



Supplementary Materials for

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

Stephen K. Gire, Augustine Goba, Kristian G. Andersen, Rachel S. G. Sealfon, Daniel J. Park, Lansana Kanneh, Simbirie Jalloh, Mambu Momoh, Mohamed Fullah, Gytis Dudas, Shirlee Wohl, Lina M. Moses, Nathan L. Yozwiak, Sarah Winnicki, Christian B. Matranga, Christine M. Malboeuf, James Qu, Adrienne D. Gladden, Stephen F. Schaffner, Xiao Yang, Pan-Pan Jiang, Mahan Nekoui, Andres Colubri, Moinya Ruth Coomber, Mbalu Fonnio, Alex Moigboi, Michael Gbakie, Fatima K. Kamara, Veronica Tucker, Edwin Konuwa, Sidiki Saffa, Josephine Sellu, Abdul Azziz Jalloh, Alice Kovoma, James Koninga, Ibrahim Mustapha, Kande Kargbo, Momoh Foday, Mohamed Yillah, Franklyn Kanneh, Willie Robert, James L. B. Massally, Sinéad B. Chapman, James Bochicchio, Cheryl Murphy, Chad Nusbaum, Sarah Young, Bruce W. Birren, Donald S. Grant, John S. Scheffelin, Eric S. Lander, Christian Happi, Sahr M. Gevaio, Andreas Gnirke, Andrew Rambaut, Robert F. Garry, Sheik Humarr Khan, Pardis C. Sabeti

Correspondence to: andersen@broadinstitute.org (K.G.A), augstgoba@yahoo.com (A.G.), psabeti@oeb.harvard.edu (P.C.S)

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S11
Captions for Tables S1 to S4
Captions for Files S1 to S4

Other Supplementary Materials for this manuscript includes the following:

Tables S1 to S4 as zipped archives
Files S1 to S4 as zipped archives

Materials and Methods

Sample collection and processing

Samples were collected using existing collection and processing protocols at Kenema Government Hospital (KGH), under the emergency response efforts established by KGH. In brief, 10 ml of whole blood was collected and plasma or serum was prepared by centrifugation at 1,200 xg for 15 min. All samples were inactivated both in aliquots Buffer AVL (Qiagen) and TRizol LS 4:1 with plasma or serum. Human samples were obtained from patients suspected with EVD and stored in -20° C freezers. Both Qiagen AVL lysis buffer and Trizol have been used extensively in the literature (20-23) and have been shown to inactivate a wide range of viruses—including EBOV specifically (24). Samples were additionally heat denatured at 56° C for 10 min. In certain cases, AVL-inactivated RNA was isolated on-site using the QIAamp Viral RNA Minikit (Qiagen) according to the manufacturer's protocol. The extracted RNA was then lyophilized using the RNAsable system (Biomatrix) following the manufacturer's protocol. Samples were shipped on dry ice to Harvard University where samples were stored at -80° C. Biomatrix samples were shipped at room temperature and stored according to the manufacturer's recommendations.

PCRs performed at KGH

Diagnostic tests for the presence of EBOV were performed on-site using the SuperScript III One-Step RT-PCR System with Platinum Taq High Fidelity DNA Polymerase (Life Technologies). Each sample was run three times on three separate assays. The 25 μ L assay mix included 2 μ L RNA, one of two primer sets at 250 nM final concentration: *KGH primer set* (fwd: GTC GTT CCA ACA ATC GAG CG, rvs: CGT CCC GTA GCT TTR GCC AT), or *FiloAB primer set* (fwd: ATC GGA ATT TTT CTT TCT CAT T, rvs: ATG TGG TGG GTT ATA ATA ATC ACT GAC ATG) as well as a control assay which contained no primers and assessed non-specific amplification (fig. S2), 12.5 μ L 2x Reaction Mix and 0.5 μ L SuperScript™ III RT/ Platinum® Taq High Fidelity Enzyme Mix. The cycling conditions were 60° C for 20 min and 94° C for 5 min, followed by 35 cycles of 94° C for 15 sec, 58° C for 15 sec and 68° C for 15 sec with a final extension at 68° C for 2 min. RT-PCR was performed on a Bio-Rad thermocycler. The samples were then run on a 2.2% agarose e-gel (Lonza) and visual results recorded. All samples were re-tested following extraction with qRT-PCR at Harvard University (methods below). All samples testing positive at the field site were found to be positive by qRT-PCR and sequencing at Harvard. Additionally, all samples testing negative for EBOV at the field also tested negative at Harvard.

PCRs performed at Harvard

EBOV RNA was quantified using the Power SYBR Green RNA-to-Ct 1-Step qRT-PCR assay (Life Technologies). The 10 μ L assay mix included 2 μ L RNA, 0.3 μ M primer ZEBOV-kga-fwd, 0.3 μ M primer ZEBOV-kga-rv, 5 μ L 2x Power SYBR Green RT-PCR Mix and 0.08 μ L RT Enzyme Mix. The cycling conditions were 48° C for 30 min and 95° C for 10 min, followed by 45 cycles of 95° C for 15 sec and 60° C for 1 min with a melt curve of 95° C for 15 sec, 60° C for 15 sec and 95° C for 15 sec. RT-PCR was performed on the ABI7900 (Applied Biosystems) instrument. Standard PCR amplicons encompassing qRT-PCR products were prepared to determine viral copy number in qRT-

PCR assays. This was done by using synthetic oligonucleotides representing a portion of the EBOV segment within the VP24 gene as a template for PCR. These amplicons were cleaned up using AMPure XP beads (Beckman Coulter Genomics) and quantified by TapeStation (Agilent). Amplicon concentrations were converted to EBOV copies per microliter for quantification.

PCR validation

A total of three EBOV-specific and Pan-filovirus assays were tested at the KGH Laboratory and validated to assess which assays performed the best for diagnosis of EVD in Sierra Leone. Two of the PCR assays were adapted from published primer sets (*FiloAB (Pan-filo)* (8) and *Kulesh* (25)). Primer sequences can be found in table S3. *FiloAB* is a traditional or qPCR-based assay, whereas the modified *Kulesh* assay was a probe-based qPCR assay that was adapted to traditional PCR by omitting the probe. These two assays were run along with an assay designed at Harvard University (referred to as *KGH primer set*) (3). A no-primer control assay was also run and determined to be essential in the panel because of non-specific primer-independent amplification that occurs in some samples. The finalized KGH EBOV panel was comprised of the KGH primer assay, the *FiloAB* assay, and a no-primer control assay. These EBOV primer sets were tested in KGH Laboratory on inactivated EBOV seed stock obtained from USAMRIID.

Carrier RNA and Host rRNA depletion

Carrier RNA and host rRNA was depleted from RNA samples using RNase H selective depletion (26). Briefly, oligo d(T) (40 nt long) and/or DNA probes complementary to human rRNA were hybridized to the sample RNA. The sample was then treated with 20 units of Hybridase Thermostable RNase H (Epicentre) for 30 min at 45° C. The complementary DNA probes were removed by bringing the reaction up to 75 µL and treating with RNase-free DNase kit (Qiagen) according to the manufacturer's protocol. rRNA-depleted samples were purified using 2.2x volumes AMPure RNA clean beads (Beckman Coulter Genomics) and eluted into 10 µL water for cDNA synthesis.

cDNA synthesis, Nextera library construction and Illumina sequencing

EBOV sample RNA from selective depletion was used for cDNA synthesis and Illumina library preparation similarly to previously published RNA-Seq methods (27) with the following additional modifications. First, controls were used to monitor our library construction process. 500 fg of one, unique synthetic RNA (ERCC, gift from M. Salit, National Institute of Standards and Technology (28)) was spiked in using a different RNA for each individual EBOV sample to aid in tracking our viral sequencing process and potential index cross-contamination. Also, libraries were prepared from 200 ng human K-562 total RNA (Ambion) with each batch as an EBOV-negative control. Second, the oligo d(T) selection step was omitted. Third, Illumina Nextera XT was used for library preparation. ~50% of the cDNA product was used for the Nextera tagmentation step and libraries were generated using 15-16 cycles of PCR. Each individual sample was indexed with a unique dual barcode and libraries were pooled equally and sequenced on the HiSeq2500 (101 bp paired-end reads; Illumina) platform.

cDNA synthesis, standard library construction and Illumina sequencing

EBOV sample RNA from depleted samples according to published RNA-Seq methods mentioned above (27). Similar to the Nextera library preparation, spike-ins were added at 50 fg for quality control. Similarly, libraries were prepared from 200 ng human HeLa total RNA (Ambion) with each batch as an EBOV-negative control. Libraries were generated using 11-18 cycles of PCR. Each individual sample was indexed with a unique dual barcode and libraries were pooled equally and sequenced on the HiSeq2500 (101 bp paired-end reads; Illumina) platform.

NuGEN Ovation RNA-Seq, Nextera library construction and Illumina sequencing

RNA amplification was done as previously described (9). Illumina library construction was performed using NexteraXT (Illumina) following the manufacturer's protocol for >500 bp input DNA. Sequencing was performed on the Illumina HiSeq2500 platform, generating paired-end 101 bp reads.

Pacific Biosciences library construction and sequencing

For 7 EBOV RNA samples, sufficient NuGEN Ovation RNA-Seq material was obtained to generate Pacific Biosciences sequencing libraries. For each sample, 500 ng to 1 µg DNA was treated with Mungbean nuclease (New England Biolabs), exonuclease I (New England Biolabs) and RiboShredder (EpiCentre) and purified using 1.8 X Ampure XP beads (Beckman Coulter Genomics). Library construction was performed using the SMRTbell Template Prep Kit 1.0 (Pacific Biosciences) using 0.6X Ampure PB bead purification between enzymatic reactions. The final products were purified using two rounds of 0.6X Ampure PB bead purification. Libraries were combined with sequencing primer and polymerase (MagBead kit, Pacific Biosciences) and the resulting complex was subjected to sequencing, followed by primary data analysis on the Pacific Biosciences RS instrument following the manufacturer's recommendations.

Library selection

An initial set of 15 samples was prepared and sequenced using all three library preparation methods described above. The cDNA synthesis coupled with Nextera library construction gave the fastest and most consistent results, and was therefore utilized for all subsequent samples and all replicates.

Demultiplexing of raw Illumina sequencing reads

Raw Illumina sequencing reads were demultiplexed using Picard v1.4. To minimize cross-contamination between samples within each multiplexed sequencing run, the default settings were changed to allow for one mismatch in the two 8 bp barcodes and a minimum quality score of Q10 in the individual bases of the index. Sequencing quality metrics were calculated using FastQC and only high-quality sequencing libraries were used in subsequent analyses.

Assembly of full-length EBOV genomes

EBOV reads were extracted from the demultiplexed Fastq files using Lastal against a custom-made database containing all full-length EBOV genomes. The reads were then *de novo* assembled using Trinity and contigs were oriented, merged and cleaned using a custom-made pipeline. Contigs were indexed and all sequencing reads from each individual sample were aligned back to its own EBOV consensus sequence using Novoalign v3 with the following parameters: -k -l 40 -g 40 -x 20 -t 160. Duplicates were removed using Picard v1.4 and alignment files were realigned using GATK v2. Consensus sequences were called from the EBOV-aligned reads using GATK v2. All generated genomes were annotated as well as manually inspected for accuracy, such as the presence of intact ORFs, using Geneious v7. Regions where depth of coverage was less < 3x were called as 'N'. Eight patients in our data set had sequences for multiple time points of collection. There were no differences in their consensus assemblies across time. Therefore only one consensus sequence per patient was reported.

Multiple sequence alignments

EBOV consensus sequences were aligned using MAFFT v6 with the following parameters: --localpair --maxiterate 1000 --reorder --ep 0.123 before being trimmed using trimAl v1.4 with the maximum likelihood specific parameter: -automated1. Alignments for the 101 available EBOV genomes and 124 ebolavirus genomes (from 1976 to 2014) were used in subsequent analyses (file S1). Genic alignments across all ebolaviruses were generated by first aligning amino acid sequences using MUSCLE (29) and then aligning the nucleotide sequence based on the amino acid alignment.

Screening for recombinant sequences

Multiple sequence alignments were screened for recombinant viral sequences using the programs RDP, GENECONV, MAXCHI, CHIMAERA, 3SEQ, BOOTSCAN and SISCAN as implemented in the RDP3 software package with default settings. Potential recombinant sequences were identified when two or more methods were in agreement with P-values of < 0.001. No recombinant sequences were identified in any of our screens as there was no evidence for phylogenetic incongruence in our datasets.

SNP Calling

Polymorphic sites were identified directly from the multiple sequence alignments of 101 available EBOV genomes (from 1976 to current). 1,303 SNPs were identified in this sample set. Annotated SNP calls are available in VCF format (file S1). Protein coding effects were computed using a custom release of SnpEff (v4.0, build 2014-07-01) provided to us by its author to handle the unusual ribosomal slippage site in the GP gene (30). SNP annotations were made using our longest assembled isolate, G3686 (accession KM034562.1), as a reference genome.

Phylogenetic tree construction

Maximum likelihood trees were constructed using RAxML v7.3 with the GTR Γ nucleotide substitution model (31). Fifty instances were run to find the best tree and statistical support for each node in the tree was calculated using the standard bootstrapping algorithm with 500 pseudoreplicates. Trees containing all ebolaviruses

were rooted using midpoint rooting, whereas trees with only EBOV sequences were rooted using either the 1976 or 2014 EBOV clade. Bayesian phylogenies were made with MrBayes v3.2 using the GTR Γ model with four gamma categories for 1 million generations until all PSRF values were within a distance of four significant figures of 1 (32). To assess temporal structure of the data, linear regression was performed on the root-to-tip distances of samples versus the date of the isolate for the maximum likelihood trees using the program Path-O-Gen v1.4 (33). Tree outputs for the above analyses are provided (file S2).

Molecular dating using BEAST

EBOV phylogenies incorporating time of sampling were estimated using Bayesian Markov Chain Monte Carlo (MCMC) as incorporated into the program BEAST v1.8 (33). The date for each individual sample was based on the time of diagnostic testing (usually the same day as sample receipt at KGH). Alignments contained only unique sequences with no ambiguous calls; all unknown positions from the three Guinea EBOV isolates were masked out in all sequences. The alignment was divided into 3 partitions comprising 1st + 2nd codon positions, 3rd codon positions and intergenic sites. The nucleotide substitution process was modeled independently for each partition with the HKY Γ with four gamma categories (34, 35). A Skygrid non-parametric coalescent model (36) and uncorrelated lognormal relaxed clock (37) were found to be the best fit to the data. These models were compared to a strict molecular clock and a constant-size population coalescent model using the path-sampling estimator of the marginal likelihood (38). An uninformative CTMC reference prior (39) was used on the rate of evolution. BEAST XML files are provided (file S3). Maximum-clade credibility trees summarizing all MCMC samples were generated using TreeAnnotator v1.8 with a burn-in rate of 10%.

Counting fixed and variable polymorphic positions for each outbreak

The number of polymorphic positions falling on different branches of the phylogenetic tree was counted. A polymorphic position was considered fixed across all outbreaks if there was no within-outbreak variation at the position for any outbreak. A position was considered fixed within a particular outbreak if it was fixed for every sequence from that outbreak with a non-ambiguous and non-gap base call, but different from every sequence from any other outbreak. A position was considered variable for an outbreak if two sequences from the outbreak differed at the position (and both are non-ambiguous and non-gap).

Intrahost variant calling and analysis

Intrahost variants (iSNVs) were identified using V-Phaser 2 on sequences obtained from the Nextera library preparation and validated with a replicate Nextera library. Variants from the two Nextera libraries were subjected to an initial set of filters: variant calls with fewer than five forward or reverse reads or more than a 10-fold strand bias were eliminated. iSNVs were also removed if there was more than a five-fold difference between the strand bias of the variant call and the strand bias of the reference call. Variant calls in the primary Nextera library were additionally subjected to a 0.5% frequency filter, but were validated by calls at any frequency.

The final list of iSNVs contains only the filtered Nextera calls at positions where at least one patient had a concordant call in the validation library. Annotated iSNV calls are available in VCF format (file S4). This file infers 100% allele frequencies for all samples at an iSNV position where there was no intra-host variation within the sample, but a clear consensus call during assembly. Annotations were computed with SnpEff in the same manner as the population SNPs.

Eight patients had multiple time points of sequence data. For these patients, there was very little change in iSNV allele frequencies over time (fig. S4), suggesting a lack of significant change in intrahost viral composition during the course of a patient's hospitalization. Most analyses were restricted to a data set where each patient's iSNV allele frequencies were an average (median) of all that patient's time points, and focused on iSNP variation alone (leaving indels out). This reduced data set is provided in tabular text format (file S4).

Metagenomic analysis of Illumina sequencing reads

Illumina Fastq files were trimmed with Trimmomatic to remove bases from the ends of the reads with phred-scaled quality scores below Q20 or with a score below Q25 over a 4 bp window. All reads shorter than 70 bp after quality trimming were discarded. Human reads, as well as reads derived from commonly used cloning vectors and contaminating bacteria (e.g. reverse-transcriptase, *E. coli* reads derived from the production of enzymes used in sequencing library preparation), were removed using BMTagger (NCBI). Duplicate reads and low complexity reads were removed using PRINSEQ. All of the reads were then *de novo* assembled using MetaVelvet followed by Trinity. Contigs of at least 200 bp were used for BLASTn or BLASTx queries of the GenBank nucleotide (NT) or protein (NR) databases (E-score cutoffs of 10^{-6} and 10^2 , respectively). In a parallel pipeline, individual reads were used for BLASTn or BLASTx queries of GenBank with the same E-score cutoff values. Taxonomic classification of assembled contigs and individual reads were performed and visualized using MEGAN v4. Samples were considered to have an organism present if MEGAN 4 'min support' was ≥ 5 for read-based classification, or ≥ 1 for contig-based classification. The 'min score' requirements were ≥ 50 for reads, and ≥ 150 for contigs.

Calculation of doubling time and outbreak growth in Sierra Leone, Liberia, and Guinea

Confirmed, probable, and suspected cases were considered. Occasional declines observed in Fig. 1B represent reclassification of patients' conditions. These data were obtained from WHO reports (4) and then combined into daily totals. Days for which the WHO reported no data were dropped from the set. To enable production of a continuous plot, remaining gaps were filled under the assumption that case numbers in a given country did not change between reports. The totals were log-transformed and then regressed linearly (using least squares) to produce an exponential fit of n (total cases) = $104.96^{(0.0199*t)}$, where t =days since initial detection of outbreak, taken as March 23. This fit was then used to infer the doubling time of the outbreak ($\log_2/0.0199 = 34.79$ days).

Supplementary Text

Ethics statement

This study has been evaluated and approved by Institutional Review Boards in Sierra Leone and at Harvard University. Both the Office of the Sierra Leone Ethics and Scientific Review Committee and the Harvard Committee on the Use of Human Subjects have granted a waiver of consent to sequence and make publically available viral sequences obtained from patient and contact samples collected during the EVD outbreak in Sierra Leone. Both committees also granted use of clinical and epidemiological data for de-identified samples collected from all suspected EVD patients receiving care during the outbreak response. Dual Use Research of Concern was considered but deemed not to be necessary.

The Sierra Leone Ministry of Health and Sanitation provided approval for non-infectious, inactivated samples originating from EVD patients to be shipped from Sierra Leone to the Broad Institute and Harvard University for viral sequencing. They additionally granted a waiver of consent for genomic studies and use of patient metadata as part of its emergency response to the 2014 EVD outbreak.

Biological safety approvals

The EBOV-related research and laboratory safety protocols are registered with the Committee of Microbiological Safety (COMS) at Harvard University, and the viral sequencing work is registered with the Institutional Biosafety Committee (IBC) at the Broad Institute. COMS and IBC both require complete documentation of any potential biohazards in the laboratory, which is reviewed by a committee prior to commencement of any research with biological materials. These organizations also provide risk assessments to help establish safe research policies and procedures.

Guinean sequence correction

For the three 2014 Guinean EBOV genomes (3), SNP calls were masked in thirteen PCR primer binding sites that were used for Sanger sequencing as well as the short overlapping regions between primers. In many of these regions, the Guinea lineages were different than the Sierra Leone sequences and other EBOV lineages, but instead matched the ancestral 1976 sequences. This pattern was not observed outside the primer binding sites or overlapping regions, suggesting that these calls may be reference bias artifacts (possibly due to imputation in regions of low confidence sequencing). The masked sequences can be obtained from our alignments (file S1).

Glycoprotein RNA editing

The RNA editing site of the glycoprotein (GP) gene consists of 7 U residues; co-transcriptional stuttering can result in transcripts with more or less A residues. The resulting frameshifts allow for the expression of distinct glycoproteins called sGP (7 A), GP (predominantly 8 A), and ssGP (predominantly 6 A). Previous studies have demonstrated that EBOV passaging results in distinct changes in the genomic editing site, which switches to 8 U in tissue culture and to 7 U in infected guinea pigs and nonhuman primates, and thereby in different ratios of edited transcripts. Deep sequencing revealed 8 U at ~1% and 7 U at ~99% (fig. S5B). This differs from the proportions previously

reported in animal models and tissue culture (8 U 20%; 7 U 80%) (12, 19) and represents the first measurement of these intrahost ratios in an unpassaged (p0) isolate and in a human outbreak setting. Caution is needed in comparing these differences, however, since the previous studies were performed using cloning-based Sanger sequencing.

Potential duplicate samples in our dataset

Preliminary metadata obtained after the completion of our manuscript suggest that three pairs of sequences may correspond to the same patients (table S2). The patient identification numbers for these three pairs are: G3679/EM096, G3682/EM098, and G3787/G3831. In all three cases, these pairs have identical consensus sequences. We cannot confirm whether these are in fact duplicates or not, but based on the data available to us, we suspect this might be the case. Importantly, the presence of these potential duplicates does not change any of the conclusions in this study. The only difference would be a count of 75 unique patients in our dataset, rather than the reported 78 patients (and 16 patients with multiple time-points and/or extraction methods instead of 13).

Supplementary Figures

Fig. S1.

The temporal spread of EVD in Sierra Leone by district. The gradient denotes the timing of spread and the arrow depicts the likely direction of spread within Sierra Leone. Key cities have also been marked, including Kailahun (where the outbreak in Sierra Leone started) and Kenema (where Kenema Government Hospital is located). Due to a lapse in district-level status reporting, the infection dates of districts in grey (Kono, Kambia, Bombali, Tonkolili, Pujehun, Moyamba, Bonthe, and Western Area Rural) can only be determined to the date range between July 23 and August 6. Infection statuses of the districts are current as of August 18, 2014.

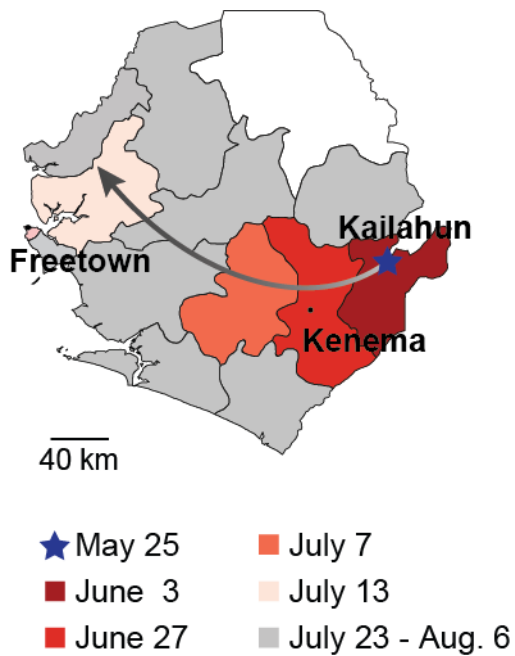


Fig. S2.

PCR validation. A total of three EBOV-specific and Pan-filovirus assays were tested at the KGH Laboratory and validated to assess which assays performed the best for diagnosis of EVD in Sierra Leone. Seed stock was serially diluted from 1:1 to 1:1000, and 2 μ l were inputted into PCR reactions, using Invitrogen's ssIII one-step RT-PCR HiFi kit, along with a negative RNA extraction control. **(A)** Both the FiloAB and Modified Kulesh primer sets could only detect EBOV seed stock at a dilution of 1:10, whereas the *KGH primer* set could readily detect EBOV seed stock at a 1:1000 dilution. **(B)** EBOV seed stock were spiked into patient samples in order to mimic patient sample conditions (at that time, no EVD patient samples were available in Sierra Leone). The *KGH primer* set was able to detect seed stock at a 1:1000x dilution in patient extracted RNA, whereas the other two published primer sets could only detect down to a 1:10 dilution under the same experimental conditions.

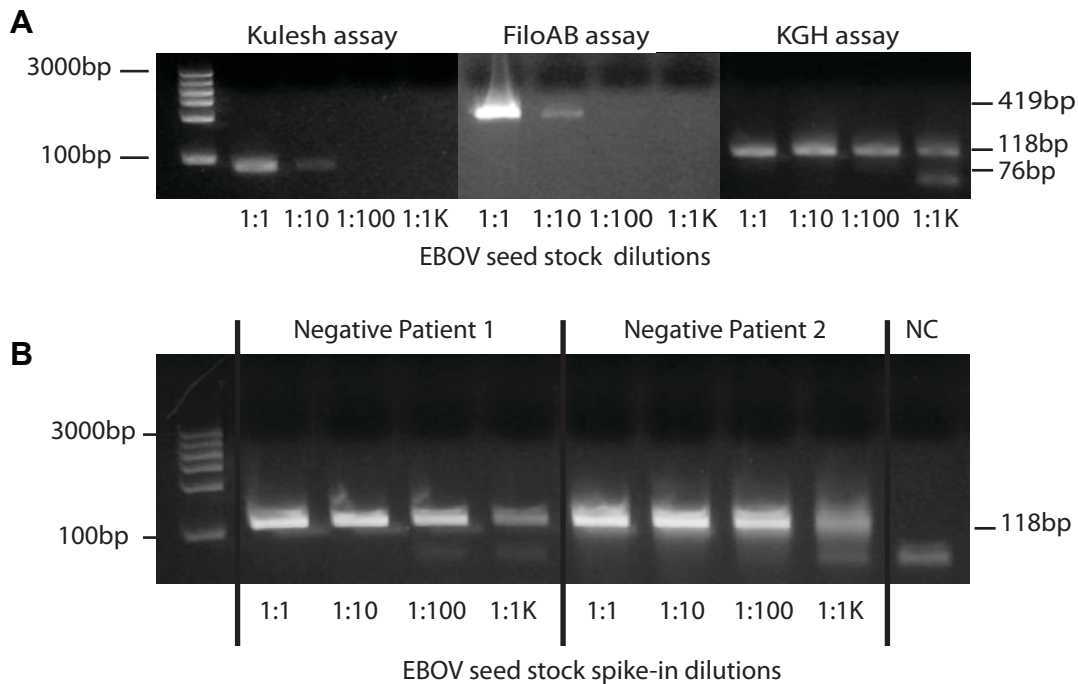


Fig. S3.

Metagenomic screening of suspected EVD patients. (A) All suspected EVD patients were screened for known viruses (excluding known contaminants and phages) and malaria (plasmodia). Several suspected cases showed evidence of other pathogens prevalent in West Africa, including plasmodia (5 cases), HIV-1 (2 cases), and Lassa Fever (1 case). These findings confirm that EVD can be easily mistaken for other common illnesses and vice versa, therefore highlighting the importance of accurate diagnostics. **(B)** Percentage of confirmed and suspected EVD cases with evidence of malarial infection. There is no obvious correlation between infection with malaria and EVD, suggesting that a positive diagnosis for malaria does not necessarily rule out EVD.

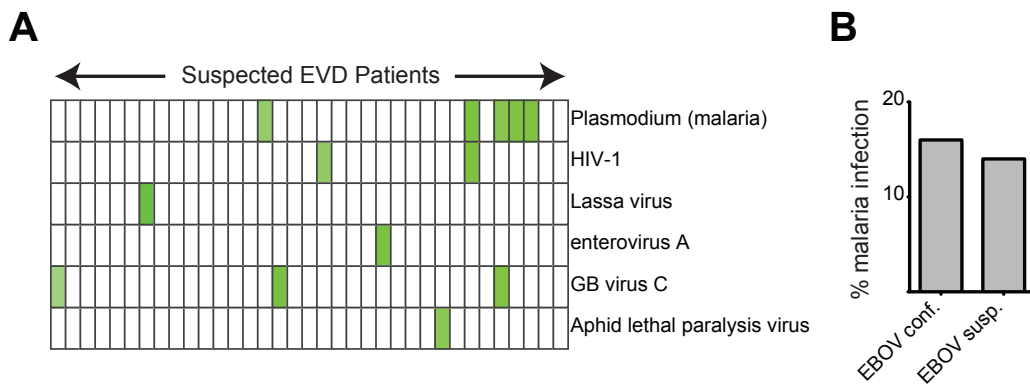


Fig. S4.

Patient time courses of iSNV frequencies and viral copy number. (A) iSNV minor allele frequencies are generally stable across time for patients where multiple time points are available. Each color represents a different variant position in the genome. Both SNP and indel intrahost variants are shown here. **(B)** Viral copy number (table S2) is shown at each time point. Note that the fourth time point of EM124 produced a consensus assembly, but did not provide sufficient read depth for the identification of iSNVs.

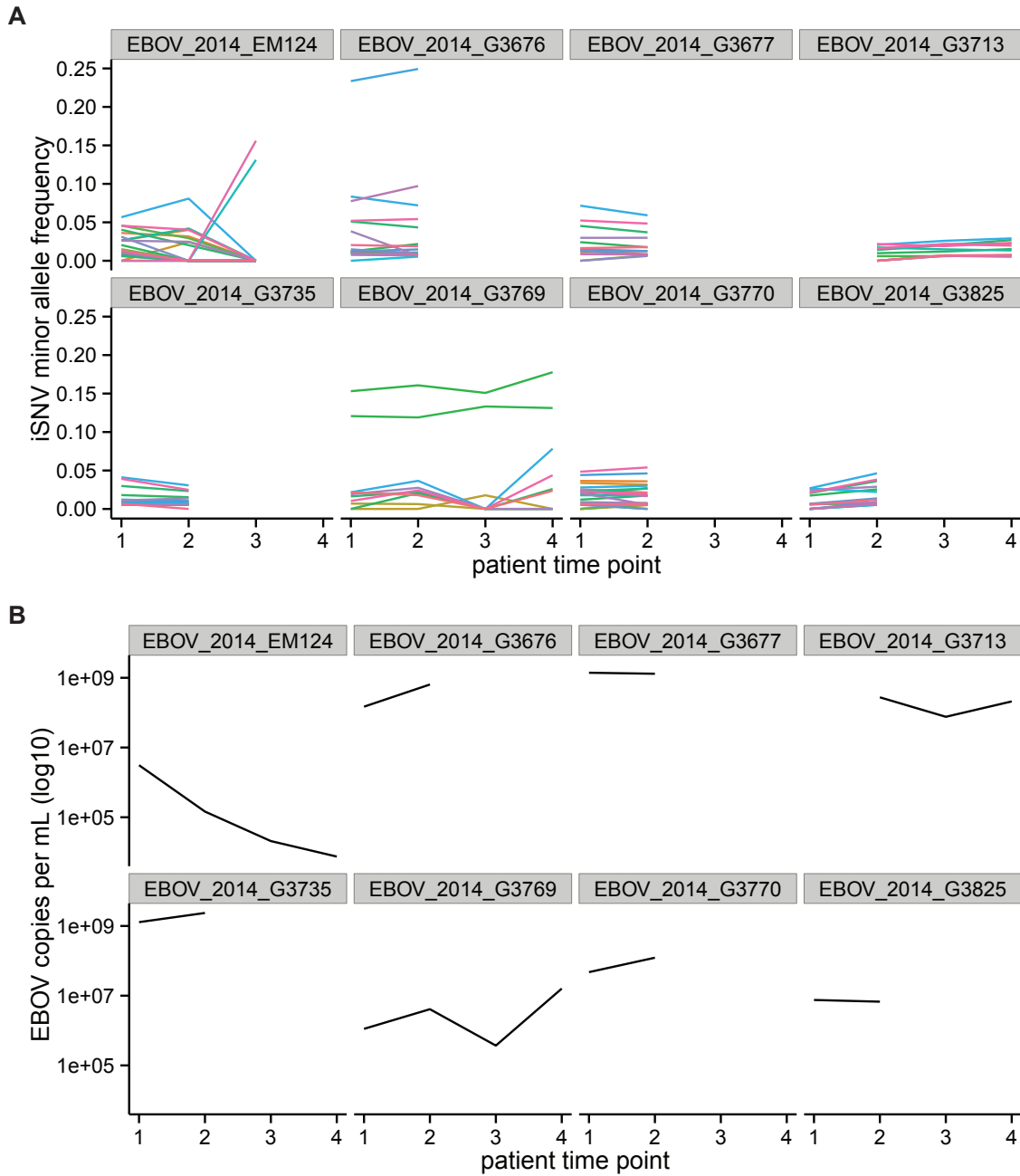


Fig. S5.

Deep-sequencing of glycoprotein polyU variants. A functionally important feature of EBOV is the synthesis of the structural glycoprotein (GP), the secreted glycoprotein (sGP) and the small soluble glycoprotein (ssGP), which is tightly regulated by a transcriptional RNA editing phenomenon (10, 11, 40). The regulation of expression of GP and sGP is suggested to play an essential role in replication and spread of EBOV (12). **(A)** This phenomenon results in the insertion or deletion of a uridine residue during transcription at the editing site, which contains 7 consecutive uridines. **(B)** The average frequency of this insertion across Sierra Leone patients was GP at 1.20% (8U) and sGP at 98.56% (7U).

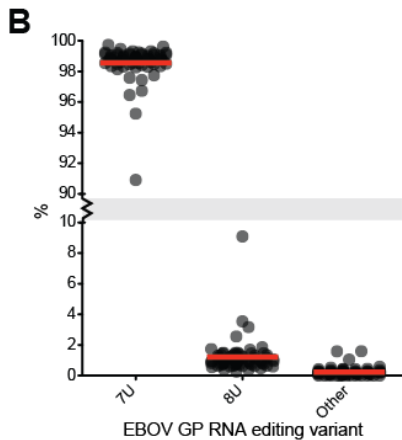
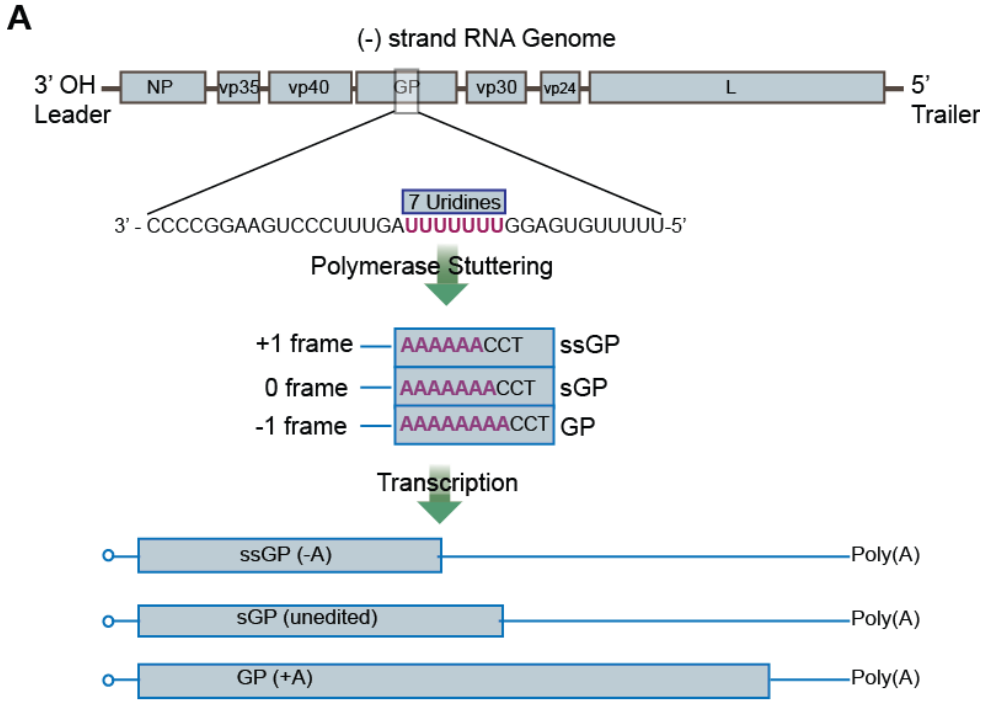


Fig. S6.

Phylogenetic trees estimating the relationship between members of the Ebolavirus genus. (A) A Bayesian phylogenetic tree created using MrBayes of all ebolaviruses places the 2014 outbreak lineages as ancestral to the rest of EBOV. Posterior support values are shown for each node. **(B)** A maximum likelihood tree created with RAxML puts the 1995 outbreak clade as ancestral to other EBOV lineages. Bootstrap values (500 pseudoreplicates) are shown for each node. **(A, B)** All trees were mid-point rooted. Bundibugyo virus=BDBV, Tai Forest virus=TAFV, Sudan virus=SUDV, and Reston virus=RESTV.

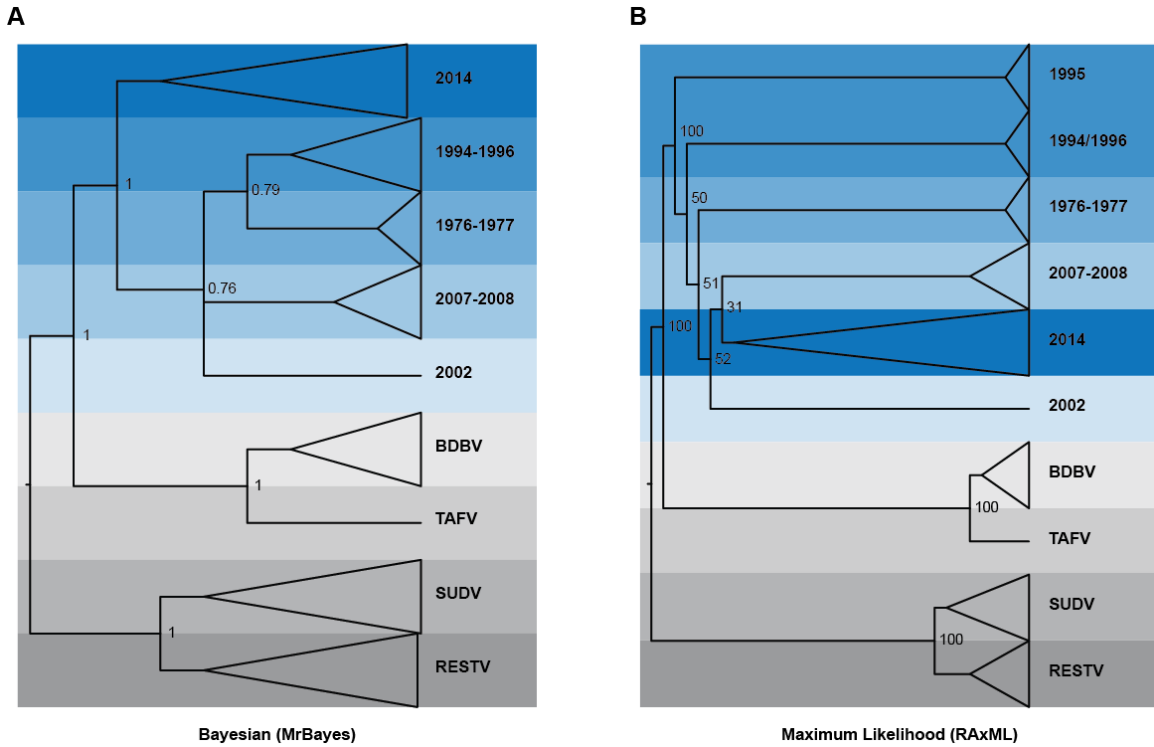


Fig. S7.

Phylogenetic trees rooted using either the 2014 or 1976 outbreak strains. Phylogenetic trees were created using RAXML and the resulting trees based on **(A)** rooting on the 1976 EVD outbreak variants of EBOV, or **(B)** the 2014 EVD outbreak variants of EBOV are displayed.



Fig. S8.

Phylogenetic tree showing the individual lineages in the 2014 EVD outbreak. A maximum likelihood tree was created using RAXML and the four main clusters (Guinea, as well as three Sierra Leone clusters) are displayed. Bootstrap values (500 pseudoreplicates) are shown for each node. Scale bar = nucleotide substitutions/site.

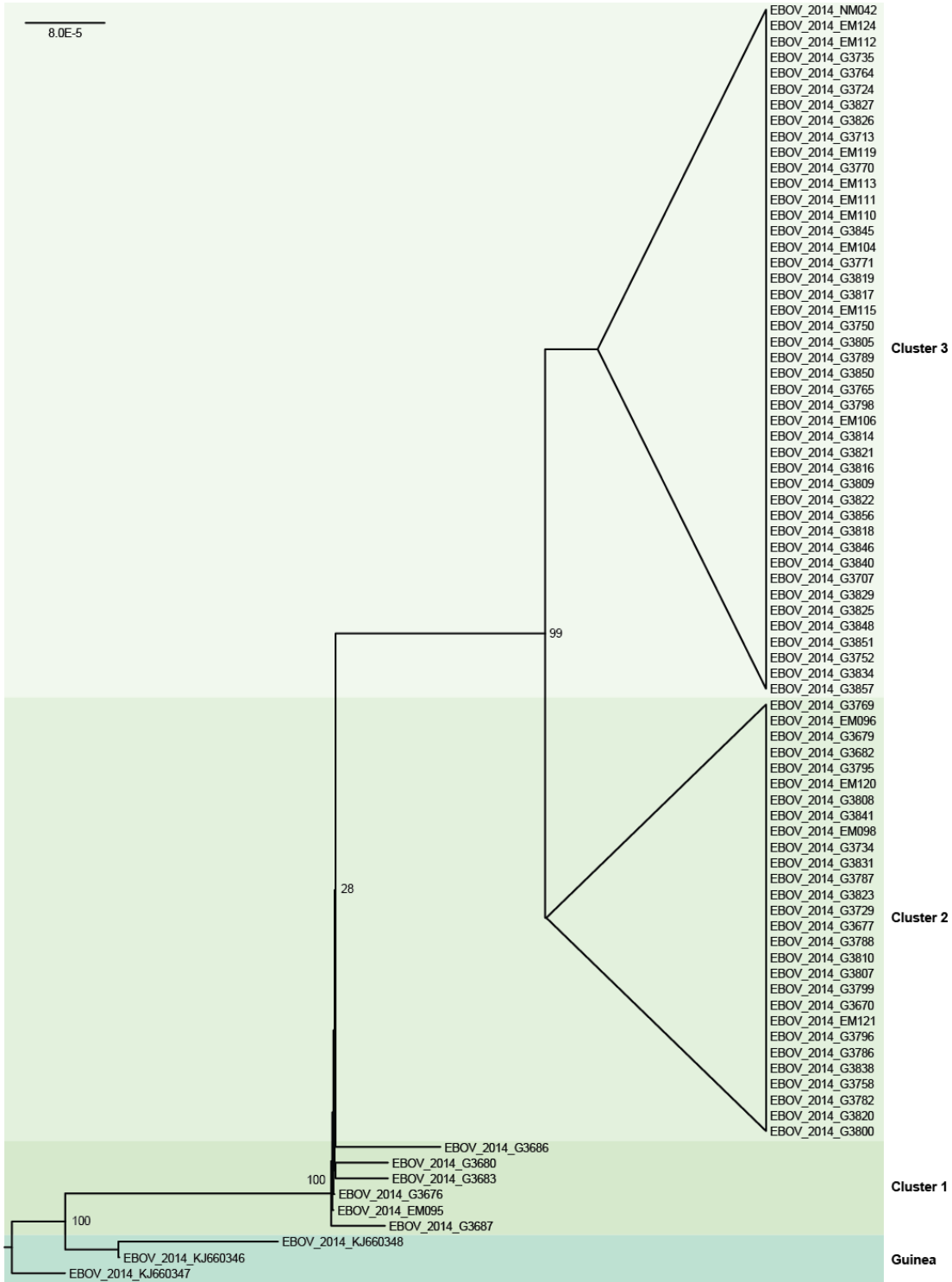
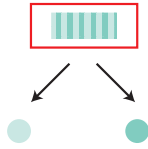


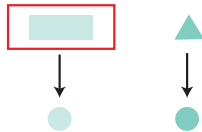
Fig. S9.

Different transmission scenarios for the first two genetic clusters within Sierra Leone. Each circle represents one of the twelve initial patients (travelers) from Sierra Leone. The deceased traditional healer (probable EVD) is represented by a rectangle. Guinean funeral attendees are represented by triangles. **Scenario A.** Travelers from Sierra Leone became infected by EBOV from two related EBOV lineages present in the body of the deceased healer at the funeral. **Scenario B.** Travelers acquire two related Guinean EBOV lineages from multiple sources at the funeral. **Scenario C.** Some travelers acquire a single lineage from Guinea, which subsequently mutates and transmits within the group.

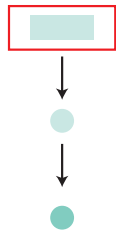
- A** Transmission of multiple distinct haplotypes from a single individual at the healer's funeral



- B** Transmission of distinct haplotypes from distinct infected individuals in Guinea



- C** Transmission of a single haplotype at the healer's funeral and subsequent mutation and transmission within the group of travellers







- Key:
-  Healer
 -  Possible transmission link
 -  Attendee at funeral (from Guinea)
 -  Attendee at funeral (from Sierra Leone)

Fig. S10.

Patients with shared variation show temporal patterns suggesting possible transmission relationships. (A) Clusters of 2-7 patients with genetically identical viruses at the consensus sequence level (this excludes clusters of identical genomes from more than 7 patients). Identical viruses often group temporally, either occurring within one to two days, suggesting infection from a common source, or spaced between six to nine days apart (median = 8 days), suggesting potential transmission between patients. (B) Patients with shared intrahost SNPs (most intrahost SNPs are unique to one patient and not displayed here). Shared intrahost SNPs either appear around the same time, or eight to fourteen days apart (median = 12 days), suggesting potential transmission between patients.

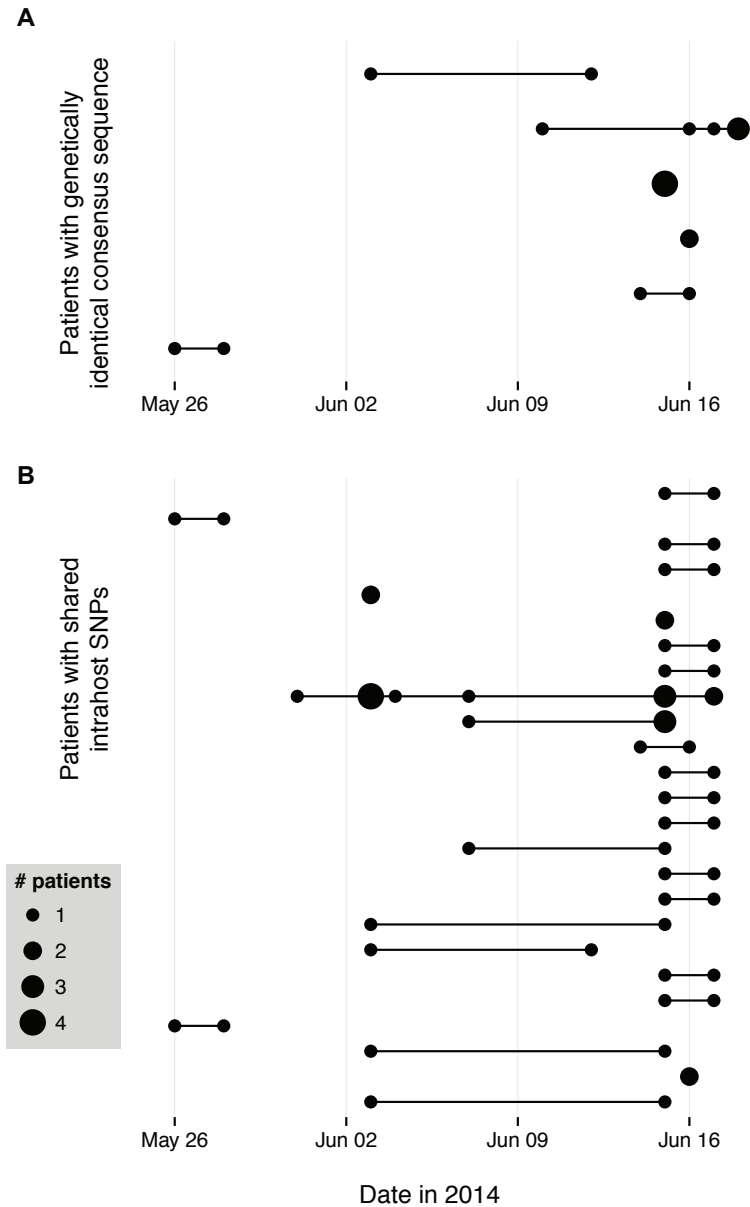
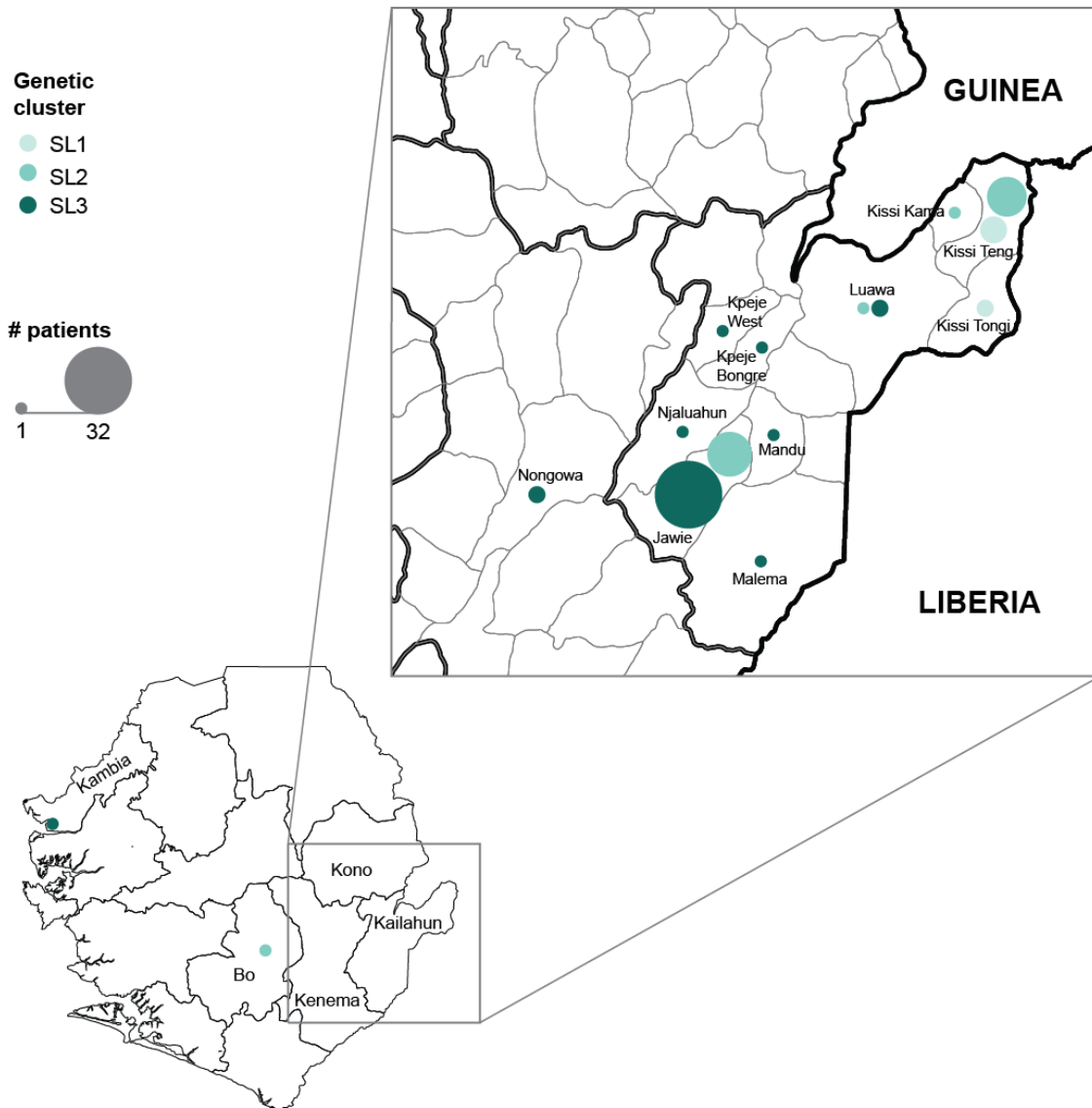


Fig. S11.

Map of 78 sequenced patients from Sierra Leone. Bottom: a map of all of Sierra Leone is shown. Five districts are labeled (Kambia, Bo, Kono, Kenema, and Kailahun). Bo and Kambia have only one case each, depicted with small green circles (shade of green corresponds to genetic cluster designation). Top: the eastern region of Sierra Leone, including the Kailahun and Kenema districts, are expanded to show more detail. Cases are plotted at the chiefdom level within each district, separately for each genetic cluster. Circles are proportional to the number of cases of that genetic cluster seen in that chiefdom. Names of chiefdoms with identified cases are shown. Kissi Teng = origin of EVD outbreak in Sierra Leone. Jawie contains the Daru Health Clinic, where the bulk of later cases occur. Nongowa is where Kenema Government Hospital (KGH) is located; two cases originate from there on the last day of our data set. All 78 cases were identified by the VHF laboratory KGH; the geography of each patient indicates their town of residence.



Supplementary Tables (compressed archives)

Table S1.

Overview of the different sequencing methods used for the first batch of EBOV samples. EBOV samples from batch 1 were sequenced using Illumina (three different library preps) and PacBio (one library prep). The samples for Illumina Nextera and Illumina Nugen are the same (n=15), whereas the samples for Illumina Standard and PacBio Nugen are subsets of these samples. The median depth of coverage including range, as well as the mean percent coverage are shown.

Table S2.

Summary of sequence data produced in this study. Sample information and sequencing statistics for all 99 samples prepared using the Nextera library preparation method. EBOV copies/ml of serum was determined using qPCR (see Material and Methods above). The dates correspond to the date that the sample was tested at the KGH Laboratory.

Table S3.

Primer Comparison. Twelve primer sets from eleven published assays (7, 8, 21, 25, 41-44) and the *KGH primer set* comprising both EBOV-specific and Pan-filovirus assays were screened against the EBOV consensus from Sierra Leone sequences using Geneious R6. Mapped primers and probes were compared to consensus sequence and nucleotide discrepancies noted. These discrepancies are shown in red italics in the table. There were a total of 9 nucleotide discrepancies in either the forward or reverse primer, or the probe. It is unknown how these discrepancies affect sensitivity and specificity of each assay. Further validation is needed, comparing these primer sets to the Guinea and Sierra Leone EBOV variants by conventional and quantitative RT-PCR methods in order to assess reaction kinetics and inform diagnostic suitability of the assays. Note that no nucleotide discrepancies were seen in primer sets with degenerate bases. This may constitute a good strategy in future assay design.

Table S4.

SNPs unique to the 2014 outbreak variant. Table of SNPs unique to the 2014 outbreak. The amino acid position, reference and alternate amino acids, BLOSUM62 score for nonsynonymous substitutions, count of sequences in the outbreak clade carrying the variant, and conservation across all ebolaviruses are given. A list is provided of amino acid sites that are polymorphic or unique in any EVD outbreak (1976-7, 1994/1996, 1995 2002, 2007-8, 2014) and otherwise completely conserved across all ebolaviruses. Additionally a list is provided of amino acid sites that are polymorphic or unique in any EVD outbreak (1976/1977, 1994/1996, 1995, 2002, 2007/2008, 2014), have a non-conservative substitution (BLOSUM62 score < 0) between the reference and alternate amino acids, and otherwise have only conservative amino acid substitutions across all ebolaviruses. Finally, two tables are provided of amino acid differences in GP between the 2014 outbreak variants and the Mayinga (Genbank accession number NC002549) and Kikwit (Genbank accession number JQ352763) variants.

Supplementary Files (compressed archives)

File S1.

Alignment and SNP calls used in this study.

File S2.

Phylogenetic trees created using MrBayes and RAxML.

File S3.

BEAST XML files used to estimate the divergence time for the 2014 EBOV lineages, as well as for all EBOV isolates.

File S4.

Intrahost variants for 78 Sierra Leone EVD patients. iSNVs are described in annotated VCF format. Tabular text formats are provided for a subset of analyses described in this paper.