**Draft: LASV phylogenetics**

1. Introduction

Lassa fever (LF) is a viral haemorrhagic disease, caused by *Lassa mammarenavirus* (LASV) which is endemic in several countries of West Africa. Estimations based on longitudinal serosurveys in Sierra Leone in late 80’s indicated that 100,000 t 300,000 cases of Lassa fever would occur annually in West Africa (McCornick, 1987). However, a recent spillover risk model predicted that up to 900,000 West-Africans might be infected by LASV every year (Basinski, 2021). The discrepancy in estimations is a result of the limited epidemiological research on Lassa fever.

Humans become infected with LASV upon contact with or inhalation of excretions from rodent species (Andersen *et al.*, 2015; Lo Iacono *et al.*, 2015; Oti, 2018). The Natal multimammate mouse (*Mastomys natalensis*) is the primary reservoir, however other rodents have been found to be infected with LASV, namely *Mastomys erythroleucus, Hylomyscus pamfi* and *Mus baoulei* (Lecompte *et al.*, 2006; Olayemi *et al.*, 2016; Yadouleton *et al.*, 2019; Forni and Sironi, 2020). Human-to-human transmission has been reported, typically associated with nosocomial outbreaks, however, it is expected that this occurs at a lower frequency (Lo Iacono *et al.*, 2015).

LASV is a segmented ssRNA- virus of the family *Arenaviridae* (Günther and Lenz, 2004; Hallam et al., 2018). The viral genome consists of two segments encompassing four non-overlapping genes. The large segment (L) encodes the RNA-dependent RNA polymerase (RdRP) and the matrix protein (Z), while the small segment (S) encodes the structural proteins: nucleoprotein (NP) and the glycoprotein complex (GPC) (MacLachlan and Dubovi, 2011). LASV has been classified into seven lineages and demonstrates spatial distribution across the endemic range. Phylogenetic analysis and molecular dating suggest a westward route of dispersal of the lineages, from the most recent common ancestor in Nigeria. (Bowen *et al.*, 2000; Andersen *et al.*, 2015; Manning, Forrester and Paessler, 2015; Olayemi *et al.*, 2016, 2020; Whitmer *et al.*, 2018).

Here we compiled a comprehensive dataset of all publicly available full-segment LASV sequences, spanning West Africa and host species, to better understand LASV phylodynamics. The collection of the associated meta-data also highlighted gaps of information that could be filled through sustainable epidemiological programmes. A better understanding of LASV phylogeography would improve control strategies in endemic regions (i.e. through optimised assays) and support implementation of measures to prevent an increase in the size of the endemic region.

1. Methods

2.1. Data Collection and Processing

LASV nucleotide and protein sequences were obtained from the National Centre for Biotechnology Information (NCBI) GenBank (<https://www.ncbi.nlm.nih.gov/>). The search query run on 2021-09-24 was for “Lassa mammarenavirus” in the organism field of the NCBI nucleotide dataset. Data was downloaded through the NCBI Entrez API and the GenBankr package and assessed using the R statistical package. Sequences with large portions of missing nucleotide data on the L- or S-segment (10% missing compared to reference sequences) were excluded.Nucleotide sequences were aligned and translated using Geneious Prime 20201.2. Visual inspection was performed, and entries were excluded when: i) contained >100 continuous ambiguous nucleotide calls ii) lack of associated metadata (collection year, host species, country, and geographical region of sampling)

2.2. Phylogenetic Analysis

Phylogenetic analysis was undertaken through Bayesian Markov Chain Monte Carlo (MCMC) method using BEAST.v1.10.4 (<https://beast.community/programs>; Suchard *et al.*, 2018). 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 (Olayemi and Fichet-Calvet, 2020; Olayemi *et al.*, 2020). 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 (<http://tree.bio.ed.ac.uk/software/figtree/>; Rambaut *et al.*, 2018).

1. Results

3.1. Compiled Dataset

The initial dataset was compiled of 2,298 sequences. Country data were missing for 134 records. For phylogenetic analysis, 680 sequences of complete S segment (NP and GPC genes) and 578 sequences of the RNA polymerase gene (L protein) were used. Accession numbers of included and excluded sequences are available in Supplementary table 1.

3.2. Descriptive Analysis

Descriptive analysis of available metadata was performed for 2,298 records. Most samples have been obtained since 2000, more than half of all samples have been obtained since 2016 (Figure 1A), 190 samples did not have an associated year of sampling. Sequences were most commonly recorded from Nigeria (53%), followed by Guinea (19%) Sierra Leone (13%), Liberia (4%) and Cote d’Ivoire (3%) with the remainder obtained from, Benin, Ghana, Mali and Togo or the country of origin not reported(Figure 1C). Regional-level data of records were available for 41% of samples (n=958), of these 54% were from Nigeria, 40% from Guinea and less than 2% from Benin, Ghana, Liberia and Sierra Leone.. The regional distribution of records is shown as points in Figure 1B with the size of the points related to the number of samples obtained from that site. .

Human-derived LASV sequences comprise the majority of available records (62%), other species providing viral sequences include *Mastomys natalensis* (27%), *Mastomys spp.* (2%), , *Mastomys erythroleucus*, *Mus baoulei*and *Hylomyscus pamfi* (all <1%) (Figure 1B). The species sampled was not documented in 8% of records.

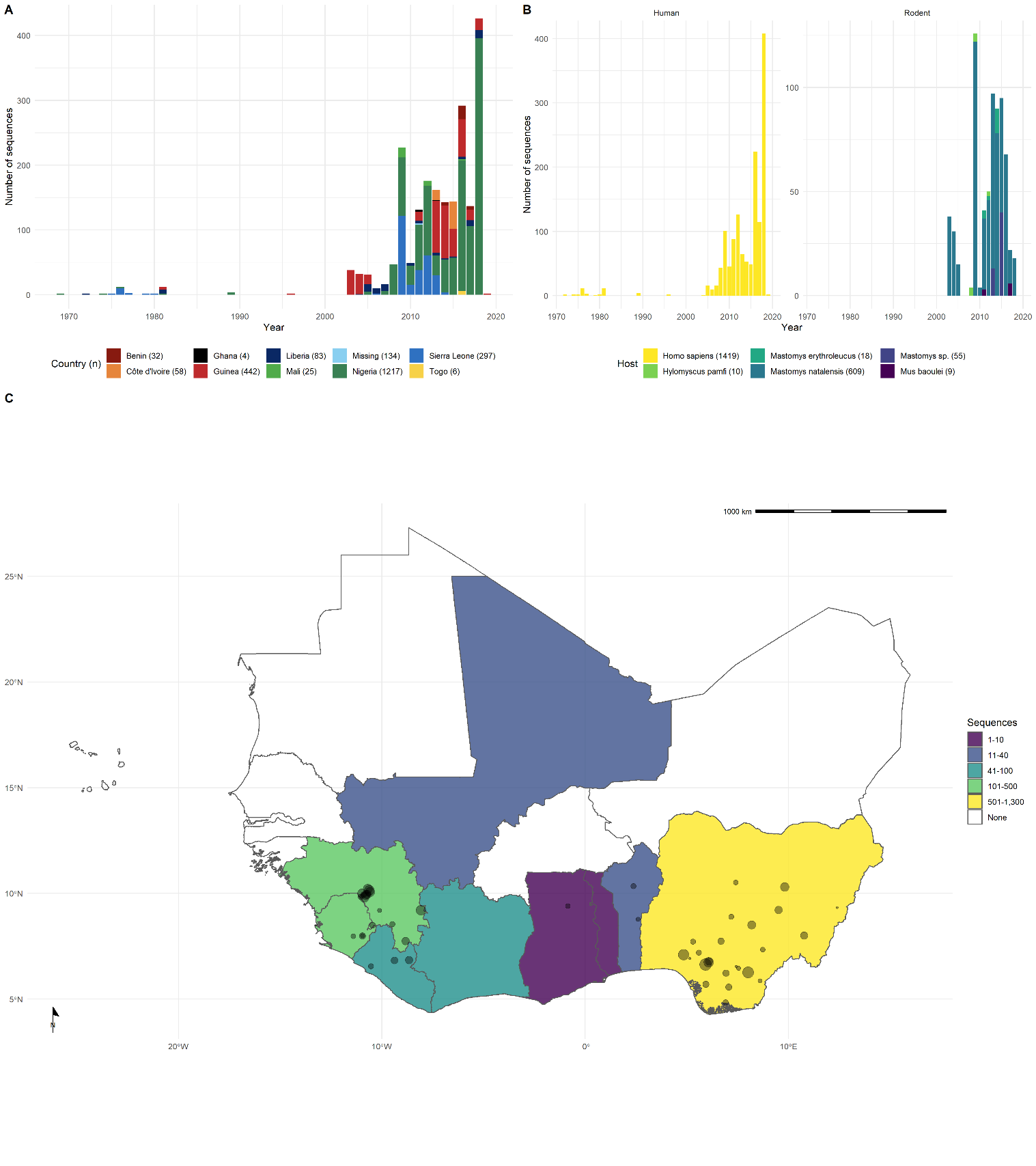


Figure 1 – **A** The year of sampling for all *Lassa mammarenavirus* sequences retrieved from NCBI GenBank with colour corresponding to country of origin (n = 2298). **B** Host species sampled LASV sequences were obtained by year of collection (n=2088). **C** Location data for the complied *Lassa mammarenavirus (LASV)* sequences as recorded in NCBI GenBank database Country of LASV sample collection is shown by country colour (n=2164). The location of the sample where available is shown by the circles on the map (n=956). The size of the circle corresponds to the number of samples obtained at that location.

3.3. Phylogenetic Analysis

Sequences for each segment of LASV showed clustering according to the described lineages I-VII which also correspond to a geographical clustering where lineages I-III and VI circulate in Nigeria, lV in Liberia, Guinea and Sierra Leone, V in Mali and VII in Togo (Figure 4). In this analysis sequences of lineage V from Cote d’Ivoire were not 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 predicted in the year 950, compared to 1361 for the S segment (Table 1).

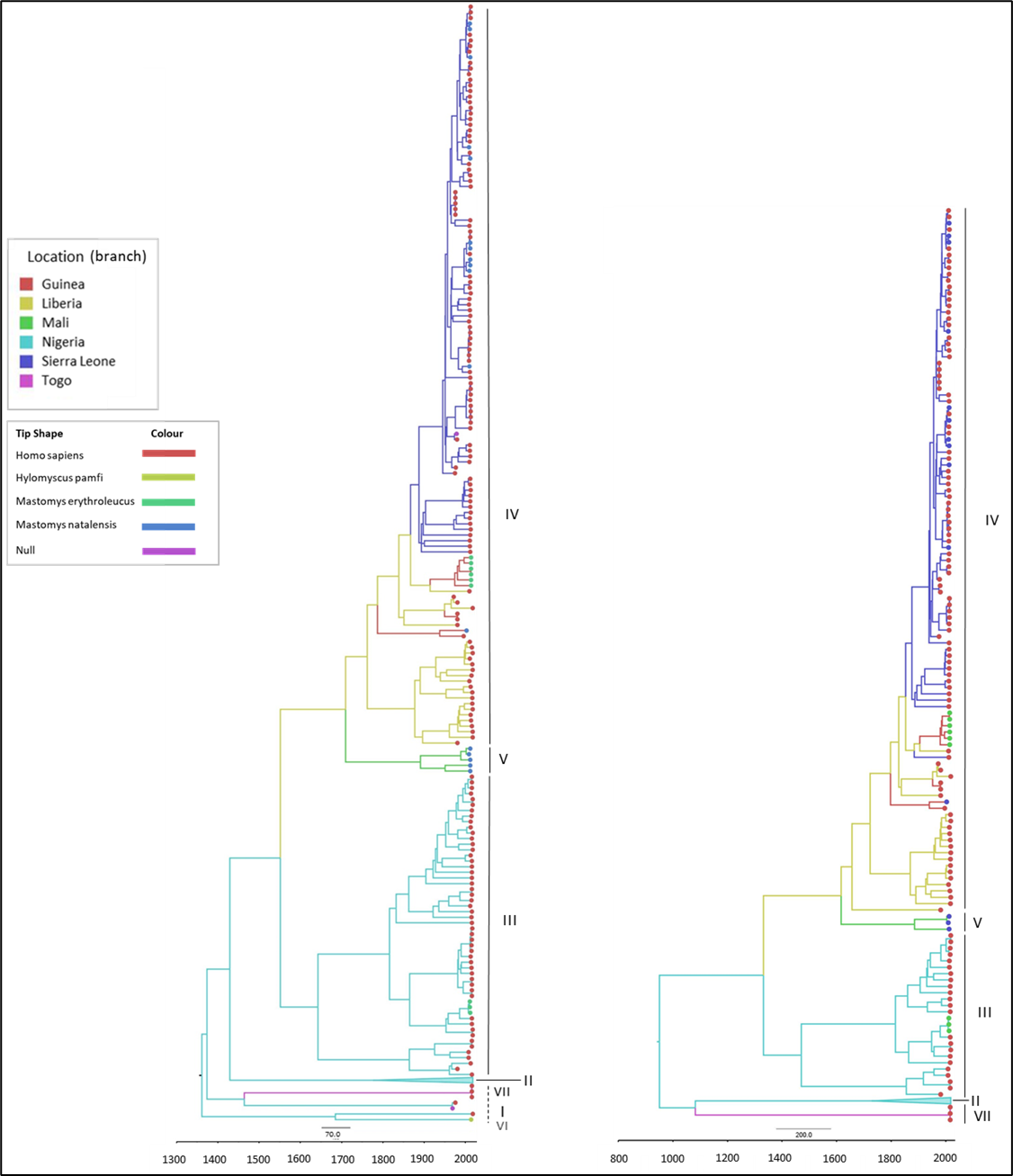


Figure 4 – Time-calibrated phylogenetic analysis of Lassa mammarenavirus (LASV) small (S) segment (left; n=650) and large (L) segment (right; n=518). Predicted emergency date (year) is indicated on the axis. Country of collection is indicated by branch colour (key upper left) and host species of collection is indicated by tip shape colour (key lower left). This analysis used Bayesian Markov Chain Monte Carlo method of inference. Lineage is indicated on the right-hand side of each tree; lineage II is collapsed and consisted of sequences from humans only.

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



There was a lack of sequence information from lineage I and VI, however, phylogeny suggests these lineages are basal to others in Nigeria (Figure 4). Lineage VII in Togo is most closely related to Nigerian isolates and potentially had diverged between 500-900 years ago. The divergence of lineage III and IV is predicted to have occurred between the years 1332-1551. Introduction to countries west of Nigeria appear to disperse initially to Liberia, followed by Guinea in the 1700s, then Sierra Leone and Mali approximately 100 years later. In this analysis, there is a lack of full segment sequences from lineage V, apparent in the time difference from when this lineage diverged from IV and its most recent common ancestor (approximately 200 years). Regional-level data was available for sequences obtained from Nigeria (Figure 5). The lineages circulating in Nigeria also tend to form regional clusters, with lineage II dispersed in the southern region and III across the central region of the country.

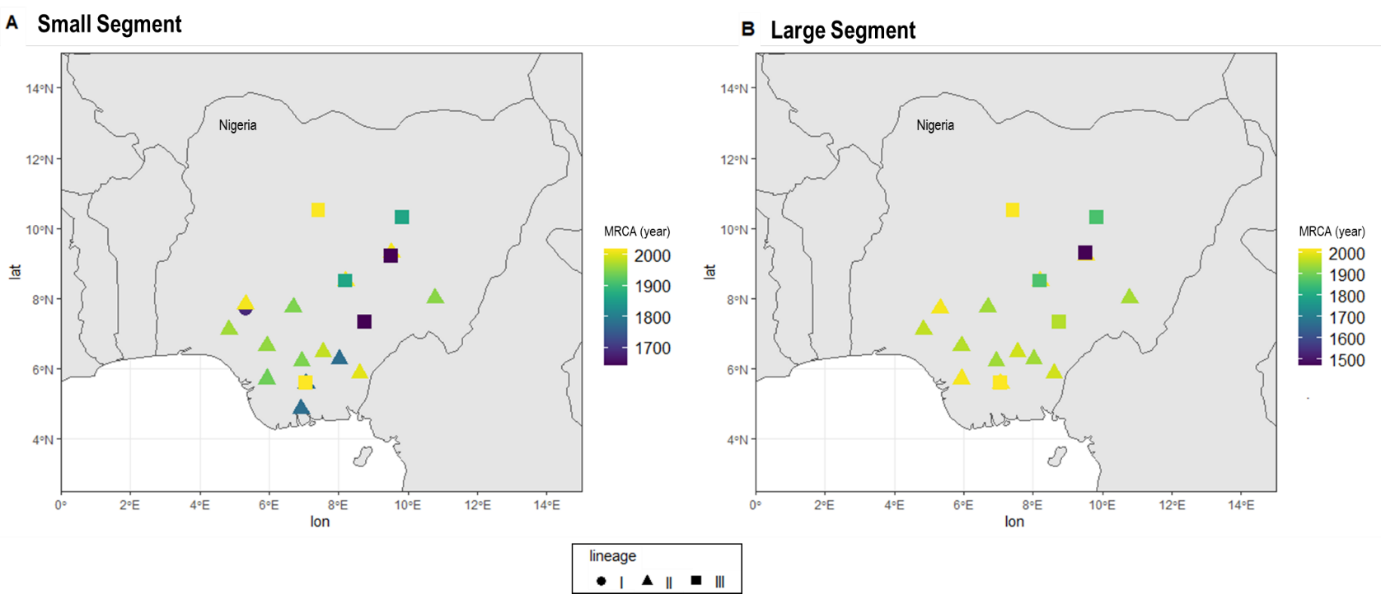


Figure 5 – Regional-level data of Lassa mammarenavirus used in phylogenetic analysis from Nigerian data, plotted longitude (lon) by latitude (lat). Lineage is according to shape and the most recent common ancestor (MRCA) according to the colour gradient. (A) small segment data and (B) large segment data.

The phylogenetic analysis of LASV according to host species appears to maintain the trend of spatial evolution, as opposed to intra-host viral evolution (Figure 4). 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 most recent common ancestor of humans and *M. natalensis* in the same country (Table 1), suggesting introduction of LASV to *M. erythroleucus* was secondary. Sequences from *M. natalensis* in Sierra Leone exhibited minimal clustering, and rather were interspersed with sequences from humans, potentially representing frequent introductory events. The most recent common ancestor of LASV sequences from *M. natalensis* suggested its presence in Sierra Leone was subsequent to other countries.

4 Discussion

Lassa fever (LF) and its causative agent Lassa virus (*Lassa mammarenavirus –* LASV) were firstly identified in Nigeria in 1969. Subsequent studies in the 70’s isolated LASV from the natal multimammate mouse *M. natalensis* which has been known as the main reservoir for this viral haemorrhagic disease. LASV lineages circulate endemically around the Mano River basin, pronominally in the West African countries of Nigeria, Sierra Leone, Guinea and Liberia, although Togo, Benin, Cote d’Ivoire, Ghana and Mali had also reported isolated cases and outbreaks of LF.

From the 2,295 LASV sequences collected from GenBank, it was observed an expressively high number of sequences from Nigerian cases (74%), followed by Sierra Leone (13%) which correlates with the location of Lassa fever research programs (Khan *et al.*, 2008; Ehichioya *et al.*, 2010; Townsend Peterson, Moses and Bausch, 2014). From 2015 there was a substantial increase in the number of LASV sequences in the repository, likely reflecting research efforts and/or Lassa fever epidemics, such as in the 2018 Nigeria Lassa fever outbreak where Edo state had the highest number of confirmed Lassa fever cases (Nigeria Centre for Disease Control, 2018). 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. Although there is a significant lack of information on country of sample collection, the phylogenetic analysis showed LASV lineages clustering geographically, supporting previous inferences (Lalis *et al.*, 2013; Manning, Forrester and Paessler, 2015; Olayemi *et al.*, 2016, 2020; Wiley *et al.*, 2019; Ibukun, 2020; Olayemi and Fichet-Calvet, 2020; Yadouleton *et al.*, 2020). Nigeria has improved their surveillance system in the past few years, leading to an overrepresentation of LASV sequences from this region on GenBank. Therefore, it was possible to evaluate the data at a regional-level corroborating previous finding that lineages II and III show clustering aligning with the progression of their ancestry from North-East to South-West (Bowen *et al.*, 2000; Ehichioya *et al.*, 2011).

There is a lower coverage of rodent derived LASV sequences, with those from the primary reservoir *M. natalensis* forming only 25% (n=518). The most recent common ancestor of LASV sequences from *M. natalensis* suggested its presence in Sierra Leone was subsequent to other countries, and similarly to Andersen et al.’s findings, also to LASV isolates from humans (2015). It is apparent that there is a low overall frequency of sampling of the other rodent species, despite the increased number of LASV sequences recorded in GenBank since 2008. Moreover, host species of LASV detection recorded at a genus level are limiting to analysis, as LASV has been detected in multiple *Mastomys* species (Figure 2; Lecompte et al., 2005, 2006; Olayemi et al., 2016). The lack of full segment sequences from rodents is limiting in this analysis, particularly from species which are not considered a primary reservoir, e.g. *H. pamfi.* In current literature, despite the initial report of LASV in *H. pamfi* in 2016, the most recent common ancestor of appears in the late 1600s (Table 1; Olayemi *et al.*, 2016). It is therefore possible lineage VI and/or *H. pamfi* as a reservoir of LASV has gone undetected due to lack of sampling.

Most of the sequences retrieved corresponded to short fragments derived from PCR products used for diagnostics purposes rather than molecular epidemiology surveillance. The phylogenetic analysis only included complete gene sequences to improve phylogenetic inferences. LASV is 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 emergency of LASV when compared to S segment analysis. This is likely due to the functionality of RNA polymerase properties of the L protein, essential for viral replication, thus expected to be conserved (Capul *et al.*, 2011). Contrastingly, for the S segment which encodes genes for structural proteins, which are subject to host immune pressure and therefore variation (Andersen *et al.*, 2015; Hastie and Saphire, 2018; Ibukun, 2020).

References

Andersen, K. G. *et al.* (2015) ‘Clinical Sequencing Uncovers Origins and Evolution of Lassa Virus’, *Cell*. doi: 10.1016/j.cell.2015.07.020.

Bowen, M. D. *et al.* (2000) ‘Genetic Diversity among Lassa Virus Strains’, *Journal of Virology*, 74(15), pp. 6992–7004. doi: 10.1128/jvi.74.15.6992-7004.2000.

Capul, A. A., de la Torre, J. C. and Buchmeier, M. J. (2011) ‘Conserved Residues in Lassa Fever Virus Z Protein Modulate Viral Infectivity at the Level of the Ribonucleoprotein’, *Journal of Virology*. doi: 10.1128/jvi.02081-10.

Centers for Disease Control and Prevention (2019) *Lassa Fever*. Available at: https://www.cdc.gov/vhf/lassa/index.html (Accessed: 18 May 2020).

Control, N. C. for D. (2018) *Lassa fever Situation Report*.

Ehichioya, D. U. *et al.* (2010) ‘Lassa fever, Nigeria, 2005-2008’, *Emerging Infectious Diseases*. doi: 10.3201/eid1606.100080.

Ehichioya, D. U. *et al.* (2011) ‘Current molecular epidemiology of Lassa virus in Nigeria’, *Journal of Clinical Microbiology*. doi: 10.1128/JCM.01891-10.

Forni, D. and Sironi, M. (2020) ‘Population Structure of Lassa Mammarenavirus in West Africa’, *Viruses*, 12(4), p. 437.

Günther, S. and Lenz, O. (2004) ‘Lassa virus’, *Critical Reviews in Clinical Laboratory Sciences*. doi: 10.1080/10408360490497456.

Hallam, S. J. *et al.* (2018) ‘Review of mammarenavirus biology and replication’, *Frontiers in Microbiology*, 9(AUG), pp. 1–8. doi: 10.3389/fmicb.2018.01751.

Hastie, K. M. and Saphire, E. O. (2018) ‘Lassa virus glycoprotein: stopping a moving target’, *Current Opinion in Virology*. doi: 10.1016/j.coviro.2018.05.002.

Lo Iacono, G. *et al.* (2015) ‘Using Modelling to Disentangle the Relative Contributions of Zoonotic and Anthroponotic Transmission: The Case of Lassa Fever’, *PLoS Neglected Tropical Diseases*. doi: 10.1371/journal.pntd.0003398.

Ibukun, F. I. (2020) ‘Inter-lineage variation of lassa virus glycoprotein epitopes: A challenge to lassa virus vaccine development’, *Viruses*. doi: 10.3390/v12040386.

Khan, S. H. *et al.* (2008) ‘New opportunities for field research on the pathogenesis and treatment of Lassa fever’, *Antiviral Research*. doi: 10.1016/j.antiviral.2007.11.003.

Lalis, A. *et al.* (2013) ‘The Impact of Human Conflict on the Genetics of Mastomys natalensis and Lassa Virus in West Africa’, *PLoS ONE*, 7(5). doi: 10.1371/journal.pone.0037068.

Lecompte, E. *et al.* (2005) ‘Molecular identification of four cryptic species of Mastomys (Rodentia, Murinae)’, *Biochemical Systematics and Ecology*. doi: 10.1016/j.bse.2004.12.015.

Lecompte, E. *et al.* (2006) ‘Mastomys natalensis and Lassa fever, West Africa’, *Emerging Infectious Diseases*. doi: 10.3201/eid1212.060812.

Leski, T. A. *et al.* (2015) ‘Sequence variability and geographic distribution of Lassa Virus, Sierra Leone’, *Emerging Infectious Diseases*. doi: 10.3201/eid2104.141469.

MacLachlan, N. J. and Dubovi, E. J. B. T.-F. V. V. (Fourth E. (eds) (2011) ‘Chapter 23 - Arenaviridae’, in. San Diego: Academic Press, pp. 385–392. doi: https://doi.org/10.1016/B978-0-12-375158-4.00023-7.

Manning, J. T., Forrester, N. and Paessler, S. (2015) ‘Lassa virus isolates from Mali and the Ivory Coast represent an emerging fifth lineage’, *Frontiers in Microbiology*. doi: 10.3389/fmicb.2015.01037.

Olayemi, A. *et al.* (2016) ‘New Hosts of The Lassa Virus’, *Scientific Reports*. doi: 10.1038/srep25280.

Olayemi, A. *et al.* (2020) ‘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*. doi: 10.3390/biology9020026.

Olayemi, A. and Fichet-Calvet, E. (2020) ‘Systematics, ecology, and host switching: Attributes affecting emergence of the Lassa virus in rodents across western Africa’, *Viruses*. doi: 10.3390/v12030312.

Oti, V. B. (2018) ‘A Reemerging Lassa Virus: Aspects of Its Structure, Replication, Pathogenicity and Diagnosis’, in Alfonso J. Rodriguez-Morales (ed.) *Current Topics in Tropical Emerging Diseases and Travel Medicine*. BoD – Books on Demand.

Rambaut, A. *et al.* (2018) ‘Posterior summarization in Bayesian phylogenetics using Tracer 1.7’, *Systematic Biology*. doi: 10.1093/sysbio/syy032.

Suchard, M. A. *et al.* (2018) ‘Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10’, *Virus Evolution*. doi: 10.1093/ve/vey016.

Townsend Peterson, A., Moses, L. M. and Bausch, D. G. (2014) ‘Mapping transmission risk of lassa fever in West Africa: The importance of quality control, sampling bias, and error weighting’, *PLoS ONE*. doi: 10.1371/journal.pone.0100711.

Whitmer, S. L. M. *et al.* (2018) ‘New lineage of lassa virus, Togo, 2016’, *Emerging Infectious Diseases*, 24(3), pp. 599–602. doi: 10.3201/eid2403.171905.

Wiley, M. R. *et al.* (2019) ‘Lassa virus circulating in Liberia: a retrospective genomic characterisation’, *The Lancet Infectious Diseases*. doi: 10.1016/S1473-3099(19)30486-4.

Yadouleton, A. *et al.* (2019) ‘Lassa virus in pygmy mice, Benin, 2016-2017’, *Emerging Infectious Diseases*. doi: 10.3201/eid2510.180523.

Yadouleton, A. *et al.* (2020) ‘Lassa fever in Benin: description of the 2014 and 2016 epidemics and genetic characterization of a new Lassa virus’, *Emerging Microbes & Infections*, pp. 1–23.