Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak

Kafetzopoulou LE[1,2,3], Pullan ST[1,2], Lemey P[4], Suchard MA[5], Ehichioya DU[3,6], Pahlmann M[3,6], Thielebein A[3,6], Hinzmann J[3,6], Oestereich L[3,6], Wozniak DM[3,6], Efthymiadis K[7], Schachten D[3], Koenig F[3], Matjeschk J[3], Lorenzen S[3], Lumley S[1], Ighodalo Y[8], Adomeh DI[8], Olokor T[8], Omomoh E[8], Omiunu R[8], Agbukor J[8], Ebo B[8], Aiyepada J[8], Ebhodaghe P[8], Osiemi B[8], Ehikhametalor S[8], Akhilomen P[8], Airende M[8], Esumeh R[8], Muoebonam E[8], Giwa R[8], Ekanem A[8], Igenegbale G[8], Odigie G[8], Okonofua G[8], Enigbe R[8], Oyakhilome J[8], Yerumoh EO[8], Odia I[8], Aire C[8], Okonofua M[8], Atafo R[8], Tobin E[8], Asogun D[8,9], Akpede N[8], Okokhere PO[8,9], Rafiu MO[8], Iraoyah KO[8], Iruolagbe CO[8], Akhideno P[8], Erameh C[8], Akpede G[8,9], Isibor E[8], Naidoo D[10], Hewson R[1,2,11,12], Hiscox JA[2,13,14], Vipond R[1,2], Carroll MW[1,2], Ihekweazu C[15], Formenty P[10], Okogbenin S[8,9], Ogbaini-Emovon E[#8], Günther S[#16,6], Duraffour S[#3,6].

Affiliations:

1. Public Health England, National Infection Service, Porton Down, UK.
2. National Institute of Health Research (NIHR), Health Protection Research Unit in Emerging and Zoonotic Infections, University of Liverpool, Liverpool, UK.
3. Bernhard Nocht Institute for Tropical Medicine, Hamburg, Germany.
4. Department of Microbiology and Immunology, Rega Institute, KU Leuven - University of Leuven, Leuven, Belgium.
5. Departments of Biomathematics, Biostatistics, and Human Genetics, University of California, Los Angeles, CA, USA.
6. German Centre for Infection Research (DZIF), partner site Hamburg, Germany.
7. Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Brussels, Belgium.
8. Irrua Specialist Teaching Hospital, Irrua, Nigeria.
9. Faculty of Clinical Sciences, College of Medicine, Ambrose Alli University, Ekpoma, Nigeria.
10. World Health Organization, Geneva, Switzerland.
11. Faculty of Infectious and Tropical Diseases, Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, London, UK.
12. Faculty of Clinical Sciences and International Public Health, Liverpool School of Tropical Medicine, Liverpool, UK.
13. Singapore Immunology Network, Agency for Science, Technology and Research (A*STAR), Singapore.
14. Institute of Infection and Global Health, University of Liverpool, Liverpool, UK.
15. Nigeria Centre for Disease Control, Abuja, Nigeria.

# Contributed equally

Abstract: The 2018 Nigerian Lassa fever season saw the largest ever recorded upsurge of cases, raising concerns over the emergence of a strain with increased transmission rate. To understand the molecular epidemiology of this upsurge, we performed, for the first time at the epicenter of an unfolding outbreak, metagenomic nanopore sequencing directly from patient samples, an approach dictated by the highly variable genome of the target pathogen. Genomic data and phylogenetic reconstructions were communicated immediately to Nigerian authorities and the World Health Organization to inform the public health response. Real-time analysis of 36 genomes and subsequent confirmation using all 120 samples sequenced in the country of origin revealed extensive diversity and phylogenetic intermingling with strains from previous years, suggesting independent zoonotic

transmission events and thus allaying concerns of an emergent strain or extensive human-to-human transmission.

Lassa fever is an acute viral hemorrhagic illness, first described in 1969 in the town of Lassa, Nigeria (1). It is contracted primarily through exposure to urine or feces of infected Mastomys spp. rodents or, less frequently, through the bodily fluids of infected humans. Lassa virus (LASV) is endemic in parts of West Africa, including Nigeria, Benin, Côte d'Ivoire, Mali, Sierra Leone, Guinea, and Liberia (2). The upsurge of Lassa fever cases during the 2018 endemic season in Nigeria—referred to here as the 2018 Lassa fever outbreak—has been the largest on record, reaching 1495 suspected cases and 376 confirmed cases and affecting more than 18 states by 18 March (fig. S1). This notably exceeds the 102 confirmed cases reported during the same period in 2017 (fig. S1) (3). The unprecedented scale of the outbreak raised fears of the emergence of a strain with a higher rate of transmission. Because of these concerns, on 28 February the Nigeria Centre for Disease Control (NCDC) and the World Health Organization (WHO) urgently requested sequencing information and preliminary results from our pilot-scale study, in which we used a metagenomic approach with the Oxford Nanopore MinION device (Oxford Nanopore Technologies) to conduct in-country, mid-outbreak viral genome sequencing. This instigated a major uptick in sequencing efforts, leading to the sequencing of 120 samples.

Nanopore sequencing is an emerging technology with great potential. The MinION is a small, robust sequencing device suited for the genetic analysis of pathogens in remote or resource limited settings (4). Nanopore sequencing of polymerase chain reaction (PCR) amplicons of Ebola virus genomes provided important data from the field in real time during the 2014–2016 Ebola virus disease outbreak in West Africa (5), and a more sophisticated multiplex amplicon sequencing methodology (6) has been used effectively during recent Zika and yellow fever outbreaks in Brazil (7, 8). However, highly variable pathogens such as LASV present a substantial challenge for this type of amplicon-based approach. Owing to an interstrain nucleic acid sequence variation of up to 32 and 25% for the L (large segment encoding the RNA polymerase and the zinc-binding protein) and S (small segment encoding the glycoprotein and the nucleoprotein) segments, respectively (9), even PCR-based laboratory diagnosis poses a serious challenge. Designing targeted whole-genome sequencing approaches, such as those using PCR amplicons or bait-and capture probes, without prior knowledge of the targeted LASV lineage is therefore cumbersome. Random reverse-transcription (RT) and amplification by sequence-independent single primer amplification (SISPA) formetagenomic sequencing to identify RNA viruses has been demonstrated to work on the MinION (10), and our previous work highlighted the feasibility of retrieving complete viral genomes directly from patient samples at clinically relevant viral titers using this approach for dengue and chikungunya viruses (11). We describe here the application of field metagenomic sequencing of LASV at the Irrua Specialist Teaching Hospital (ISTH), Edo State, during the 2018 Lassa fever season.

A total of 120 LASV-positive samples were sequenced during a 7-week mission; these were selected on the basis of cycle threshold value and location of the 341 cases reported by ISTH between 1 January and 18 March 2018 (figs. S1 and S2). The majority of samples originated from Edo State followed by Ondo and Ebonyi (fig. S2). Selected samples covered the wide range of clinical viral loads observed, including several samples testing negative in one of the two realtime RT-PCR assays used (fig. S3 and data S1). Up to six samples were run in multiplex per MinION flow cell, along with a negative control. To produce high-confidence consensus sequences for phylogenetic inference, we chose to map both base called reads and raw signal data to a reference sequence and call variants using Nanopolish software, as developed for the West African Ebola virus disease outbreak (5); basecalled reads were then remapped to the consensus and a further round of correction was applied (fig. S4). Owing to the diversity of LASV, selection of an individual reference genome for read alignment was required for each sample. To select the closest existing LASV reference genome, nonhuman reads from each sample were assembled de novo using Canu (12). A notable proportion of reads generated per sample were LASV at an average frequency of 4.26% with a maximum of 42.9%, allowing for sufficient genomic sequence (>70%) for phylogenetic comparison of at least one segment in 91 of the samples tested (figs. S3 to S6).

Additionally, sequences were validated by Illumina resequencing of 14 SISPA preparations, which matched with their Oxford Nanopore counterparts with little to no divergence, confirming the accuracy of the Oxford Nanopore approach (table S1).

Metagenomic classification using the Centrifuge software system (13) identified 0.10% of reads from sample 110 as originating from hepatitis A virus, providing 74% genome coverage at 20-fold depth. LASV accounted for 0.83% of reads in the same sample, providing 96% genome coverage. These findings demonstrate the potential of this simple approach to identify multiple RNA viruses, including those present as co-infections. In all other samples tested, LASV was the sole pathogen identified despite a small number of reads classified as other viruses (fig. S7 and data S1).

To dissect the molecular epidemiology of the 2018 Lassa fever outbreak in Nigeria, we performed phylogenetic analysis of all newly generated LASV sequences together with unpublished sequences from previous years (data S2) and sequences available in GenBank. We used this as a frame of reference to document how the genomic data generated in real time (made publicly available at virological.org) provided valuable epidemiological insights into the unfolding outbreak dynamics.

Maximumlikelihood phylogenetic reconstruction of the S segment sequences indicates that all 2018 viruses fall within the Nigerian LASV diversity, specifically within genotypes II and III, and they are phylogenetically interspersed with Nigerian LASV sequences from previous years (Fig. 1). This phylogenetic pattern is mimicked by the L segment reconstruction (fig. S8). Only seven viruses in the entire genome dataset (n = 348) were identified as clustering significantly differently in the L and S segments (supplementary methods), which is in line with the small number of potential LASV reassortments identified previously (9). The phylogenetic pattern implicates independent spillover from rodent hosts as the major driver of Lassa fever incidence during the outbreak (Fig. 1 and fig. S8).

However, a number of sequences from the 2018 outbreak clustered as pairs in the phylogenetic reconstructions, raising concerns over human-to-human transmission. We illustrate such cluster pairs in a Bayesian time-measured tree estimated from genotype II S (Fig. 2) and L segment sequences (fig. S9). These analyses resulted in highly similar evolutionary rate estimates for both segments (mean, ~1.2 × 10–3 substitutions per site per year) (Fig. 2 and figs. S9 and S10), in agreement with previous estimates (9). We used these rate estimates together with an estimate of the time between successive cases in a transmission chain to assess how many substitutions can be expected between directly linked infections. We compared conservative to more liberal expectations, the latter accommodating an independent upper estimate of potential sequencing errors (Fig. 2 and fig. S9). In the S segment, for example, more than two substitutions between sequences from directly linked infections is highly unlikely ($P<0.01$ and $P=0.03$, respectively, for the conservative and liberal probability estimates). This expectation is consistent with the low number of substitutions observed in the coding region of human-to-human LASV transmission (14). Four clusters of sequences showing ≤4 and ≤12 nucleotide differences in the S and Lsegments, respectively, were identified (035-045, 035-058, 137-138, and 053-089-106; for some of them, only the S or L segment sequence was available). Retrospective tracing revealed that the sequences for pairs 137-138 and 035-058 were derived from the same patients. Epidemiological investigation of the remaining clusters did not provide evidence for transmission chains, though direct linkage cannot be excluded. Even when applying liberal assumptions for the number of mutations during human-to human transmission, the vast majority of cases during the 2018 outbreak resulted from spillover from the natural reservoir.

Fig. 1 Phylogenetic reconstruction of the S segment data.

The circular tree includes 96 sequences from 2012 to 2017, 88 sequences from 2018, and sequences available from GenBank. The rectangular tree focuses on the genotype II clade (in blue in the circular tree), which includes most of the 2018 sequences. The six genotypes are indicated with different colors and roman numerals. Bootstrap support >90% is indicated with a small gray circle at the middle of their respective branches. The color strip highlights the human LASV sequences obtained from previous years (light gray); sequences obtained from rodent samples (dark gray); and, for 2018, the first seven sequences generated in Nigeria (light pink), the remaining 28 sequences analyzed on-site (medium pink), and the remaining sequences finalized in Europe (dark pink). The same color code is used in the genotype II rectangular tree. Bootstrap values >80% are shown for the major genotype II lineages.

Fig. 2 Assessing the potential for direct linkage between pairs of 2018 sequences in the S segment. The maximum clade credibility tree summarizes a Bayesian evolutionary inference for the genotype II sequences in the S segment. A time scale and a marginal posterior distribution for the time to the most recent common ancestor are shown to the left. The size of the internal node circles reflects posterior probability support values. 2018 sequences clustering as pairs are indicated in dark pink; the number of substitutions between them is indicated at their respective tips. A posterior estimate of the evolutionary rate and probability distributions for observing a given number of substitutions during a human-to-human transmission event are shown as insets. The distribution represented by gray bars is based on the mean evolutionary rate estimate and a mean estimate for the generation time, whereas the light blue distribution is based on upper estimates and also incorporates an upper estimate for the MinION sequencing error (supplementary methods). At the bottom of the tree, clusters of sequences for which human-to-human transmission cannot be excluded according to the upper estimates of generation time are indicated. A pair of identical sequences (137-138) that was retrospectively found to be derived from the same patient is marked with a gray box. One pair (096-115) was disregarded as a potential transmission chain because of 21 differences in the L segment (fig. S9). The temporal signal before BEAST inference was explored in fig. S10.

Supplementary

Fig. S1. Epidemiology of the Lassa fever outbreak and timeline of sequencing in Nigeria.

(A) Epidemiological curve for 2018. ISTH confirmed 341 of the 376 Lassa fever cases reported by Nigeria Centre of Disease Control (NCDC) between 1st January and 18th March 2018. The epidemiological curve shows the 341 confirmed cases according to patient outcome. (B) Number of cases diagnosed and reported by ISTH from 2015 through 2018. (C) Number of samples sequenced per epidemiological week in 2018. (D) Timeline of sequencing efforts. Equipment and consumables for sequencing of ~50 samples and the computer hardware were deployed at ISTH with the aim of testing and troubleshooting on-site sequencing capacity. Sequencing data were requested by the NCDC on the 28th of February. The alarming increase in cases effectuated an upscale in efforts leading to sequencing of 120 samples on-site.

Fig. S2. Annotated map of confirmed Lassa fever cases between 1st January and 18th March 2018.

(A) Affected States; (B-C) geographical origin of patients from whom samples were sequenced (orange markers).

Fig. S3. Correlation between Ct values from Altona and Nikisins real-time RT-PCR assays.

Seven samples tested negative in the Nikisins assay and one tested negative in the Altona assay, demonstrating the importance of combined use of both assays for diagnosis of acute Lassa fever and subsequent sequencing. Negative results have been assigned a Ct value of 45 to facilitate visualization. Values of samples from survivors are plotted in grey, and those from deceased patients in black.

Fig. S4. Workflow of consensus sequence generation.

Summary of the steps performed during the bioinformatics pipeline for consensus generation.

Fig. S5. Percentage of reads mapping to LASV depending on Ct value in Altona and Nikisins real-time RT-PCR assay.

Negative results have been assigned a Ct value of 45 to facilitate visualization. Values of samples from survivors are plotted in grey and those from deceased patients in black.

Fig. S6. Percentage of genome coverage (20×) per segment depending on Ct value in Altona and Nikisins real-time RT-PCR assay.

Negative results have been assigned a Ct value of 45 to facilitate visualization. Values of samples from survivors are plotted in grey and those from deceased patients in black.

Fig. S7. Classification of MinION reads depending on Ct value in (A) Altona and (B) Nikisins real-time RT-PCR assay.

Reads were classified by Centrifuge software as either *Arenaviridae* or other viruses. The analysis allowed for identification of a coinfection in sample 110 with 0.1% reads classifying as Hepatitis A virus. In all other samples, the distribution of reads classified within the other viruses did not include a sufficient proportion of specific origin to suggest the presence of a virus other than LASV. *na*, not applicable as samples were not tested with the respective RT-PCR.

Fig. S8. Phylogenetic reconstruction of the L segment data.

The circular tree includes 64 new sequences from 2012 to 2017 (PRJNA482054 and PRJNA482058), 79 new sequences from 2018 (PRJNA482058), and sequences available from GenBank. The rectangular tree focuses on the genotype II clade (in blue in the circular tree) which includes most of the 2018 sequences. In the circular tree, the 6 genotypes are indicated with different colors and roman numerals. Bootstrap support >90% is indicated with a small grey circle at the middle of their respective branches. The color strip highlights the human LASV sequences obtained from previous years (light grey), sequences obtained from rodent samples (dark grey) and 2018 sequences (light pink/medium pink/dark pink). The first 7 sequences generated and analyzed in Nigeria are represented by a light pink color. The additional 28 sequences that were analyzed on-site are marked with a medium pink. All other 2018 sequences analyzed upon return to Europe are marked in dark pink. The same color code is used in the genotype II rectangular tree. Bootstrap values >80% are shown for the major lineages.

Fig. S9. Assessing the potential for direct linkage between pairs of 2018 sequences in the L segment.

The maximum clade credibility tree summarizes a Bayesian evolutionary inference for the genotype II sequences in the L segment. A time scale and a marginal posterior distribution for the time to the most recent common ancestor are shown to the left. The size of the internal node circles reflects posterior probability support values. Sequences clustering as pairs are indicated in dark pink; the number of substitutions between them is indicated at their respective tips. A summary for the posterior estimate of the evolutionary rate as well as the probability distributions for observing a number of substitutions for directly linked infections are shown as inset. The distribution represented by grey bars is based on the mean evolutionary rate estimate and a mean estimate for the generation time whereas the light blue distribution is based on upper estimates and also incorporates an upper estimate for the MinION sequencing error (Supplementary Methods). At the bottom, a pair of sequences (035/058) that was retrospectively found to be derived from the same patient is marked with a grey box. For one pair of sequences with four substitutions (053-106), a link cannot be excluded. One pair (096-115), for which direct linkage could not be excluded based on the S segment data with 2 nucleotide differences (Fig. 2), was disregarded as potential transmission chain due to 21 differences in L segment. The temporal signal was explored prior to BEAST inference (Fig. S10).

Fig. S10. Root-to-tip divergence of LASV genotype II sequences as a function of sampling time for the S (A-B) and L (C-D) segments.

The temporal signal was explored prior to the BEAST inference using regression analysis, i.e. root-to-tip divergence as a function of sampling time. (A) Complete genotype II data set (*n* = 238) and (B) Major subclade (*n* = 202, 85%) of the S segment. (C) Complete genotype II data set (*n* = 206) and (D) Major subclade (*n* = 176, 85%) of the L segment.

Table S1. Comparison between Nanopore and Illumina consensus sequences.

A total of 14 samples were randomly selected for re-sequencing using Illumina technology. Nucleotide disagreements between the Illumina and Nanopore derived sequences are listed for each segment. Disagreements within the coding regions, which were used in phylogenetic analysis, are highlighted in red. Ten of 14 had zero differences in either S or L segment coding regions, whilst four had 1-3 nucleotide disagreements in total across the combined S and L coding regions. Visual inspection of these regions suggested the basecall was consistent with the read alignment for both the Illumina and Nanopore data and so do not appear to be the result of the extra "noise" within the Nanopore signal.

Captions Supplementary Data S1 and S2

Data S1. Metadata of the 120 samples sequenced.

Data S2. Identifiers and Bioproject numbers of the 2012 to 2017 LASV sequences.

Materials and Methods

Sample collection

Samples from suspected Lassa fever patients were routinely tested for the presence of Lassa virus (LASV) RNA at the Institute of Lassa Fever Research and Control (ILFRC) at Irrua Specialist Teaching Hospital (ISTH), Irrua, Edo State, Nigeria, using two real-time reverse transcription PCR (RT-PCR) assays, the commercially available Altona kit (RealStar® Lassa Virus RT-PCR Kit 1.0 CE, Altona Diagnostics, Hamburg, Germany) targeting the S segment along with an in-house version of the previously described Nikisins RT-PCR targeting the L segment *(17)*. The latter has been optimized by using the SuperScript™ III Platinum™ One-Step qRT-PCR reagents (Invitrogen) according to manufacturer instructions (without magnesium sulfate; reaction volume of 25 µl). The temperature profile was identical to that of the Altona assay, while primer and probe sequences and concentrations were used as described (*17*). Both Altona and Nikisins 15 real-time RT-PCR assays have been implemented and extensively evaluated in terms of analytical and clinical characteristics by the authors and were found to have good performance in diagnosing acute Lassa fever when used in combination. Therefore, all samples were generally tested in both assays, although the Cycle threshold (Ct) values obtained with the two assays may differ (Fig. S3). A total of 120 plasma, breast milk, or cerebrospinal fluid samples identified as LASV-positive by one or both real-time PCRs were selected for direct sequencing based on Ct value (inversely correlated with viral load) and/or geographical information on the sample origin. The use of diagnostic leftover specimen and corresponding patient data was approved by the ISTH Research and Ethics Committee (approval ISTH/HREC/20171208/45).

Nucleic acid extraction and metagenomic library preparation

Extraction and metagenomic library preparation were performed as described in detail previously (*11*). Briefly, 70 µl of each sample was manually extracted using the QIAamp viral RNA kit (Qiagen); nucleic acid extracts were then subjected to a DNAse digest (TURBO DNase, Thermo Fisher Scientific), randomly reverse-transcribed, and amplified using a Sequence Independent Single Primer Amplification (SISPA) approach.

MinION library preparation and sequencing

Barcoded MinION sequencing libraries were prepared using the Ligation Sequencing kit 1D (SQK-LSK108) and Native Barcoding Kit (EXP-NBD103) (Oxford Nanopore Technologies [ONT]). Up to 6 samples plus one negative control, consisting of a water blank sample included in each batch of extractions, were included per multiplex library. Libraries were sequenced for 48 h on FLO-MIN106 flow cells using a Mark 1B MinION device (Oxford Nanopore Technologies).

Data handling

An overview of the data analysis workflow used can be found in Fig. S4. Raw reads were basecalled using the ONT Albacore sequencing pipeline software v2.2.7, and output basecalled fastq files were concatenated and demultiplexed using Porechop v0.2.3 (https://github.com/rrwick/Porechop). SeqTK (https://github.com/lh3/seqtk) was then used to trim 30 bp from both ends to eliminate primer sequences and resulting fastq files were mapped to the human genome (human_g1k_v37; 1000 genomes). Mapped reads were excluded from the subsequent *de novo* assembly, which allowed for LASV reference identification. CANU v1.6 (*12*) was used for *de novo* assembly with the following settings: corOutCoverage=1000, genomeSize=10000, minReadLength=400, minOverlapLength=200. Canu genomeSize and minReadLength parameters were lowered for samples that did not assemble any LASV contigs with the specified values. Assemblies were then used in a blastn search against the NCBI database to identify the closest LASV reference genome available. BWA MEM v0.7.15 (*18*) using -x ont2d mode allowed for read alignment to the reference genome identified (see supplementary data file S1 for references used for each sample). Nanopolish (v0.9.0) variants (*5*, *19*) using --snps mode was used to detect single-nucleotide polymorphisms (SNPs) with respect to the reference genome and Nanopolish output vcf file was used as input to the margin_cons.py script (*6*) (https://github.com/zibraproject/zika-pipeline) to filter out low-quality or low-coverage candidate SNPs and compute a consensus. Reads were then re-aligned to the consensus and a second round of correction performed with consensus bases called at a minimum support fraction of 70%. Samtools v1.7 was used to compute percentage reads mapped along with coverage depth and Bedtools v2.27.1 was used to calculate genome coverage at 20×. Taxonomic classification of the data was performed using Centrifuge v1.0.4 (*13)* and the provided "Bacteria, Archaea, Viruses, Human (compressed)" indexes (update version 12/06/2016).

Illumina library preparation, sequencing, and analysis

Nextera XT v2 Kit (Illumina) sequencing libraries were prepared using 1 ng of SISPA amplified cDNA, according to the manufacturer's instructions, with a total of 14 cycles in the library amplification PCR. Samples were multiplexed in batches of maximum 12 per run and sequenced on a 2 × 300 bp Illumina MiSeq run. BWA MEM v0.7.15 (*18*) was used with default settings to align reads to the references. Mapping consensus sequences for Illumina were generated using QuasiBam (*20*).

Data set compilation, alignment, and reassortment/recombination analyses

The sequence data sets were assembled by combining the newly generated MinION sequences plus Illumina controls from 2018 (NCBI Bioproject PRJNA482058), new sequences generated from 2012-2017 LASV isolates (PRJNA482054 and PRJNA482058; D. U. Ehichioya, unpublished), and LASV genomic sequences available on GenBank. Only sequences with less than 30% missing bases per segment were included, which resulted in a total of 352 sequences for the L segment (79 MinION sequences from 2018 and 14 Illumina controls, and 64 new sequences from previous years) and 425 sequences for the S segment (88 MinION sequences from 2018 and 14 Illumina controls, and 96 new sequences from previous years). The L and S segment alignments were compiled separately by concatenating the Z and L (polymerase) gene sequences and the glycoprotein and nucleoprotein gene sequences, respectively, and aligning them using Muscle (*21*). Ambiguously aligned regions in the polymerase gene of the L segment were removed. Potential segment reassortment and phylogenetic inconsistency within each segment was examined using RDP4 (*22*) based on a significant result for more than three of the following recombination detection methods: RDP, GENECONV, Chimaera, MaxChi, Bootscan, SiScan, and 3Seq. We used a 0.05 as highest acceptable *P* value for each method and a Bonferroni correction for multiple testing. Among the strains for which both an L and S segment sequence was available, this analysis revealed 7 potential reassortants (5 among previously obtained genomes and two new ones). Relative short phylogenetically inconsistent stretches in the L (*n* = 3) and S (*n* = 1) segment 5 sequences identified using the same procedure were masked as unobserved characters ('N') prior to phylogenetic analyses and BEAST inference.

Maximum likelihood phylogenetic reconstruction and Bayesian time-measured phylogenetic inference.

We used RAxML (*23*) to infer maximum likelihood (ML) phylogenetic trees under a general time-reversible (GTR) substitution model with gamma-distributed among-site rate heterogeneity. Upon evaluating node support using 1000 bootstrap replicates, we employed a thorough tree search using relatively exhaustive subtree pruning and regrafting (SPR) moves to search for the ML tree.

We used plots of root-to-tip divergence as a function of sampling time summarized by TempEst (*24*) from the ML trees to examine the temporal signal in the genotype II L & S data sets (prior to fitting a dated tip model in subsequent Bayesian analyses). This revealed that the pattern of divergence accumulation over the sampling time range was obfuscated by rate variation between the relatively divergent sub-clusters within genotype II for the S segment (Fig. S10), as was previously observed in different viral example with a similar time to the most recent common ancestor (*25*). Focusing on the major sub-cluster in this genotype however (*n* = 202, 85%), allows identifying a reasonable temporal signal (Fig. S10). The rate variability along deep branches appears to be less pronounced in the L segment, but in this case, there is considerably more root-to- tip divergence variability for the 2018 genotype II sequences (both in the complete genotype II data set and the major sub-cluster, Fig. S10).

Bayesian time-measured evolutionary histories were reconstructed for the genotype II sequences using BEAST v1.10 (*26*). Specifically, we estimated a posterior distribution of time measured trees for the genotype II S segment data set. Identical sequences from previous years were reduced to a single representative sequence. We specified six partitions, one for each codon position in both the glycoprotein and nucleoprotein genes (constraining relative substitution rates to sum to 3 separately in both genes) and a separate GTR model of substitution with gamma distributed rate variation among sites for each partition, but with hierarchical prior distributions over the different GTR substitution rates to share information across partitions (*27*). We used a flexible Bayesian skyride tree prior (*28*) and an uncorrelated lognormal relaxed 5 molecular clock model to allow for rate variation among lineages (*29*). When exact sampling dates were not available, tip ages were integrated over their appropriate uncertainty (month or year). In addition to standard Markov-chain Monte Carlo (MCMC) transition kernel, we employed an adaptive multivariate normal kernel on the substitution model parameters (*30*). Multiple independent MCMC chains were run until the continuous parameters in the combined posterior sample achieved sufficiently high effective sample sizes (ESSs >100). We summarized continuous parameters using mean estimates and 95% highest posterior density (HPD) intervals. Trees were summarized as a maximum clade credibility (MCC) trees using TreeAnnotator. Alignments, ML trees, BEAST xml files, and MCC trees are available at: https://github.com/ISTH-BNITM-PHE/LASVsequencing (DOI: 10.5281/zenodo.1481015) (*16*).

Assessing the potential for direct linkage among the 2018 samples

To investigate whether pairs of 2018 sequences could represent directly linked infections (human-to-human transmission), we calculated the Poisson probability distribution to observe a number of substitutions conditioning on (i) the BEAST estimate of the mean nucleotide substitution rate and (ii) an estimate of the Lassa fever generation time (~ the time between successive cases in a transmission chain). For the latter, we used three weeks based on a mean estimate of 10 days for the incubation period and a mean time to hospital presentation after disease onset of 8 to 10 days. So, we obtain the rate parameter ($\lambda 1$) for the Poisson probability distribution by dividing the substitution rate (per genome segment per time unit) by the generation time in the same time units. We also calculated a second version of this probability distribution taking caution not to reject direct linkage for large numbers of substitutions. To this purpose, we calculated a Poisson rate parameter ($\lambda 2$) based on the 95% upper HPD interval estimate for the rate of evolution and a longer generation time of 4 weeks. In addition, we incorporated a liberal estimate of the basecalling differences observed between the subset of MinION sequences and the corresponding Illumina controls by fitting a Poisson distribution to the substitutions to estimate the sequencing error rate (λe). For the second more liberal version of the Poisson probability distribution we thus take as rate the sum of the per-generation substitution rate parameter ($\lambda 2$) and twice the sequencing error (λe) because we model differences between two sequences generated 5 using MinION sequencing

References

1. J. D. Frame, J. M. Baldwin Jr, D. J. Gocke, J. M. Troup, Lassa fever, a new virus disease of man from West Africa. I. Clinical description and pathological findings. Am. J. Trop. Med. Hyg. 19, 670–676 (1970).

2. D. A. Asogun et al., Molecular diagnostics for lassa fever at Irrua specialist teaching hospital, Nigeria: lessons learnt from two years of laboratory operation. PLoS Negl. Trop. Dis. 6, e1839 (2012).

3. WHO | Lassa Fever – Nigeria (2018) (available at http://www.who.int/csr/don/23-march- 2018-lassa-fever-nigeria/en/).

4. M. Jain, H. E. Olsen, B. Paten, M. Akeson, The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 17, 239 (2016).

5. J. Quick et al., Real-time, portable genome sequencing for Ebola surveillance. Nature. 530, 228–232 (2016).

6. J. Quick et al., Multiplex PCR method for MinION and Illumina sequencing of Zika and 15 other virus genomes directly from clinical samples. Nat. Protoc. 12, 1261–1276 (2017).

7. N. R. Faria et al., Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature. 546, 406–410 (2017).

8. N. R. Faria et al., Genomic and epidemiological monitoring of yellow fever virus transmission potential. bioRxiv (2018), p. 299842.

9. K. G. Andersen et al., Clinical Sequencing Uncovers Origins and Evolution of Lassa Virus. Cell. 162, 738–750 (2015).

10. A. L. Greninger et al., Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome Med. 7, 99 (2015).

11. L. E. Kafetzopoulou et al., Assessment of Metagenomic MinION and Illumina sequencing as an approach for the recovery of whole genome sequences of chikungunya and dengue viruses directly from clinical samples Euro Surveill. 23, 1800228 (2018).

12. S. Koren et al., Canu: scalable and accurate long-read assembly via adaptive -mer weighting and repeat separation. Genome Res. 27, 722–736 (2017).

13. D. Kim, L. Song, F. P. Breitwieser, S. L. Salzberg, Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 26, 1721–1729 (2016).

14. Whitmer SLM, Strecker T, Cadar D, Dienes HP, Faber K, Patel K, Brown SM, Davis WG, Klena JD, Rollin PE, Schmidt-Chanasit J, Fichet-Calvet E, Noack B, Emmerich P, Rieger T, Wolff S, Fehling SK, Eickmann M, Mengel JP, Schultze T, Hain T, Ampofo W, Bonney K, Aryeequaye JND, Ribner B, Varkey JB, Mehta AK, Lyon GM 3rd, Kann G, De Leuw P, Schuettfort G, Stephan C, Wieland U, Fries JWU, Kochanek M, Kraft CS, Wolf T, Nichol ST, Becker S, Ströher U, Günther S. New Lineage of Lassa Virus, Togo, 2016. Emerg Infect Dis. 2018 24(3):599-602. (2018)

15. Nigeria Centre for Disease Control, (available at https://ncdc.gov.ng/news/121/early-resultsSubmitted of-lassa-virus-sequencing-%26-implications-for-current-outbreak-response-in nigeria).

16.P.Lemey, ISTH-BNITM-PHE/LASVsequencing: LASVrelease, Zenodo (2018); http://doi.org/10.5281/zenodo.1481015.

17. S. Nikisins, T. Rieger, P. Patel, R. Müller, S. Günther, M. Niedrig, International external quality assessment study for molecular detection of Lassa virus. *PLOS Negl. Trop. Dis.* **9**, e0003793 (2015).

18. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN] (26 May 2013).

19. N. J. Loman, J. Quick, J. T. Simpson, A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* **12, 733–735** (2015).

20. A. R. Penedos, R. Myers, B. Hadef, F. Aladin, K. E. Brown, Assessment of the Utility of Whole Genome Sequencing of Measles Virus in the Characterisation of Outbreaks. *PLOS ONE* **10**, e0143081 (2015).

21. R. C. Edgar, MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).

22. D. P. Martin, B. Murrell, M. Golden, A. Khoosal, B. Muhire, RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol.* **1**, vev003 (2015).

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

24. A. Rambaut, T. T. Lam, L. Max Carvalho, O. G. Pybus, Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016).

25. N. S. Trovão, G. Baele, B. Vrancken, F. Bielejec, M. A. Suchard, D. Fargette, P. Lemey, Host ecology determines the dispersal patterns of a plant virus. *Virus Evol.* **1**, vev016 (2015).

26. M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, A. Rambaut, Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).

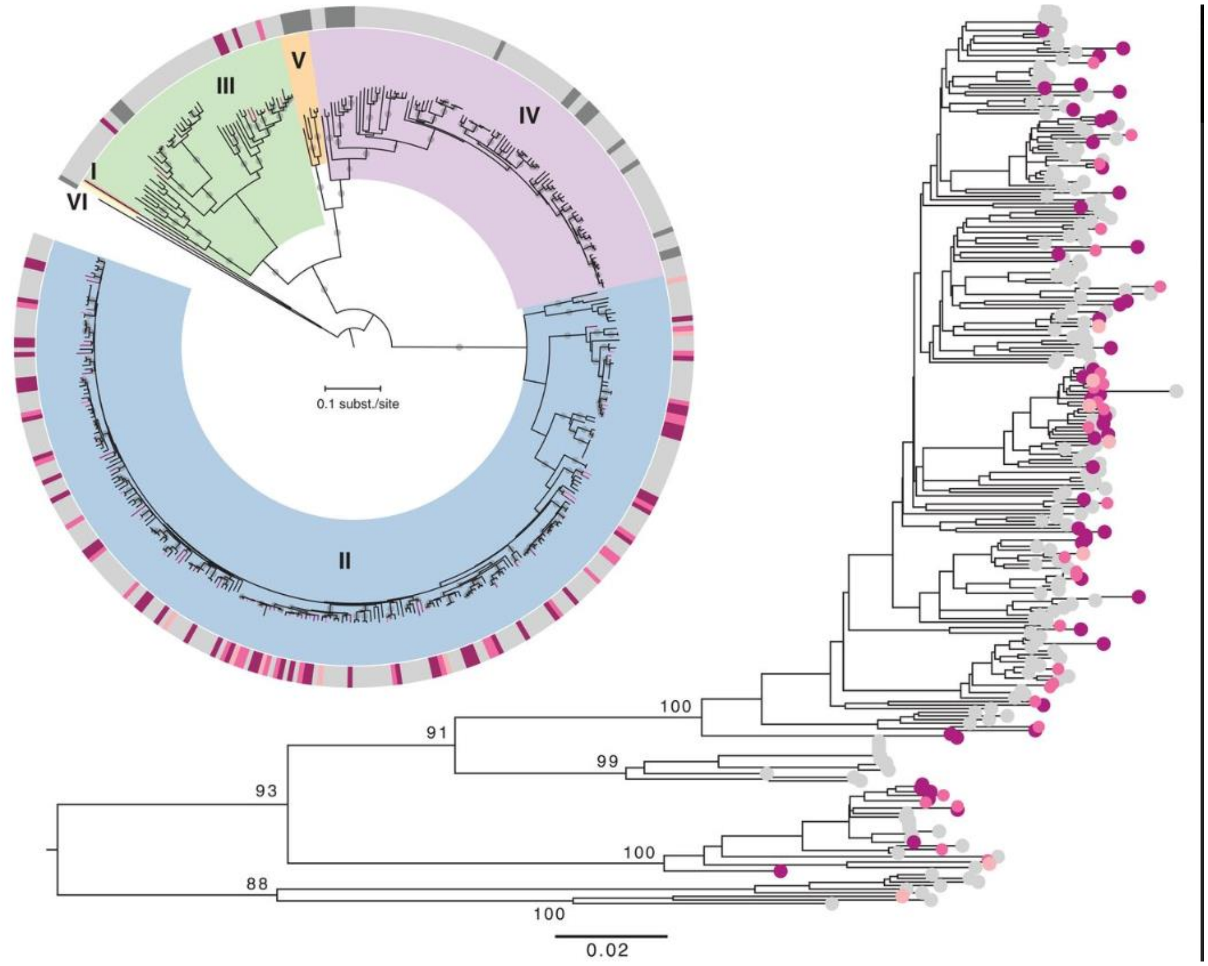
27. D. Edo-Matas, P. Lemey, J. A. Tom, C. Serna-Bolea, A. E. van den Blink, A. B. van 't Wout, H. Schuitemaker, M. A. Suchard, Impact of CCR5delta32 host genetic background and disease progression on HIV-1 intrahost evolutionary processes: Efficient hypothesis testing through hierarchical phylogenetic models. *Mol. Biol. Evol.* **28**, 1605–1616 (2011).

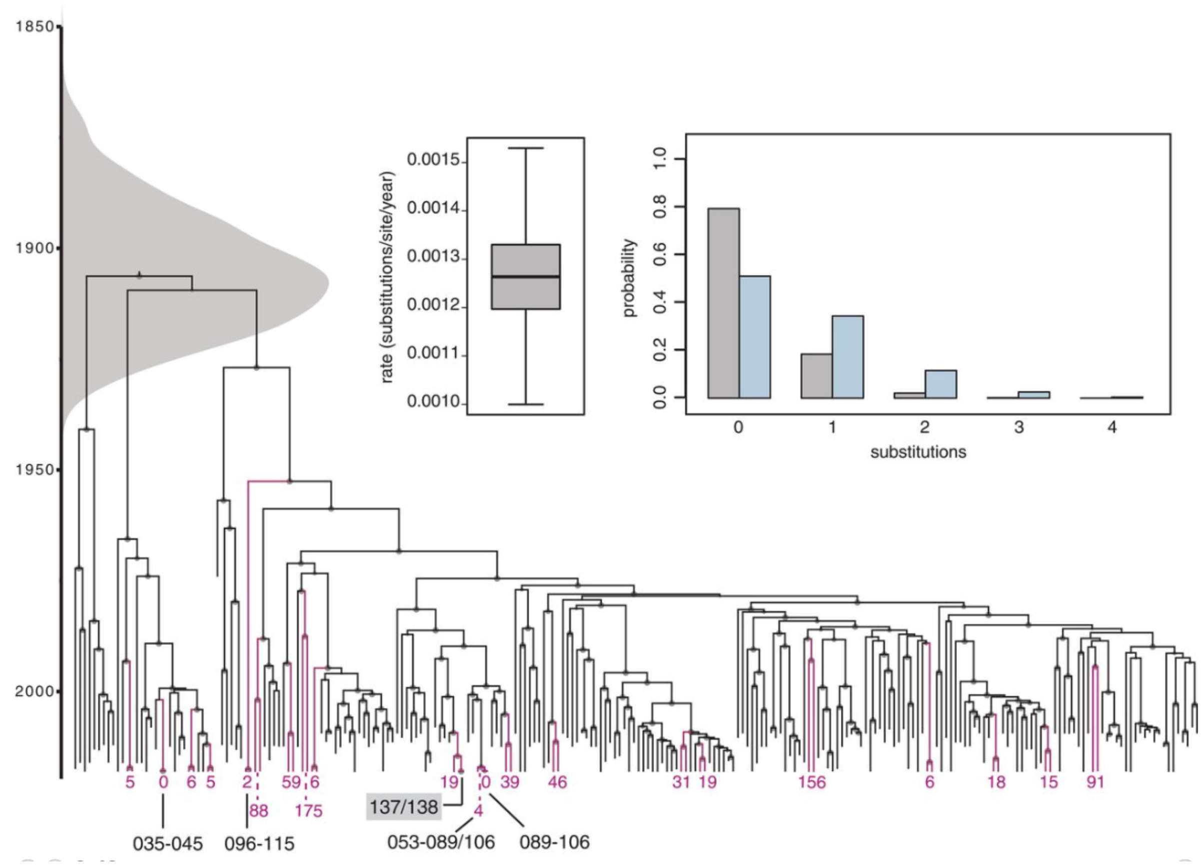
28. V. N. Minin, E. W. Bloomquist, M. A. Suchard, Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol.* **25**, 1459–1471 (2008).

29. A. J. Drummond, S. Y. W. Ho, M. J. Phillips, A. Rambaut, Relaxed phylogenetics and dating with confidence. *PLOS Biol.* **4**, e88 (2006).

30. G. Baele, P. Lemey, A. Rambaut, M. A. Suchard, Adaptive MCMC in Bayesian phylogenetics: An application to analyzing partitioned data in BEAST. *Bioinformatics* **33**, 1798–1805 (2017).

Figures





Supplementary figs

Figure S1

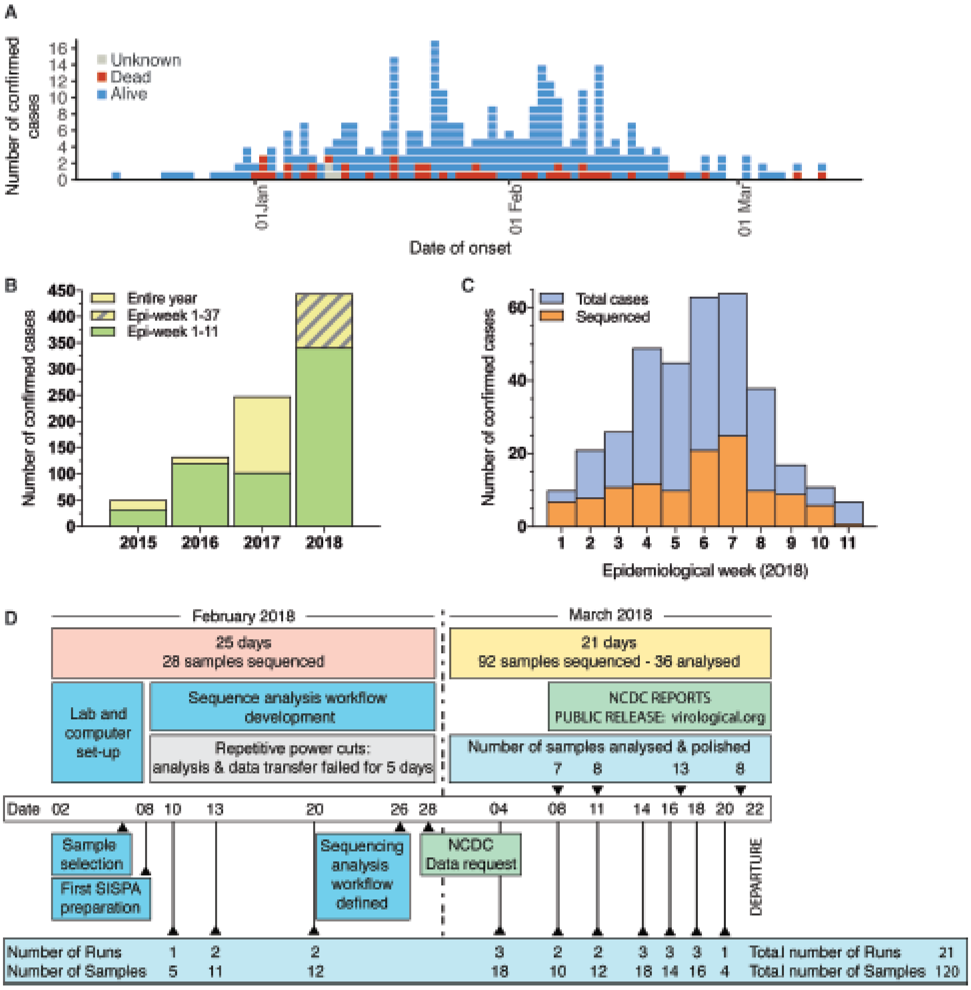


Figure S2

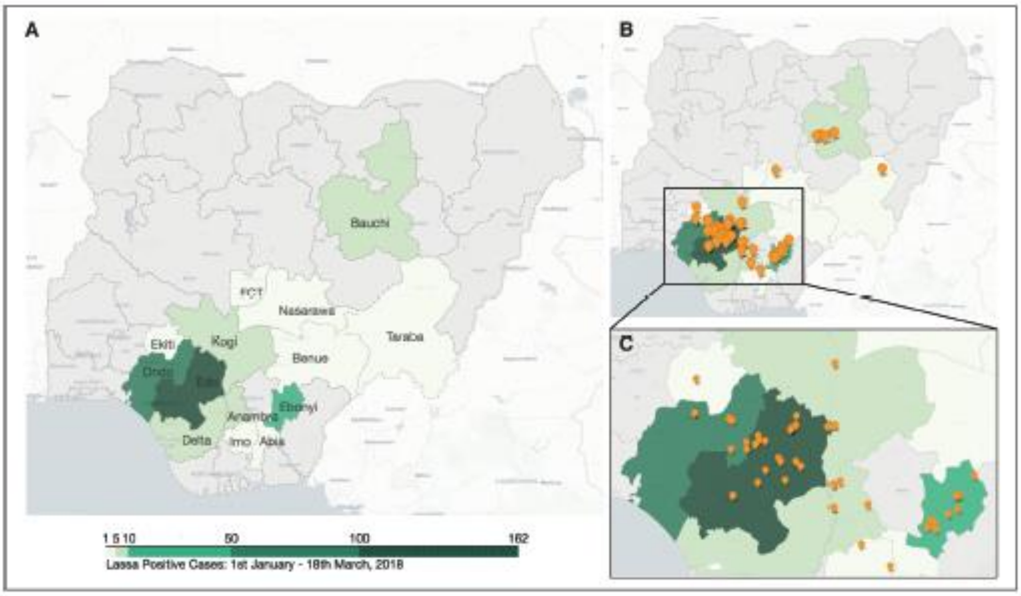


Figure S3

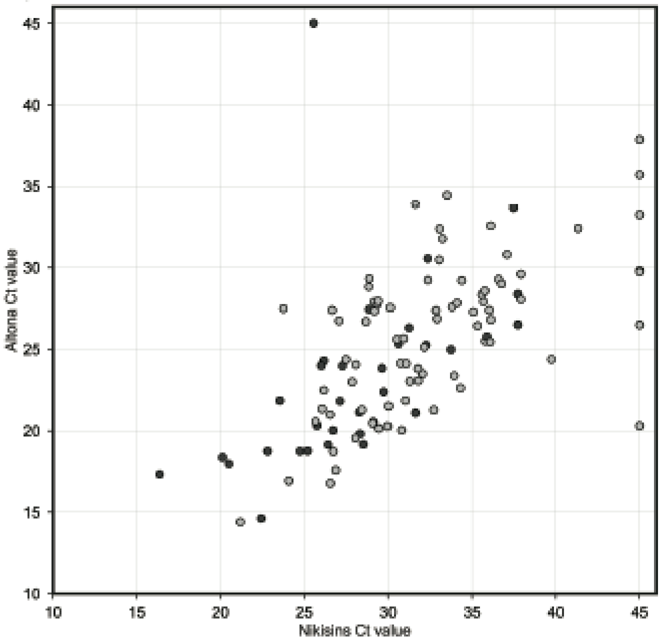
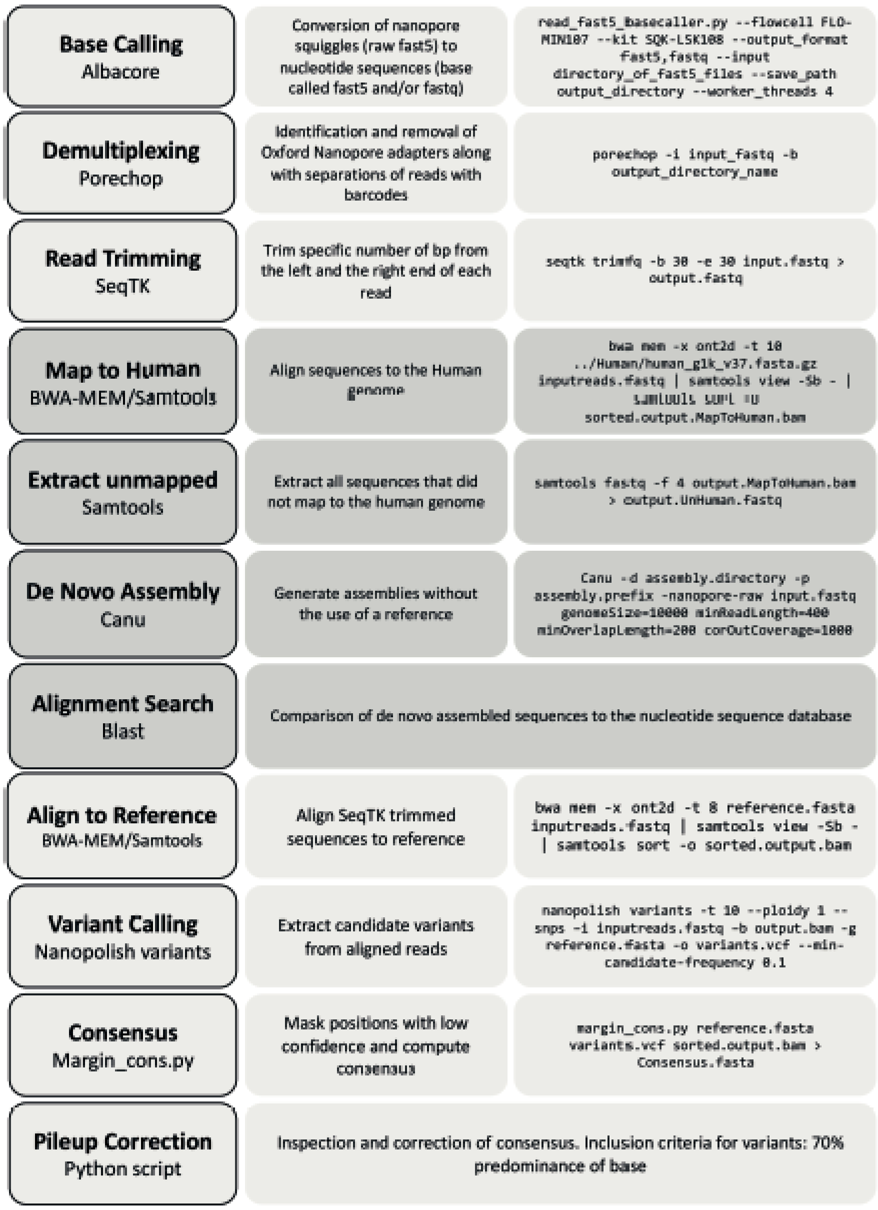

Altona Ct value
Nikisins Ct value
10
15
20
25
30
35
40
45
10
15
20
25
30
35
40
45

Figure S4



Base Calling Albacore	Conversion of nanopore squiggles (raw fast5) to nucleotide sequences (base called fast5 and/or fastq)	`read_fast5_basecaller.py --flowcell FLO-MIN107 --kit SQK-LSK108 --output_format fast5,fastq --input directory_of_fast5_files --save_path output_directory --worker_threads 4`
Demultiplexing Porechop	Identification and removal of Oxford Nanopore adapters along with separations of reads with barcodes	`porechop -i input_fastq -b output_directory_name`
Read Trimming SeqTK	Trim specific number of bp from the left and the right end of each read	`seqtk trimfq -b 30 -e 30 input.fastq > output.fastq`
Map to Human BWA-MEM/Samtools	Align sequences to the Human genome	`bwa mem -x ont2d -t 10 ../Human/human_g1k_v37.fasta.gz inputreads.fastq | samtools view -Sb - | samtools sort -o sorted.output.MapToHuman.bam`
Extract unmapped Samtools	Extract all sequences that did not map to the human genome	`samtools fastq -f 4 output.MapToHuman.bam > output.UnHuman.fastq`
De Novo Assembly Canu	Generate assemblies without the use of a reference	`Canu -d assembly.directory -p assembly.prefix -nanopore-raw input.fastq genomeSize=10000 minReadLength=400 minOverlapLength=200 corOutCoverage=1000`
Alignment Search Blast	Comparison of de novo assembled sequences to the nucleotide sequence database	
Align to Reference BWA-MEM/Samtools	Align SeqTK trimmed sequences to reference	`bwa mem -x ont2d -t 8 reference.fasta inputreads.fastq | samtools view -Sb - | samtools sort -o sorted.output.bam`
Variant Calling Nanopolish variants	Extract candidate variants from aligned reads	`nanopolish variants -t 10 --ploidy 1 --snps -i inputreads.fastq -b output.bam -g reference.fasta -o variants.vcf --min-candidate-frequency 0.1`
Consensus Margin_cons.py	Mask positions with low confidence and compute consensus	`margin_cons.py reference.fasta variants.vcf sorted.output.bam > Consensus.fasta`
Pileup Correction Python script	Inspection and correction of consensus. Inclusion criteria for variants: 70% predominance of base	

Figure S5

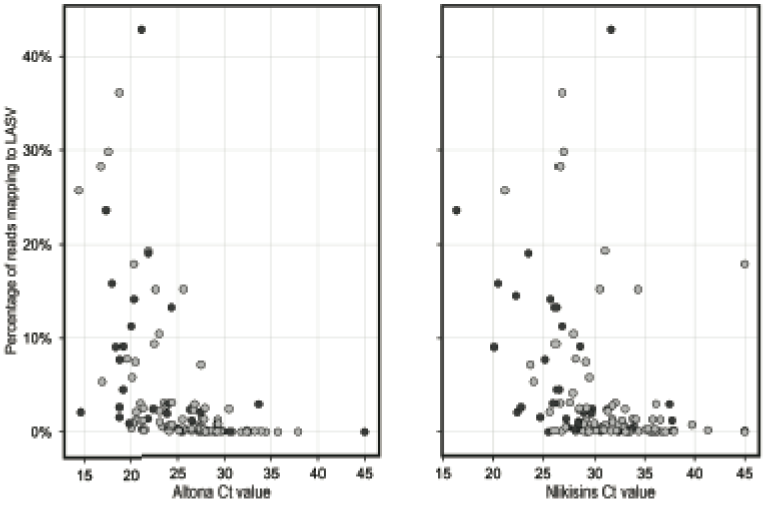


Figure S6

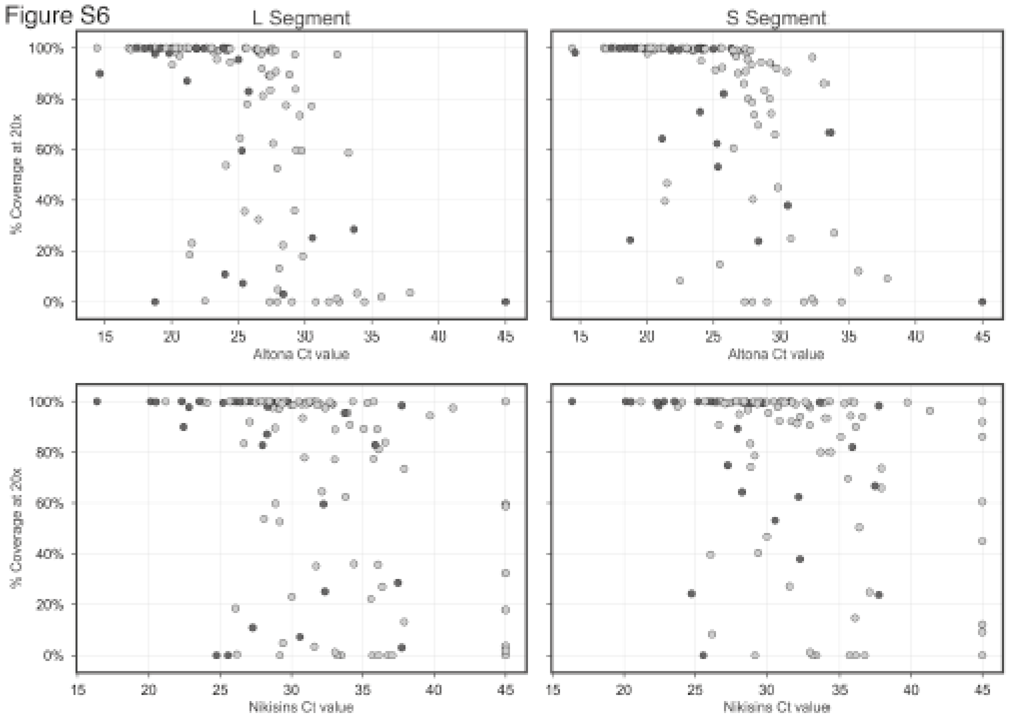


Figure S7

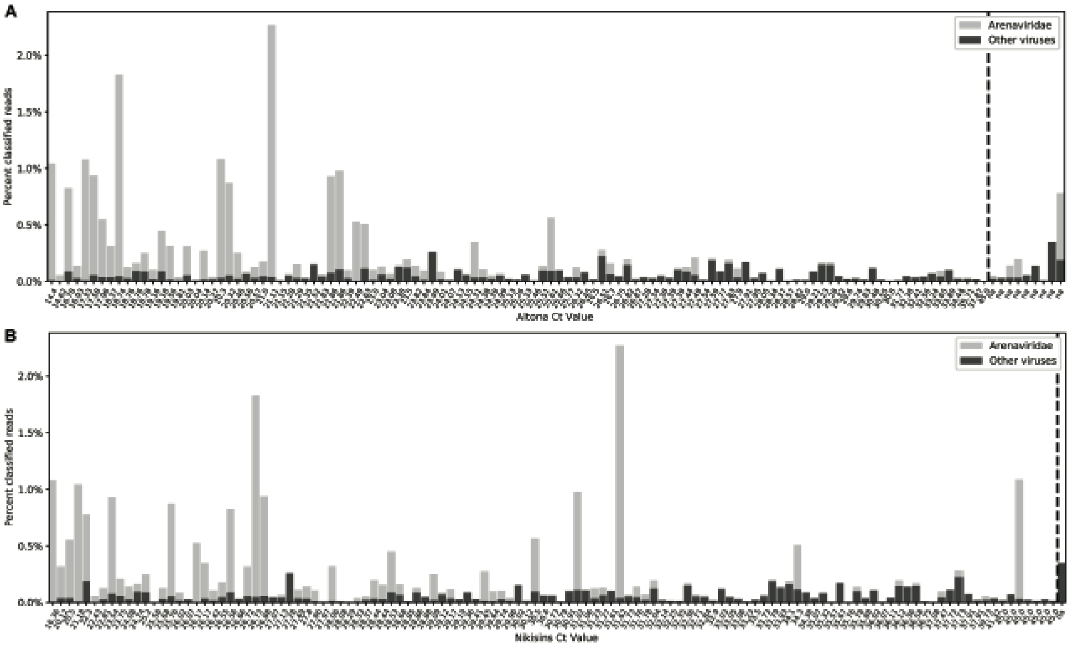


Figure S8

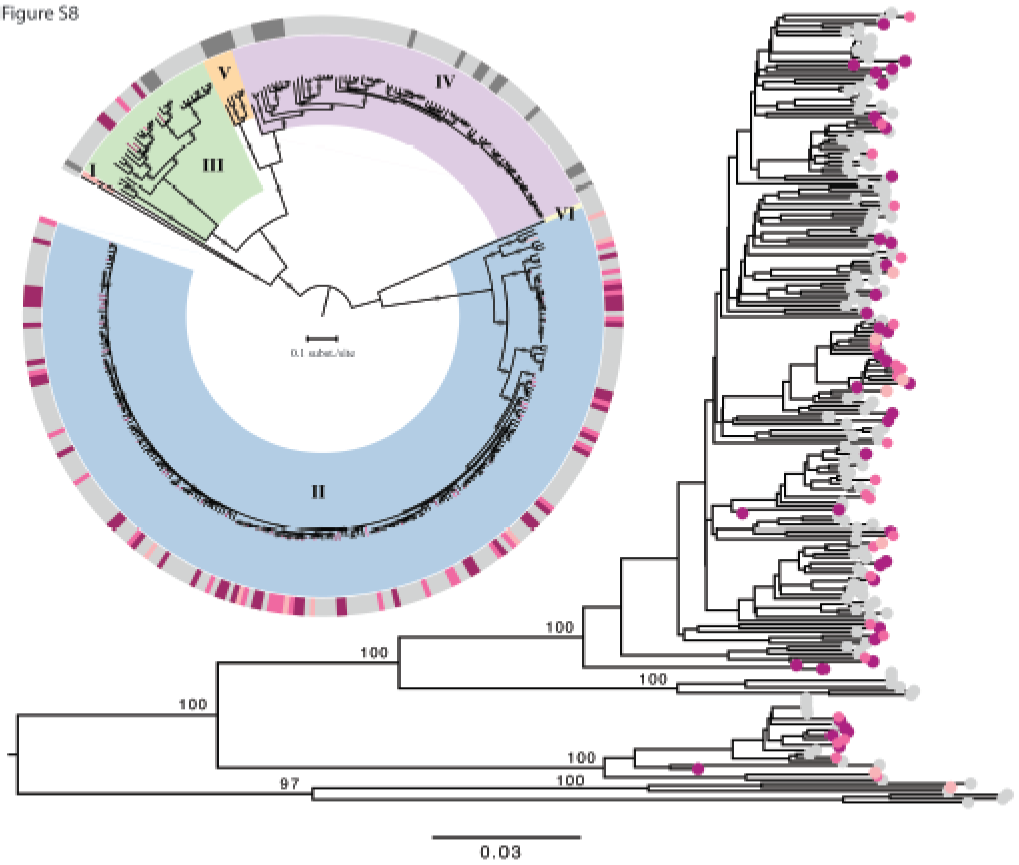


Figure S9

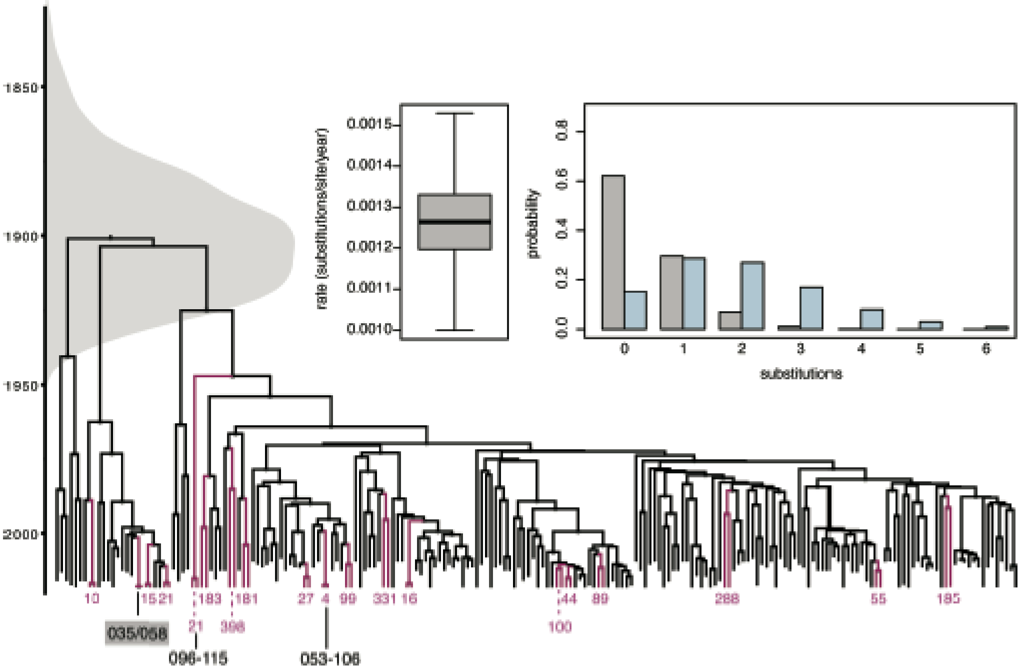
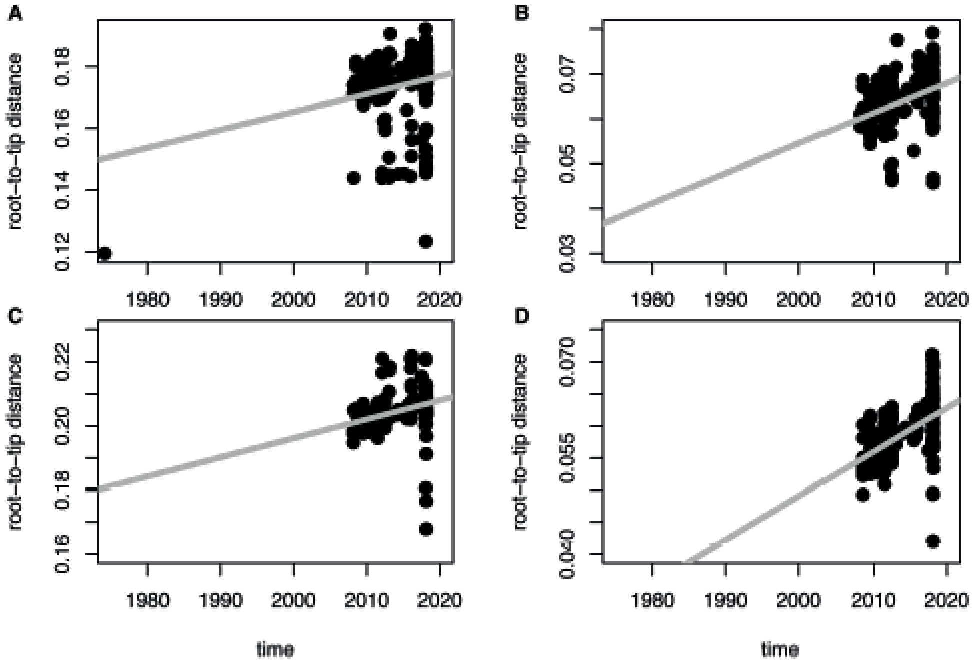


Figure S10


A
B
C
D
root-to-tip distance
time