A Genomic Landscape of Haplotype Diversity and Signatures of Phylogeographic Distribution in Zaire Ebolavirus during the 2014 EVD Epidemic

Yue Teng, Yan Yu, Yuan Jin, Xiaoping An and
Dan Feng

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/63247

Abstract

The Ebola virus (EBOV) disease epidemic from 2013 to 2015 is the largest in history, affecting multiple countries in West Africa. Genome sequencing of EBOV has revealed extensive genetic variation and mutation rate. The evolution and the variations among genotypes of EBOV observed remain low, which suggests that the viral haplotypes may be common in this transmission. To address this hypothesis, we investigated the genomic portrait of haplotype diversity in EBOV from 1976 to the 2014 outbreaks. We obtained 176 haplotypes in 305 gene-coding sequences of EBOV and found that the Hap8 in multiple viral haplotypes is the major epidemic lineage in the 2014 Sierra Leone outbreak. The phylogeographic analysis of EBOV transmission in Sierra Leone during 2014 outbreaks indicated that the genetic flow in EBOV was no more likely to occur within or without populations and the correlation between genetic and geographical distance is not significant. Our study first detected the diversity of viral haplotypes with systematic calculation of phylogeographic distribution in EBOV. This observation highlighted how Ebola virus is substantially different in virulence or transmissibility in comparison to the virus lineages associated with 2014 outbreaks in Sierra Leone, which provides a clue to understand the 2014 EBOV spreading.

Keywords: Ebolavirus, Genome sequencing, Evolutionary, Haplotype diversity, Phylogeographic distribution

1. Introduction

The recent Zaire Ebolavirus (EBOV) epidemic (2013–2015) was the largest in history, affecting multiple countries in West Africa.

The Zaire Ebolavirus (EBOV) is an unprecedented epidemic in West Africa during 2013 to 2015. Started from December 2013 in Guinea [1], the current EBOV outbreak spread into Liberia in March, into Sierra Leone in May, into Nigeria in late July, into Senegal in late August, and into Democratic Republic of Congo in early September [2]. Ebola virus disease (EVD) has become a global concern in 2014.

Genome sequencing of EBOV samples isolated from affected individuals during this outbreak revealed extensive viral genetic variation and mutations.

As the Ebola epidemic sweeps through West Africa, the genetic data of EBOV will answer some questions regarding the evolution dynamic of EBOV. Recently, *Tong et al*. published 175 genomes of EBOV [3], collected in Sierra Leone from September to November 2014 and *Hoene et al*. reported 4 viral genomes obtained in Mali from October to November 2014 [4]. Gire *et al*. published 99 genomes from 78 patients infected in or around Kenema, Sierra Leone during May and June of 2014 [5]. Additionally, Baize *et al*. released three EBOV genomes, collected in Guinea during March 2014 [1]. Those previous studies indicated that the EBOV is evolving slowly and is not undergoing rapid evolution in humans during the current outbreak [3, 4]. However, the detailed information of the lineage of virus, the viral mutational pattern, and the dispersal route among districts are still largely unknown, which should show the signatures of transmission and is of strategic importance to find the acquiring mutations evading diagnostic tests or vaccines [6–8].

Recent studies have investigated the genomic 'portrait' of haplotype diversity in EBOV outbreaks from 1976 onwards, and the recent major outbreak presented a unique opportunity to understand genetic variation and evolution with regards to EBOV.

Currently, the patterns of EBOV genomes variation within and between hosts have shown that human-to-human transmission can involve two or more viral haplotypes. It is possible to use geographic, temporal, and epidemiological metadata work together with the transmission clustering inferred from viral genetic data. Thus, we use the coding sequences of EBOV genomes from 1976 to 2014 to investigate haplotypes of EBOV and apply phylogeographic analysis on them [1, 3–5].

The aim of this chapter is to describe the methodology for the investigation of the genomic landscape of haplotype diversity and signatures of phylogeographic distribution in Ebola virus.

This study aims to identify the major haplotype of EBOV and the snapshot of transmission with viral genetic distance during this outbreak, which may shed light on the underlying mechanisms of EBOV spreading with a clear landscape of the 2014 outbreak.

2. Methods of investigating EBOV genetic differences

The diversity of viral haplotypes can be systematically calculated using a phylogeographic distribution approach. The following techniques that are required to understand the genetics of a viral outbreak will be described.

2.1. Sequencing

We obtained the available full-length genome sequences of EBOV, resulting 305 whole genomic sequences in total at this study, from Ebolavirus Resource in NCBI (http://www.ncbi.nlm.nih.gov/genome/viruses/variation/ebola/). The analysed EBOV genomes included 19 EBOV genomes from previous outbreaks before 2014, 99 genomes from Sierra Leone from May to June 2014, 6 sequences from Guinea in March 2014, 4 sequences from Mali 2014, 1 from Liberia in 2014, 1 from United Kingdom in 2014 and 175 newly viral genomic data during 25 September to 11 November from Sierra Leone. The sequences of EBOV-coding gene (NP, VP35, VP40, GP, VP30, VP24, L) were aligned using MAFFT v7.05.

2.2. Phylogenetic tree reconstruction

Phylogenetic trees were constructed with MrBayes v3.2 under the GTR model of nucleotide substitution and gamma-distributed rates among sites for 10 million generations. We sampled every 1000 steps and the first 25% of the samples were removed as burn-in. The convergence was checked when average standard deviation of split frequency was below 0.01 and all potential scale reduction factor (PSRF values approached 1.00). We also inferred maximum likelihood phylogenies (1000 bootstrap replicates) of EBOV under GTR+gamma model using RAxML v8.1.6.

2.3. Root-to-tip distance estimation

Root-to-tip distance of the 2014 EBOV was estimated using Path-O-Gen v1.4. The maximum likelihood tree inferred in the previous step was used as the input tree file and the root was placed on the common ancestral branch of Guinea isolates. The estimated root-to-tip divergence of each sample and the corresponding isolation date were projected to the same coordinate system.

2.4. Population analyses of haplotype networks based on specific towns or geographical distances

The populations were defined based on the town and geographical distances manually. The two towns within 40 km were considered as one population. DnaSP v.5.10.01 was used to generate haplotype data files and to calculate haplotype and nucleotide diversity for each population. The NETWORK software v.4.6.1.3 served to create a haplotype network using median-joining method. Population structure was assessed by calculating pairwise FST values between populations and by AMOVA, as implemented in the software ARLEQUIN v3.5. Significance levels were obtained with 10,000 permutations. Data were tested for the presence of isolation by distance (IBD) by regressing natural logarithm-transformed geographical distances between sampling sites (in km) against Slatkin's linearized FST (FST/(1-FST)). Statistical significance was assessed using a Mantel test with 10,000 permutations in ARLEQUIN. The six-bar mutational spectra (C•G→A•T, C•G→G•C, C•G→T•A, T•A→A•T, T•A→C•G and T•A→G•C) of haplotype were calculated using in-house script (available upon request). Each haplotype was compared with the sequence of earliest sample in 2014.

2.5. Phylogeographic analysis using RASP software

Phylogenetic tree of the haplotype was estimated using RAxML v8 with the GTR + gamma evolutionary model and 500 bootstrap replicates. The ML tree from RAxML analysis was then used as a starting tree for BEAST 1.8.1. The earliest time of individuals of each haplotype was used as tip date for dating. The analysis was done using the Yule Process for the tree prior and the uncorrelated lognormal (UCLD) relaxed clock model. We performed 200 million generations, sampling every 10,000 and generating 20,000 trees. The condensed tree was generated using TreeAnnotator v1.8.1 with a burn-in of 4000 trees. For the phylogeographic analyses, we used the S-DIVA (Statistical dispersal-vicariance analysis) method implemented in RASP software to analyse the ancestral geographic ranges of EBOV lineages. In S-DIVA analysis, 1000 random trees were generated from trees data set discarded the first 4000 trees. One hundred and 10,000 alternative reconstructions were kept for random trees and final tree, respectively. The alternative reconstruction with the maximal S-DIVA value was used for further analysis. The total number of dispersal and vicariance, dispersal curve and the dispersal among and within each district were calculated based on the best reconstruction in RASP.

3. Key findings regarding genetics of the 2013-2015 EBOV outbreak

Hap8 in multiple viral haplotypes is the major epidemic lineage in the 2014 Sierra Leone outbreak. This observation highlights how Ebola virus is substantially different in virulence or transmissibility in comparison to the virus lineages associated with 2014 outbreaks in Sierra Leone, which provides a clue to understand the 2014 EBOV spreading.

We first performed viral whole genomes of EBOV alignment using 305 available genomic data from 1976 to 2014, and then we did pairwise comparison of the gene coding regions (NP, VP35, VP40, GP, VP30, VP24, L) in the 305 viral genomes, which were used for phylogenetic tree analysis by MrBayes [9]. The condensed Bayesian tree reconstructed from this data matrix showed strongly supported relationships consistent with the results from previous reports [5]. In the **Figure S1A**, the phylogenetic comparison to all 19 genomes (Yellow) during earlier outbreaks before 2014 suggests that the EBOV during 2014 EVD epidemic likely spread from the central Africa within the past decade. These EBOV from Sierra Leone during September to November 2014 (Red) follow the same patterns as observed in individual EBOV sequences during the early outbreak from March to August 2014 (Blue), which are derived from 2014 Sierra Leone 2 and 2014 Sierra Leone 3 as described in early publication [3–5]. The distribution map represented the geographical distribution of the 175 newly sequenced individual samples of EBOV during September to November 2014 with different colours at five districts (47 samples in Western Urban, 67 in Western Rural, 47 in Port Loko, 5 in Kambia, and 9 in Bombali) (**Figure S1B**), which were further used in the phylogeographic analysis. Phylogenetic tree constructed using the maximum likelihood method [10] showed a similar topology with Bayesian tree [9] (**Figures S2** and **S3**). Since the evolution and the variations among genotypes of EBOV was suggested with an observed low rate in early publications [3, 4], we investigated the genomic portrait of haplotype diversity in EBOV from 1976 to 2014 [11]. In total, 176 different viral haplotypes were identified from the 305 viral coding sequences (**Figure 1A** and **Table S1**). The haplotype frequency distribution was strongly skewed, with the vast majority of haplotypes found only once (139 out of 176) and restricted to a single viral genome. Hap144,

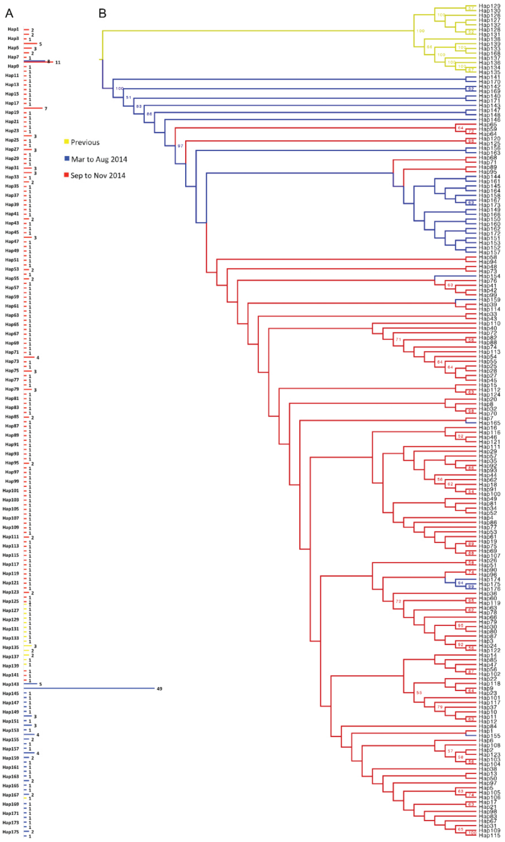

Figure 1. Multiple viral haplotypes of EBOV coexisted during the 2014 EVD outbreak. (A) The number of viral genomes in multiple viral haplotypes of EBOV (Histogram); (B) A maximum likelihood tree created with RAxML puts the 176 multiple viral haplotypes. The green values on the branch are the bootstrap values for corresponding nodes, only bootstrap values greater than 50 were shown.

which is the most common haplotype, includes 49 individuals from June 2014. The Hap8 including 19 viral sequences is the second common haplotype, which contains 8 individuals in the early outbreaks of 2014 and 11 individuals in the late outbreaks of 2014, respectively. A maximum likelihood tree of 176 haplotypes were represented in **Figure 1B**. Bootstrap values higher than 50% are shown for each node. The un-rooted ML tree revealed that there were no distinct haplotype groups with high bootstrap support.

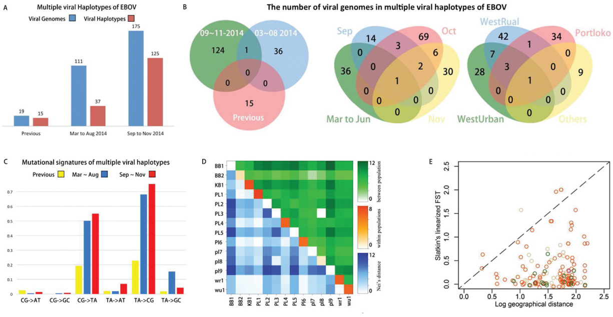

Figure 2. The mutational signatures and phylogeography analysis in multiple haplotypes of EBOV. (A) The 176 different viral haplotypes were identified in the 305 analysed genomes of EBOV; (B) The multiple viral haplotypes of EBOV with the temporal and spatial distribution (Venn); (C) The mutational signatures of multiple viral haplotypes; (D) Nei's distance (lower triangle) and Average pairwise distance (upper triangle) within and between the populations (BB, Bombali; KB, Kambia; PL, Port Loko; WR, Western Rural; WU, Western Urban.); (E) Isolation by distance plots of pairwise population values for log geographic distance (km) and genetic distance. Genetic distance is given by Slatkin's linearized FST (FST /(1- FST) Geographic distance is given in km.

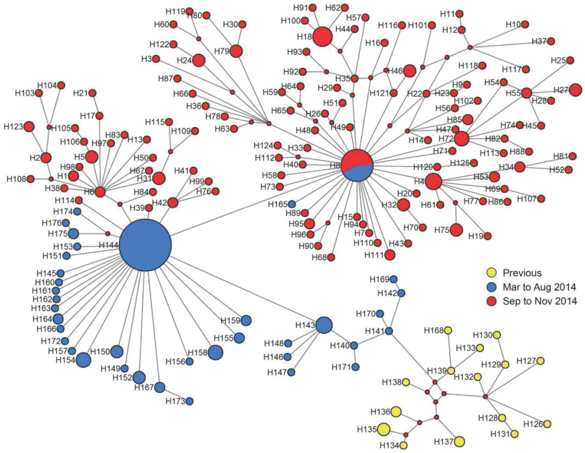

Figure 3. Median-joining haplotype network of multiple viral haplotypes in EBOV. The median-joining haplotype network covered the 176 multiple viral haplotypes of EBOV from 1976 to 2014; each circle represents a unique haplotype, and its size is proportional to its frequency.

In the **Figure 2A**, we found 15 different viral haplotypes in the 19 viral coding sequences in the previous outbreaks before 2014. In the 111 viral coding sequences during the earlier outbreak of 2014 (March to August, 2014), 37 different viral haplotypes were detected. Particularly, 125 different viral haplotypes were identified from the 175 viral coding sequences during September to November 2014. Interestingly, the Hap8 is the only one viral haplotype with the viral sequences that covers both early and late 2014 outbreaks in Sierra Leone (**Figure 2B**, the left Venn picture), which means that the Hap8 differentiated from the progenitor virus from June 2014 and maintained renewal to keep the population. Within the 11 individuals of Hap 8 from September to November 2014 (**Figure 2B**, the middle Venn picture), we found that 3 viral genomes from Western Urban, 2 viral genomes from Port Loko, 5 viral genomes from Bombali and one viral genome from Western Rural (**Figure 2B**, the right Venn picture), while all individuals of Hap144 are only represented in early outbreaks of 2014. This result implies that the Hap8 may be the major viral haplotype during 2014 Sierra Leone EVD epidemic.

3.1. Mutation types: EBOV mutations are typically C•G→T•A and T•A→C•G

The analyses of nucleotide substitutions as six-bar mutational spectra (C•G→A•T, C•G→G•C, C•G→T•A, T•A→A•T, T•A→C•G and T•A→G•C) have been proven useful in showing how mutational spectra can be specific to viral type and related to viral coding sequences. The analysis based on viral haplotypes has shown significant mutational signatures among their sequences (**Figure 2C**, **Figure S4** and **Table S2**). The results in **Figure 2C** indicated that the mutations of Ebola viruses are mainly C•G→T•A and T•A→C•G, and the nucleotide mutation rate during 2014 outbreak was much higher than the rate in previous.

3.2. Genetic characteristics are not correlated with geographical distance

We are interested in the relationship between the genetic distance in viral multiple haplotypes and the geographic distance (**Figure S1B**). In **Figure 2D**, we compared Nei's distance [12] (lower triangle) and Average pairwise distance [13] (upper triangle) within and between the populations (**Table S3**) during September to November 2014. The dark parts of lower triangle were mainly represented from PL2 to PL9, which suggested that the populations in Port Loko have higher genetic divergence than others. Population KB1, PL1, PL4, PL6, WR1 and WU1 represented high genetic distance within each of them, which suggested that these populations have higher genetic divergence within them. The isolation by distance (IBD) analyses [14] detected no positive correlation between genetic distance (Slatkin's linearized FST) and log geographical distance (**Figure 2E**) [15], and the *p* value of analyses mantel test is 0.459. This finding was consistent when we restrained the analysis within Port Loko, which included most populations. We also compared the pairwise distances from the sequences of 126 different viral haplotypes from September to November 2014 (**Figure S5**). Most of the haplotypes represent a low pairwise distances (less than 15) except Hap109 and Hap115. Both of the two haplotypes were represented in the tip of the haplotypes tree in **Figure 1B**, which indicated that they divided recently. The date of sampling for the two haplotypes was November 9th 2014, which support their position in evolutionary dynamic [3].

3.3. The evolution and variation between EBOV genotypes was low, which suggest that the viral haplotypes may be common in this transmission

In order to infer the connections of multiple viral haplotypes, we reconstructed a median joining haplotype network [16], which covered the 176 multiple viral haplotypes of EBOV (**Figure 3** and **Table S4**). The star-like network was characterized by a few common viral haplotypes, surrounded by many viral haplotypes mostly present in only 1–5 individuals (except Hap18). Most viral haplotype were divided between two haplotypes at opposite ends of the network. Thorough spatial mixing was evident, with the central haplotype (Hap8) being shared by 19 different viral coding sequences from March to November 2014 representing the major haplotype during the outbreak 2014.

In the **Figure 4A**, the haplotype network showed the main haplotypes, which contain more than 3 individuals in each. The relationship between the haplotype network and four subclades in 2014 outbreaks was identified in the phylogenetic tree, which was reconstructed

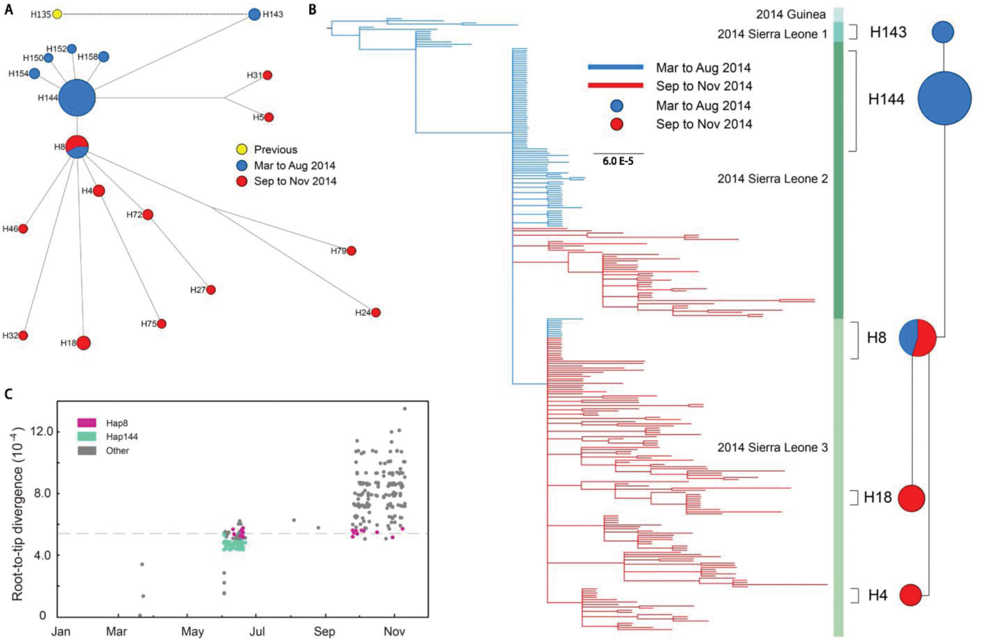

Figure 4. Main viral haplotypes in median joining haplotype network and phylogenetic analysis with outbreak in 2014. (A) Main haplotypes and those containing more than three viral genomes. Lines represent base pair changes between two haplotypes, with the length proportional to the number of base pair changes. Haplotypes and those containing island samples are labelled by name. (B) A Bayesian phylogenetic tree of EBOV lineages created by MrBayes with the main outbreak of viral haplotypes in 2014 on the left. (C) Root-to-tip divergence of each sample plotting with corresponding collection date. The estimates were based on ML tree and the root was placed on the common ancestral branch of Guinea strains. Strains of Hap 8, Hap 144 and other haplotypes were coloured magenta, cyan and grey.

by 286 viral sequences in 2014 outbreak (**Figure 4B**). The viral genomes during September to November 2014 mainly belong to Sierra Leone 3, despite some located in the Sierra Leone 2. The circles and lines alongside the tree tips indicate the corresponding common haplotypes and their relationships in the median joining haplotype network. Importantly, combining this network, we found that the 49 individuals of Hap144 from June 2014 were only located in the sub-lineage, Sierra Leone 2, while all 19 individuals of Hap8 from June to November 2014 were located in the sub-clade Sierra Leone 3. We did root-to-tip analysis by Maximum Likelihood tree rooted on the common ancestor of Guinea branches. The result of root-to-tip distances suggests that the viral genomes during September to November 2014 displayed high divergence and there appeared to be an increase in the viral diversity in **Figure 4C**. Despite the high-diversified viral sequences, the Hap8 still kept its stable rate and circulated nearly half the year. Thus, the Hap8 maintained a stable state from June to November 2014, which supported our hypothesis that the Hap8 is the major lineage in 2014 Sierra Leone EVD outbreak.

3.4. Analyses showed that the Western rural region was the dispersal and differentiation centre for EBOV in Sierra Leone

Based on the phylogenetic relationship of viral haplotypes, we analysed the phylogeographic distribution [17, 18] of the EBOV during September to November 2014 in Sierra Leone with their detailed geographic information. The viral haplotypes tree with optimal reconstruction of highest S-DIVA Value is shown in **Figure 4A** [19, 20]. Pie charts reflect probability of the respective area and areas are colour-coded as well as the legends allowing for single or combined distributions. The reconstruction reveals at least 63 dispersals and 46 variances to explain the present distribution pattern of the EBOV. The results indicated that the ancestor of the outbreak from September to November 2014 had originated in the Western Rural and Western Urban (75% support, node 249). The tree then divided into two major clades, one originated in Western Rural (node 248) and the other originated in Western Urban (node 153). The red highlight clade in **Figure 5A** represented the dispersal routes of Hap 8, which starts from Western Rural, and then disperses to Port Loko and other districts. The dispersal routes of all samples in Hap 8 are represented in **Figure S6**, which support dispersal routes of Hap 8. The diagram of dispersal events (blue line) and node density (green line) is represented in **Figure 5B**. The total number of immigration, emigration and the divergence of haplotype of each province were represented in **Figure 5C**, which revealed that the emigration and divergence in Western Rural were much higher than other place, suggesting that Western Rural is likely to have been the source of the EBOV, which was in accordance with the epidemiological findings. The circle in **Figure 5D** showed the number of dispersals among Bombali, Kambia, Port Loko, Western Rural and Western Urban. It is interesting that The EBOV in Kambia came from Western Urban and Western Rural directly and there is no migration between Kambia and Port Loko or Kambia and Bombali. The green points in geographical map (**Figure 5D**) represented the distribution of Hap 8, which had a wide distribution in Bombali, Port Loko, Western Urban and Western Rual.

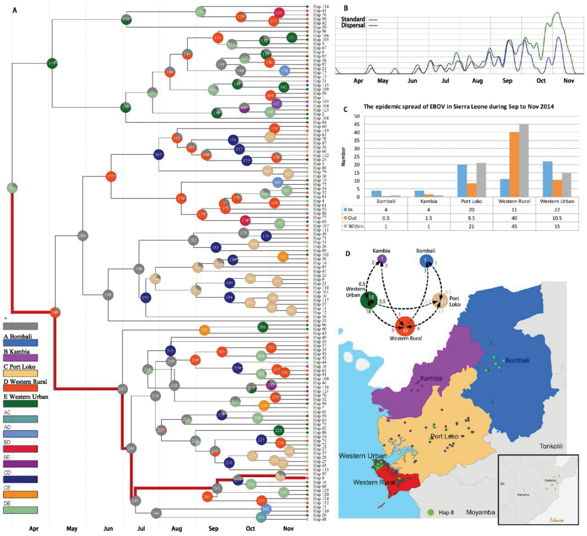

Figure 5. Phylogeographic analysis of the epidemic spread of Ebola virus in Sierra Leone. (A) Results of the biogeographic analysis of haplotypes using S-DIVA method. Pie charts show the probability values of the ancestral areas reconstructed at each node. Areas are colour-coded as in the legends allowing for single or combined distributions. (B) The diagram of dispersal events (blue line) and node density (green line) (C) The total number of immigration (in), emigration (out) and the divergence (within) of haplotype of each districts (D) The circle represented the number of dispersals among Bombali, Kambia, Port Loko, Western Rural and Western Urban. Map to show the geographical distribution of the districts of EBOV from September to November 2014. The distribution of Hap8 was highlighted with green points. We obtained the boundary data of the District from the GADM database of Global Administrative Areas version 2.0 (http://www.gadm.org), and the map was created in ArcGIS 9.2 software (ESRI Inc., Redlands, CA, USA).

4. Conclusions

Previously characterized EBOV strains were found to be substantially different in terms of virulence and transmissibility compared to the 2014 Sierra Leona lineage.

The EBOV epidemic from 2013 is having a devastating impact in West Africa. Collectively, in the previous study, they found the mutation sites of the sequenced EBOV genomes during March to November 2014 [1, 3–5], which generated genetically distinct sub-clades in Sierra Leone outbreak, following the emergence of multiple novel lineages of EBOV. Based on them, the sequencing genomes of EBOV has revealed extensive genetic variation [3,21], leading to speculation that the viral genomes have multiple viral haplotype in EBOV with the signatures of phylogeographic distribution during the 2014 EVD epidemic in Sierra Leone. It is implied that a genetically diverse multiple viral haplotypes of EBOV governs the 2014 EVD epidemic.

The lack of functional distinction between the previous and 2014 Ebola outbreaks emphasizes the importance of the diversity multiple viral haplotypes [5, 22]. Here we analysed the multiple viral haplotypes in the coding sequences of EBOV from 1976 to 2014 outbreaks. In the 176 multiple viral haplotypes during the 2014 outbreak (**Figure 1**), although the Hap144 is the most common haplotype, the Hap8 is the major viral haplotype with the long temporal and extensive spatial distribution during 2014 EVD outbreaks in Sierra Leone (**Figure 2** and **4**). The reconstructed median join network of multiple viral haplotypes displayed a "star-like" shape with no deep branching among haplotypes (**Figure 3**), which shows the similarity topology in the each outbreak with the most common haplotypes in the centre. It demonstrates the high degree of similarity of the progenitor virus of each outbreak, which suggests that future transmissions of similarly virulent potential are highly likely [3, 5, 23].

We also found evidence that the multiple viral haplotypes in EBOV have the significant mutational signatures (**Figure 2C** and **S4**). Interestingly, the mutations associated with smoking-related damage in lung cancers are mainly C•G→A•T transversions [24], whereas mutations associated with ultraviolet (UV) radiation exposure in skin cancers comprise predominantly C•G→T•A transitions [25]. Strategies for containing EBOV in West Africa have been suggested, but are predicated on lack of adaptation of the virus. These results indicate that the mutational signatures in multiple viral haplotypes of EBOV may be cause by the chemomorphosis and the environment of ultraviolet (UV) radiation.

Viral epidemics can develop strong growth heterogeneity even though the temporal and spatial scales of its initial outbreak are short.

In the 2014 Ebola epidemic, we have identified a genetic variant that has a substantially higher growth rate than its progenitor lineage [4, 21, 26]. We conclude that a viral epidemic can develop strong growth heterogeneity even on the limited temporal and spatial scales of its initial outbreak. If that heterogeneity has a genetic cause, our analysis suggests that selection can shape a fast-evolving pathogen on the time scales of a single epidemic. However, the viral genomes in the Hap8 had a stable growth rate. If this growth heterogeneity remains stable, it will generate major shifts in multiple viral haplotypes frequencies and influence the overall epidemic dynamics on time scales within the current outbreak. Thus, the long temporal distribution and stable evolution of the Hap8 indicated that it replaced Hap144 and further became the major epidemic haplotype in the year.

The difference between the 2014 outbreak and those that have occurred previously is the establishment of infections in relatively densely populated areas compared with previous outbreaks [27]. The genetic flow in EBOV was no more likely to occur between populations than among populations (**Figure 2D**). Overall, IBD analyses showed that there are no detected positive correlation between genetic and log geographical distance (**Figure 2E**). Our findings suggest that population growth, urbanization and immigration along the main road in Sierra Leone have created efficient pathways for EBOV transmission. The phylogeographic analyses showed the ancestor distribution of the outbreak from March to November 2014. Western Rural is the dispersal centre and differentiation centre of the Ebola virus, which not only spread virus to all other districts, but also generated the most haplotype (**Figure 4C**). In **Figure 4B**, the peaks of dispersal curve are (blue) below the node curve (red) most of the time except in the beginning of July and the middle of September 2014. This result indicated that there may have been small-scale outbreaks during these periods.

Further analysis of these differences may help to explain how the 2013-2015 outbreak spread so rapidly and widely. Genetic data will be able to yield insights into the evolutionary dynamics of EBOV.

In summary, our study first detected the diversity of multiple viral haplotypes in EBOV with systematic calculation of phylogeographic distribution. We found that the haplotype, Hap8, is the major epidemic lineage in the 2014 Sierra Leone outbreak. The mutations of Ebola viruses are mainly C•G→T•A and T•A→C•G, and the nucleotide mutation rate during 2014 outbreak was much higher than the rate in previous outbreaks. Moreover, the continuously increasing genetic diversity of the 2014 EBOV were also found in our result. The genetic flow in EBOV was no more likely to occur within or without populations and the correlation between genetic and log geographical distance is not significant. However, Western Rural is the dispersal centre and differentiation centre of the Ebola virus in Sierra Leone. Our method is based on simple summary statistics of multiple viral haplotypes, which can be inferred from genealogical trees with an underlying lineage-specific model of the infection dynamics. However, all analysis of haplotype diversity and phylogeographic distribution starting from the initial phase of an epidemic are probabilistic extrapolations; they are based on limited data and subject to confounding factors such as variation in sampling density [28]. As more sequence data emerge, updated haplotype diversity and phylogeographic distribution will suggest targets for detailed epidemiological investigation and provide predictive insight into the dynamics of the epidemic.

5. Limitations

Analyses of haplotype diversity and phylogeographic distribution are probabilistic extrapolations based on limited data as a result of the intrinsic difficulty in collecting samples.

A range of confounding factors is also implicit, such as variations in sampling density.

Acknowledgements

This work was supported by a grant from the State Key Laboratory of Pathogen and BioSecurity Program (No. SKLPBS1408 and No. SKLPBS1451).

Authors' contributions

Y.T., Y.Y., and Y.J. characterized the materials, under the supervision of Y.T., Y.Y., and D.F. wrote the manuscript with further contributions from J.Y. and X.P.A. analysed the data. All authors reviewed the manuscript.

Additional information

Competing financial interests and the authors declare no competing financial interests.

Author details

Yue Teng[1*], Yan Yu[2,3], Yuan Jin[1], Xiaoping An[1] and Dan Feng[4*]

*Address all correspondence to: yueteng@me.com and fddd@263.net

1 State Key Laboratory of Pathogen and Biosecurity, Beijing, China

2 Key Laboratory of Bio-Resources and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Sichuan Province, China

3 Department of Biology, Duke University, Durham, USA

4 Division of Standard Operational Management, Institute of Hospital Management, Chinese PLA General Hospital, Beijing, China

References

[1] Baize S., *et al*. Emergence of Zaire Ebola Virus Disease in Guinea. *N Engl J Med*. 371, 1418–1425 (2014)

[2] World Health Organization. One year into the Ebola epidemic: a deadly, tenacious and unforgiving virus. http://www.who.int/csr/disease/ebola/one-year-report/introduction/en/

[3] Tong Y.G., *et al*. Genetic Diversity and Evolutionary Dynamics of Ebola Virus in Sierra Leone. *Nature*. (Accepted, doi:10.1038/nature14490, 2015)

[4] Hoenen T., et al. Mutation rate and genotype variation of Ebola virus from Mali case sequences. *Science*. 348, 117–119 (2015)

[5] Gire S.K., *et al*. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*. 345, 1369–1372 (2014)

[6] Agnandji S.T., et al. Phase 1 Trials of rVSV Ebola Vaccine in Africa and Europe – Preliminary Report. *N Engl J Med*. 2015 Apr 1. [Epub ahead of print]. DOI: 10.1056/NEJMoa1414216

[7] Regules J. A., et al. A Recombinant Vesicular Stomatitis Virus Ebola Vaccine - Preliminary Report. *N Engl J Med*. 2015 Apr 1. [Epub ahead of print]. DOI: 10.1056/NEJMoa1414216

[8] Marzi A., et al. An Ebola whole-virus vaccine is protective in nonhuman primates. *Science*. 2015 Mar 26. [Epub ahead of print]. DOI: 10.1126/science.aaa4919

[9] Ronquist F., Huelsenbeck J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*. 19, 1572–1574 (2003)

[10] Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 30, 1312–1313 (2014)

[11] Kortenhoeven C. Virus genome dynamics under different propagation pressures: reconstruction of whole genome haplotypes of west Nile viruses from NGS data. *BMC Genomics*. 16, 118 (2015)

[12] Nei M. Molecular Evolutionary Genetics. Columbia University Press, New York, NY, USA. (1987)

[13] Tajima F. Evolutionary relationship of DNA sequences in finite populations. *Genetics*. 105, 437–460 (1983)

[14] Peel A. J., *et al*. Continent-wide panmixia of an African fruit bat facilitates transmission of potentially zoonotic viruses. *Nature communications*. 4, 2770 (2013)

[15] Slatkin M. A measure of population subdivision based on microsatellite allele frequencies. *Genetics*. 139,457–462 (1995)

[16] Bandelt H. J., Forster P., Röhl A. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol*. 16, 37–48 (1999)

[17] Excoffier L., Lischer H. E. L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour*. 10, 564–567 (2010).

[18] Rozas J., et al. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*. 19, 2496–2497 (2003)

[19] Yu Y., et al. RASP (Reconstruct Ancestral State in Phylogenies): a tool for historical biogeography. *Mol Phylogenet Evol*. 87, 46–49 (2015)

[20] Yu Y., Harris A. J., He X. S-DIVA (Statistical Dispersal-Vicariance Analysis): A tool for inferring biogeographic histories. *Mol Phylogenet Evol*. 56, 848–850 (2010)

[21] Vogel G. Infectious Diseases. A reassuring snapshot of Ebola. *Science*. 347, 1407 (2015)

[22] Scarpino S. V. Epidemiological and viral genomic sequence analysis of the 2014 Ebola outbreak reveals clustered transmission. *Clin Infect Dis*. 60, 1079–1082 (2015)

[23] Dowall S. D., et al. Elucidating variations in the nucleotide sequence of Ebola virus associated with increasing pathogenicity. *Genome Biol*. 15, 540 (2014)

[24] Helleday T. Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet*. 15, 585–598 (2014)

[25] Pfeifer G. P., You Y. H., Besaratinia A. Mutations induced by ultraviolet light. *Mutat Res*. 571, 19–31 (2005).

[26] Alizon S. Quantifying the epidemic spread of Ebola virus (EBOV) in Sierra Leone using phylodynamics. *Virulence*. 5, 825–827 (2014)

[27] Goeijenbier M. Ebola virus disease: a review on epidemiology, symptoms, treatment and pathogenesis. *Neth J Med*. 72, 442–448 (2014)

[28] Kugelman J. R. Evaluation of the potential impact of Ebola virus genomic drift on the efficacy of sequence-based candidate therapeutics. *MBio*. 6 (2015)