PLOS GLOBAL PUBLIC HEALTH

RESEARCH ARTICLE

Current sampling and sequencing biases of Lassa mammarenavirus limit inference from phylogeography and molecular epidemiology in Lassa fever endemic regions

Liã Bárbara Arruda[1¤a]*, Hayley Beth Free[2☯¤b], David Simons[2☯], Rashid Ansumana[3], Linzy Elton[1], Najmul Haider[2¤c], Isobella Honeyborne[1], Danny Asogun[4], Timothy D. McHugh[1], Francine Ntoumi[5,6], Alimuddin Zumla[1,7], Richard Kock[2]

1 Division of Infection and Immunity, Centre for Clinical Microbiology, University College London, London, United Kingdom, **2** The Royal Veterinary College, University of London, Hatfield, United Kingdom, **3** School of Community Health Sciences, Njala University, Bo, Sierra Leone, **4** Ekpoma and Irrua Specialist Teaching Hospital, Ambrose Alli University, Irrua, Nigeria, **5** Fondation Congolaise Pour la Recherche Médicale (FCRM), Brazzaville, Republic of Congo, **6** Institute for Tropical Medicine, University of Tübingen, Tübingen, Germany, **7** NIHR Biomedical Research Centre, UCL Hospitals NHS Foundation Trust, London, United Kingdom

☯ These authors contributed equally to this work.
¤a Current address: Wellcome Connecting Science, Hinxton, United Kingdom
¤b Current address: Oxford Brookes University, Oxford, United Kingdom
¤c Current address: Faculty of Natural Sciences, School of Life Sciences, Keele University, Staffordshire, United Kingdom
* libarbara@gmail.com

Abstract

Lassa fever (LF) is a potentially lethal viral haemorrhagic infection of humans caused by *Lassa mammarenavirus* (LASV). It is an important endemic zoonotic disease in West Africa with growing evidence for increasing frequency and sizes of outbreaks. Phylogeographic and molecular epidemiology methods have projected expansion of the Lassa fever endemic zone in the context of future global change. The Natal multimammate mouse (*Mastomys natalensis*) is the predominant LASV reservoir, with few studies investigating the role of other animal species. To explore host sequencing biases, all LASV nucleotide sequences and associated metadata available on GenBank (n = 2,298) were retrieved. Most data originated from Nigeria (54%), Guinea (20%) and Sierra Leone (14%). Data from non-human hosts (n = 703) were limited and only 69 sequences encompassed complete genes. We found a strong positive correlation between the number of confirmed human cases and sequences at the country level ($r = 0.93$ (95% Confidence Interval = 0.71–0.98), $p < 0.001$) but no correlation exists between confirmed cases and the number of available rodent sequences ($r = -0.019$ (95% C.I. -0.71–0.69), $p = 0.96$). Spatial modelling of sequencing effort highlighted current biases in locations of available sequences, with increased sequencing effort observed in Southern Guinea and Southern Nigeria. Phylogenetic analyses showed geographic clustering of LASV lineages, suggestive of isolated events of human-to-rodent transmission and the emergence of currently circulating strains of LASV from the year 1498 in Nigeria. Overall, the current study highlights significant geographic

OPEN ACCESS

Citation: Arruda LB, Free HB, Simons D, Ansumana R, Elton L, Haider N, et al. (2023) Current sampling and sequencing biases of Lassa mammarenavirus limit inference from phylogeography and molecular epidemiology in Lassa fever endemic regions. PLOS Glob Public Health 3(11): e0002159. https://doi.org/10.1371/journal.pgph.0002159

Editor: Vijaykrishna Dhanasekaran, University of Hong Kong, HONG KONG

Received: June 16, 2023

Accepted: October 16, 2023

Published: November 8, 2023

Data Availability Statement: All data used in these analyses are publicly available from GenBank. The accession numbers of records used are available as supplementary material. Code to reproduce the metadata analyses are available as an archived Git release on Zenodo (https://doi.org/10.5281/zenodo.6340162).

Funding: Linzy Elton, Timothy D McHugh, Francine Ntoumi, and Alimuddin Zumla acknowledge support from EDCTP-Central Africa and East African Clinical Research Networks (CANTAM-3, EACCR-3). Sir Zumla is an NIHR Senior Investigator, a Mahathir Science Award, Sir Patrick Manson Medal and EU-EDCTP Pascoal Mocumbi Prize laureate. Liã Bárbara Arruda, David Simons, Rashid Ansumana, Linzy Elton, Najmul Haider, Isobella Honeyborne, Danny Asogun, Timothy D McHugh, Francine Ntoumi, Alimuddin Zumla and Richard Kock acknowledge support from the Pan-African Network for Rapid Research, Response and Preparedness for Infectious Diseases Epidemics – PANDORA-ID-NET, funded through the European and Developing Countries Clinical Trials Partnership (EDCTP) (grant number RIA2016E-1609). David Simons is supported by a PhD studentship from the UK Biotechnology and Biological Sciences Research Council (BB/M009513/1).

Competing interests: The authors have declared that no competing interests exist.

limitations in LASV surveillance, particularly, in non-human hosts. Further investigation of the non-human reservoir of LASV, alongside expanded surveillance, are required for precise characterisation of the emergence and dispersal of LASV. Accurate surveillance of LASV circulation in non-human hosts is vital to guide early detection and initiation of public health interventions for future Lassa fever outbreaks.

1 Introduction

Lassa fever (LF) is a lethal zoonotic viral haemorrhagic disease of humans, caused by *Lassa mammarenavirus* (LASV). The estimated westward route of dispersal of the seven lineages of this arenavirus has been used to project the potential for LF to extend beyond the current endemic zone [1], currently concentrated into eight West African countries: Benin, Ghana, Guinea, Liberia, Mali, Sierra Leone, Togo and Nigeria (S1 Fig) [2].

Epidemiological data on LF is limited and constrained by availability of current testing and reporting in the endemic region, making accurate estimates of its true burden challenging [3]. Many individuals infected with LASV do not seek healthcare with up to 80% of infections assumed asymptomatic or presenting as mild illness [4]. Estimates based on longitudinal serological surveys in Sierra Leone in the early 1980's indicated that 100,000 to 300,000 infections of LF occurred annually in West Africa, with more recent estimates being up to 900,000 infections/year [4,5]. Identification of symptomatic cases is further confounded by overlapping symptoms with other diseases (e.g., malaria) and lack of available diagnostic methods [6–9].

Although human-to-human transmission of LF has been reported–typically associated with nosocomial outbreaks–these are rare events when compared with spillover from rodent hosts [10]. Humans become infected with LASV upon contact with or inhalation of excretions from the rodent species [11,12]. The Natal multimammate mouse (*Mastomys natalensis*) is considered the primary reservoir of LASV. Despite 11 other rodent species having been found to be acutely infected or have seropositivity to LASV including; *Mastomys erythroleucus*, *Hylomyscus pamfi*, *Mus baoulei* and *Rattus rattus* there are a limited number of studies investigating the role in LASV transmission in these other rodents [13–18].

Access to diagnostic tests varies spatially, and the increased availability at centres of excellence in LF treatment and research such as the Irrua Specialist Teaching Hospital, Nigeria and Kenema General Hospital, Sierra Leone results in a spatial bias of reported cases from these locations [19]. This results in uneven LF surveillance initiatives across host species and endemic regions. The consequent paucity of LASV genomic data and associated metadata leads to spatial biases playing a potentially important role in understanding of the disease epidemiology and ecology. In order to identify surveillance gaps and quantify some of the current biases in genomic surveillance, we performed a study of LASV nucleotide sequences available from the National Centre for Biotechnology Information (NCBI) GenBank, using associated metadata to spatially model sequencing effort, adjusted for the number of suspected and confirmed human LF cases to determine potential biases in locations of available sequences or significant geographic limitations in LASV surveillance, particularly, in non-human hosts.

2 Methods

2.1 Data collection and processing

LASV nucleotide sequences from both S and L segments of the viral genome were obtained from the NCBI GenBank [20]. The search query run on 24 Sep 2021 was for "Lassa mammarenavirus" in the organism field of the NCBI nucleotide dataset. Data were obtained using the NCBI Entrez API with analysis conducted using the "genbankr" package within the R statistical programming language [20–22]. Associated citations were manually retrieved to identify missing metadata for sequences including hosts and geographic location of samples. Sequences with large portions (10% missing compared to reference sequences, NC_004296.1 and NC_004297.1 for S and L segments respectively) of missing nucleotide data on the L- or S-segment or lacking associated metadata (collection year, host species, country, and geographical region of sampling) were excluded from phylogenetic analysis. Nucleotide sequences were aligned using the 'map to reference' tool on Geneious Prime 20201.2 (https://www.geneious.com). Alignment, visual inspection and manual editing were performed, and entries that contained >100 continuous ambiguous nucleotide calls were excluded (S1 Data). We first summarise available metadata on year of collection, the species of origin of the LASV sequence and the country of sampling.

Obtained sequences are associated with the level-1 administrative region of sampling. For sequences with no sub-national location we enriched available metadata by extracting location data from associated publications. We geocoded the provided location (e.g., longitude and latitude, or town name) using the Google Geocoding API and the "ggmap" R package [23]. These produced longitude and latitude locations were associated with level-1 administrative regions using the "sf" R package, shapefiles were obtained from GADM 4.0. and accessed through the "geodata" R package [24,25]. Additional cleaned and enriched was archived on Zenodo (https://doi.org/10.5281/zenodo.6340162).

2.2 Sequencing bias

We compared the number of sequences returned in GenBank from each country with the number of Lassa fever (LF) cases reported from a country between 2008 and February 2023 to test the correlation between availability of sequences and reported LF cases.

The number of reported LF cases aggregated to country level was used as a measure of national LASV transmission intensity. This period was selected by the study team due to increasing standardisation of case reporting since 2008 [26,27]. Prior to 2008 cases were only sporadically reported as individual outbreaks in academic literature or outbreak reports [28]. The extension of the time-period to February 2023 (compared to the search being run in September 2021) was to incorporate more recent data from the continually improving surveillance in the endemic region. The case dataset (S2 Data) was compiled from a systematic review of the literature, public health, outbreak response and ministry of health resources. The references of data sources used to compile the reported case dataset are included in S2 Data.

We report the Pearson's product-moment correlation coefficient (r), the number of degrees of freedom, the 95% confidence interval (95% C.I.) and the p-value for a null-hypothesis of no correlation between the number of sequences obtained from a country and the national transmission intensity of LASV. We then tested the association between the number of available sequences obtained from a) humans or b) rodent hosts with LASV transmission intensity to test the null-hypothesis that the number of both human derived and rodent derived sequences from a country was not associated with national LASV transmission intensity. The "stats" R package was used for all correlation tests [22].

To explore the bias of sequenced samples at a sub-national level the origin of a sequenced sample was geocoded using the Google Geocoding API and the "ggmap" R package [23]. Reported sequence locations were associated with level-1 administrative regions using the "sf" R package, shapefiles were obtained from GADM 4.0.2. and accessed through the "geodata" R package [24,25,29]. We stratified sequences into human and rodent sourced samples to visualise the spatial heterogeneity of sampling within rodent hosts and infected humans.

To estimate the relative sampling effort bias at the sub-national level (i.e., level-1 administrative regions), the number of samples obtained within a region was associated with the centroid of the region. As in the prior analysis, the number of reported confirmed LF cases within a region was used as a measure of regional LASV transmission intensity. To standardise this to the population size of the region, the number of reported cases within a region was divided by the human population count to produce the number of confirmed cases per 100,000 individuals. Regional population counts were obtained from WorldPop 2020 raster data accessed through the "geodata" R package [25]. GADM shapefiles of level-1 administrative regions were used to extract and aggregate human population counts from each raster cell to level-1 administrative regions using the "terra" R package [30]. The relative sequencing effort at a location is derived from the smoothed two-dimensional coefficient of a spatial Generalised Additive Model using the number of sequences as the response variable, with geographic coordinates (Longitude = X, Latitude = Y) and regional LASV transmission intensity (i.e., reported human cases per 100,000 individuals) as covariates (Eq 1). This model was constructed using the "mgcv" R package [31].

$$LASV\ sequences \sim s(X*Y) + s(Reported\ cases\ per\ 100,000) \tag{1}$$

The smooth terms (s(X * Y) and s(Reported human cases per 100,000)) were specified as thin plate regression splines with a basis dimension (*k*) of 102 for the spatial smooth term (Longitude by Latitude) and 10 for the smooth term of LASV transmission intensity. We assessed the appropriateness of the selected basis dimensions by comparing to the estimated degrees of freedom (EDF) of the smooth terms [31]. This model is used to test the null hypothesis that there is no association between sequencing effort and LASV transmission intensity across the endemic region. We present the two-dimensional relative sequencing effort at coordinates within the study area reporting the EDF and *p*-value for the spatial smooth and LASV transmission intensity smooth terms.

Phylogenetic analysis

Phylogenetic analysis was undertaken through Bayesian Markov Chain Monte Carlo (MCMC) method using BEAST.v1.10.4 [32]. In BEAUTi, the parameters were a substitution model as a generalised time reversible plus gamma site heterogeneity, with codon partition positions 1, 2, 3. A strict clock and a coalescent tree prior with a constant size population was used. Each analysis consisted of 20 million MCMC steps and trees were sampled every 20,000 generations. Sample collection dates from the metadata were used as tip dates to fit to a molecular clock, and country of sample collection was incorporated as a discrete state [33,34]. To assess the log files of the output TRACER.v.1.7.1 was used. Maximum-clade credibility trees were generated through TreeAnnotator v1.8.4 and visualised in FigTree.v1.4.4 [35].

3 Results

3.1 Compiled dataset

The initial dataset comprised 2,298 records (from samples obtained 1969–2019), including nucleotide sequences and associated metadata. Incomplete gene sequences and sequences lacking metadata information (n = 1,045) were removed from phylogenetic analyses. Therefore, 680 sequences of complete S segment and 573 sequences of partial L segment (L protein only) were used. Accession numbers of included and excluded sequences are available in S1 Data.

3.2 Descriptive analysis

Year of collection was available for 2,108 records, with the oldest sequence dating from 1969 and latest from 2019. Among these records, most sequences (n = 1,936, 92%) have been obtained since 2008. Human-derived LASV sequences comprised most of the available records (67%), other host species include *Mastomys natalensis* (29%) and *Mastomys spp*. (3%), while *Mastomys erythroleucus* (n = 18), *Mus baoulei* (n = 9) and *Hylomyscus pamfi* (n = 10) represent < 1% each. The species sampled was not documented in 107 records. Country of collection was available for 2,238 records. Most sequences were produced from samples collected in Nigeria (54%), followed by Guinea (20%), Sierra Leone (14%), Liberia (4%) and Cote d'Ivoire (3%) with the remainder obtained from, Benin, Ghana, Mali and Togo (Fig 1).

Sequences for human derived samples with sub-national location data (n = 1328, 63%) were clustered in Edo State, Nigeria (n = 519, 39%), Ondo State, Nigeria (n = 220, 17%) and Eastern Province, Sierra Leone (n = 159, 12%) with 430 samples from the remaining endemic regions. Sequences from rodent samples with sub-national location data (n = 527, 25%) were most commonly obtained from Faranah, Guinea (n = 210, 39%) and Eastern Province, Sierra Leone (n = 107, 20%) with 210 samples from the regions.

3.3 Sequencing bias

We observed a strong positive correlation between the number of confirmed human cases between 2008–2023 and the number of GenBank deposited sequences at country level (Correlation coefficient (r) (degrees of freedom = 7) = 0.93, 95% Confidence Interval (95% C.I.) = 0.71–0.98, $p < 0.001$). When analysed by species source no correlation with the number of confirmed cases and the number of available rodent derived sequences was observed ($r(6)$ = -0.019 95% C.I. -0.71–0.69, $p = 0.96$). There remained a strong positive correlation between the number of confirmed cases and human derived sequences ($r(6) = 0.99$, 95% C.I. 0.998–0.999, $p = < 0.001$).

When combining both human and rodent-derived samples at the level-1 administrative region to explore spatial sampling biases, we found differences in relative sequencing effort across the study region. The spatial smooth term identified sampling effort as greatest in Southwest Nigeria, centred over Edo State, the Faranah and Nzérékoré regions of Guinea, Eastern Province of Sierra Leone and Nimba district of Liberia (estimated degrees of freedom (EDF) = 29.2, $p < 0.001$) (Fig 2). There was a positive, non-linear association between LASV transmission intensity (measured as number of reported cases per 100,000 individuals) and the number of available rodent and human derived LASV sequences at level-1 administrative region (EDF = 3.96, p < 0.001).

3.4 Phylogenetic analysis

Sequences for each segment of LASV showed clustering according to previously documented lineages I-VII alongside geographical clustering with lineages I-III and VI present in Nigeria, lV in Liberia, Guinea and Sierra Leone, V in Mali and VII in Togo (S2 Fig). In this analysis only L segment sequences of lineage V from Cote d'Ivoire were included due to quality control exclusion criteria. The phylogeny of the L segment indicates an older emergence of LASV in the human population, with the most recent common ancestor (MRCA) predicted in the year 1498 in Nigeria, inference based on the S segment indicates the emergence in the year 1681 (Table 1).

There was a lack of sequence information from lineage I and VI, however, phylogeny suggests these lineages are basal to others in Nigeria (S2 Fig). Lineage VII in Togo is most closely related to Nigerian isolates and potentially diverged between 500–900 years ago. The divergence of lineage III and II was predicted to have occurred approximately between mid-900

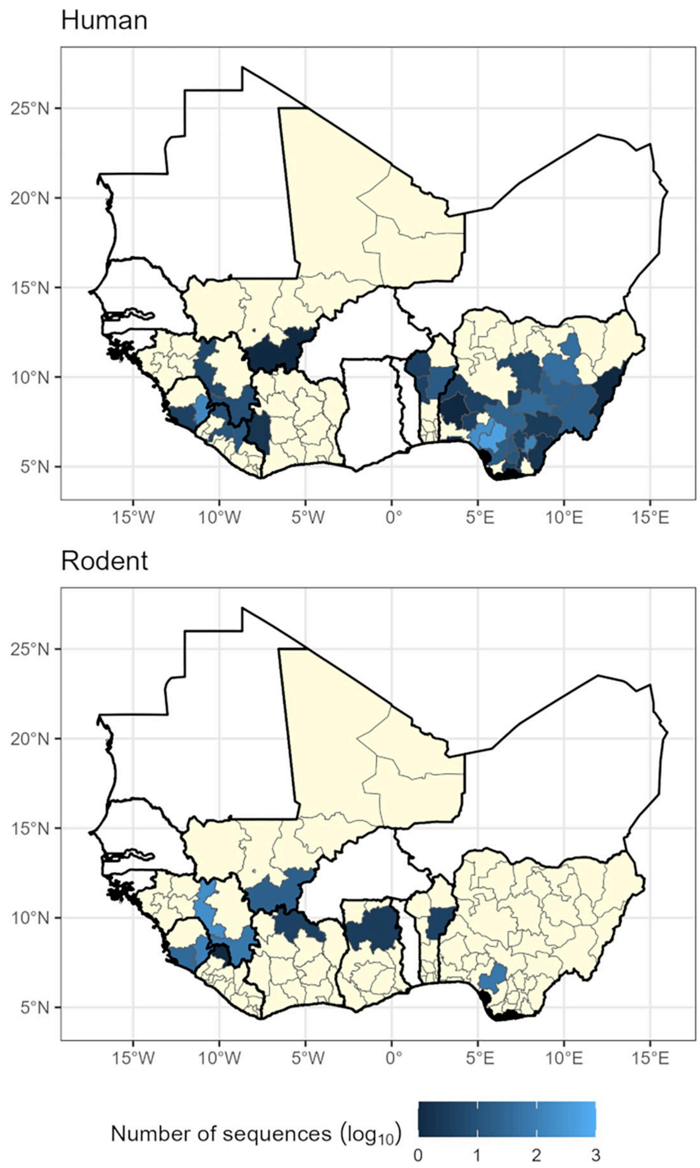

Fig 1. The number of sequences, shown on a $\log_{10}$ scale, retrieved from NCBI GenBank with associated sub-national sampling location and host for human samples (top, n = 1,328) and rodent samples (bottom, n = 527). Yellow regions represent level-1 administrative areas with no sequences within countries that have at least one available sequence. White countries are West African countries with no available LASV sequences. See S1 Fig for country names. Shapefiles for basemap layer obtained from GADM 4.0.2 (www.gadm.org) [29].

https://doi.org/10.1371/journal.pgph.0002159.g001

and mid-1400 according to the phylogenetics trees for L and S segments respectively (S2 Fig). Introduction to countries west of Nigeria appears to be by dispersal initially to Liberia, followed by Guinea in the 1700s, followed by Sierra Leone and Mali approximately 100 years later. A lack of full segment sequences from lineage V limits calculation of divergence from the most recent common ancestor from lineage IV (approximately 200 years).

4 Discussion

There are several important aspects of our study and findings. First, we studied a comprehensive dataset of publicly available full-segment LASV sequences, spanning West Africa and host

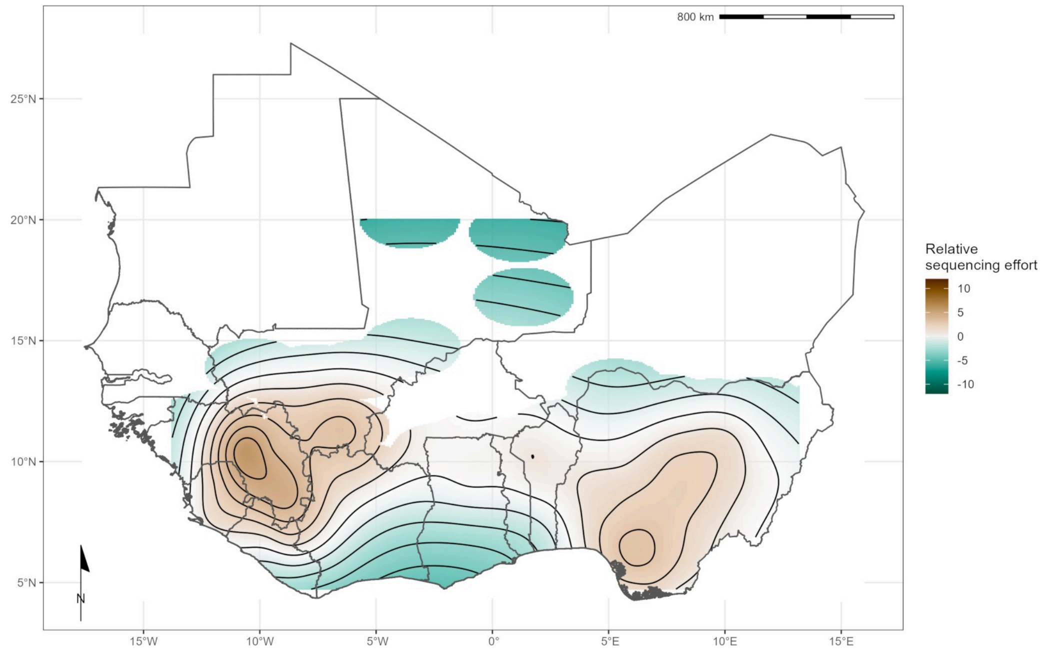

Fig 2. Modelled relative sequencing effort derived from both human and rodent samples. Greatest sequencing effort coincides with areas where sampling in humans (Edo, Nigeria and Kenema, Sierra Leone) and rodents (Faranah, Guinea) have historically been focussed. Shapefiles for basemap layer obtained from GADM 4.0.2 (www.gadm.org) [29].

https://doi.org/10.1371/journal.pgph.0002159.g002

species, to inform our understanding of the phylogeny of LASV dispersal. Second, we identified substantial variability in the origin of available sequences and completeness of records. Third, we showed strong geographic clustering among lineages supporting prior hypotheses of radiation from both Nigeria and a subsequent introduction into Liberia [1]. Fourth, the synthesis of available metadata highlights important gaps in currently available data, including spatial bias in the sequencing of samples we propose that this should be used to inform the design of epidemiological programmes going forward.

Our analyses of 2,298 LASV sequences obtained from GenBank highlights the spatial biases in the availability of sequence data that may limit our understanding of the current and

Table 1. The most recent common ancestor (MRCA) stratified by host and country of collection of Lassa mammarenavirus (LASV) S and L segments. Samples were collected between 1969–2018.

Host species	*Country*	*S segment MRCA*	*L segment MRCA*
Homo sapiens (n = 1181)	Benin	1995	1989
	Guinea	1895	1871
	Liberia	1895	1627
	Nigeria	1681	1498
	Sierra Leone	1901	1874
	Togo	2016	2014
Hylomyscus pamfi (n = 2)	Nigeria	1681	1498
Mastomys erythroleucus (n = 18)	Guinea	1975	2010
	Nigeria	2008	2006
Mastomys natalensis (n = 36)	Guinea	1938	1997
	Mali	1951	2007
	Sierra Leone	1909	1979

https://doi.org/10.1371/journal.pgph.0002159.t001

historic dispersal of LASV lineages in West Africa. First, sequence data was typically obtained from three of the eight endemic countries: Nigeria, Guinea and Sierra Leone. We found a strong association between LASV transmission intensity, measured as the number of reported human cases, and the number of available sequences (Fig 1). When stratifying by host species this trend did not remain; with rodent derived samples showing no association with LASV transmission intensity. LASV sequences obtained from rodents were relatively over-sampled from low transmission intensity settings; this may indicate a substantial lack of human case reporting in these locations. In high LASV transmission intensity settings rodents appeared to be relatively undersampled compared to human derived sequences, this is potentially an important source of bias when attempting to infer phylogeography within the reservoir host of this zoonotic pathogen. Sequence data from other countries, and more regions within them, across West Africa are required to increase confidence in the timelines of the currently inferred westward expansion. Greater focus needs to be placed on acquiring sequences from the rodent host to understand viral genetic diversity within the primary reservoir species. Comparing rodent derived sequences with those acquired from spillover into human populations may also allow identification of genetic drivers of transmission [36].

We used the number of reported cases between 2008–2023 as a measure of transmission intensity of LASV within a country over the study period. This may not be suitable due to substantial biases in the reporting of Lassa fever cases due to different surveillance sensitivity between countries and over time [3]. While unlikely, it is possible, that the magnitude of viral transmission within a country may not be directly related to the number of observed cases Therefore, in the absence of an unbiased estimate of transmission intensity we believe the number of reported cases is an acceptable proxy to model sequencing effort. We mapped the relationship between where sequences have been obtained as a function of this transmission intensity to produce a measure of relative sequencing effort to identify regions where increased LASV sequencing are required to counteract current sequencing biases (Fig 2).

Second, geographic clustering of LASV lineages, suggest isolated events of human-to-rodent transmission (S2 Fig) and the emergence of LASV dating from 1498 in Nigeria (Table 1). Similarly, Olayemi *et al*. report evidence of earlier emergence of the virus in humans than in rodents in Nigeria [34]. Comparatively limited data from non-human hosts with limited genome coverage, (69/703 sequences encompassed complete genes) produce important uncertainty around the observation of human-to-rodent transmission. Taken together, this data highlight limited surveillance among animal species, necessitating further investments in data acquisition and sharing to accurately define the spatiotemporal expansion of LASV in West Africa.

The phylogenetic analysis of LASV stratified by host species supports spatial evolution, in addition to intra-host viral evolution (S2 Fig). For instance, LASV sequences from *M. erytholeucus* sampled in Nigeria and Guinea clustered within lineages III and IV, respectively. Interestingly, these isolates appear to occur after the emergence of the most recent common ancestor virus circulating among humans and *M. natalensis* in these countries (Table 1), suggesting introduction of LASV into *M. erythroleucus* populations was a consequence of pathogen circulation in human and *M. natalensis* populations. Sequences from *M. natalensis* in Sierra Leone exhibit minimal clustering, and were interspersed with sequences from humans, potentially representing isolated events of pathogen introduction into human populations with spillback into commensal rodent populations (i.e., reverse zoonosis). The most recent common ancestor of LASV sequences from *M. natalensis* in Sierra Leone suggest a later emergence of the virus in this country. Our findings corroborate those of Olayemi et al., that within Sierra Leone LASV appears to have emerged in human hosts before rodents [34]. However, this data must be caveated by the limited information from rodent species in these locations.

There is a lower coverage of rodent-derived LASV sequences, with those from the primary reservoir *M*. *natalensis* forming fewer than one-third of all sequences (n = 642, 28%), with substantially lower sampling of other possible rodent hosts, including other *Mastomys* species. Rodent sampling has not increased at the same rate as human samples despite increased sampling effort since 2008 [15,16,37]. There is substantial heterogeneity in the locations in which rodent and human samples are available. For example, a relatively high number of rodent samples (n = 429) have been obtained from Guinea while few human sequences (n = 20) are available from these locations. The inverse is true of Nigeria where most human derived sequences are obtained (n = 1,147) but only 85 rodent sequences are available, and all of these from a single state (Edo, Nigeria). The number of suspected and reported cases was found to be positively but non-linearly associated with the number of available sequences. This is suggestive of a consolidation of research and focus of sampling in areas historically with high numbers of human cases but there remains to a paucity of sequences from elsewhere in the endemic region. The limited number of full segment sequences from rodents, from few geographic locations, limits our understanding of viral radiation in rodent hosts, particularly from species which are not considered the primary reservoir, e.g., *H*. *pamfi*. The most recent common ancestor for the viral sequence obtained from *H*. *pamfi* is estimated to be in the late 1600s; it is therefore possible lineage VI and/or *H*. *pamfi* as a reservoir of LASV has gone undetected due to lack of sufficient sampling [16].

Interpreting available LASV sequences is challenging for several reasons. A large proportion of available sequences (70%) have been obtained within Lassa fever research programs, representing spatial ascertainment bias [38–40]. In addition to these spatial biases temporal biases are apparent. Since 2016 there has been a substantial increase in the number of LASV sequences available in NCBI GenBank, reflecting increasing research effort, availability of sequencing platforms and increased data collection during Lassa fever epidemics, such as in the 2018 Nigeria Lassa fever outbreak [41–43]. There are notably fewer recorded sequences of LASV from Benin, Togo, and Ghana, suggesting a potential a gap in surveillance and research capacity in these locations or a lack of circulating LASV, despite several reported outbreaks [44–46]. Phylogenetic analysis on 60% of our initial dataset, following removal of sequences due to incompleteness or missing geographic and year of collection information (n = 1,045) demonstrated geographic clustering of LASV lineages, supporting prior analyses [16,33,34,44,47–50]. Increased data availability from Nigeria following increased LASV surveillance allowed regional analysis of phylogeny for lineages II and III supporting previous findings of expansion of these lineages from North-East Nigeria to the South-West of the country [51–53].

A substantial number (n = 869) of the sequences retrieved corresponded to short fragments (< 1 Kb) probably derived from PCR products used for diagnostic purposes rather than for viral genomic surveillance. LASV is a segmented virus, and it was not possible to identify complete genome sequences since both S and L segments are reported separately on the sequence's repository. The molecular clock analyses from L protein indicated an earlier emergence of LASV when compared to S segment analysis (1498 and 1681 respectively), potentially because the viral RNA polymerase (L protein) is less affected by selective pressure than the S segment [11,47,54]. In arenavirus LCMV populations the number of mutations in NP and GP regions were more abundant that in the polymerase region, suggesting that mutations in the L region could be less tolerated [55]. In addition, investigations on LASV intra-host evolution, both in humans and rodents, showed that the substitution ratios varied widely across LASV genes, with the GPC gene, encoded in the S segment showing higher within-host diversity. Furthermore, this same study suggested that despite the short duration of LASV infection, B and T cells response seem to positively select escape mutations GPC [11]. Despite these challenges, this study has synthesised currently available data on LASV sequences to investigate the location and period of sampling to reconstruct the dispersal of viral lineages across the endemic region. Despite the regionalisation of LF being driven by rodent-to-human transmission, there remains scarce LASV genomic data from non-human hosts. We have mapped the locations of relative under sampling to guide targeted efforts to counteract biases in currently available data for both rodent and human derived sequences. Expanded sampling of LASV from animal species within the endemic region will improve our current understanding of LASV evolution and ecology and improve confidence in current estimates of westward expansion of Lassa fever in humans. Further understanding of the viral evolution dynamics of LASV and spatial expansion of current lineages will be vital to ensure adequate diagnostic tools are available to respond to the expected sporadic outbreaks of Lassa Fever across the region.

Supporting information

S1 Fig. Map of West Africa. displays a map of West Africa with country names for reference with Figs 1 and 2. Shapefiles for mapping obtained from GADM 4.0.2 [29].
(TIF)

S2 Fig. Time-calibrated phylogeny for both the small segment (S) and large segment (L) from included LASV sequences.
(TIF)

S1 Data. GenBank accession number of analysed sequences. This dataset includes available data about host, country, region, year, sequence length, genome segment (L or S) and predicted MRCA.
(XLSX)

S2 Data. Dataset on confirmed Lassa fever cases. This presents the number of confirmed cases of Lassa fever reported from countries between 2008 and 2023 at a subnational level that were used to calculate the number of cases per 100,000 people. References for the reports used to produce this dataset are included.
(CSV)

Acknowledgments

The authors would like to thank the researchers who submitted LASV sequences and associated metadata to GenBank. We hope our work encourages other researchers to do the same.

Author Contributions

Conceptualization: Liã Bárbara Arruda, Hayley Beth Free, David Simons.

Data curation: Liã Bárbara Arruda, Hayley Beth Free, David Simons.

Formal analysis: Liã Bárbara Arruda, Hayley Beth Free, David Simons.

Funding acquisition: Rashid Ansumana, Timothy D. McHugh, Francine Ntoumi, Alimuddin Zumla, Richard Kock.

Investigation: Liã Bárbara Arruda, Hayley Beth Free, David Simons, Linzy Elton, Najmul Haider.

Methodology: Liã Bárbara Arruda, Hayley Beth Free, David Simons, Linzy Elton, Najmul Haider.

Project administration: Liã Bárbara Arruda.

Software: David Simons.

Supervision: Liã Bárbara Arruda, Isobella Honeyborne, Timothy D. McHugh, Francine Ntoumi, Alimuddin Zumla.

Validation: Liã Bárbara Arruda, Isobella Honeyborne, Danny Asogun, Timothy D. McHugh, Francine Ntoumi, Alimuddin Zumla, Richard Kock.

Visualization: Liã Bárbara Arruda, David Simons.

Writing – original draft: Liã Bárbara Arruda, Hayley Beth Free, David Simons, Najmul Haider, Timothy D. McHugh, Alimuddin Zumla, Richard Kock.

Writing – review & editing: Liã Bárbara Arruda, Hayley Beth Free, David Simons, Rashid Ansumana, Linzy Elton, Najmul Haider, Isobella Honeyborne, Danny Asogun, Timothy D. McHugh, Francine Ntoumi, Alimuddin Zumla, Richard Kock.

References

1. Klitting R, Kafetzopoulou LE, Thiery W, Dudas G, Gryseels S, Kotamarthi A, et al. Predicting the evolution of the Lassa virus endemic area and population at risk over the next decades. Nature communications. 2022; 13: 5596. https://doi.org/10.1038/s41467-022-33112-3
2. WHO. Lassa fever factsheet. In: WHO [Internet]. World Health Organization; 2020 [cited 21 Oct 2020]. Available: http://www.who.int/csr/don/archive/disease/lassa_fever/en/.
3. Simons D. Lassa fever cases suffer from severe underreporting based on reported fatalities. International Health. 2022. https://doi.org/10.1093/inthealth/ihac076
4. McCormick JB, Webb PA, Krebs JW, Johnson KM, Smith ES. A prospective study of the epidemiology and ecology of Lassa fever. J Infect Dis. 1987; 155: 437–44. https://doi.org/10.1093/infdis/155.3.437 PMID: 3805771
5. Basinski AJ, Fichet-Calvet E, Sjodin AR, Varrelman TJ, Remien CH, Layman NC, et al. Bridging the gap: Using reservoir ecology and human serosurveys to estimate Lassa virus spillover in West Africa. Wesolowski A, editor. PLoS Comput Biol. 2021; 17: e1008811. https://doi.org/10.1371/journal.pcbi.1008811 PMID: 33657095
6. Asogun DA, Gunther S, Akpede GO, Ihekweazu C, Zumla A. Lassa Fever: Epidemiology, Clinical Features, Diagnosis, Management and Prevention. [Review]. Infectious Disease Clinics of North America. 2019; 33: 933–951. https://doi.org/10.1016/j.idc.2019.08.002
7. Takah NF, Brangel P, Shrestha P, Peeling R. Sensitivity and specificity of diagnostic tests for Lassa fever: a systematic review. BMC Infectious Diseases. 2019; 19: 647. https://doi.org/10.1186/s12879-019-4242-6 PMID: 31324229
8. Nnaji ND, Onyeaka H, Reuben RC, Uwishema O, Olovo CV, Anyogu A. The deuce-ace of Lassa Fever, Ebola virus disease and COVID-19 simultaneous infections and epidemics in West Africa: clinical and public health implications. Tropical Medicine and Health. 2021; 49: 102. https://doi.org/10.1186/s41182-021-00390-4 PMID: 34965891
9. Ashcroft JW, Olayinka A, Ndodo N, Lewandowski K, Curran MD, Nwafor CD, et al. Pathogens that Cause Illness Clinically Indistinguishable from Lassa Fever, Nigeria, 2018. Emerging Infectious Diseases. 2022; 28: 994–997. https://doi.org/10.3201/eid2805.211153 PMID: 35226800
10. Lo Iacono G, Cunningham AA, Fichet-Calvet E, Garry RF, Grant DS, Khan SH, et al. Using Modelling to Disentangle the Relative Contributions of Zoonotic and Anthroponotic Transmission: The Case of Lassa Fever. PLoS Neglected Tropical Diseases. 2015. https://doi.org/10.1371/journal.pntd.0003398 PMID: 25569707
11. Andersen KG, Shapiro BJ, Matranga CB, Sealfon R, Lin AE, Moses LM, et al. Clinical Sequencing Uncovers Origins and Evolution of Lassa Virus. Cell. 2015. https://doi.org/10.1016/j.cell.2015.07.020 PMID: 26276630
12. Oti VB. A Reemerging Lassa Virus: Aspects of Its Structure, Replication, Pathogenicity and Diagnosis. In: Alfonso J. Rodriguez-Morales, editor. Current Topics in Tropical Emerging Diseases and Travel Medicine. BoD–Books on Demand; 2018.
13. Bangura U, Buanie J, Lamin J, Davis C, Bongo GN, Dawson M, et al. Lassa Virus Circulation in Small Mammal Populations in Bo District, Sierra Leone. BIOLOGY-BASEL. ST ALBAN-ANLAGE 66, CH-

4052 BASEL, SWITZERLAND: MDPI; 2021. https://doi.org/10.3390/biology10010028 PMID: 33466234

14. Forni D, Sironi M. Population Structure of Lassa Mammarenavirus in West Africa. Viruses. 2020; 12: 437. https://doi.org/10.3390/v12040437 PMID: 32294960

15. Lecompte E, Fichet-Calvet E, Daffis S, Koulémou K, Sylla O, Kourouma F, et al. Mastomys natalensis and Lassa fever, West Africa. Emerging Infectious Diseases. 2006. https://doi.org/10.3201/eid1212.060812 PMID: 17326956

16. Olayemi A, Cadar D, Magassouba N, Obadare A, Kourouma F, Oyeyiola A, et al. New Hosts of The Lassa Virus. Scientific Reports. 2016. https://doi.org/10.1038/srep25280 PMID: 27140942

17. Wulff H, Fabiyi A, Monath TP. Recent isolations of Lassa virus from Nigerian rodents. Bull World Health Organ. 1975; 52: 609–613. PMC2366652. PMID: 1085216

18. Yadouleton A, Agolinou A, Kourouma F, Saizonou R, Pahlmann M, Bedié SK, et al. Lassa virus in pygmy mice, Benin, 2016–2017. Emerging Infectious Diseases. 2019. https://doi.org/10.3201/eid2510.180523 PMID: 31365854

19. Naidoo D, Ihekweazu C. Nigeria's efforts to strengthen laboratory diagnostics—Why access to reliable and affordable diagnostics is key to building resilient laboratory systems. African Journal of Laboratory Medicine. 2020; 9: 1–5. https://doi.org/10.4102/ajlm.v9i2.1019 PMID: 32934913

20. National Center for Biotechnology Information. National Center for Biotechnology Information. 2022 [cited 3 Feb 2022]. Available: https://www.ncbi.nlm.nih.gov/.

21. Becker G, Lawrence M. genbankr: Parsing GenBank files into semantically useful objects. 2021.

22. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available: https://www.R-project.org/.

23. Kahle D, Wickham H. ggmap: Spatial Visualization with ggplot2. The R Journal. 2013; 5: 144–161. https://doi.org/10.32614/RJ-2013-014

24. Pebesma E. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal. 2018; 10: 439–446. Available: https://journal.r-project.org/archive/2018/RJ-2018-009/index.html.

25. Hijmans RJ, Barbosa M, Ghosh A, Mandel A. geodata: Download Geographic Data. 2023. Available: https://CRAN.R-project.org/package=geodata.

26. Agbonlahor DE, Akpede GO, Happi CT, Tomori O. 52 Years of Lassa Fever outbreaks in Nigeria, 1969–2020: An epidemiologic analysis of the temporal and spatial trends. The American Journal of Tropical Medicine and Hygiene. 2021; 1. https://doi.org/10.4269/ajtmh.20-1160 PMID: 34460421

27. Shaffer JG, Schieffelin JS, Gbakie M, Alhasan F, Roberts NB, Goba A, et al. A medical records and data capture and management system for Lassa fever in Sierra Leone: Approach, implementation, and challenges. PLoS ONE [Electronic Resource]. 2019; 14: e0214284. https://doi.org/10.1371/journal.pone.0214284 PMID: 30921383

28. Garry RF. Lassa fever—the road ahead. Nat Rev Microbiol. 2023; 21: 87–96. https://doi.org/10.1038/s41579-022-00789-8 PMID: 36097163

29. Database of Global Administrative Areas. GADM. 2022 [cited 25 Apr 2021]. Available: https://gadm.org/index.html.

30. Hijmans RJ. terra: Spatial Data Analysis. 2022. Available: https://rspatial.org/terra/.

31. Wood SN. Generalized Additive Models: An Introduction with R. 2nd ed. Chapman and Hall/CRC; 2017.

32. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evolution. 2018. https://doi.org/10.1093/ve/vey016

33. Olayemi A, Fichet-Calvet E. Systematics, ecology, and host switching: Attributes affecting emergence of the Lassa virus in rodents across western Africa. Viruses. 2020. https://doi.org/10.3390/v12030312 PMID: 32183319

34. Olayemi A, Adesina AS, Strecker T, Magassouba N, Fichet-Calvet E. Determining ancestry between rodent-and human-derived virus sequences in endemic foci: Towards a more integral molecular epidemiology of lassa fever within West Africa. Biology. 2020. https://doi.org/10.3390/biology9020026 PMID: 32046182

35. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Systematic Biology. 2018. https://doi.org/10.1093/sysbio/syy032 PMID: 29718447

36. Whitlock AOB, Bird BH, Ghersi B, Davison AJ, Hughes J, Nichols J, et al. Identifying the genetic basis of viral spillover using Lassa virus as a test case. R Soc Open Sci. 2023; 10: 221503. https://doi.org/10.1098/rsos.221503 PMID: 36968239

37. Lecompte E, Brouat C, Duplantier JM, Galan M, Granjon L, Loiseau A, et al. Molecular identification of four cryptic species of Mastomys (Rodentia, Murinae). Biochemical Systematics and Ecology. 2005. https://doi.org/10.1016/j.bse.2004.12.015

38. Townsend Peterson A, Moses LM, Bausch DG. Mapping transmission risk of lassa fever in West Africa: The importance of quality control, sampling bias, and error weighting. PLoS ONE. 2014. https://doi.org/10.1371/journal.pone.0100711 PMID: 25105746

39. Ehichioya DU, Hass M, Ölschläger S, Becker-Ziaja B, Onyebuchi Chukwu CO, Coker J, et al. Lassa fever, Nigeria, 2005–2008. Emerging Infectious Diseases. 2010. https://doi.org/10.3201/eid1606.100080 PMID: 20507773

40. Khan SH, Goba A, Chu M, Roth C, Healing T, Marx A, et al. New opportunities for field research on the pathogenesis and treatment of Lassa fever. Antiviral Research. 2008. https://doi.org/10.1016/j.antiviral.2007.11.003 PMID: 18241935

41. Maxmen A. Deadly Lassa-fever outbreak tests Nigeria's revamped health agency. Nature. 2018; 555: 421–422. https://doi.org/10.1038/d41586-018-03171-y

42. Siddle KJ, Eromon P, Barnes KG, Mehta S, Oguzie JU, Odia I, et al. Genomic Analysis of Lassa Virus during an Increase in Cases in Nigeria in 2018. New England Journal of Medicine. 2018; 379: 1745–1753. https://doi.org/10.1056/NEJMoa1804498 PMID: 30332564

43. Ilori EA, Frank C, Dan-Nwafor CC, Ipadeola O, Krings A, Ukponu W, et al. Increase in Lassa Fever Cases in Nigeria, January–March 2018. Emerging Infectious Diseases. 2019;25. https://doi.org/10.3201/eid2505.181247

44. Yadouleton A, Picard C, Rieger T, Loko F, Cadar D, Kouthon EC, et al. Lassa fever in Benin: description of the 2014 and 2016 epidemics and genetic characterization of a new Lassa virus. Emerging Microbes & Infections. 2020; 1–23. https://doi.org/10.1080/22221751.2020.1796528 PMID: 32723007

45. World Health Organisation. Lassa Fever–Togo. 21 Nov 2022 [cited 24 Nov 2022]. Available: https://www.who.int/emergencies/disease-outbreak-news/item/2022-DON362.

46. Ghana Health Services. Lassa Fever Press Release Ghana 2023. 2023 [cited 19 Apr 2023]. Available: https://osf.io/ft2gy/.

47. Ibukun FI. Inter-lineage variation of lassa virus glycoprotein epitopes: A challenge to lassa virus vaccine development. Viruses. 2020. https://doi.org/10.3390/v12040386 PMID: 32244402

48. Lalis A, Leblois R, Lecompte E, Denys C, ter Meulen J, Wirth T. The Impact of Human Conflict on the Genetics of Mastomys natalensis and Lassa Virus in West Africa. PLoS ONE. 2013;7. https://doi.org/10.1371/journal.pone.0037068

49. Manning JT, Forrester N, Paessler S. Lassa virus isolates from Mali and the Ivory Coast represent an emerging fifth lineage. Frontiers in Microbiology. 2015. https://doi.org/10.3389/fmicb.2015.01037 PMID: 26483768

50. Wiley MR, Fakoli L, Letizia AG, Welch SR, Ladner JT, Prieto K, et al. Lassa virus circulating in Liberia: a retrospective genomic characterisation. The Lancet Infectious Diseases. 2019. https://doi.org/10.1016/S1473-3099(19)30486-4 PMID: 31588039

51. Bowen MD, Rollin PE, Ksiazek TG, Hustad HL, Bausch DG, Demby AH, et al. Genetic Diversity among Lassa Virus Strains. Journal of Virology. 2000; 74: 6992–7004. https://doi.org/10.1128/jvi.74.15.6992-7004.2000 PMID: 10888638

52. Ehichioya DU, Hass M, Becker-Ziaja B, Ehimuan J, Asogun DA, Fichet-Calvet E, et al. Current molecular epidemiology of Lassa virus in Nigeria. Journal of Clinical Microbiology. 2011. https://doi.org/10.1128/JCM.01891-10 PMID: 21191050

53. Naidoo D, Ihekweazu C. Nigeria's efforts to strengthen laboratory diagnostics–Why access to reliable and affordable diagnostics is key to building resilient laboratory systems. African Journal of Laboratory Medicine. 2020;9. https://doi.org/10.4102/ajlm.v9i2.1019 PMID: 32934913

54. Hastie KM, Saphire EO. Lassa virus glycoprotein: stopping a moving target. Current Opinion in Virology. 2018. https://doi.org/10.1016/j.coviro.2018.05.002 PMID: 29843991

55. Grande-Pérez A, Sierra S, Castro MG, Domingo E, Lowenstein PR. Molecular indetermination in the transition to error catastrophe: systematic elimination of lymphocytic choriomeningitis virus through mutagenesis does not correlate linearly with large increases in mutant spectrum complexity. Proc Natl Acad Sci U S A. 2002; 99: 12938–12943. https://doi.org/10.1073/pnas.182426999 PMID: 12215495