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

Authors: L. E. Kafetzopoulou[1,2,3], S. T. Pullan[1,2], P. Lemey[4], M. A. Suchard[5], D. U. Ehichioya[3,6], M. Pahlmann[3,6], A. Thielebein[3,6], J. Hinzmann[3,6] , L. Oestereich[3,6], D. M. Wozniak[3,6], K. Efthymiadis[7], D. Schachten[3], F. Koenig[3], J. Matjeschk[3], S. Lorenzen[3], S. Lumley[1], Y. Ighodalo[8], D. I. Adomeh[8], T. Olokor[8], E. Omomoh[8], R. Omiunu[8], J. Agbukor[8], B. Ebo[8], J. Aiyepada[8], P. Ebhodaghe[8], B. Osiemi[8], S. Ehikhametalor[8], P. Akhilomen[8], M. Airende[8], R. Esumeh[8], E. Muoebonam[8], R. Giwa[8], A. Ekanem[8], G. Igenegbale[8], G. Odigie[8], G. Okonofua[8], R. Enigbe[8], J. Oyakhilome[8], E. O. Yerumoh[8], I. Odia[8], C. Aire[8], M. Okonofua[8], R. Atafo[8], E. Tobin[8], D. Asogun[8,9], N. Akpede[8], P. O. Okokhere[8,9], M. O. Rafiu[8], K. O. Iraoyah[8], C. O. Irolagbe[8], P. Akhideno[8], C. Erameh[8], G. Akpede[8,9], E. Isibor[8], D. Naidoo[10], R. Hewson[1,2], J. A. Hiscox[2,11,12], R. Vipond[1,2], M. W. Carroll[1,2], C. Ihekweazu[13], P. Formenty[10], S. Okogbenin[8,9], E. Ogbaini-Emovon[8]†, S. Günther[3,6,*]†, S. Duraffour[3,6]†.

Affiliations:

[1] Public Health England, National Infections Service, Porton Down, UK.

[2] National Institute of Health Research (NIHR), Health Protection Research Unit in Emerging and Zoonotic Infections, University of LiverpooI, 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, USA.

[6] German Center for Infection Research (DZIF), partner site Hamburg-Lübeck-Borstel-Riems, 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]Singapore Immunology Network, Agency for Science, Technology and Research (A*STAR), Singapore.

[12]Institute of Infection and Global Health, University of Liverpool, Liverpool L69 7BE, UK.

[13] Nigeria Centre for Disease Control, Abuja, Nigeria.

*Corresponding author: Stephan Günther, Bernhard-Nocht-Institute for Tropical Medicine, Bernhard-Nocht-Str. 74, 20359 Hamburg, Germany, Phone: +49 40 42818 940, Fax: +49 40 42818 931, email: guenther@bni.uni-hamburg.de

† These authors contributed equally to this work

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 WHO to inform the public health response. Real-time analysis of 36 genomes, and subsequent confirmation using all 120 samples sequenced in-country, revealed extensive diversity and phylogenetic intermingling with strains from previous years, suggesting independent zoonotic transmission events; allaying concerns of an emergent strain or extensive human-to-human transmission.

Main Text:

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 1,495 suspected cases, 376 confirmed cases, and affecting over 18 states by March 18th (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. Due to these concerns, on February 28th the Nigeria Centre for Disease Control (NCDC) and the WHO urgently requested sequencing information and preliminary results from our pilot-scale study, which employed in-country, mid-outbreak, viral genome sequencing directly from clinical samples using a metagenomic approach on the Oxford Nanopore MinION device (Oxford Nanopore Technologies, UK). This instigated a major upscale in sequencing efforts, leading to sequencing of 120 samples.

Nanopore sequencing is an emerging technology with significant potential. The MinION is a small and robust sequencing device suited for the genetic analysis of pathogens in remote or resource-limited settings (*4*). Nanopore sequencing of 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 to great effect during recent Zika and Yellow fever outbreaks in Brazil (*7*, *8*). However such an amplicon-based approach is extremely challenging for highly variable pathogens such as LASV. Due to an inter-strain 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 significant challenge. Designing targeted whole-genome sequencing approaches, such as PCR amplicons or bait/capture probes, without prior knowledge of the targeted LASV lineage is therefore cumbersome. Random reverse-transcription and amplification by Sequence-Independent Single Primer Amplification (SISPA) for metagenomic 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 seven-week mission, selected based on Cycle threshold (Ct) value and location of the 341 cases reported by ISTH between 1st January and 18th March 2018 (Figs. S1 and S2). The majority of samples originated from Edo state followed by Ondo and Ebonyi (Fig. S2). Samples selected covered the wide range of clinical viral loads observed, including several samples testing negative in one of the two real-time 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 basecalled 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, non-human reads from each sample were assembled *de novo* using Canu (*12*). A significant 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. S3-6).

Additionally, sequences were validated by Illumina re-sequencing of 14 SISPA preparations which matched with their Nanopore counterpart with little to no divergence between them confirming the accuracy of the 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% in the same sample, providing 96% genome coverage. This demonstrates 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 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 use this as a frame of reference to document how the genomic data generated in real-time (made publicly available as posted on virological.org) provided valuable epidemiological insights into the unfolding outbreak dynamics.

Maximum likelihood phylogenetic reconstruction of the S segment sequences indicates that all 2018 viruses fall within the Nigerian LASV diversity, specifically within genotype 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 complete genome data set (n = 348) were identified as clustering significantly differently in the L and S segment (Supplementary Methods), in line with the small number of potential LASV reassortments identified previously (*9*). The phylogenetic pattern clearly implicates independent spill-over from rodent hosts as the major driver of Lassa fever incidence during the outbreak (Figs. 1 and S8).

However, a number of sequences from the 2018 outbreak clustered as pairs in the phylogenetic reconstructions, raising concerns over human-to-human transmissions. 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 around $1.2\text{x}10^{-3}$ subst./site/year, Figs. 2 and S9-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 compare conservative to more liberal expectations, the latter accommodating an independent upper estimate of potential sequencing errors (Figs. 2 and S9). In the S segment, for example, more than 2 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 S and L segment, respectively, were identified (035-045, 035-058, 137-138, and 053-089-106; for some of them only S or L segment sequence was available). Retrospective tracing revealed that the sequences for pairs 137-138 and 035-058, respectively, were in fact derived from the same patients. Epidemiological investigation of the remaining clusters did not provide evidence for transmission chains, though direct linkage cannot be excluded. In conclusion, 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 spill-over from the natural reservoir.

A request for information on circulating strains was made on 28th February at the height of the outbreak, within 10 days our pilot research study was expedited, and the initial analysis completed. The fact that the 2018 outbreak was fueled by the circulating LASV diversity and not by transmission of a new or divergent lineage was already evident from the first seven genomes generated by 10th March (Fig. S1). This information was promptly communicated to the NCDC forming the basis of their report "Early Results of Lassa Virus Sequencing & Implications for Current Outbreak Response in Nigeria" released on March 12th 2018 (*15).* While this small sample was restricted to genotype II, the final collection of 36 LASV genome sequences generated on-site also included a representative of genotype III (Figs. 1 and S9), further supporting the spill-over of longstanding LASV diversity in the outbreak. The conclusions drawn from the first set of genome sequences immediately removed fears of extensive human-to-human transmission and allowed public health resources to be allocated appropriately. The response was focused on intensified community engagement on rodent control, environmental sanitation, and safe food storage. Further research is needed to evaluate if improved diagnostics and disease awareness and/or ecological and climate factors promoting transmission are the drivers behind the changing epidemiology of Lassa fever in Nigeria.

Portable metagenomic sequencing of genetically diverse RNA viruses on the MinION, direct from patient samples without the need to export material outside of the country of origin and with no pathogen-specific enrichment, is shown to be a feasible methodology enabling a real-time characterization of potential outbreaks in the field.

REFERENCES AND NOTES

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 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. Science. 361, 894-899 (2018)

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. bioRxiv (2018), p. 355560.

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. S. L. M. Whitmer et al., New Lineage of Lassa Virus, Togo, 2016. Emerg. Infect. Dis. 24, 599–602 (2018).

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

16. S. Nikisins et al., International external quality assessment study for molecular detection of Lassa virus. PLoS Negl. Trop. Dis. 9, e0003793 (2015).

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

18. N. J. Loman, J. Quick, J. T. Simpson, A complete bacterial genome assembled de novo using only nanopore sequencing data (2015), , doi:10.1101/015552.

19. 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).

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

21. 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).

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

23. 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).

24. N. S. Trovão et al., Host ecology determines the dispersal patterns of a plant virus. Virus Evol. 1, vev016 (2015).

25. M. A. Suchard et al., Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 4, vey016 (2018).

26. D. Edo-Matas et al., 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).

27. 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).

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

29. 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).

ACKNOWLEDGMENTS

We thank the health authorities of Nigeria for their excellent cooperation during the outbreak response. **Funding**: LK, STP, RH, RV, MWC, JAH acknowledge funding by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Emerging and Zoonotic Infections at University of Liverpool in partnership with Public Health England (PHE), in collaboration with Liverpool School of Tropical Medicine. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health or Public Health England. LEK has received travel expenses and accommodation from Oxford Nanopore to speak at conferences regarding this work. LEK has received some reagents free of charge from Oxford Nanopore in support of her PhD project. MWC has received reagents free of charge from Oxford Nanopore in support of previous projects not related to the work presented in this manuscript. LEK and MWC have not receive other financial compensation or hold shares. PL and MAS acknowledge funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement no. 725422-ReservoirDOCS) and from the Wellcome Trust Collaborative Award, 206298/Z/17/Z. PL acknowledges support by the Special Research Fund, KU Leuven ('Bijzonder Onderzoeksfonds', KU Leuven, OT/14/115), and the Research Foundation -- Flanders ('Fonds voor Wetenschappelijk Onderzoek – Vlaanderen', G066215N, G0D5117N and G0B9317N). MAS acknowledges support under National Science Foundation grant DMS 1264153. This study was supported by the German Federal Ministry of Health through support of the WHO Collaborating Centre for Arboviruses and Hemorrhagic Fever Viruses at the Bernhard-Nocht-Institute for Tropical Medicine (agreements ZMV I 1-2517WHO005 and ZMV I 1-2517WHO010) and through the Global Health Protection Program (agreement number ZMVI1-2517-GHP-704), the German Federal Ministry for Economic Cooperation and Development through the Rapid Deployment Expert

Group to Combat Threats (SEEG), the European Union's Horizon 2020 research and innovation program to SG (grant agreement no. 653316-EVAg), and the German Research Foundation (DFG) to SG and DUE (GU 883/4-1). DUE acknowledges fellowships from Alexander von Humboldt Foundation and Kirmser Foundation. The funders had no role in the design and interpretation of the data and preparation of the manuscript. **Author contributions:** LEK, SG, SD, STP, PL, SG conceptualized the study; LEK, STP, PL set up the methodology; LEK, JH, AT, SD, DUE performed sequencing and data validation; LEK, PL, MAS, STP, DS, FK, JM, SLo performed the formal sequencing data analysis; LEK, SD, JH, AT, MP, LO performed sample selection, data collection and organisation of sequencing datasets; DMW, KE, DS, FK, JM set up and assisted with bioinformatics pipeline; MAS, DUO, MP, LO, YI, DIA, TO, EO, RO, JAg, BE, JAi, PE, BO, SE, PA, MA, RE, EM, RG, AE, GI, GOd, GOk, RE, JO, EOY, IO, CA, MO, RA, ET, DA, NA, POO, MOR, KOI, COI, PA, CE, GA, EI performed diagnostic analysis; LEK, STP, PL, SD visualized data presentation; LEK, STP, PL, SD wrote the manuscript; all authors reviewed and edited the manuscript; SG, MWC, JAH, RH, RV supervised the study; MP, RV, AT, CI, PF, DN, SO, EOE, SG, SD, SLu performed project administration and implementation; SG, PL, MWC, RV, RH, JAH, LEK, DUE were involved in funding acquisition. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** LASV sequences from 2018 are deposited in GenBank, under BioProject PRJNA482058 (Data S1); sequences from 2012 to 2017 are deposited under BioProjects PRJNA482054 and PRJNA482058 (Data S2, to be published elsewhere). Alignments, trees, and BEAST xml files are available at: https://github.com/ISTH-BNITM-PHE/LASVsequencing.

SUPPLEMENTARY MATERIALS:

Materials and Methods

Figures S1-S10

Supplementary Data S1 and S2

References (16-29)

LEGENDS TO FIGURES

Figure 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) that includes most of the 2018 sequences. The six genotypes are indicated with different colors and roman symbols. 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 as light pink for the first seven sequences generated in Nigeria, magenta for the remaining 28 sequences analysed on-site, and purple for the remaining finalised in Europe. The same color code is used in the genotype II rectangular tree. Bootstrap values >80% are shown for the major genotype II lineages.

Figure 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 purple; 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 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, 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 grey box. One pair (096-115) was still disregarded as potential transmission chain due to 21 differences in L segment (Fig. S9). The temporal signal prior to BEAST inference was explored in Fig. S10.

Figure 1

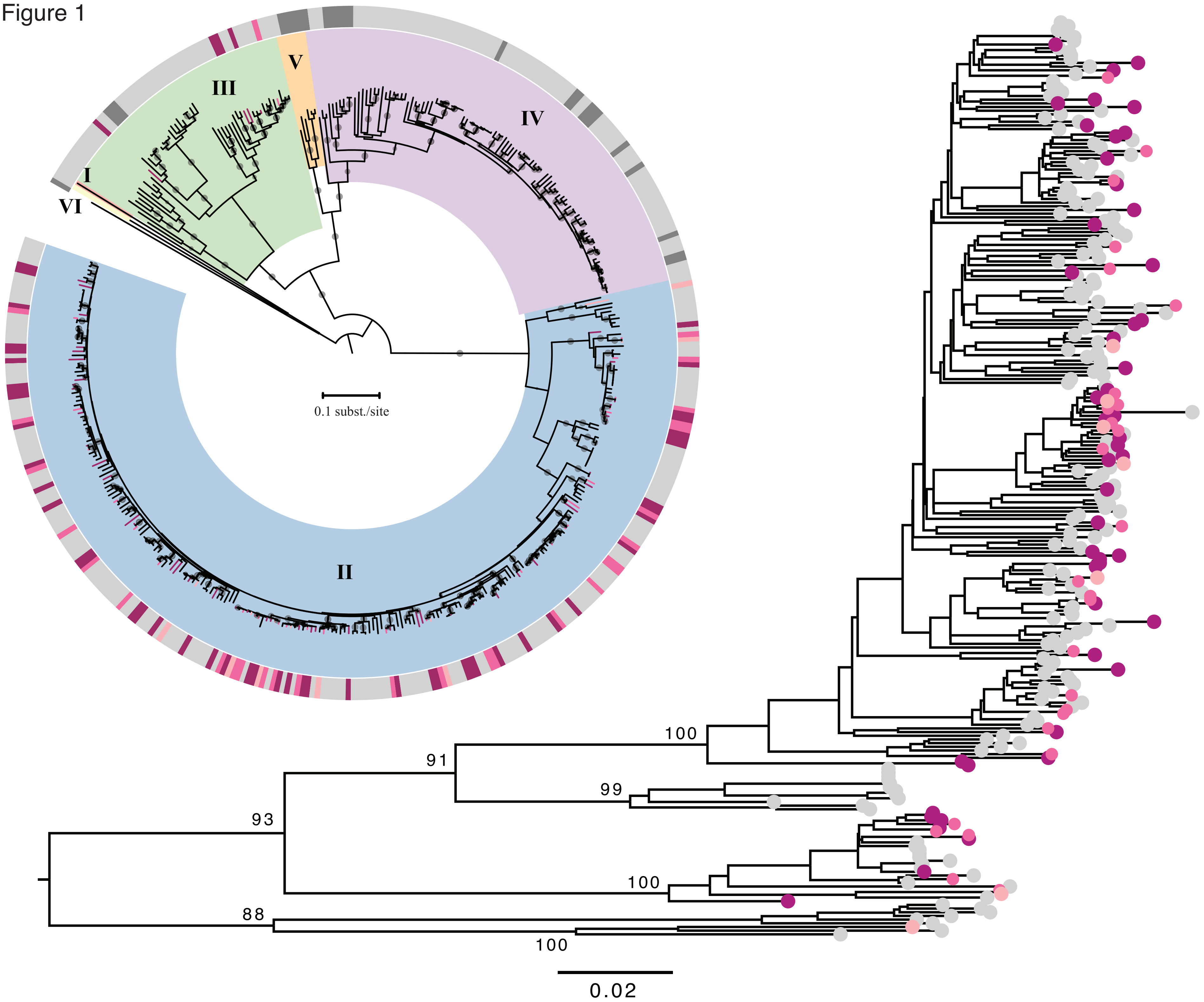
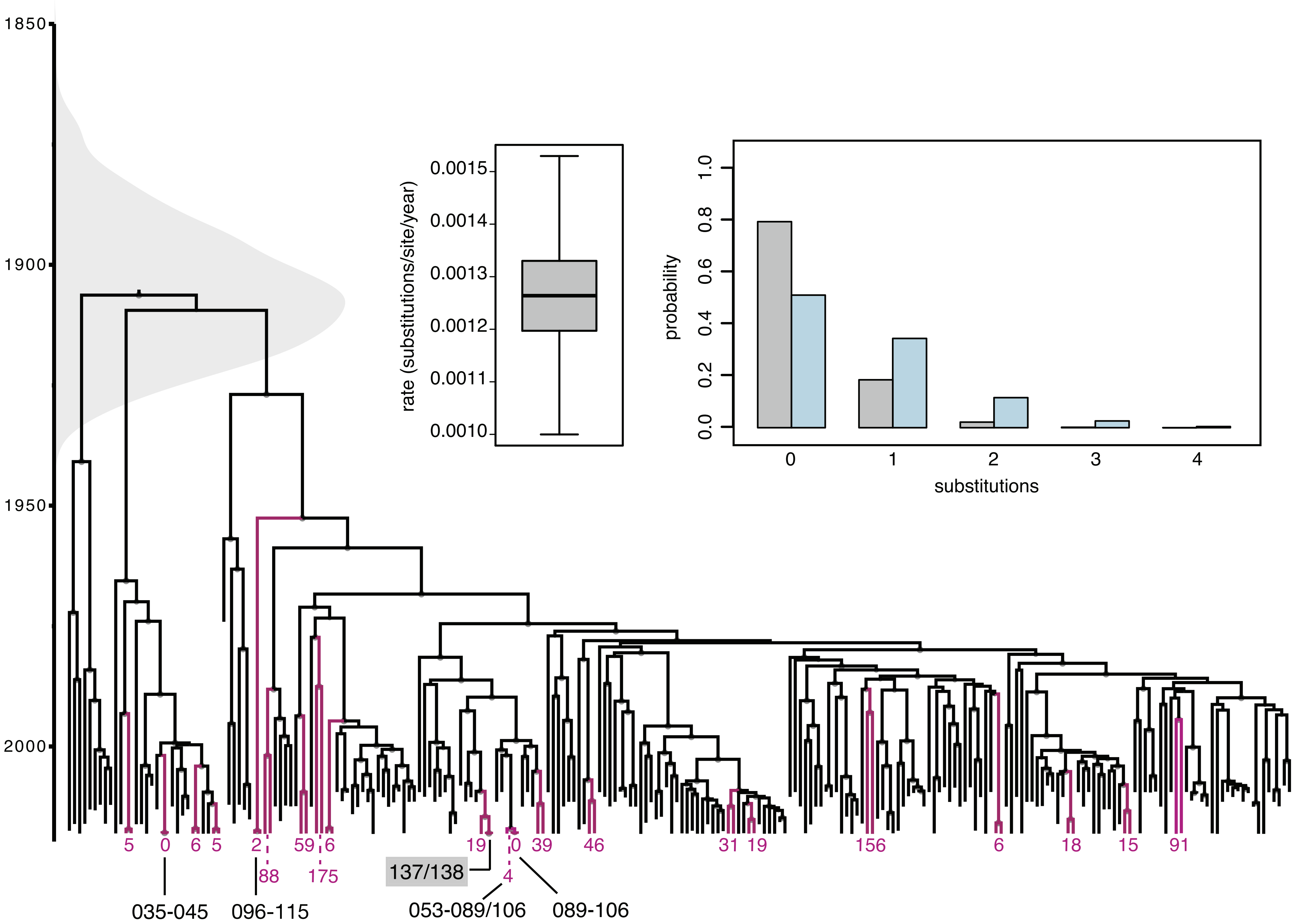


Figure 2


1850
1900
1950
2000
rate (substitutions/site/year)
0.0010
0.0011
0.0012
0.0013
0.0014
0.0015
probability
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
substitutions
5
0
6
5
2
88
59
175
6
19
137/138
4
0
39
46
31
19
156
6
18
15
91
035-045
096-115
053-089/106
089-106

Science
AAAS

Supplementary Materials for

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

Authors: L. E. Kafetzopoulou[1,2,3], S. T. Pullan[1,2], P. Lemey[4], M. A. Suchard[5], D. U. Ehichioya[3,6], M. Pahlmann[3,6], A. Thielebein[3,6], J. Hinzmann[3,6], L. Oestereich[3,6], D. M. Wozniak[3,6], K. Efthymiadis[7], D. Schachten[3], F. Koenig[3], J. Matjeschk[3], S. Lorenzen[3], S. Lumley[1], Y. Ighodalo[8], D. I. Adomeh[8], T. Olokor[8], E. Omomoh[8], R. Omiunu[8], J. Agbukor[8], B. Ebo[8], J. Aiyepada[8], P. Ebhodaghe[8], B. Osiemi[8], S. Ehikhametalor[8], P. Akhilomen[8], M. Airende[8], R. Esumeh[8], E. Muoebonam[8], R. Giwa[8], A. Ekanem[8], G. Igenegbale[8], G. Odigie[8], G. Okonofua[8], R. Enigbe[8], J. Oyakhilome[8], E. O. Yerumoh[8], I. Odia[8], C. Aire[8], M. Okonofua[8], R. Atafo[8], E. Tobin[8], D. Asogun[8,9], N. Akpede[8], P. O. Okokhere[8,9], M. O. Rafiu[8], K. O. Iraoyah[8], C. O. Irolagbe[8], P. Akhideno[8], C. Erameh[8], G. Akpede[8,9], E. Isibor[8], D. Naidoo[10], R. Hewson[1,2], J. A. Hiscox[2,11,12], R. Vipond[1,2], M. W. Carroll[1,2], C. Ihekweazu[13], P. Formenty[10], S. Okogbenin[8,9], E. Ogbaini-Emovon[8]†, S. Günther[3,6,*]†, S. Duraffour[3,6]†.

Correspondence to: guenther@bni.uni-hamburg.de

This PDF file includes:

Materials and Methods
Figs. S1 to S11
Table S1
Captions for Data S1 and S2

Other Supplementary Materials for this manuscript include the following:

Data S1 and S2

Materials and Methods

Sample collection

Samples from suspected Lassa fever patients were routinely tested for 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 *(16)*. 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 (*16*). Both Altona and Nikisins 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 (unpublished data). Therefore, all samples were generally tested in both assays, although the 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 (*17*) 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*, *18*) 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 (*17*) was used with default settings to align reads to the references. Mapping consensus sequences for Illumina were generated using QuasiBam (*19*).

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 (*20*). 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 (*21*) 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 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 (*22*) 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 (*23*) 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 (*24*). 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 (*25*). 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 (*26*). We used a flexible Bayesian skyride tree prior (*27*) and an uncorrelated lognormal relaxed molecular clock model to allow for rate variation among lineages (*28*). 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 (*29*). 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.

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 (λ_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 (λ_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 (λ_e). 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 (λ_2) and twice the sequencing error (λ_e) because we model differences between two sequences generated using MinION sequencing.

Figure S1

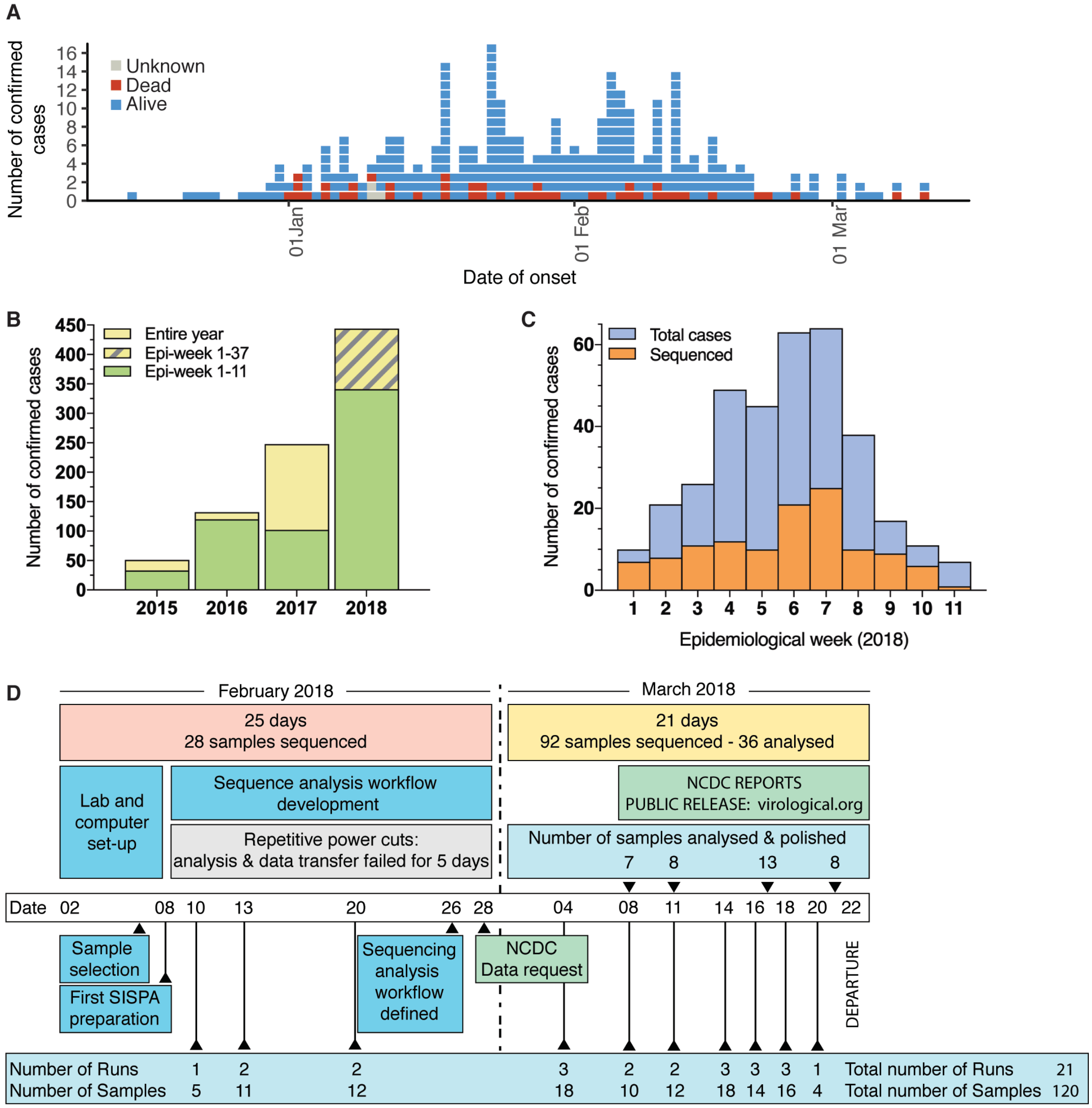


Fig. S1. Epidemiology of the Lassa fever outbreak and timeline of sequencing in Nigeria. (A) Epidemiological curve for 2018. ISTH confirmed 341 of the 376 Lassa fever cases reported by Nigeria Center of Disease Control (NCDC) between 1st January and 18th March 2018. The epidemiological curve shows the 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.

Figure S2

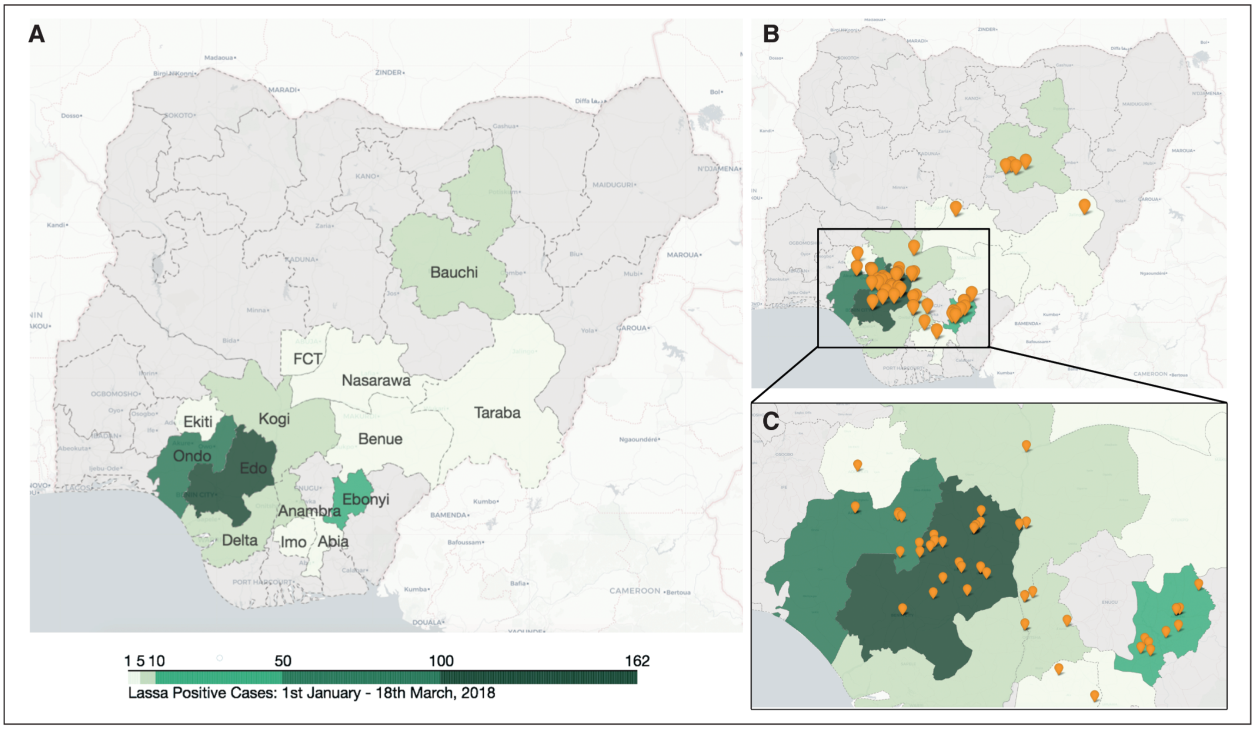


Fig. S2. Annotated map of confirmed Lassa fever cases between 1st January and 18th March 2018.

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

Figure S3

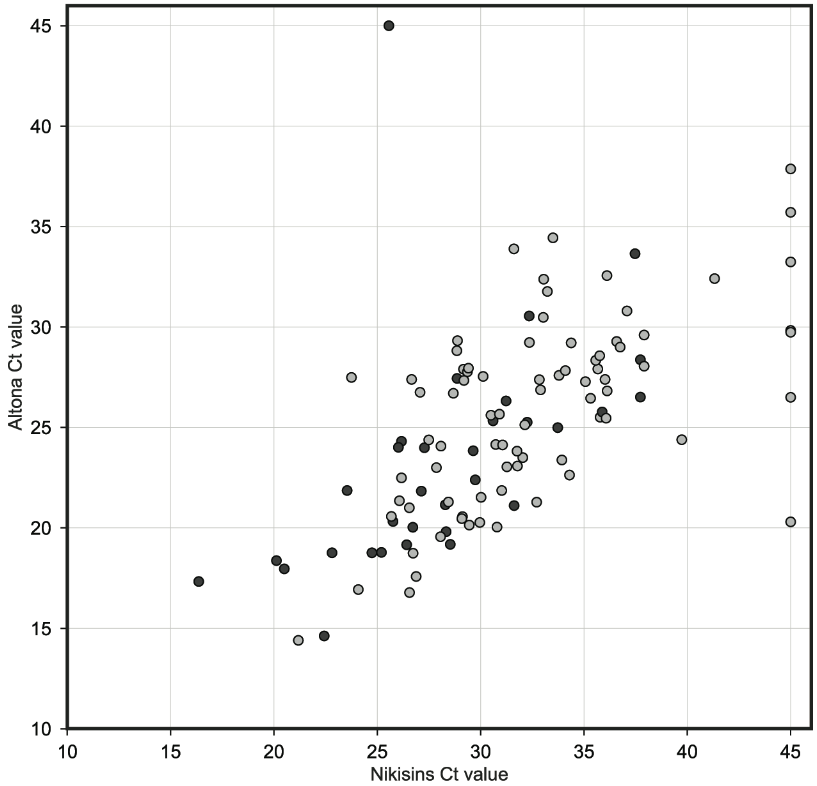
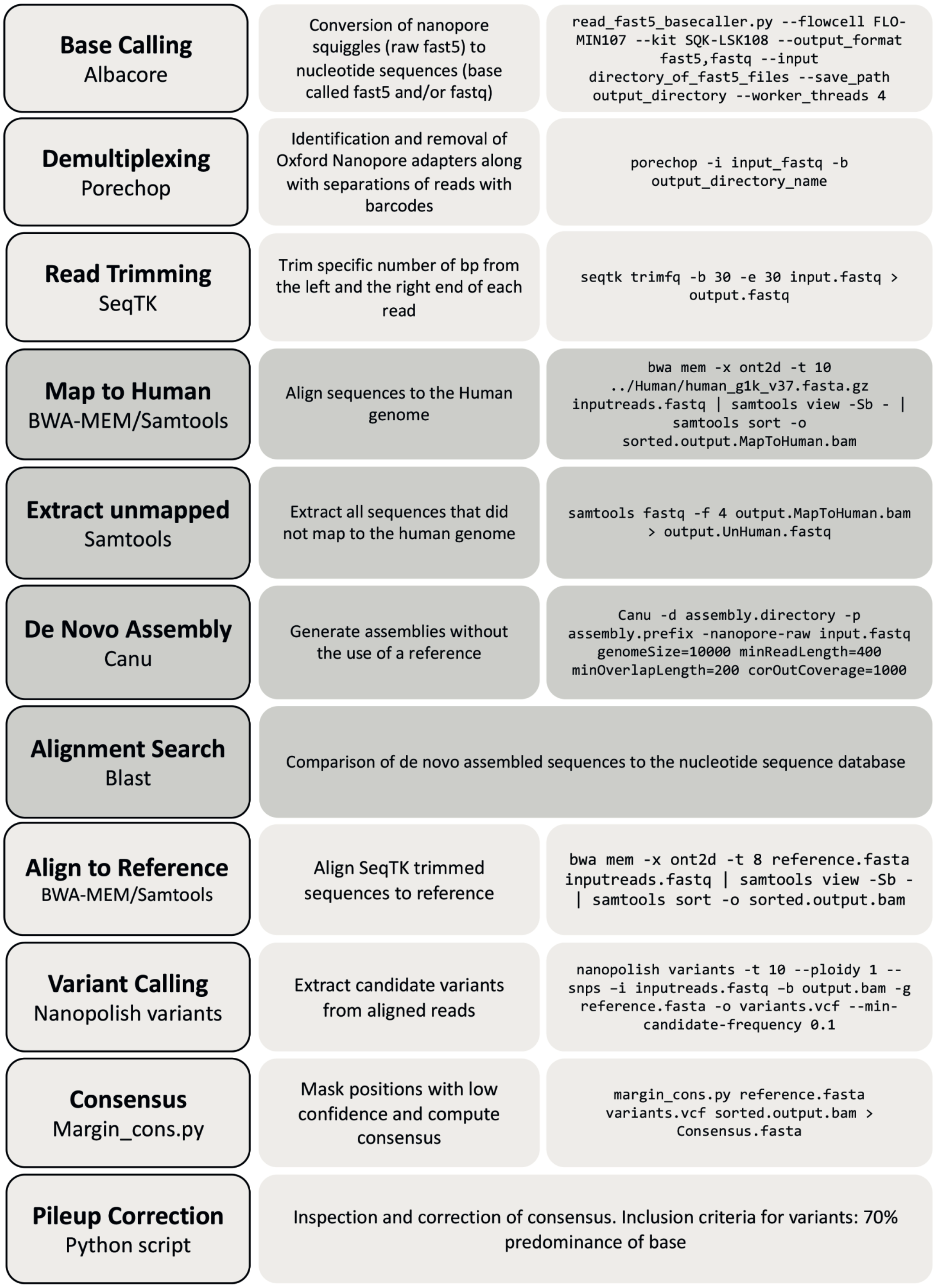


Fig. S3. Correlation between Ct values from Altona and Nikisins real-time RT-PCR assays. Seven samples tested negative in the Nikisins assay and one tested negative in the Altona assay, 5 demonstrating the importance of combined use of both assays for diagnosis of acute Lassa fever and subsequent sequencing. 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.

Figure S4



<table>
<tr><td><b>Base Calling</b>
Albacore</td><td>Conversion of nanopore squiggles (raw fast5) to nucleotide sequences (base called fast5 and/or fastq)</td><td><code>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</code></td></tr>
<tr><td><b>Demultiplexing</b>
Porechop</td><td>Identification and removal of Oxford Nanopore adapters along with separations of reads with barcodes</td><td><code>porechop -i input_fastq -b output_directory_name</code></td></tr>
<tr><td><b>Read Trimming</b>
SeqTK</td><td>Trim specific number of bp from the left and the right end of each read</td><td><code>seqtk trimfq -b 30 -e 30 input.fastq > output.fastq</code></td></tr>
<tr><td><b>Map to Human</b>
BWA-MEM/Samtools</td><td>Align sequences to the Human genome</td><td><code>bwa mem -x ont2d -t 10 ../Human/human_g1k_v37.fasta.gz inputreads.fastq | samtools view -Sb - | samtools sort -o sorted.output.MapToHuman.bam</code></td></tr>
<tr><td><b>Extract unmapped</b>
Samtools</td><td>Extract all sequences that did not map to the human genome</td><td><code>samtools fastq -f 4 output.MapToHuman.bam > output.UnHuman.fastq</code></td></tr>
<tr><td><b>De Novo Assembly</b>
Canu</td><td>Generate assemblies without the use of a reference</td><td><code>Canu -d assembly.directory -p assembly.prefix -nanopore-raw input.fastq genomeSize=10000 minReadLength=400 minOverlapLength=200 corOutCoverage=1000</code></td></tr>
<tr><td><b>Alignment Search</b>
Blast</td><td colspan="2">Comparison of de novo assembled sequences to the nucleotide sequence database</td></tr>
<tr><td><b>Align to Reference</b>
BWA-MEM/Samtools</td><td>Align SeqTK trimmed sequences to reference</td><td><code>bwa mem -x ont2d -t 8 reference.fasta inputreads.fastq | samtools view -Sb - | samtools sort -o sorted.output.bam</code></td></tr>
<tr><td><b>Variant Calling</b>
Nanopolish variants</td><td>Extract candidate variants from aligned reads</td><td><code>nanopolish variants -t 10 --ploidy 1 --snps –i inputreads.fastq –b output.bam -g reference.fasta -o variants.vcf --min-candidate-frequency 0.1</code></td></tr>
<tr><td><b>Consensus</b>
Margin_cons.py</td><td>Mask positions with low confidence and compute consensus</td><td><code>margin_cons.py reference.fasta variants.vcf sorted.output.bam > Consensus.fasta</code></td></tr>
<tr><td><b>Pileup Correction</b>
Python script</td><td colspan="2">Inspection and correction of consensus. Inclusion criteria for variants: 70% predominance of base</td></tr>
</table>

Fig. S4. Workflow of consensus sequence generation.

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

Figure S5

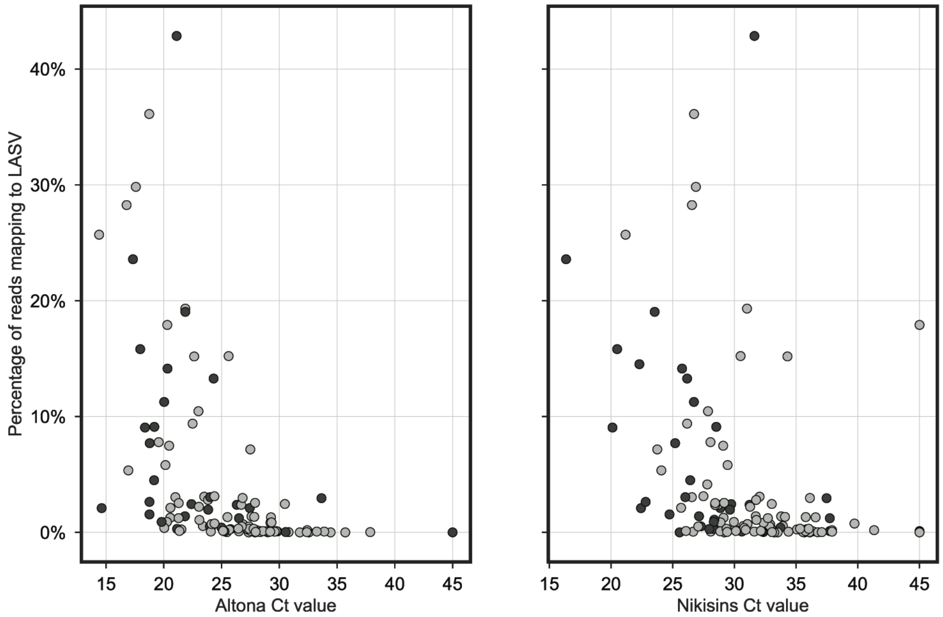


Fig. S5. Percentage of reads mapping to LASV depending on Ct value in Altona and Nikisins real-time RT-PCR assay.

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.

Figure S6

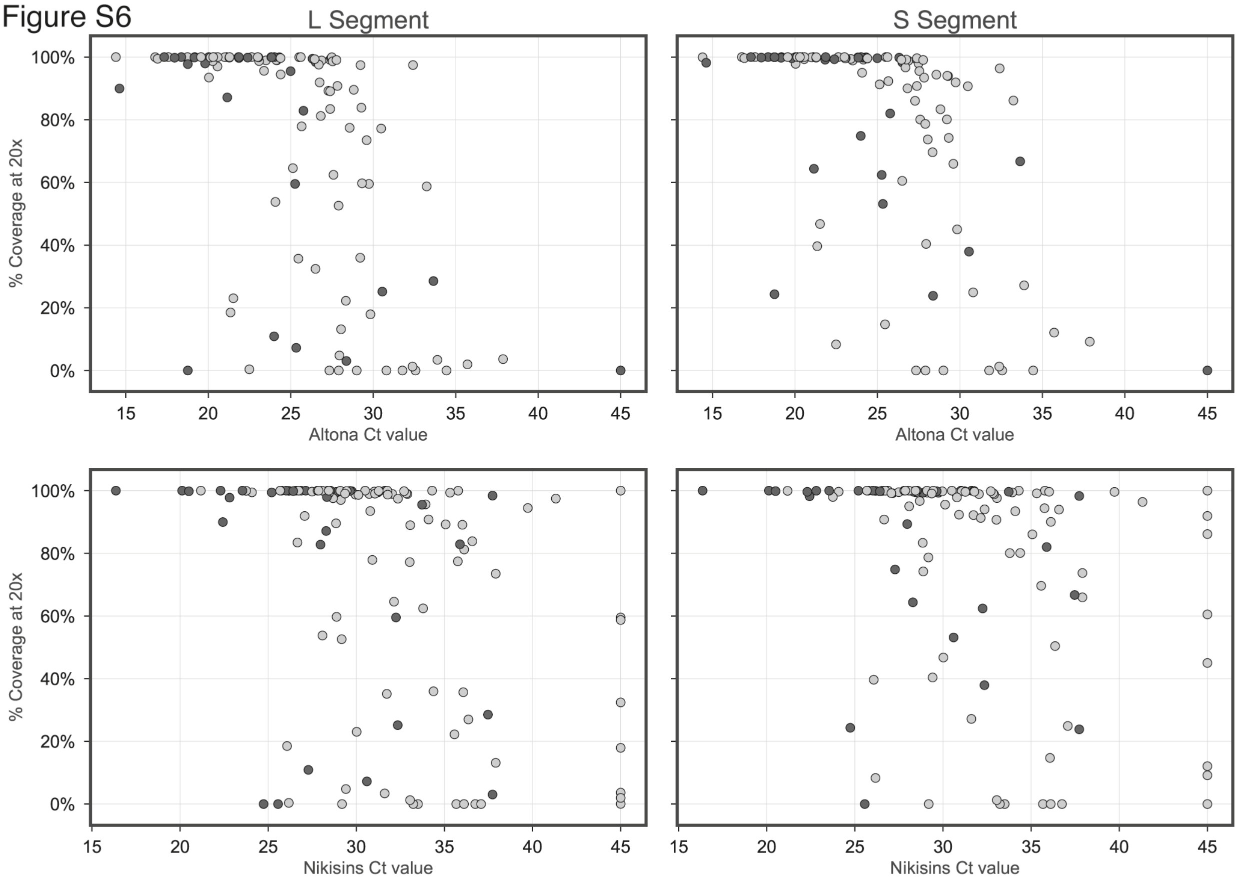


Fig. S6. Percentage of genome coverage (20×) per segment depending on Ct value in Altona and Nikisins real-time RT-PCR assay.

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.

Figure S7

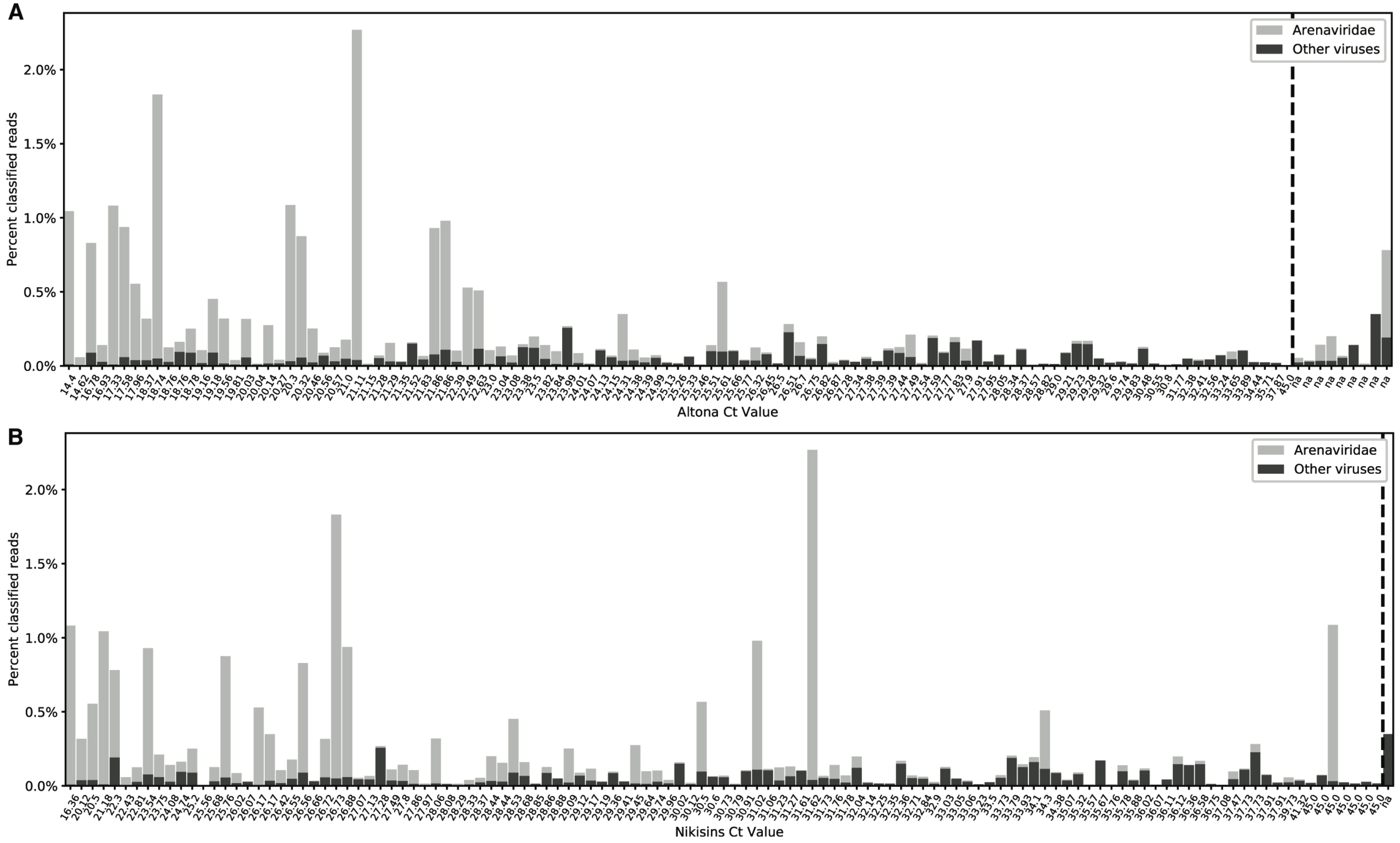


Fig. S7. Classification of MinION reads depending on Ct value in (A) Altona and (B) Nikisins real-time RT-PCR assay. Reads were classified by Centrifuge software as either *Arenaviridae* or other viruses. The analysis allowed for identification of a co-infection in sample 110 with 0.1% reads classifying 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. 5
na, not applicable as samples were not tested with the respective RT-PCR.

Figure S8

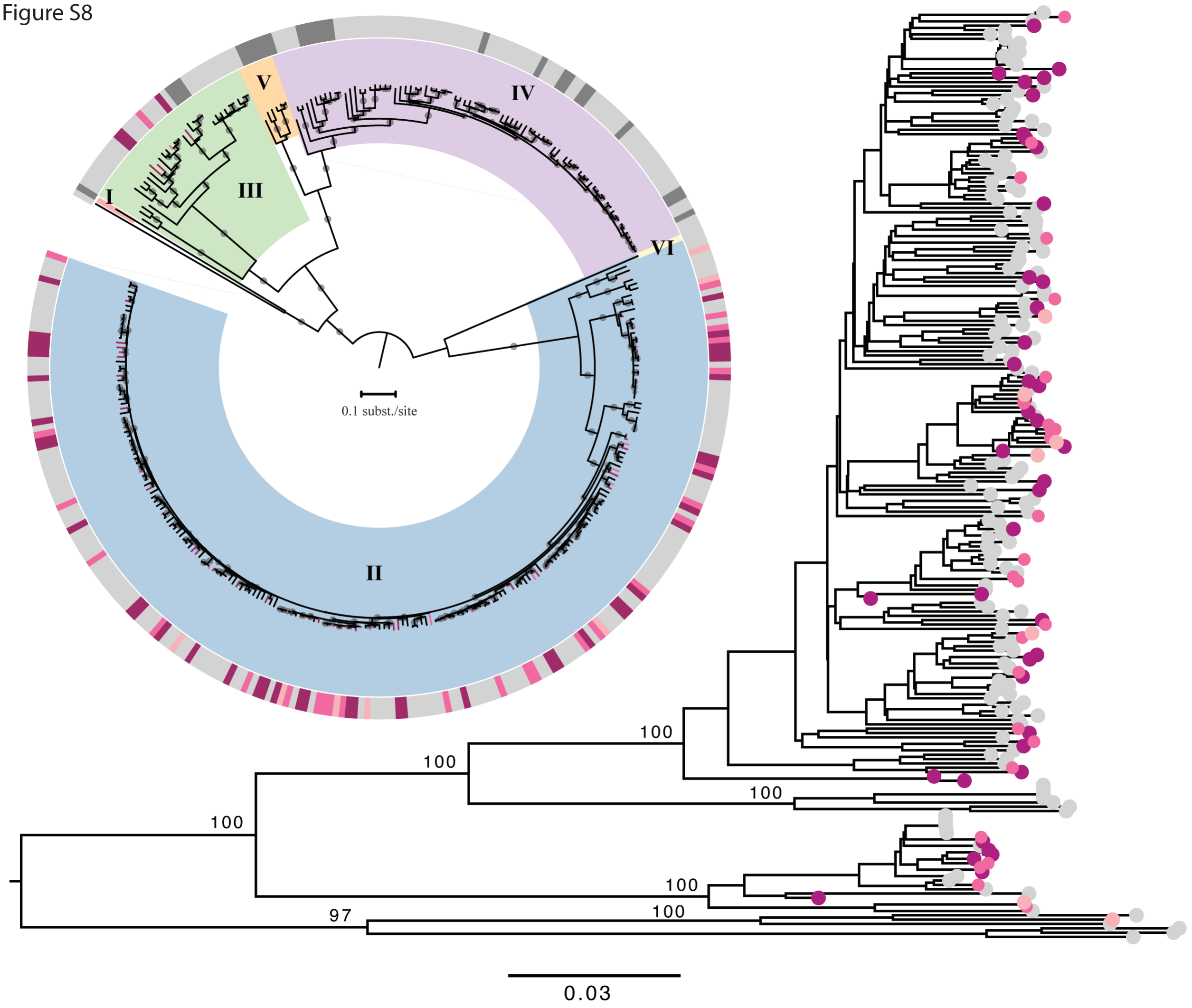


Fig. S8. Phylogenetic reconstruction of the L segment data.

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) that includes most of the 2018 sequences. In the circular tree, the 6 genotypes are indicated with different colors and roman symbols. 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 (pink/magenta/purple). 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 magenta color. All other 2018 sequences analyzed upon return to Europe are marked in purple. The same color code is used in the genotype II rectangular tree. Bootstrap values >80% are shown for the major lineages.

Figure S9

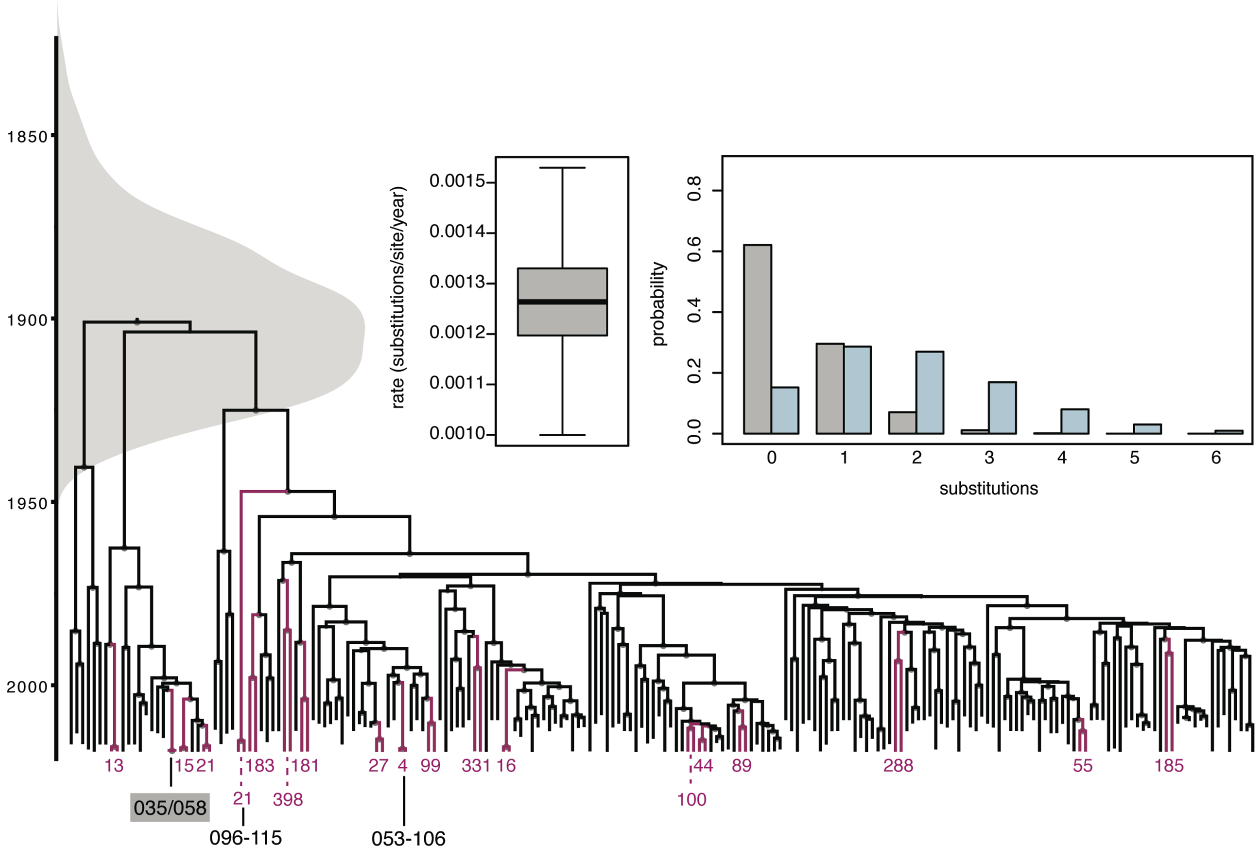


Fig. S9. Assessing the potential for direct linkage between pairs of 2018 sequences in the L segment.

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 purple; 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).

Figure S10

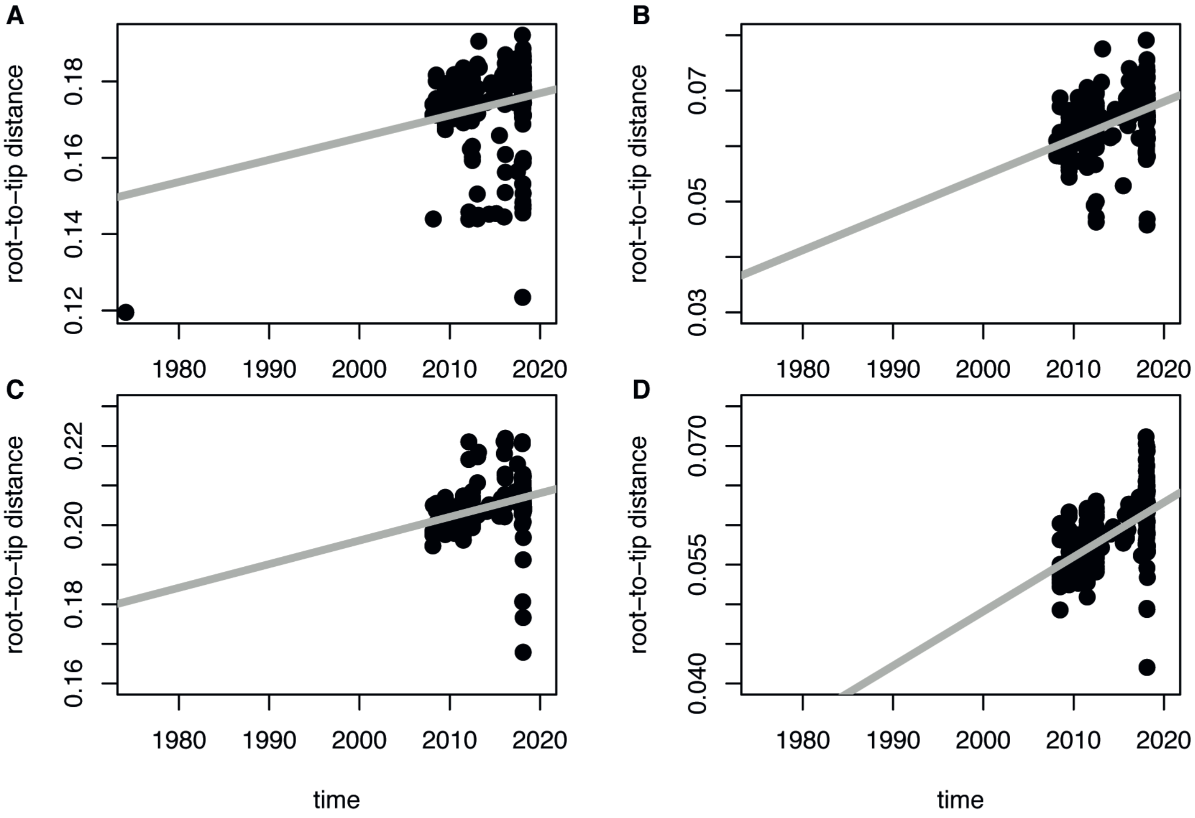


Fig. S10. Root-to-tip divergence of LASV genotype II sequences as a function of sampling time for the S (A-B) and L (C-D) segments.

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.

Sample ID	Segment Length	Total Disagreements	In coding regions	Positions (Illumina > Nanopore, non-coding/coding)
S segment				
001	3490	0	0	
013	3367	0	0	
014	3407	1	0	1587T>G
021	3385	0	0	
036	3387	0	0	
066	3393	0	0	
072	3398	6	0	1554C>G, 1568C>G, 1569G>A, 1570C>A,1572G>A, 1573A>G
073	3406	1	0	3422C>G
074	3367	0	0	
075	3412	0	0	
115	3387	0	0	
119	3403	2	2	596C>A, 597A>C
126	3407	1	0	2C>G
131	3389	0	0	
L segment				
001	7196	2	0	2A>T, 416T>G
013	7230	0	0	
014	7245	1	0	445G>A
021	7237	0	0	
036	7261	1	0	7209T>C
066	7135	1	1	274G>T
072	7250	0	0	
073	7260	4	3	5792G>T, 5793A>G, 5795G>A, C>T7258.
074	7238	0	0	
075	7256	0	0	
115	7183	3	3	727C>T, 728T>C, 731C>T
119	7258	2	0	404C>T, 407C>T
126	7245	0	0	
131	7238	0	0	

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.