Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak

Title: Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak

Authors: Stephen K. Gire[1,2†], Augustine Goba[3†*], Kristian G. Andersen[1,2†*], Rachel S. G. Sealfon[2,4†], Daniel J. Park[2†], Lansana Kanneh[3], Simbirie Jalloh[3], Mambu Momoh[3,5], Mohamed Fullah[3,5§], Gytis Dudas[6], Shirlee Wohl[1,2,7], Lina M. Moses[8], Nathan L. Yozwiak[1,2], Sarah Winnicki[1,2], Christian B. Matranga[2], Christine M. Malboeuf[2], James Qu[2], Adrianne D. Gladden[2], Stephen F. Schaffner[1,2], Xiao Yang[2], Pan-Pan Jiang[1,2], Mahan Nekoui[1,2], Andres Colubri[1], Moinya Ruth Coomber[3], Mbalu Fonnie[3§], Alex Moigboi[3§], Michael Gbakie[3], Fatima K. Kamara[3], Veronica Tucker[3], Edwin Konuwa[3], Sidiki Saffa[3], Josephine Sellu[3], Abdul Azziz Jalloh[3], Alice Kovoma[3], James Koninga[3], Ibrahim Mustapha[3], Kandeh Kargbo[3], Momoh Foday[3], Mohamed Yillah[3], Franklyn Kanneh[3], Willie Robert[3], James L. B. Massally[3], Sinéad B. Chapman[2], James Bochicchio[2], Cheryl Murphy[2], Chad Nusbaum[2], Sarah Young[2], Bruce W. Birren[2], Donald S. Grant[3], John S. Scheiffelin[8], Eric S. Lander[2,7,9], Christian Happi[10], Sahr M. Gevao[11], Andreas Gnirke[2‡], Andrew Rambaut[6,12,13‡], Robert F. Garry[8‡], S. Humarr Khan[3§‡], Pardis C. Sabeti[1,2‡*]

Affiliations:

1. Harvard University, Center for Systems Biology, Department of Organismic and Evolutionary Biology, Cambridge, MA, 02138
2. The Broad Institute of MIT and Harvard, Cambridge, MA 02142
3. Kenema Government Hospital, Kenema, Sierra Leone
4. Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, 02139
5. Eastern Polytechnic College, Kenema, Sierra Leone
6. Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK
7. Harvard Medical School, Systems Biology, Boston, MA, 02115
8. Tulane University Medical Center, New Orleans, LA
9. Massachusetts Institute of Technology, Department of Biology, Cambridge, MA, 02139
10. Redeemer's University, Ogun State, Nigeria
11. University of Sierra Leone, Freetown, Sierra Leone
12. Fogarty International Center, National Institutes of Health, Bethesda, MD 20892
13. Centre for Immunity, Infection and Evolution, University of Edinburgh, Edinburgh, UK

† Authors contributed equally to this work.

‡ Authors jointly supervised this work.

§ Deceased.

* Corresponding author.

Abstract:
In its largest outbreak, Ebola virus disease is spreading through Guinea, Liberia, Sierra Leone, and Nigeria. We sequenced 99 Ebola virus genomes from 78 patients in Sierra Leone to ~2,000x coverage. We observed a rapid accumulation of interhost and intrahost genetic variation, allowing us to characterize patterns of viral transmission over the initial weeks of the epidemic. This West African variant likely diverged from Middle African lineages ~2004, crossed from Guinea to Sierra Leone in May 2014, and has exhibited sustained human-to-human transmission subsequently, with no evidence of additional zoonotic sources. Since many of the mutations alter protein sequences and other biologically meaningful targets, they should be monitored for impact on diagnostics, vaccines, and therapies critical to outbreak response.

Main Text:
Ebola virus (EBOV; formerly Zaire ebolavirus), one of five ebolaviruses, is a lethal human pathogen, causing Ebola virus disease (EVD) with an average case fatality rate of 78% (*1*). Previous EVD outbreaks were confined to remote regions of Middle Africa; the largest, in 1976, had 318 cases (*2*) (Fig. 1A). The current outbreak started in February 2014 in Guinea, West Africa (*3*) and spread into Liberia in March, Sierra Leone in May, and Nigeria in late July. It is the largest known EVD outbreak and is expanding exponentially with a doubling period of 34.8 days (Fig. 1B). As of August 19th, 2,240 cases and 1,229 deaths have been documented (*4, 5*). Its emergence in the major cities of Conakry (Guinea), Freetown (Sierra Leone), Monrovia (Liberia), and Lagos (Nigeria) raises the specter of increasing local and international dissemination.

In an ongoing public health crisis, where accurate and timely information is crucial, new genomic technologies can provide near real-time insights into the pathogen's origin, transmission dynamics, and evolution. We used massively parallel viral sequencing to understand how and when EBOV entered human populations in the 2014 West African outbreak, whether the outbreak is continuing to be fed by new transmissions from its natural reservoir, and how the virus changed, both before and after its recent jump to humans.

In March 2014, Kenema Government Hospital (KGH) established EBOV surveillance in Kenema, Sierra Leone, near the origin of the 2014 outbreak (Figs. 1C, S1) (*6*). Following standards for field-based tests in previous (*7*) and current (*3*) outbreaks, KGH performed conventional PCR-based EBOV diagnostics (*8*) (Fig. S2); all tests were negative through early May. On May 25, KGH scientists confirmed the first case of EVD in Sierra Leone. Investigation by the Ministry of Health and Sanitation (MoHS) uncovered an epidemiological link between this case and the burial of a traditional healer who had treated EVD patients in Guinea. Tracing led to 13 additional cases—all females who attended the burial. We obtained ethical approval from MoHS, the Sierra Leone Ethics and Scientific Review Committee, and our U.S. institutions to sequence patient samples in the U.S. using approved safety standards (*6*).

We evaluated four independent library preparation methods and two sequencing platforms (*9*) (table S1) for our first batch of 15 inactivated EVD samples from 12 patients. Nextera library construction and Illumina sequencing provided the most complete genome assembly and reliable intrahost single nucleotide variant (iSNV, frequency >0.5%) identification (*6*). We used this combination for a second batch of 84 samples from 66 additional patients, performing two independent replicates from each sample (Fig. 1D). We also sequenced 35 samples from

suspected EVD cases that tested negative for EBOV; genomic analysis identified other known pathogens, including Lassa virus, HIV-1, enterovirus A and malaria parasites (Fig. S3).

In total, we generated 99 EBOV genome sequences from 78 confirmed EVD patients, representing over 70% of the EVD patients diagnosed in Sierra Leone in late May to mid June; we employed multiple extraction methods or timepoints for 13 patients (table S2). Median coverage was >2,000x, spanning more than 99.9% of EBOV coding regions (Fig. 1D,E, table S2).

We combined the 78 Sierra Leonean sequences with 3 published Guinean samples (*3*) (correcting 21 likely sequencing errors in the latter (*6*)) to obtain a dataset of 81 sequences. They reveal 341 fixed substitutions between the 2014 EBOV and all previously published EBOV (35 nonsynonymous, 173 synonymous, 133 noncoding), with an additional 55 single nucleotide polymorphisms (SNPs) (fixed within individual patients) within the West African outbreak (15 nonsynonymous, 25 synonymous, 15 noncoding). Notably, the Sierra Leonean genomes differ from PCR probes for five separate assays used for EBOV and pan-filovirus diagnostics (table S3).

Deep-sequence coverage allowed identification of 263 iSNVs (73 nonsynonymous, 108 synonymous, 70 noncoding, and 12 frameshift) in the Sierra Leone patients (*6*). For all patients with multiple time points, consensus sequences were identical and iSNV frequencies remained stable (Fig. S4). One notable intrahost variation is the RNA editing site of the glycoprotein (GP) gene (Fig. S5A) (*10-12*), which we characterize in patients (*6*).

Phylogenetic comparison to all 20 genomes from earlier outbreaks suggests the 2014 West African virus likely spread from Middle Africa within the last decade. Rooting the phylogeny using divergence to other ebolavirus genomes is problematic (Figs. 2A, S6) (*6, 13*). However, rooting the tree on the oldest outbreak reveals a strong correlation between sample date and root-to-tip distance, with a substitution rate of 8×10^{-4}/site/year (Figs. 2B, S7) (*13*). This suggests that the lineages of the three most recent outbreaks all diverged from a common ancestor at roughly the same time c. 2004 (Fig. 2C, 3A), supporting the hypothesis that each outbreak represents an independent zoonotic event from the same genetically diverse viral population in its natural reservoir.

Genetic similarity across the sequenced 2014 samples suggests a single transmission from the natural reservoir, followed by human-to-human transmission during the outbreak. Molecular dating places the common ancestor of all sequenced Guinea and Sierra Leone lineages around late February 2014 (Fig. 3B), three months after the earliest suspected cases in Guinea (*3*); this coalescence would be unlikely had there been multiple transmissions from the natural reservoir. Thus, in contrast to some previous EVD outbreaks (*14*), continued human-reservoir exposure is unlikely to have contributed to the growth of this epidemic in areas represented by available sequence data.

Our data suggest the Sierra Leone outbreak stemmed from the introduction of two genetically distinct viruses from Guinea around the same time. Samples from 12 of the first EVD patients in Sierra Leone, all believed to have attended the funeral of an EVD case from Guinea, fall into two distinct clusters (clusters 1 and 2) (Figs. 4A, S8). Molecular dating places the divergence of these

two lineages in late April (Fig. 3B), pre-dating their co-appearance in Sierra Leone in late May (Fig. 4B), suggesting the funeral attendees were most likely infected by two lineages then circulating in Guinea, possibly at the funeral (Fig. S9). All subsequent diversity in Sierra Leone accumulated on the background of those two lineages (Fig. 4A), consistent with epidemiological information from tracing contacts.

Patterns in observed intrahost and interhost variation provide important insights about transmission and epidemiology. Groups of patients with identical viruses or with shared intrahost variation show temporal patterns suggesting transmission links (fig. S10). One iSNV (position 10,218) shared by twelve patients is later observed as fixed within 38 patients, becoming the majority allele in the population (Fig. 4C) and defining a third Sierra Leone cluster (Figs. 4A, 4D, S8). Repeated propagation at intermediate frequency suggests that transmission of multiple viral haplotypes may be common. Geographic, temporal, and epidemiological metadata supports the transmission clustering inferred from genetic data (Figs. 4D, 4E, S11) (*6*).

The observed substitution rate is roughly twice as high within the 2014 outbreak as between outbreaks (Fig. 4F). Mutations are also more frequently nonsynonymous during the outbreak (Fig. 4G). Similar findings have been seen previously (*15*) and are consistent with expectations from incomplete purifying selection (*16*-*18*). Determining whether individual mutations are deleterious, or even adaptive, would require functional analysis; however, the rate of nonsynonymous mutations suggests that continued progression of this epidemic could afford an opportunity for viral adaptation (Fig. 4H), underscoring the need for rapid containment.

As in every EVD outbreak, the 2014 EBOV variant carries a number of genetic changes distinct to this lineage; our data do not address whether these differences are related to the severity of the outbreak. However, the catalog of 395 mutations, including 50 fixed nonsynonymous changes with 8 at positions with high levels of conservation across ebolaviruses, provide a starting point for such studies (table S4).

To aid in relief efforts and facilitate rapid global research, we immediately released all sequence data as generated. Ongoing epidemiological and genomic surveillance is imperative to identify viral determinants of transmission dynamics, monitor viral changes and adaptation, ensure accurate diagnosis, guide research on therapeutic targets, and refine public-health strategies. It is our hope that this work will aid the multidisciplinary, international efforts to understand and contain this expanding epidemic.

References and Notes:

1. J. Kuhn *et al.*, *Biosec Bioterr* **9**, 361-371 (2011).
2. J. Burke, 1976. *Bull. World Health Org* **56**, 271-293 (1978).
3. S. Baize *et al.*, *New Eng J Med* (2014).
4. WHO (2014), http://www.who.int/csr/don/archive/disease/ebola/en/
5. O. Reynard, V. Volchkov, C. Peyrefitte, *Med Sci* **30**, 671-673 (2014).
6. Materials and methods are available as supplementary material on *Science* Online.
7. J. Towner, T. Sealy, T. Ksiazek, S. Nichol, *J Inf Dis* **196 Suppl 2**, S205-212 (2007).
8. M. Panning *et al.*, *J Inf Dis* **196 Suppl 2**, S199-204 (2007).
9. C. Malboeuf *et al.*, *Nuc Acids Res* **41**, e13 (2013).
10. A. Sanchez, B. Mahy, C. Peters, *Proc Nat Acad Sci*, (1996).
11. V. Volchkov *et al.*, *Virology* **214**, 421-430 (1995).
12. V. Volchkova, O. Dolnik, M. Martinez, O. Reynard, V. Volchkov, *J Inf Dis* **204 Suppl 3**, S941-946 (2011).
13. G. Dudas, A. Rambaut, *PLoS Curr* **6**, (2014).
14. J. Kuhn, *Arch Vir Supp* **20**, 13-360 (2008).
15. M. Schreiber *et al.*, *J Virol* **83**, 4163-4173 (2009).
16. J. Wertheim, S. Kosakovsky *Mol Bio Evol* **28**, 3355-3365 (2011).
17. S. Ho, M. Phillips, A. Cooper, A. Drummond, *Mol Bio Evol* **22**, 1561-1568 (2005).
18. E. Holmes, *J Virol* **77**, 11296-11298 (2003).
19. J. Kugelman *et al.*, *PloS One* **7**, e50316 (2012).
20. S. Gunther *et al.*, *Antivir Res* **63**, 209-215 (2004).
21. G. Grard *et al.*, *J Inf Dis* **204 Suppl 3**, S776-784 (2011).
22. G. Kobinger *et al.*, *J Inf Dis* **204**, 200-208 (2011).
23. T. Hoenen, S. Jung, A. Herwig, A. Groseth, S. Becker, *Virol* **403**, 56-66 (2010).
24. J. Blow, C. Mores, J. Dyer, D. Dohm, *J Virol Methods* **150**, 41-44 (2008).
25. A. Trombley *et al.*, *Am J Trop Med Hyg* **82**, 954-960 (2010).
26. J. Morlan, K. Qu, D. Sinicropi, *PloS One* **7**, e42882 (2012).
27. X. Adiconis *et al.*, *Nat Methods* **10**, 623-629 (2013).
28. L. Jiang *et al.*, *Genome Res* **21**, 1543-1551 (2011).
29. R. Edgar, *Nuc Acids Res* **32**, 1792-1797 (2004).
30. P. Cingolani *et al.*, *Fly* **6**, 80-92 (2012).
31. A. Stamatakis, T. Ludwig, H. Meier, *Bioinf* **21**, 456-463 (2005).
32. F. Ronquist, J. Huelsenbeck, *Bioinf* **19**, 1572-1574 (2003).
33. A. Drummond, M. Suchard, D. Xie, A. Rambaut, *Mol Bio Evol* **29**, 1969-1973 (2012).
34. M. Hasegawa, H. Kishino, T. Yano, *J Mol Evol* **22**, 160-174 (1985).
35. Z. Yang, *J Mol Evol* **39**, 306-314 (1994).
36. M. Gill *et al.*, *Mol Bio Evol* **30**, 713-724 (2013).
37. A. Drummond, S. Ho, M. Phillips, A. Rambaut, *PLoS Bio* **4**, e88 (2006).
38. G. Baele, P. Lemey, S. Vansteelandt, *BMC Bioinf* **14**, 85 (2013).
39. M. Ferreira *et al.*, *Nat Genet* **40**, 1056-1058 (2008).
40. M. Mehedi *et al.*, *J Virol* **85**, 5406-5414 (2011).
41. T. Gibb, D. Norwood, N. Woollen, E. Henchal, *J Clin Microbio* **39**, 4125-4130 (2001).
42. J. Morvan *et al.*, *Microbes Inf* **1**, 1193-1201 (1999).
43. A. Sanchez *et al.*, *J Inf Dis* **179 Suppl 1**, S164-169 (1999).
44. M. Weidmann, E. Muhlberger, F. Hufert, *J Clin Virol* **30**, 94-99 (2004).

Acknowledgments: We thank the Sierra Leone MoHS (Hon. Minister M. Kargbo, B. Kargbo, M.A. Vandi, A. Jambai), the Kenema District Health Management Team and Lassa fever program for their efforts in outbreak response. We thank P. Cingolani, Y.-C. Wu, M. Lipsitch, S. Günther, S. Baize, N. Wauquier, J. Bangura, V. Lungay, L. Hensley, J. Johnson, M. Voorhees, A. O'Hearn, and R. Schoepp, L. Gaffney, J. Kuhn, S.C. Sealfon, J.B. Shapiro, C. Edwards, Sabeti lab members for technical support and feedback. This project has been funded in part by NIH 1DP2OD006514-01 and NIAID HHSN272200900049C. RS is supported by NSF GRFP, SW by NIH GM080177; CH by NIH 1U01HG007480-01 and the World Bank; AR by EU FP7/2007-2013 278433-PREDEMICS and ERC 260864; and GD by NERC D76739X. Sequence data are available at NCBI (NCBI BioGroup: PRJNA257197). Sharing of RNA samples used in this study requires approval from the Sierra Leone Ministry of Health and Sanitation. Tragically, four co-authors, who contributed greatly to public-health and research efforts in Sierra Leone, contracted EVD in the course of their work and lost their battle with the disease before this manuscript could be published. We wish to honor their memory.

Supplementary Materials:

Materials and Methods

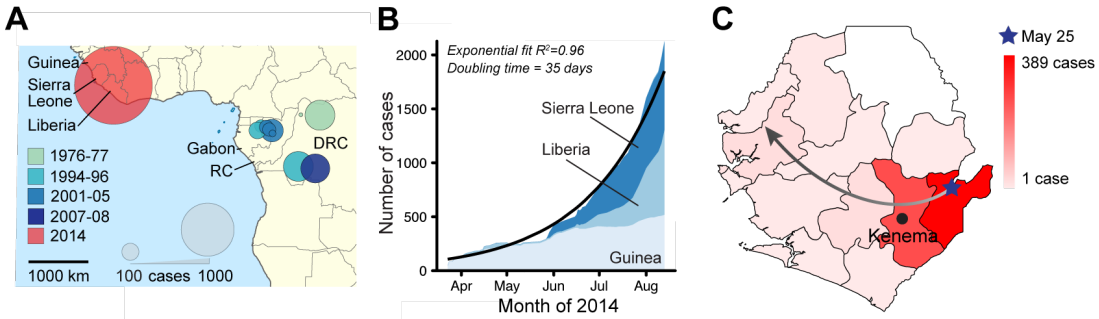
Supplementary Text

Figures S1-S11

Tables S1-S4

Files S1-S4

Fig. 1. **Ebola outbreaks, historical and current.** (**A**) Historical EVD outbreaks, colored by decade. Circle area represents total number of cases (RC = Republic of the Congo; DRC = Democratic Republic of Congo). (**B**) 2014 outbreak growth (confirmed, probable and suspected cases). **(C)** Spread of EVD in Sierra Leone by district. The gradient denotes number of cases and the arrows depict likely direction. (**D**) EBOV samples from 78 patients were sequenced in two batches, totaling 99 viral genomes (Replication = technical replicates (*6*)). Mean coverage and median depth of coverage with range are shown. (**E**) Combined normalized (to the sample average) coverage across sequenced EBOV genomes.



D

	# Total	# Patients	% Coverage	x Coverage
Batch 1	15	12	99.77%	555 [16-23,042]
Batch 2	84	66	99.91%	2,186 [6-18,623]
Combined	99	78	99.88%	2,048 [6-23,042]
Replication	140	78	99.39%	2,109 [8-17,537]

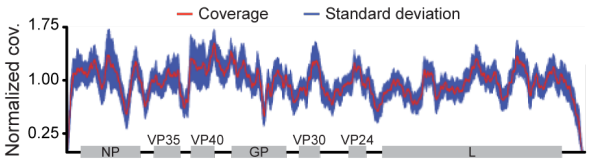


Fig. 2. Relationship between outbreaks. (**A**) Unrooted phylogenetic tree of EBOV samples; each major clade corresponds to a distinct outbreak (scale bar = nucleotide substitutions/site). (**B**) Root-to-tip distance correlates better with sample date when rooting on the 1976 branch (R^2=0.92, top) than on the 2014 branch (R^2=0.67, bottom). (**C**) Temporally rooted tree from (A).

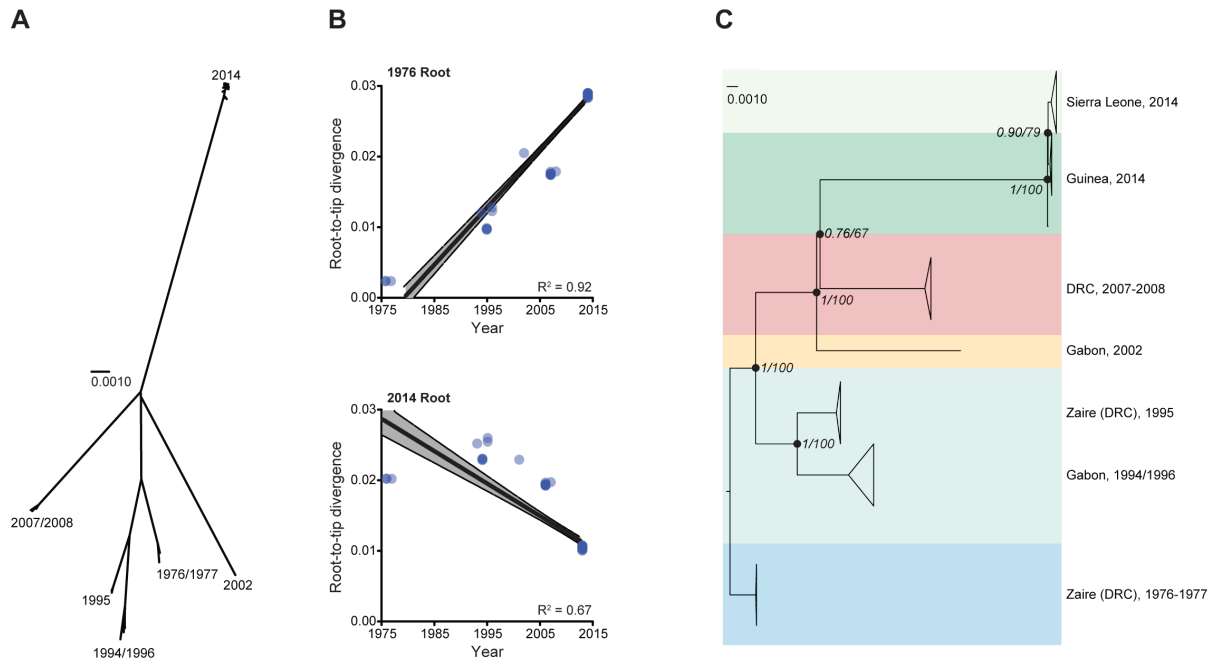


Fig. 3. Molecular dating of the 2014 outbreak. (**A**) BEAST dating of the separation of the 2014 lineage from Middle African lineages (SL = Sierra Leone; GN = Guinea; DRC = Democratic Republic of Congo; tMRCA: Sep 2004, 95% HPD: Oct 2002 - May 2006). (**B**) BEAST dating of the tMRCA of the 2014 West African outbreak (tMRCA: Feb 23, 95% HPD: Jan 27 - Mar 14) and the tMRCA of the Sierra Leone lineages (tMRCA: Apr 23, 95% HPD: Apr 2 - May 13); probability distributions for both 2014 divergence events overlayed below. Posterior support for major nodes is shown.

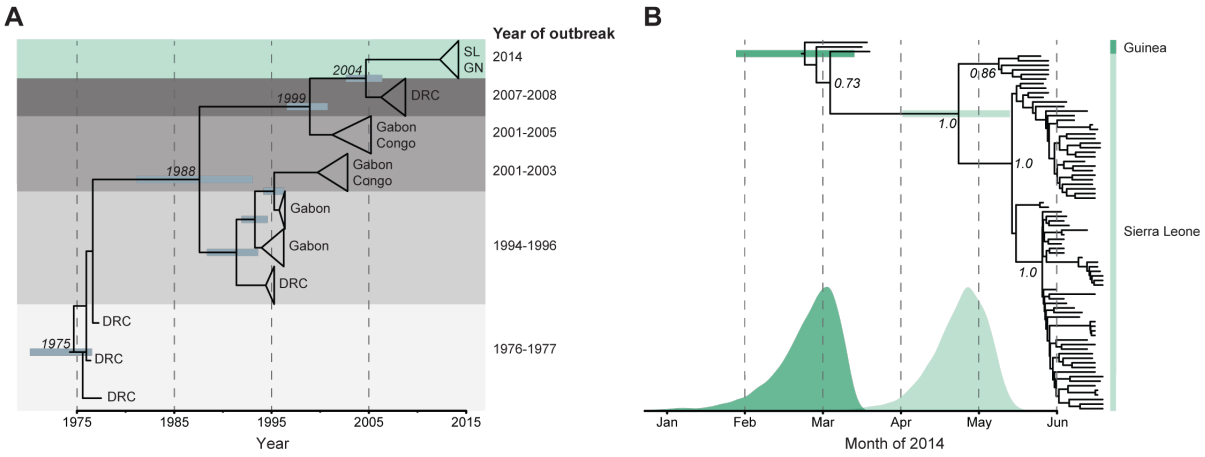


Fig. 4. Viral dynamics during the 2014 outbreak. (**A**) Mutations, one patient sample per row; beige = identical to Kissidougou Guinean sequence (accession KJ660346). The top row shows the type of mutation (green: synonymous, pink: nonsynonymous, intergenic: gray), with genomic locations indicated above. Clusters assignments are shown on left. **(B)** Number of EVD-confirmed patients per day, colored by cluster (arrow: first appearance of the derived allele at position 10,218, distinguishing clusters 2 and 3). (**C**) Intrahost frequency of SNP 10,218 in all 78 patients (absent in 28 patients, polymorphic in 12, fixed in 38). (**D & E**) 12 patients carrying iSNV 10,218 cluster geographically and temporally (HCW-A = unsequenced health care worker, Driver drove HCW-A from Kissi Teng to Jawie, then continued alone to Mambolo, HCW-B treated HCW-A). (**F**) Substitution rates within the 2014 outbreak and between all EVD outbreaks. (**G**) Proportion of nonsynonymous changes observed on different time scales (green=synonymous; pink=nonsynonymous). (**H**) Acquisition of genetic variation over time. 50 mutational events (short dashes) and 29 new viral lineages (long dashes) were observed (intrahost variants not included).