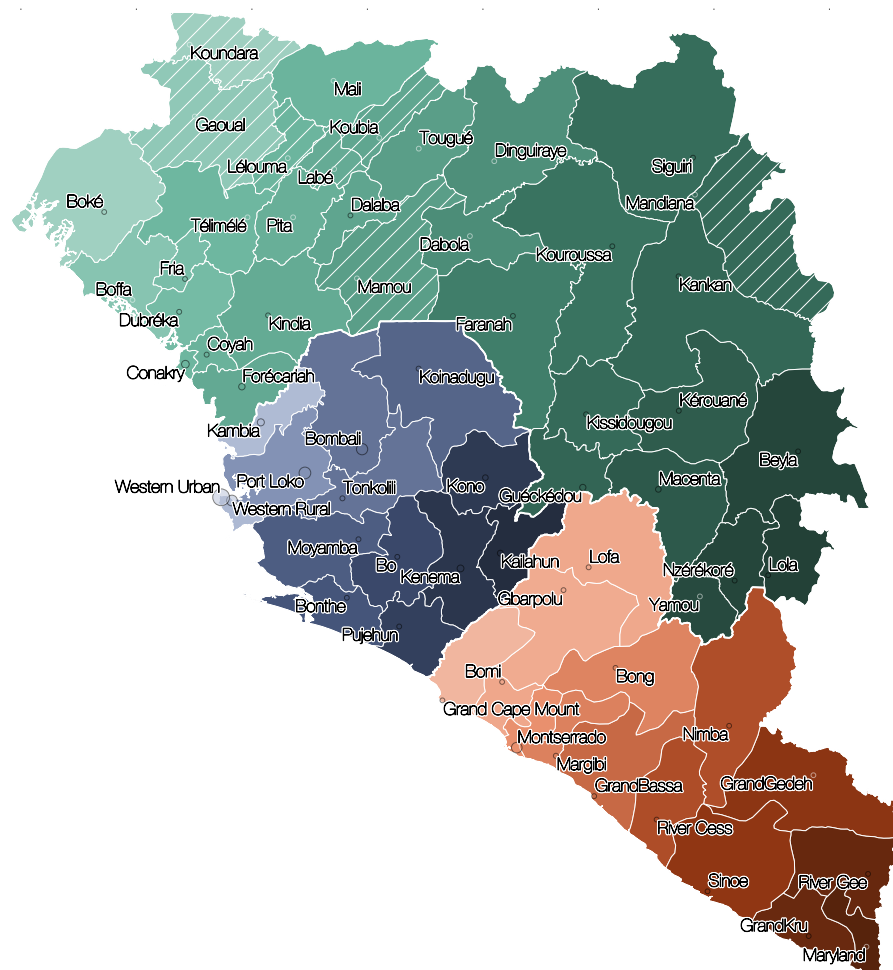


Figure S2. Tree location legend. Regions used in the analysis coloured by country (Liberia in red, Guinea in green and Sierra Leone in blue) and position of each region within the country along a south-east (dark) to north-west (light) gradient. Hatched regions had no recorded EVD cases.

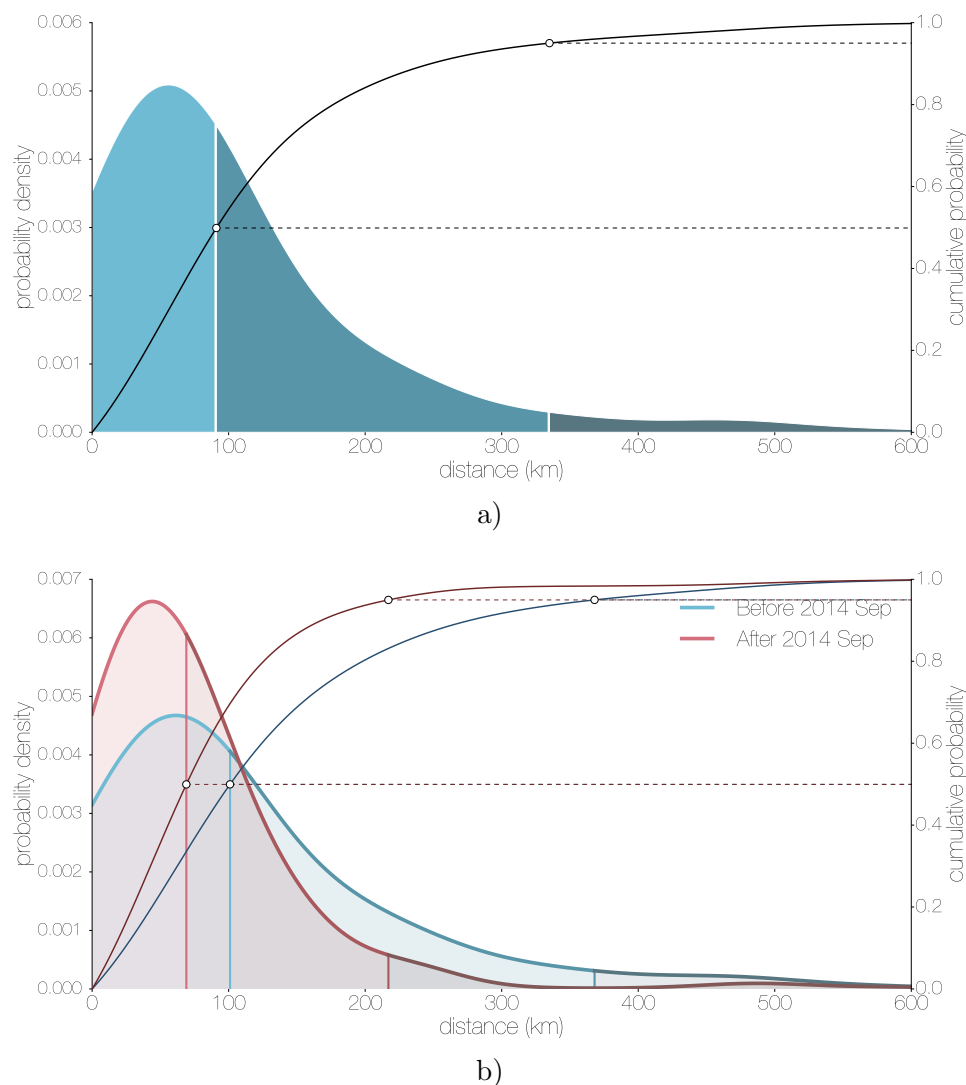


Figure S3. Kernel density estimate of migration distances across the posterior distribution of migrations with a bandwidth of 0.5. a) Kernel density estimate for all inferred migrations: 50% occur over distances ≤ 100 km and 5% occur over distances ≤ 340 km. b) Kernel density estimates for migrations occurring before (blue) and after (red) September 2014, the inferred epoch change point (see Fig 3). After September 2014, half of migrations occur at a distance ≤ 89 km (compared to ≤ 118 km before September) and 5% happen over distances ≤ 308 km (≤ 367 km before September).

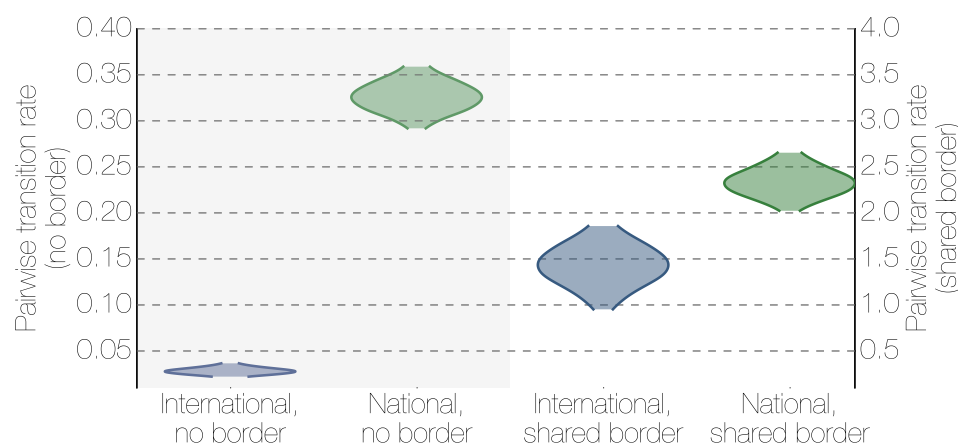


Figure S4. The effect of borders on EBOV migration rates between regions. Posterior densities of the migration rates between locations that share a geographical border (left) and those that don't (right) for international migrations and national migrations. Where two regions share a border, national 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.

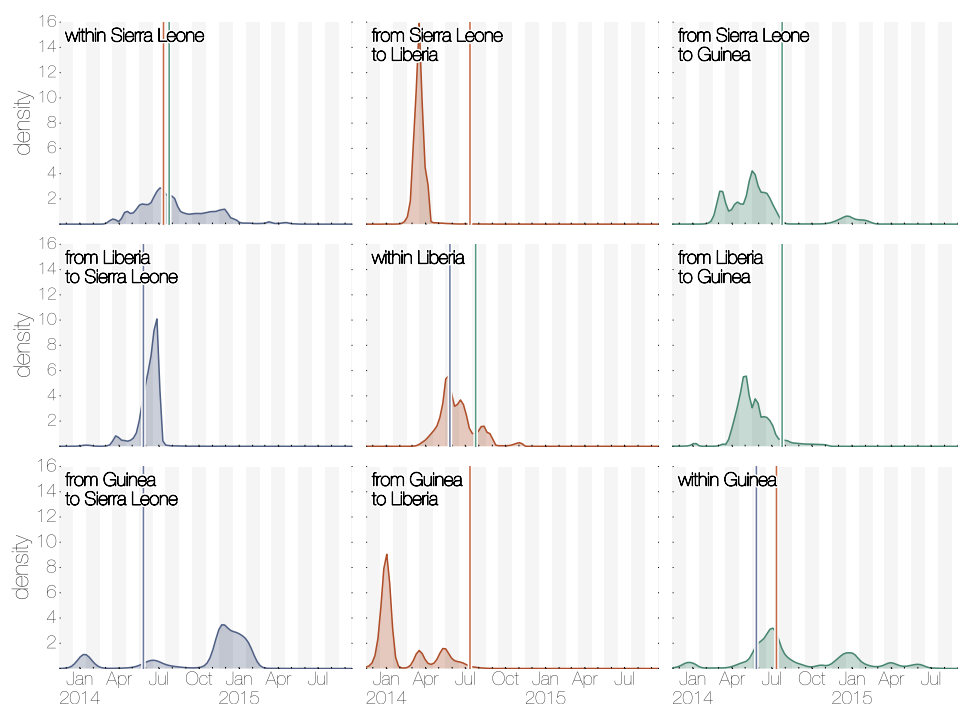


Figure S5. Summary of migration intensity over time in the region. Each cell shows the posterior probability density of temporal migration intensity. Vertical lines within each cell indicate the dates of declared border closures by each of the three countries: 11 June 2014 in Sierra Leone (blue), 27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

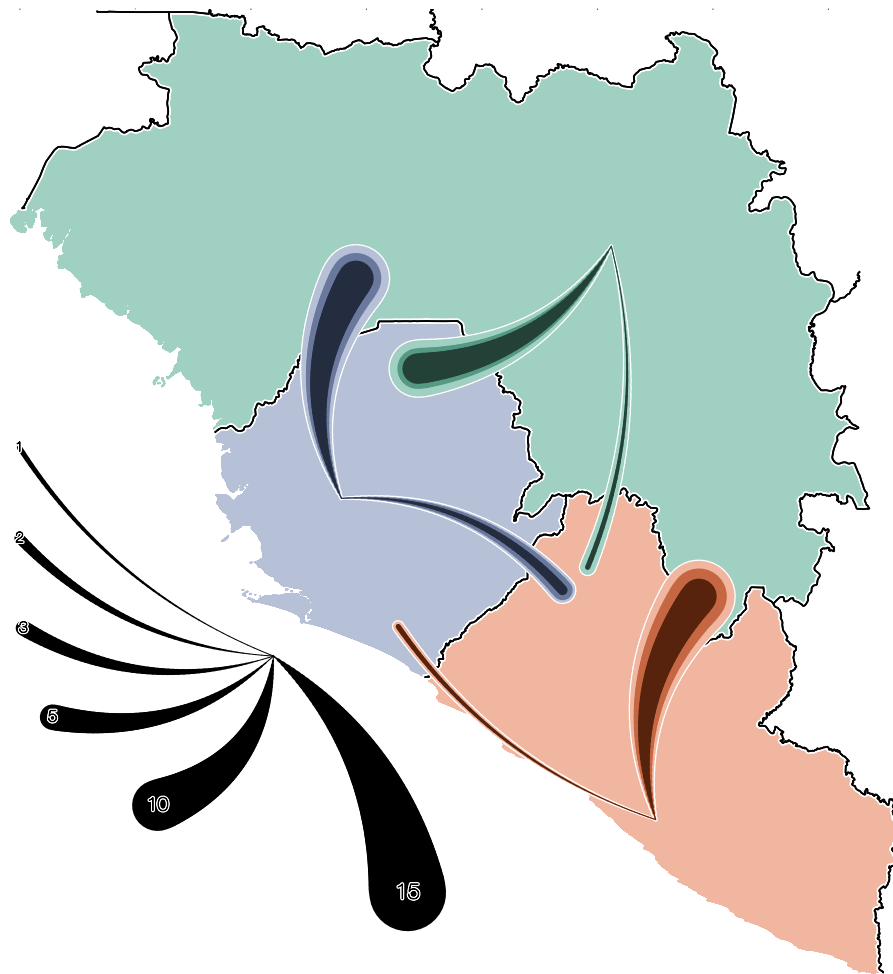


Figure S6. Summary of inferred international migrations. Lines connecting countries indicate the inferred magnitude of viral migration throughout the epidemic. Intermediate darkness contour indicates the median number of inferred migrations, lower and upper 95% highest posterior density intervals correspond to the darkest and lightest contours, respectively.

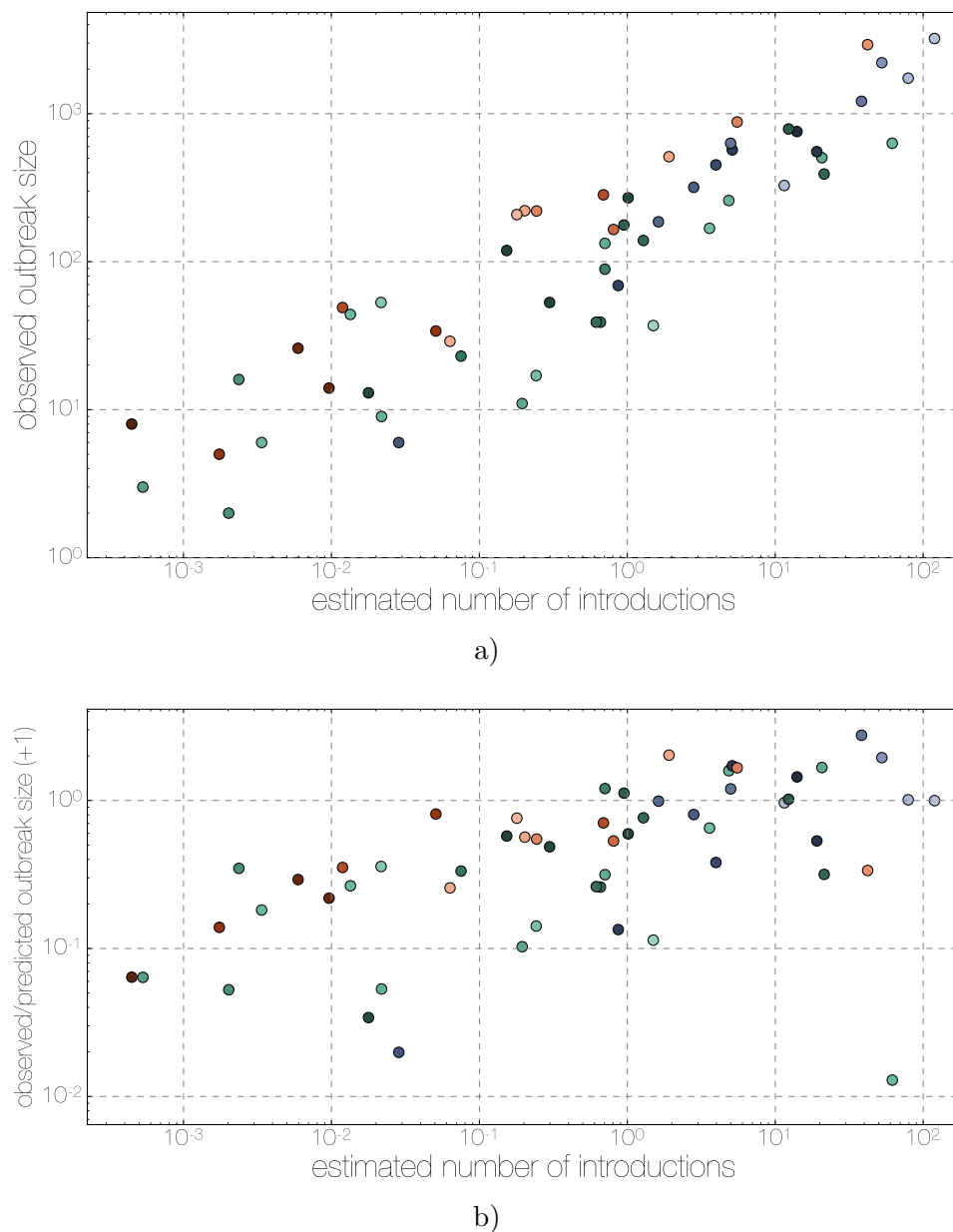


Figure S7. b) The relationship of the number of cases in each region to the phylogeographically number of introductions across the posterior distribution. The regions with the most cases have also had the most introductions (Spearman rank correlation coefficient = 0.91). a) The discrepancy between predicted and actual log cumulative case numbers (log ratio) for each region is also correlated with the estimated number of introductions. The positive relationship (Spearman rank correlation coefficient = 0.61) suggests that regions that had more cases than expected based on the demographic and climatic predictors were "seeded" more.