Machine Learning Reveals Key Glycoprotein Mutations and Rapidly Assigns Lassa Virus Lineages

Richard Olumide Daodu 1,2,***, Jens Uwe-Ulrich** 1**, Denise Kühnert** 1,*,

1. Center for Artificial Intelligence in Public Health Research, Robert Koch Institute, Wildau, Germany
2. Department of Mathematics and Computer Science, Freie Universität Berlin, Takustr. 9, 14195, Berlin, Germany

* daodur@rki.de, kuehnertd@rki.de

Abstract

Lassa fever, caused by the Lassa virus (LASV), has led to numerous fatalities in West Africa and cases exported intercontinentally since its discovery in 1969. Currently, there are no approved vaccines, with recent research focusing on immunotherapy. Lassa virus is grouped into different lineages that circulate in specific geographical areas, elicit varying immune responses, and display distinct pathophysiological effects. Therefore, investigating the genetic differences between these lineages is crucial. Here, we analyzed the LASV glycoprotein, the only surface protein, using statistics, machine learning, and phylogenetics to identify key differences between Nigerian lineages and those endemic to other West African countries. We found that amino acid positions near the stable signal peptide cleavage site and sites impacting immune recognition, such as those between positions 59 and 76, were highly variable among the lineages. Additionally, we discovered that Lineage II and Lineage III sequences are one codon shorter than Lineage IV sequences, due to a codon insertion in positions 178-180, corresponding to amino acid position 60. This may explain the structural and phenotypical differences between the lineages. To quickly identify which lineages cause emerging outbreaks or exported infections, we also developed a highly accurate lineage classification tool based on machine learning.

Introduction

Since its discovery in 1969, Lassa fever (LF) constitutes a major public health threat, with 500,000 cases and around 5,000 deaths annually in West Africa (Garry, 2023; Macher & Wolfe, 2006; Richmond & Baglole, 2003). Although the overall fatality rate in hospitalized population is relatively low (~15 %), it is alarmingly high among pregnant women (80%) and fetuses (up to 100%) (Agboeze et al., 2019; Salami et al., 2022; Simons, 2023). Currently, no approved vaccine nor any appropriate therapeutic measures exist, although efforts are ongoing (Garry, 2023). The infectious agent causing Lassa fever is the Lassa Virus (LASV). Its main reservoir host is *M. Natalensis*, with several additional hosts with the potential to spread the virus having been discovered recently (Happi et al., 2023; Olayemi et al., 2016). Experimental research involving LASV is confined to Biosafety Level 4 (BSL4) laboratories, and the disease is currently on the World Health Organization's (WHO) top priority list (WHO, n.d.).

Lassa fever, a viral hemorrhagic disease first identified in Northern Nigeria, has caused global concern due to its international spread and pandemic potential. By 1973, notable cases were reported, including the first case introduced to Britain (Woodruff et al., 1973) and a hospital epidemic in Liberia (Mertens et al., 1973; Monath et al., 1973). Between 1974 and 1977, a special isolation unit was created in London (Woodruff, 1975), and an imported case was reported in the United States involving a Peace Corps worker (Zweighaft et al., 1977). By the 1980s, cases were reported globally, including Israel (Shlaeffer et al., 1988), Japan (Yanase et al., 1989), and Canada (Mahdy et al., 1989). Fatal cases involving travelers from West Africa to Germany were reported in 2000 and 2016 (Günther et al., 2000; Wolff et al., 2016). A recent review has documented cases worldwide (Wolf et al., 2020), including in Sweden and the Netherlands, emphasizing the need for robust biosecurity measures, rapid diagnosis, effective treatments and vaccines.

To date, there are seven known LASV lineages circulating in distinct geographical locations (Figure 1A). Lineages I, II and III circulate in non-overlapping parts of Nigeria (Ehichioya et al., 2019), while lineage IV circulates in Sierra Leone, Guinea, and Liberia (Andersen et al., 2015), lineage V in Mali and Cote d'Ivoire (Manning et al., 2015), and lineage VII in Togo and Benin (Whitmer et al., 2018; Yadouleton et al., 2020). The strain found in the host *H. pamfi* in Nigeria was designated lineage VI (Olayemi et al., 2016; Whitmer et al., 2018).

Currently, treatment for Lassa fever relies primarily on ribavirin, a drug with severe side effects (De Franceschi et al., 2000; Salam et al., 2022). Immunotherapy is a promising treatment option. Early attempts at passive immunization included the use of convalescent plasma and immune serum, which showed some success (Clayton, 1977; Leifer et al., 1970). Most recently, a monoclonal antibody cocktail, which is composed of three neutralizing antibodies targeting the LASV glycoprotein complex (GPC), has shown to be efficient in non-human primates against representatives of lineages I-IV (Mire et al., 2017). It is now apparent that even this cocktail may fail in some variants or new lineages, as recently reported in lineage VII (Woolsey et al., 2024). This may be due to mutations in the GPC, which may prevent binding of specific antibodies (Enriquez et al., 2022; Robinson et al., 2016). Thus, it is crucial to determine which lineages carry such mutations and to what extent. While this was attempted in previous work (Ibukun,

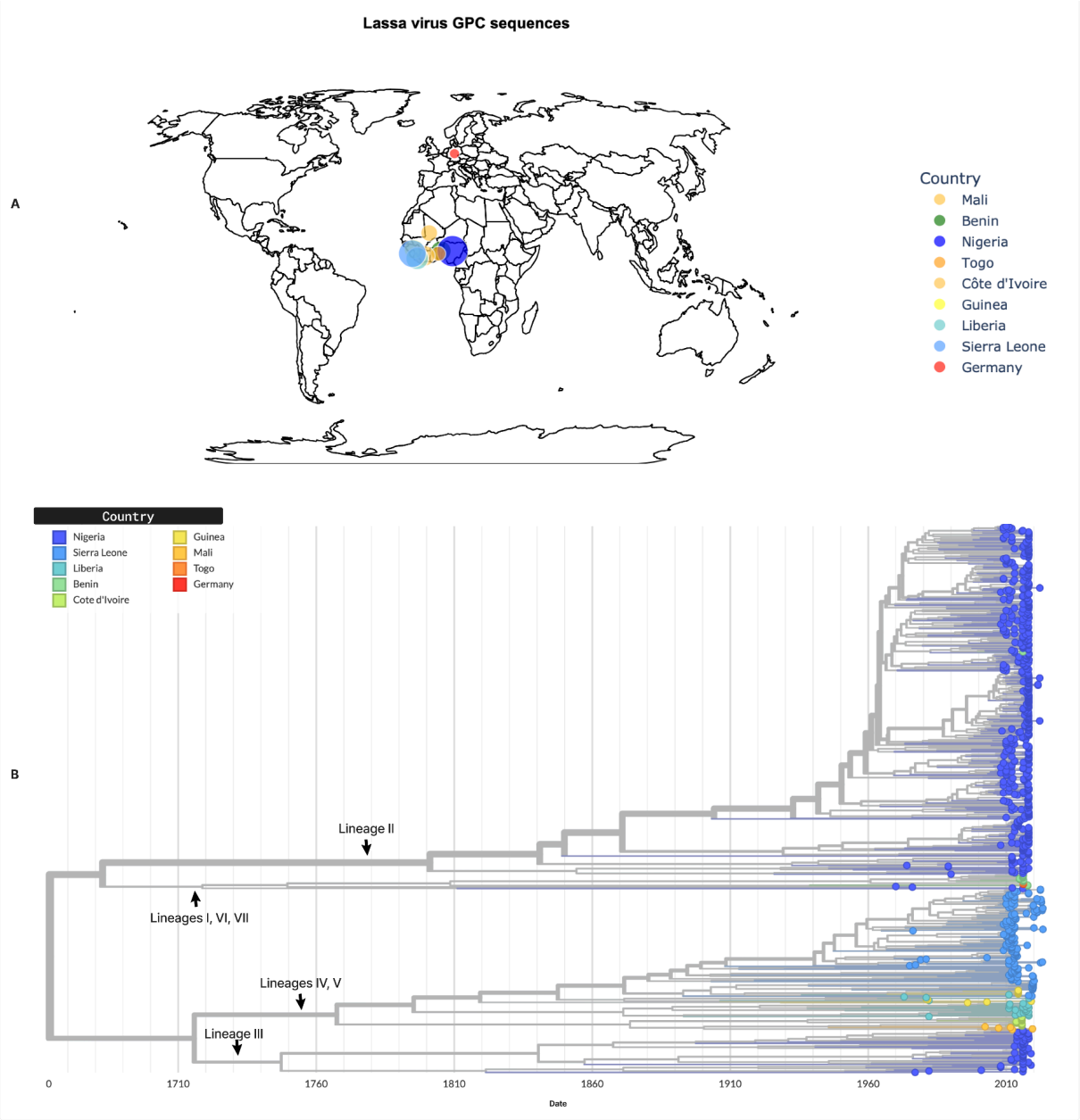
2020), it was based on a somewhat arbitrary selection of "representative" strains from various lineages for comparative analysis, instead of including all data available.

The Lassa virus genome is bi-segmented, coding for four proteins: the glycoprotein complex precursor, nucleoprotein (N), matrix (Z), and RNA-dependent RNA polymerase (L). The GPC is post-translationally cleaved into the Stable Signal Peptide (SSP), Glycoprotein subunit 1 (GP1) and Glycoprotein subunit 2 (GP2). Altogether, the GPC is crucial for cell entry, making it a key target for vaccine and therapeutic development (Garry, 2023; Katz et al., 2022). Despite its importance, specific information on LASV GPC, especially related to specific amino acid (AA) differences between lineages, is limited.

Knowing which lineage an emerging LASV outbreak strain belongs to would instantly shed light on the geographical origin of an outbreak, since distinct lineages circulate in distinct locations. Moreover, since the lineages are known to have different immunological (Buck et al., 2022) and pathophysiological effects (Andersen et al., 2015), quick lineage assignment from a patient's sample at the point of care would be beneficial. With the WHO supporting a significant scale-up of the genomic sequencing capabilities in Africa during the Covid-19 pandemic (Akande et al., 2023), genomic pathogen sequencing can become an integral part of the outbreak response.

In this study, we harness robust data, statistics, phylogenetics, and machine learning techniques to determine clade-defining amino acid mutations of possible clinical and evolutionary importance on the LASV GPC. We also provide a rapid lineage assignment tool to enable timely responses to national and international Lassa fever threats.

Figure 1: Description of the data in this study



(A) Geographical distribution of the final glycoprotein (GPC) sequences data in this study. The majority of the data is from Nigeria, followed by Sierra Leone. The only non-West African sequence is from a 2016 local spread of an imported case of Lassa fever to Germany from Togo (Wolff et al., 2016). The counts have been log-normalized for clarity.

(B) Phylogenetic tree of the nucleotide sequences annotated by lineage (black arrows). Leaf nodes are coloured by geographical location. The tree topology closely resembles the GPC tree described by (Whitmer et al., 2018)

Results

Random forest classification reveals key LASV glycoprotein mutations

To investigate LASV GPC amino acid sites of differences, we downloaded all available LASV nucleotide sequences released before December 1, 2023, and extracted the GPC region (supplementary figure 1). After quality filtering (see Methods), the final dataset contained 753 sequences, including 542 sequences from Nigeria, 141 from Sierra Leone, 11 from Guinea, 1 from Germany, 3 from Togo, 24 from Liberia, 5 from Mali, 13 from Côte d'Ivoire, and 13 from Benin. We translated the nucleotide sequences into amino acid sequences and verified the accuracy of the translation. We then grouped the sequences into two categories: those from Nigeria and those from other countries - consistent with Andersen et al. (Andersen et al., 2015).

We selected the Random Forest (RF) model due to its robustness in handling imbalanced and multi-feature datasets and based on previous work on similar tasks involving viral sequence classification (Kim et al., 2021). Before fitting the model, we encoded the alignment using a one-hot approach, representing each AA as a 21-dimensional vector (20 amino acids plus a gap). The model achieved precision, recall, and F1 scores of 100% across the two groups on the test set, which aligns with the literature indicating that LASV is generally location-bound (Andersen et al., 2015; Ehichioya et al., 2019). Subsequently, the aggregated feature importance was extracted from the model, revealing that many amino acid positions contributed minimally to the differences between Nigerian and other sequences (Figure 2, supplementary figure 2). This suggests reasonable conservation across the two groups. We ranked the positions and selected the top 15 AA positions for further analysis.

To verify the positions implicated by the RF method, we employed two additional techniques: Manhattan Distance (MD) and Pearson correlation (Pcorr). In this context, the MD is the sum of the absolute differences between AA positions across the two groups. We counted the occurrences of amino acids and gaps per position across the groups, normalized by sample number due to data imbalance, and computed the MD between corresponding position vectors. We ranked the positions and selected the top 15 for further analysis.

Finally, we applied Pearson correlation to assess variant similarity across the groups. Highly correlated positions suggest conservation, while low correlation indicates dissimilarity. We ranked the positions from least to most correlated and selected the top 15 for further analysis.

The application of these three methods revealed that 11 out of the top 15 positions were consistently implicated across all three methods (supplementary figure 3). The positions, sorted from highest to lowest by cumulative ranking of all three sets, were AA 273, 60, 61, 28, 44, 76, 74, 31, 421, 324, and 482. We hypothesize that these AA positions play a role in immunological (Buck et al., 2022) and pathophysiological processes (Andersen et al., 2015). Indeed, positions 74 and 76, for example, have been implicated as important for the functioning of antibodies 20.10C and 36.1F (Enriquez et al., 2022). This suggests that the virus might be attempting to evade these antibodies by varying these positions. In fact, deletion of amino acid positions 60-75, which

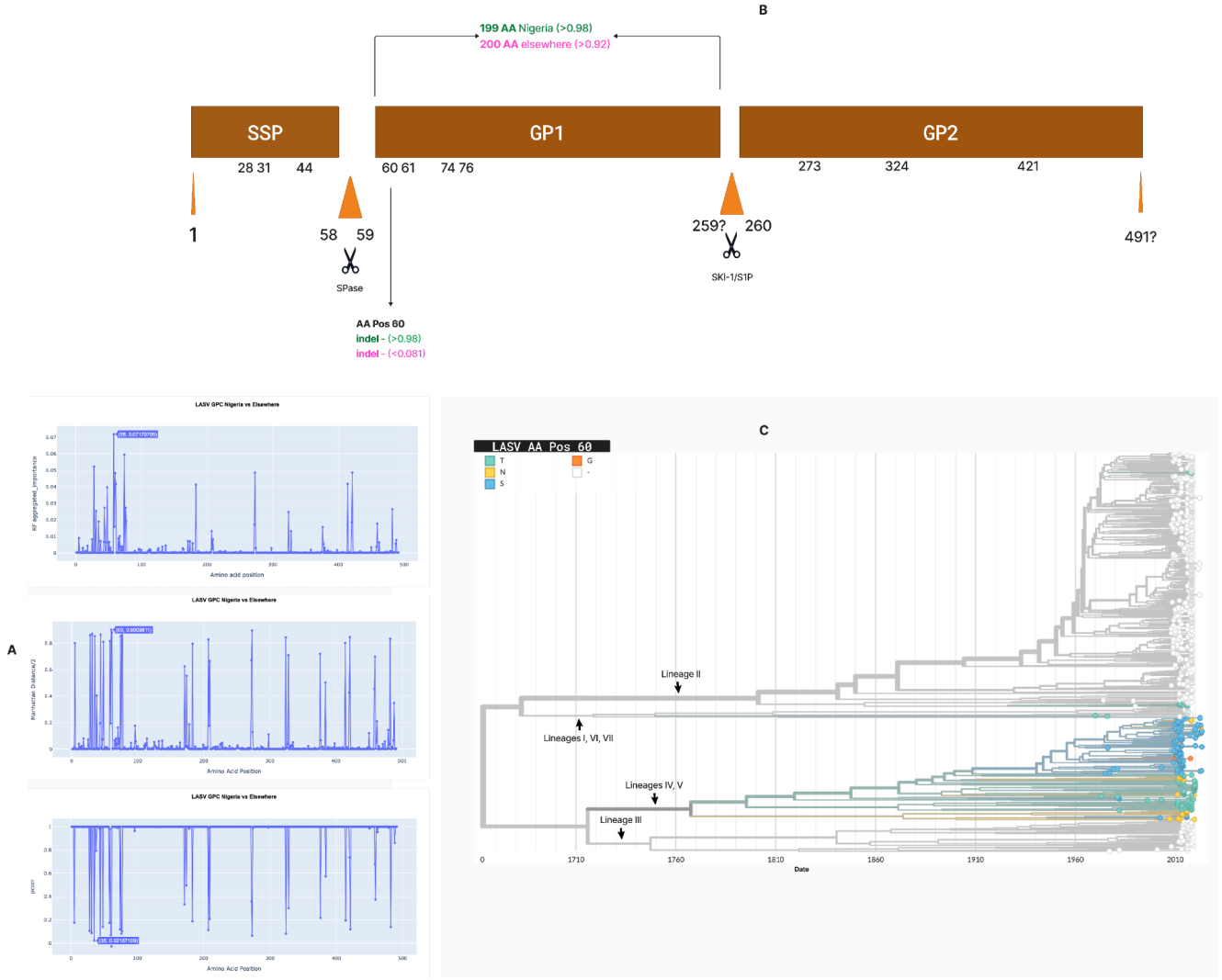
coincidentally falls within the range of 3 out of the 11 most variable positions across the LASV GPC as implicated by our tests, has been reported to disrupt the functioning of all known anti-GPC antibodies (Robinson et al., 2016).

An indel at amino acid position 60 shortens 98% of Nigerian GP1 sequences to 199AAs

The Manhattan Distance analysis highlighted position 60 as the most divergent between the groups (Figure 1C), and this position was consistently ranked highly by all three methods. Normalized value counts revealed that 98% of sequences from Nigeria have a gap at position 60, compared to less than 8.1% of sequences from other regions. This indel is particularly noteworthy as it is located in the GP1 region, in the second position after the SSP cleavage site (Figure 2B). Consequently, at least 98% of Nigerian GPC sequences are shortened by one codon (nucleotide positions 178-180) at the nucleotide level and by one amino acid at the protein level compared to most sequences from other regions. This results in the GP1 of Nigerian sequences being 199 amino acids long compared to most sequences from elsewhere, which belong mainly to lineage IV.

This discrepancy in length may affect protein folding and structure, potentially leading to functional and immunological variations. Additionally, the length inconsistency may cause issues with sequencing and assembly, as well as with polymerase chain reactions (PCR) targeted at the S segment or the GPC.

Figure 2: Implicated regions of variation on the LASV GPC



(A) Divergent regions implicated by the random forest (top), Manhattan distance (middle), and Pearson correlation (Pcorr, bottom) are shown, respectively. The results of the three methods largely overlap. The Pcorr plot appears inverted because a higher correlation indicates conservation across the group as opposed to the other two tests. The highest-ranking positions are as follows: position 58 by random forest, position 60 by Manhattan distance, and position 35 by Pcorr absolute value (reversed).

(B) A representation of the 11 highly ranked positions jointly implicated by all three tests on the LASV GPC amino acid sequence. Positions 60 and 61 are observed to be close to the SSP cleavage site on the GP1. Position 60 was ranked highest as the position with the greatest distance between Nigerian amino acid sequences and those from other countries. A normalized count of variants at that position revealed that over 98% of Nigerian sequences lack an amino acid at position 60 (supplementary figure 4), whereas more than 92% of sequences from other countries—denoted as 'elsewhere'—have an amino acid at this position. Given that the Nigerian sequences constitute the largest dataset, this raises questions about the general belief that the LASV GPC is 491 amino acids in total and that GP1 is 200 amino acids.

(C) Time calibrated tree showing that the missing position on Nigerian sequences is actually an insertion that predates the emergence of the Sierra Leonean strain. This insertion is seen to be present in all sequences in Lineage IV and Lineage V.

Phylogenetic analysis suggests that position 60 constitutes an insertion that predates the emergence of Lineage IV, the Sierra Leonean strain

To investigate the nature of the indel at position 60 and its potential public health implications, we reconstructed a time-scaled phylogenetic tree of the LASV GPC using the preprocessed data. The clock rate was estimated by Treetime (Sagulenko et al., 2018) at 8.20e-4 substitutions per site per year.

The results revealed that the indel at position 60 is an insertion in Lineage IV, dating back to when this lineage first diverged from the Nigerian lineages (Figure 2C). The insertion is estimated to have occurred in the 18th century. All sequences on the major branch leading to lineages IV and V possess this insertion. The amino acids observed at this position include the Josiah genotype S60, with variants such as S60T, S60N and S60G. Only two sequences from Sierra Leone have the S60G variant (OQ919514 and KM821773).

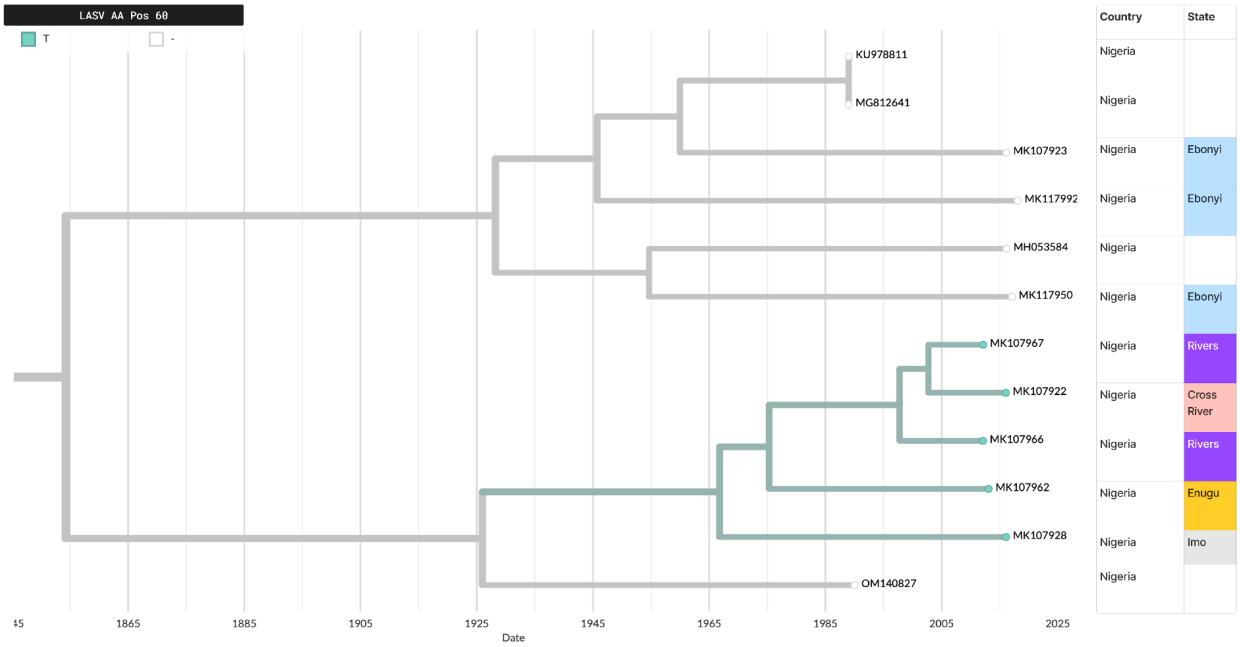
On the Nigerian side, lineages II and III almost entirely lack any amino acid at this position - one feature common to their GPC. This is consistent with the understanding that the Nigerian lineages are believed to be ancestral to the Sierra Leonean lineage (Andersen et al., 2015). Lineage VII, the Togo strain, which also includes sequences from the Benin Republic and a 2016 imported case to Germany, also lacks this insertion. However, a sequence closely related to Lineage VI (MK107927) is seen to have the insertion at position 60 (supplementary figure 5).

Lineage I, on the other hand, has the insertion (a T60), but instead appears to have an amino acid deletion at position 62. Interestingly, differences in immune response have been reported for lineage I (Buck et al., 2022).

The insertion appears to have recently reemerged in lineage II, particularly in sublineage IIb (Ehichioya et al., 2019) which is reported to circulate in the southern part of Nigeria (Figure 3). Meanwhile, the same insertion is also seen in GenBank ID MH887782, MH053506, which clusters closely with sublineage lineage 2g, a sublineage with growing incidence in Edo and Ondo States in Nigeria (Happi et al., 2023) (supplementary figure 6). These insertions were estimated to have happened around the last century.

The persistence of the insertion at this AA position 60 over the years and across multiple lineages suggest it may offer a fitness advantage for the virus. In the structural and functional aspects of the glycoprotein, this could have significant implications for virus transmission and virulence, potentially affecting public health measures and strategies for vaccine development. The independent recurrence of insertions at this specific position may suggest that it offers the virus a selective advantage. It could also indicate potential adaptation to an intermediary host.

Figure 3: Re-emergence of insertion on position 60 in lineage II



The most recent common ancestor (MRCA) that possessed this insertion emerged in the last century in the southern part of Nigeria. Viruses circulating in this region have been named sub-lineage IIb by Ehichioya et al. The implicated states in Nigeria share close geographical boundaries.

Rapid LASV Lineage Classification from GPC sequence

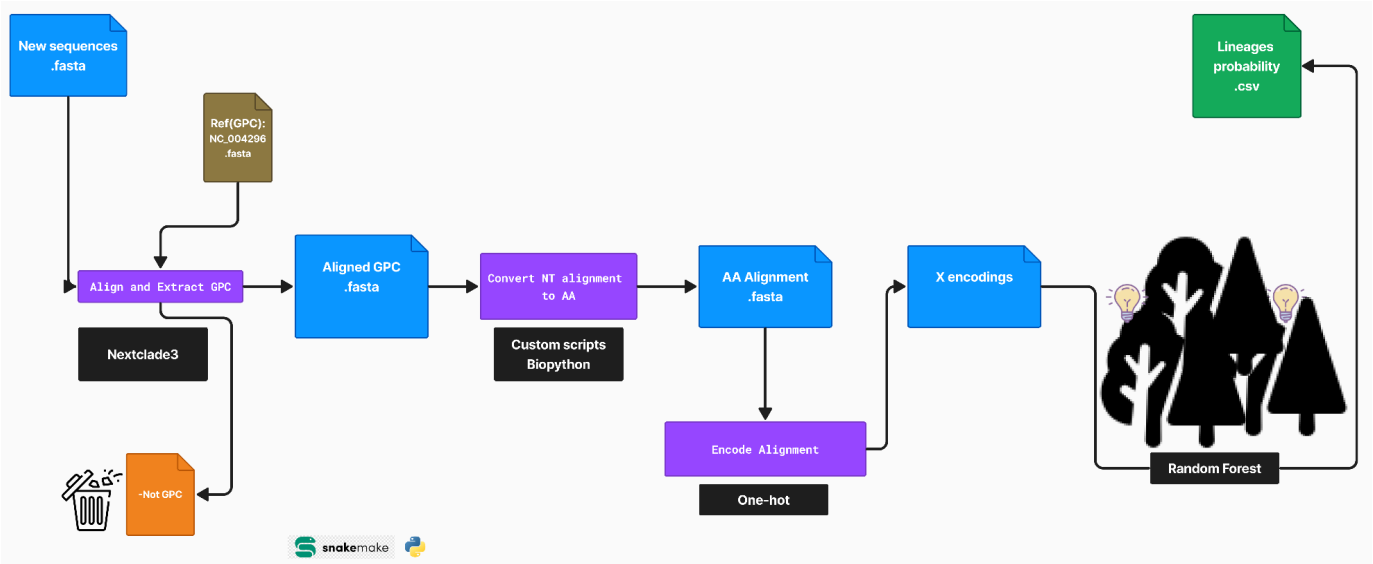
Given that different LASV lineages circulate in specific regions and possess distinct clinical properties, the ability to efficiently assign lineages to unknown sequences offers immense clinical and public health benefits. Consequently, we developed a machine learning-based pipeline for LASV lineage assignment.

We trained a Random Forest (RF) model to classify sequences into four groups: lineages II, III, IV & V, and VII, as permitted by existing data. On the initial test set, which was preprocessed similarly to the training set, the model achieved 100% across all test metrics (Precision, Recall, F1-score).

The model was subsequently exported into a pipeline, which was validated using recently published data by Adesina et al. (2023) and Bangura et al. (2024) (Adesina et al., 2023; Bangura et al., 2024). These datasets were not included in the training or test datasets, and the sequences from Adesina et al. (2023) did not cover the length of the GPC. The pipeline accurately assigned the sequences from Bangura et al. (2024) and Adesina et al. (2023) with precisions of 100% (Figure 5).

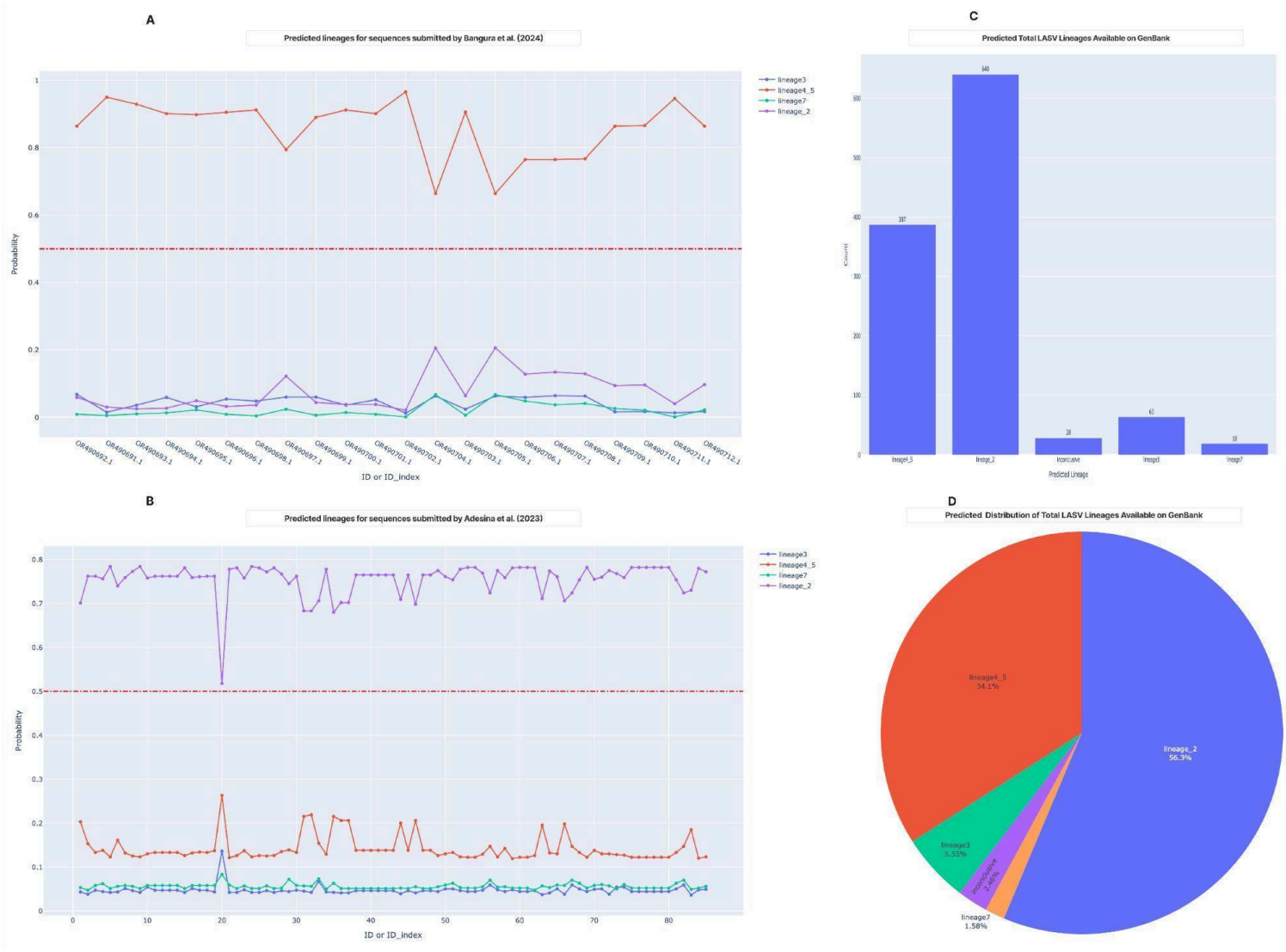
To test the pipeline's speed, we downloaded all publicly available LASV sequences and processed them through the pipeline. On an average computer using five cores, it took less than two minutes to process and predict the lineages of all LASV GPC sequences in the dataset (supplementary GitHub code).

Figure 4: Pipeline for LASV lineage prediction



The pipeline relies on the Snakemake engine for workflow management and Nextclade3, a part of Nextstrain (see Methods), as the GPC extractor and aligner. The aligned nucleotide sequences are converted into amino acid sequences and encoded similarly to the model's training process. Subsequently, the encoded sequences are run through the trained random forest model, which predicts the probability per lineage. The highest probability is taken as the final prediction, provided it is greater than 0.5. Sequences with no probability above 0.5 are classified as inconclusive.

figure 5: The pipeline performs excellently on verifiable datasets



(A) The pipeline accurately and precisely predicts sequences provided by Bangura et al. as belonging to Lineage 4 or 5.

(B) The pipeline accurately and precisely predicts sequences provided by Adesina et al. as belonging to Lineage 2, despite the sequences being around half the size of the GPC.

(C, D) The predicted distribution of LASV lineages in GenBank as of 06/07/2024. The inconclusive predictions include KM822128 (Pinneo), which belongs to lineage I. This reinforces the belief that lineage I is divergent from other lineages and emphasizes the pipeline's ability to avoid false positives (supplementary GitHub code).

These figures are auto-generated for every run, enabling rapid evaluation of the resulting lineage classification.

Discussion

In this paper, we provide novel information on the amino acid positions that are most variable by geographical location and lineage. We have used a novel application of mathematical methods, which could be used further in the field for motif finding, even beyond pathogens, especially in cancer. Our results also suggest that a recurring codon insertion may confer a natural selection advantage to LASV. Finally, we provide a rapid, accurate, and easy-to-use lineage assignment tool for LASV sequences.

Andersen et al. (Andersen et al., 2015) reported the phenomenon of differing pathophysiology and host codon usage in the LASV circulating in different regions. We have extended this work by showing differing positions of the LASV GPC at the amino acid level. Interestingly, some of the positions are around the cleavage site of the stable signal peptide by the signal peptidase, a protein reported to control the maturation of the GPC (Bederka et al., 2014). The variation in this region may lead to differences in viral replication and maturation.

Although a gap at AA position 60 of the LASV GPC was previously reported (Buck et al., 2022; Perrett et al., 2023), its prevalence was never investigated or discussed. Here, we have shown that this gap is present in over 98 percent of sequences from Nigeria, including all lineage III sequences and almost all lineage II sequences in our dataset. Consequently, Nigerian sequences are shorter than Sierra Leonean sequences (Lineage IV). To the best of our knowledge, this is the first time this has been categorically shown in the literature. Although the majority of Lassa fever cases occur in Nigeria, it is common to use the lineage IV Josiah strain as a reference for genomic analyses. The GPC length of 491 amino acids is widely cited in the literature, and it is important to clarify that the Nigerian lineages, which are the most prevalent, are one codon (nt 178 - 180) shorter. This may have profound implications for understanding the structural and phenotypical differences between the lineages.

Generally, the effects of mutations vary based on several factors, including the position of the mutation, the importance of the protein to the organism, the type of mutation, and more. In multicellular eukaryotes, many traits (or phenotypes) are controlled by multiple genes, known as polygenic traits. A mutation in one of these genes may not have a noticeable effect. However, this is not the case for monogenic traits, where mutations can have more severe consequences. For example, in the case of sickle cell anemia, a single nucleotide substitution leads to a single amino acid substitution, which can cause catastrophic consequences if both gene copies inherited by an individual are affected (Inusa et al., 2019).

In viruses, which have fewer proteins, mutations can have even more profound effects. This impact may be exacerbated when the affected protein is the only surface protein and the mutation is an indel. Hence, the insertion at position 60 may be a major contributor to the differences in pathophysiology and immune effects observed among LASV lineages. This mutation occurs on GP1, the subunit responsible for binding both LAMP1 and alpha-dystroglycan receptors of the host during viral entry (Cao et al., 1998; Jae et al., 2014). This could cause differences in cell

entry between Nigerian lineages and those from other regions, particularly the Sierra Leone strain. Interestingly, Andersen et al. (Andersen et al., 2015) reported that the Sierra Leonean strain's genome abundance in patients is significantly higher than that of Nigerian lineages. Since viruses rely on host cellular mechanisms for their life cycle, the increased cell entry - potentially influenced by this insertion in GP1 - might be one of the underlying reasons for the difference in genome abundance.

Given the codon insertion at position 178-180 of the GPC, it is advisable to avoid targeting this region with PCR primers, as this could result in inconsistent results across different LASV lineages due to variability in binding affinity depending on the reference sequence used for primer design. Similarly, the design of ELISA (Enzyme Linked Immunosorbent Assay) test kits should take into account the variability at this amino acid position (LASV GPC 60) to avoid inaccuracies in detecting the virus. The presence of this insertion can affect the epitope structure, potentially leading to reduced recognition by antibodies used in ELISA, which are designed based on a reference sequence that may not account for such variations. If targeting these positions is unavoidable, it is essential to design test kits that are specific to the respective lineages to ensure accurate and reliable results.

We also showed that the indel at AA position 60 is an insertion that predates the emergence of the Sierra Leonean lineages from the Nigerian lineages. Additionally, there has been an independent re-emergence of this insertion at this position in lineage II in recent times. The sustenance of this mutation is strongly suggestive of natural selection fitness. This finding will foster a greater understanding of the inter-lineage differences and the evolution patterns of LASV, which is important for surveillance, vaccine, and immunotherapy development. Moreover, an unknown intermediary host could also be responsible for this insertion - a prospect for future research. We hope that future work will clarify the implications of this insertion.

Our finding that positions 60, 61 and 74 are among the top positions varying between lineages, is supported by their role in immune evasion (Robinson et al., 2016). Supported by reports from Enriquez et al. (Enriquez et al., 2022) and Carr et al. (Carr et al., 2024), monoclonal antibody 21.10C may be ineffective in Southern Nigeria due to the high prevalence of the immune-resistant mutation at position 76 (supplementary figure 7). It is essential to monitor these positions closely, along with others such as position 273, which have also been identified as immunologically significant (Ibukun, 2020; Robinson et al., 2016) and have been consistently ranked as highly variable in our tests.

Our machine learning-based pipeline offers a novel ability for researchers and clinicians to quickly assign lineages to LASV sequences. Given its built-in cleaning processes, speed and precision, and ease of use, we believe this pipeline will be invaluable for managing endemic cases of Lassa fever in West Africa. Also, this will assist in management of imported cases of Lassa fever, and for global public health efforts. However, as more data emerges, the model will need to be updated using the workflow we have provided. This tool is a significant contribution to bioinformatics capacity on the African continent, a capacity that several African researchers have been calling for in recent times (Nembaware et al., 2023; Olono et al., 2024; Sharaf et al., 2023).

We highlight the need for more sequencing to properly study LASV evolution and to develop additional tools. The establishment of standard guidelines for new LASV lineage naming, as suggested by (Whitmer et al., 2018), would help make tools like this pipeline as reproducible and reusable as possible.

Methods

General preprocessing

LASV sequences released between December 1905 and December 1, 2023, along with their accompanying metadata, were downloaded from NCBI Virus (available at https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/). To ensure the inclusion of only field samples, the 'exclude lab strain' filter was applied. Using the GPC gene from the reference ID NC_004296, we extracted and aligned the GPC regions from all sequences using LAST (Kiełbasa et al., 2011) and MAFFT (Katoh et al., 2002), available at https://mafft.cbrc.jp/alignment/server/specificregion-last.html, resulting in 1021 GPC sequences. Sequences with more than 5% gaps and ambiguous nucleotides of the total alignment length were removed, reducing the dataset to 808 sequences. The final stop codon position was removed from the alignment because the stop codon signals the termination of translation, and it does not encode an amino acid. Sequences lacking sampling dates and locations were excluded, leaving 753 sequences.

Alignment visualization and exploration were conducted using Aliview (Larsson, 2014). Manual curation was performed to ensure codon consistency (Happi et al., 2023). Specifically, misalignment involving codons disrupted by 3 gaps was adjusted by moving a single nucleotide to match the two others to ensure proper translation. Nucleotide to amino acid translation was done using Aliview.

All data analyses were conducted in Python using standard libraries such as pandas (McKinney, 2010) (available at https://pandas.pydata.org/), scikit-learn (Pedregosa et al., 2011), and Biopython (Cock et al., 2009). The workflow was implemented using Jupyter Notebook in Google Colab (Bisong, 2019).

Phylogenetic analysis of GPC

Using a pipeline based on Snakemake (Köster & Rahmann, 2012), Nextstrain (Hadfield et al., 2018), and Augur (Huddleston et al., 2021), we reconstructed both a Maximum Likelihood tree and a time tree (supplementary Figure 9). The pipeline begins by accepting an alignment and metadata, which includes sampling dates, country, and host information. A Maximum Likelihood tree was built using IQ-TREE (Nguyen et al., 2015) with the application programming interface (API) provided by Augur. Similarly using Augur API, the tree was processed by TreeTime (Sagulenko et al., 2018), along with the metadata, to generate a time-calibrated tree.

The trees and accompanying metadata were parsed using the Augur export command into a JSON file, which was subsequently visualized using Auspice (available at https://auspice.us/), a part of the Nextstrain (Hadfield et al., 2018) toolkit. All phylogenetic images were generated using Auspice - which have been edited for clarity in Figma (available at https://www.figma.com/).

Sequence annotation

Sampling locations were extracted from the accompanying metadata for each sequence from GenBank. Sequences were annotated based on the country column to analyze motif differences between Nigerian sequences and those from other countries. The sequences were then grouped into those from Nigeria (542 sequences) and those from other countries (211 sequences).

For lineage annotation, clades in our reconstructed phylogenetic tree were annotated based on information from the literature (Andersen et al., 2015; Ehichioya et al., 2019; Manning et al., 2015; Olayemi et al., 2016; Whitmer et al., 2018; Yadouleton et al., 2020) (see supplementary GitHub data). The leaves of each clade were then collected. Specifically, the sequences were categorized into the following lineages: 480 sequences for lineage II, 59 sequences for lineage III, 194 sequences for lineages IV and V combined, and 16 sequences for lineage VII. Lineages I and VI were excluded due to insufficient data. Lineages IV and V were grouped together based on the recommendation of Whitmer et al. (Whitmer et al., 2018), who pointed out that the distance between the two lineages is similar to those between other sublineages.

Machine learning: Random forest

Sequence encoding

The amino acid alignment, comprising 753 sequences and 491 positions, was loaded using the Biopython alignment class. Gaps preceding and following any real amino acids were converted to 'unknowns', as these gaps typically indicate either short sequences or sequencing issues and therefore lack biological significance.

To prevent any biases, all features and targets were one-hot encoded. Each base of every amino acid sequence in the alignment was encoded into a 21-position vector. Gaps between bases were similarly encoded, while 'unknowns' were encoded with zero values. The encoded features were then flattened, with each position in the vector transformed into an individual column, resulting in a data matrix of dimensions 753 x 10,311.

The targets were encoded using one-hot variables, facilitated by pandas' `get_dummies()` method. We trained two models with different targets: country and lineage.

Model training

The dataset was split into training and testing sets using scikit-learn with a random state of 42, ensuring that the test set comprised 20% of the total dataset and stratification was based on the target variable. Using the training set, we trained Random Forest (RF) models with the scikit-learn package. The RF model was chosen for its robustness in handling multidimensional and imbalanced data. The random state for the RF model was set to 80, and the seed state was set

to 42 to ensure reproducibility. A total of 100 decision trees were used in the binary country classification, while a total of 1,000 decision trees were used in the lineage classification.

Model Evaluation

The models' performance was evaluated using the test dataset, which comprised 20% of the entire dataset. Standard metrics from the scikit-learn package, such as precision, F-score, and recall, were used to assess the models' performance through the `classification_report()` function.

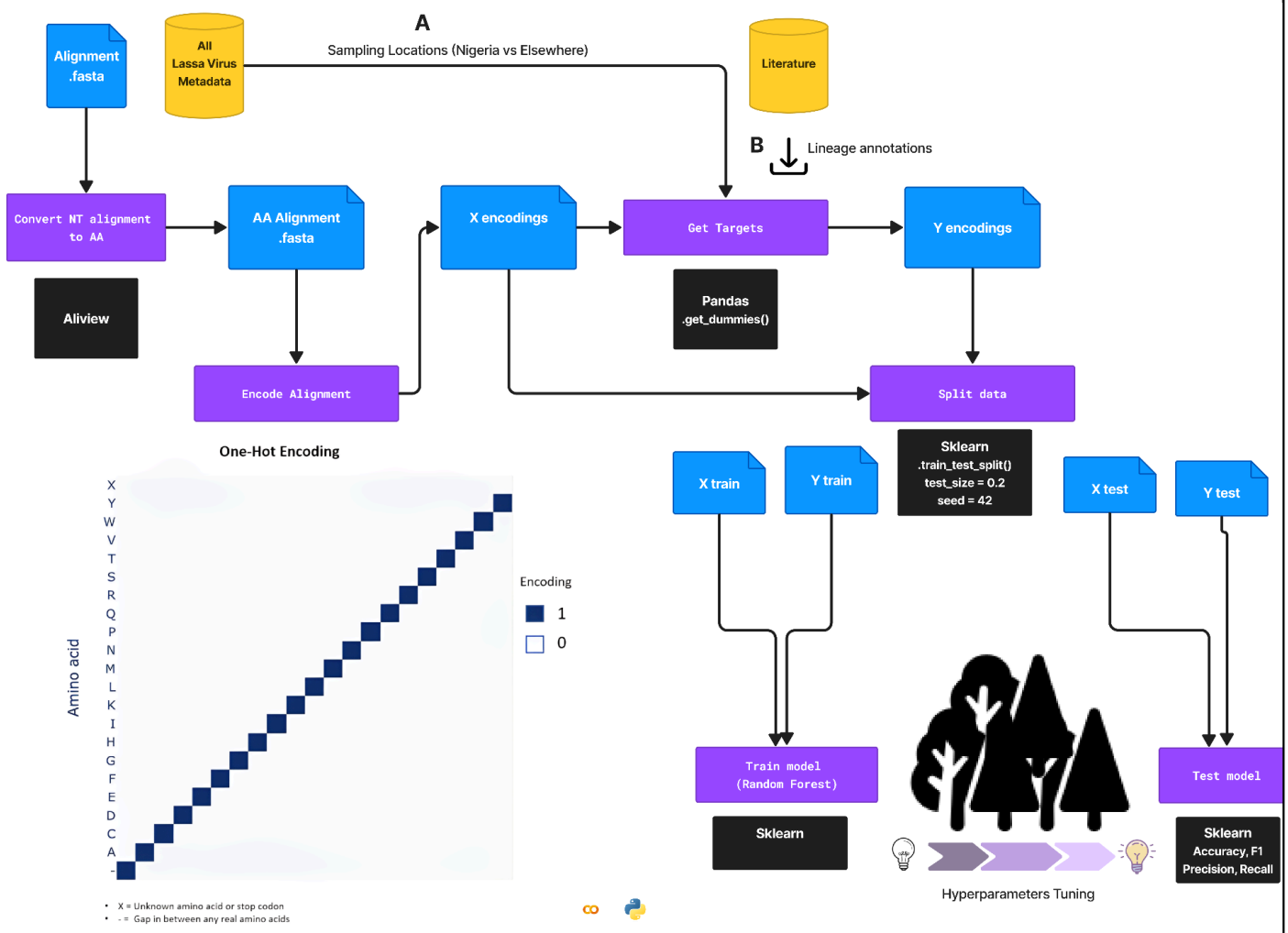
Since amino acid sequences can be identical between samples due to extreme conservation, we also investigated possible overlapping data points between the training and test sets. Approximately twenty percent of the test data was found to be present in the training set. However, this overlap does not affect the results, as the model achieved 100% across all test metrics.

Feature importance selection

During the encoding process, each position was represented by a 21-vector, and the resulting matrix was flattened, creating 491 x 21 columns (totaling 10,311 columns). Therefore, each position in the amino acid sequence is represented by 21 features.

The Random Forest algorithm outputs a large matrix containing the column index and the feature importance (10,311 rows by 2 columns). To determine the exact position of an amino acid in the alignment, each column index is divided by 21, and the integer part of the quotient is taken and incremented by 1. The weights of each position were combined into an aggregated feature importance score. Each feature corresponds to an exact position in the GPC amino acid sequence. The positions were subsequently ranked using the aggregated feature importance score from highest to lowest.

Figure 6: Machine Learning training workflow



An overview of the machine learning workflow demonstrates how the two Random Forest models were created using a similar process. (A) The sampling locations were annotated using metadata. (B) The lineages of each sequence in our dataset were annotated using insights from the literature. To achieve optimal performance from the models, hyperparameter tuning, primarily involving varying the number of decision trees, was conducted.

Manhattan Distance and Pearson Correlation

Manhattan Distance

Given an amino acid alignment with sequences belonging to either of two categorical variables, it is possible to investigate positions of differences this way:

Group the AA alignment into two categories based on metadata: Nigeria (542 sequences) and Elsewhere (211 sequences).

Calculate amino acid counts and gaps per position for both groups: Resulting in 21 by 491 matrices for each group (which is the 20 amino acids + gap (21) by length of the alignment (491))

Preprocessing:

- Convert gaps before and after any real amino acids to "Unknowns".
- Add the "Unknowns" count per position to the most frequent variant in that position.

Normalize each matrix: Divide by the number of samples.

Compute Manhattan distance: Calculate the absolute distance between the groups based on the normalized matrices as follows:

$$\text{Distance}_j = \frac{1}{2} \sum_{i=1}^{21} |A_{ij} - B_{ij}|$$

where A is the normalized amino acid count matrix of Nigerian sequences, B is the normalized amino acid count of sequences from other countries (Elsewhere), and 2 is the maximum absolute distance per position. The resulting vector contains the sum of the absolute differences for each position which was then ranked from highest to lowest (supplementary figure 8).

Pearson correlation

The normalized amino acid counts of the Nigeria and Elsewhere groups were correlated using the normalized matrices from the previous step. Using the `pandas.corr()` method, we calculated the Pearson correlation between each corresponding column of the Nigeria and Elsewhere matrices. This calculation produced a correlation vector representing each amino acid position.

The resulting vector, containing the correlation score for each position, was then ranked from least to highest by the absolute correlation score. This ranking highlights the positions with the most significant differences in variability between the two groups.

Data and code availability

Supplementary GitHub Data: https://github.com/JoiRichi/LASV_ML_manuscript_data

Supplementary GitHub code (Machine learning):
https://github.com/JoiRichi/LASV_lineage_pred

Supplementary Figures:
https://docs.google.com/document/d/1jxM6Mvun7LEop9DHgRsuRElbt7HdzDTCVnd_xtKwu-s/edit

Acknowledgement

We especially and wholeheartedly thank the Nextstrain main developers (https://nextstrain.org/team) for providing open-source code for phylogenetic analysis and for supervising the modification for Lassa virus. We also thank John Huddleston, Joseph Prescott, Zewen Yang, Sodiq Ayobami Hameed, Manfred Weiss, Laila Benz, and Knut Reinert for their helpful comments.

References

Adesina, A. S., Oyeyiola, A., Obadare, A., Igbokwe, J., Abejegah, C., Akhilomen, P., Bangura, U., Asogun, D., Tobin, E., Ayodeji, O., Osoniyi, O., Davis, C., Thomson, E. C., Pahlmann, M., Günther, S., Fichet-Calvet, E., & Olayemi, A. (2023). Circulation of Lassa virus across the endemic Edo-Ondo axis, Nigeria, with cross-species transmission between multimammate mice. *Emerging Microbes & Infections*, *12*(1), 2219350. https://doi.org/10.1080/22221751.2023.2219350

Agboeze, J., Nwali, M. I., Nwakpakpa, E., Ogah, O. E., Onoh, R., Eze, J., Ukaegbe, C., Ajayi, N., Nnadozie, U. U., Orji, M.-L., Ojide, K. C., Unigwe, S. U., Chika-Igwenyi, N., Nwidi, D. U., Clement, C. U., Kalombo, C., Makwe, C., & Tshiang, J. (2019). Lassa fever in pregnancy with a positive maternal and fetal outcome: A case report. *International Journal of Infectious Diseases*, *89*, 84–86. https://doi.org/10.1016/j.ijid.2019.08.023

Akande, O. W., Carter, L. L., Abubakar, A., Achilla, R., Barakat, A., Gumede, N., Guseinova, A., Inbanathan, F. Y., Kato, M., Koua, E., Leite, J., Marklewitz, M., Mendez-Rico, J., Monamele, C., Musul, B., Nahapetyan, K., Naidoo, D., Ochola, R., Ozel, M., … Samaan, G. (2023). Strengthening pathogen genomic surveillance for health emergencies: Insights from the World Health Organization's regional initiatives. *Frontiers in Public Health*, *11*, 1146730. https://doi.org/10.3389/fpubh.2023.1146730

Andersen, K. G., Shapiro, B. J., Matranga, C. B., Sealfon, R., Lin, A. E., Moses, L. M., Folarin, O. A., Goba, A., Odia, I., Ehiane, P. E., Momoh, M., England, E. M., Winnicki, S., Branco, L. M., Gire, S. K., Phelan, E., Tariyal, R., Tewhey, R., Omoniwa, O., … Sabeti, P. C. (2015). Clinical sequencing uncovers origins and evolution of Lassa virus. *Cell*, *162*(4), 738–750. https://doi.org/10.1016/j.cell.2015.07.020

Bangura, U., Davis, C., Lamin, J., Bangura, J., Soropogui, B., Davison, A. J., Nichols, J., Vucak, M., Dawson,

M., Ansumana, R., Sondufu, D., Cadar, D., Rieger, T., Thomson, E., Sahr, F., Magassouba, N., Ghersi, B., Bird, B. H., & Fichet-Calvet, E. (2024). Spatio-temporal spread of Lassa virus and a new rodent host in the Mano River Union area, West Africa. *Emerging Microbes & Infections*, *13*(1), 2290834. https://doi.org/10.1080/22221751.2023.2290834

Bederka, L. H., Bonhomme, C. J., Ling, E. L., & Buchmeier, M. J. (2014). Arenavirus Stable Signal Peptide Is the Keystone Subunit for Glycoprotein Complex Organization. *mBio*, *5*(6), e02063-14. https://doi.org/10.1128/mBio.02063-14

Bisong, E. (2019). Google Colaboratory. In E. Bisong (Ed.), *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners* (pp. 59–64). Apress. https://doi.org/10.1007/978-1-4842-4470-8_7

Buck, T. K., Enriquez, A. S., Schendel, S. L., Zandonatti, M. A., Harkins, S. S., Li, H., Moon-Walker, A., Robinson, J. E., Branco, L. M., Garry, R. F., Saphire, E. O., & Hastie, K. M. (2022). Neutralizing Antibodies against Lassa Virus Lineage I. *mBio*, *13*(4), e0127822. https://doi.org/10.1128/mbio.01278-22

Cao, W., Henry, M. D., Borrow, P., Yamada, H., Elder, J. H., Ravkov, E. V., Nichol, S. T., Compans, R. W., Campbell, K. P., & Oldstone, M. B. A. (1998). Identification of α-Dystroglycan as a Receptor for Lymphocytic Choriomeningitis Virus and Lassa Fever Virus. *Science*, *282*(5396), 2079–2081. https://doi.org/10.1126/science.282.5396.2079

Carr, C. R., Crawford, K. H. D., Murphy, M., Galloway, J. G., Haddox, H. K., Matsen, F. A., Andersen, K. G., King, N. P., & Bloom, J. D. (2024). Deep mutational scanning reveals functional constraints and antibody-escape potential of Lassa virus glycoprotein complex. *Immunity*. https://doi.org/10.1016/j.immuni.2024.06.013

Clayton, A. J. (1977). Lassa immune serum. *Bulletin of the World Health Organization*, *55*(4), 435–439.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: Freely available Python tools for

computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

De Franceschi, L., Fattovich, G., Turrini, F., Ayi, K., Brugnara, C., Manzato, F., Noventa, F., Stanzial, A. M., Solero, P., & Corrocher, R. (2000). Hemolytic anemia induced by ribavirin therapy in patients with chronic hepatitis C virus infection: Role of membrane oxidative damage. *Hepatology (Baltimore, Md.)*, *31*(4), 997–1004. https://doi.org/10.1053/he.2000.5789

Ehichioya, D. U., Dellicour, S., Pahlmann, M., Rieger, T., Oestereich, L., Becker-Ziaja, B., Cadar, D., Ighodalo, Y., Olokor, T., Omomoh, E., Oyakhilome, J., Omiunu, R., Agbukor, J., Ebo, B., Aiyepada, J., Ebhodaghe, P., Osiemi, B., Ehikhametalor, S., Akhilomen, P., … Günther, S. (2019). Phylogeography of Lassa Virus in Nigeria. *Journal of Virology*, *93*(21), e00929-19. https://doi.org/10.1128/JVI.00929-19

Enriquez, A. S., Buck, T. K., Li, H., Norris, M. J., Moon-Walker, A., Zandonatti, M. A., Harkins, S. S., Robinson, J. E., Branco, L. M., Garry, R. F., Saphire, E. O., & Hastie, K. M. (2022). Delineating the mechanism of anti-Lassa virus GPC-A neutralizing antibodies. *Cell Reports*, *39*(8), 110841. https://doi.org/10.1016/j.celrep.2022.110841

Garry, R. F. (2023). Lassa fever—The road ahead. *Nature Reviews Microbiology*, *21*(2), 87–96. https://doi.org/10.1038/s41579-022-00789-8

Günther, S., Emmerich, P., Laue, T., Kühle, O., Asper, M., Jung, A., Grewing, T., ter Meulen, J., & Schmitz, H. (2000). Imported lassa fever in Germany: Molecular characterization of a new lassa virus strain. *Emerging Infectious Diseases*, *6*(5), 466–476.

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., & Neher, R. A. (2018). Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics*, *34*(23), 4121–4123. https://doi.org/10.1093/bioinformatics/bty407

Happi, A. N., Ogunsanya, O. A., Ayinla, A. O., Sijuwola, A. E., Saibu, F. M., Akano, K., Nwofoke, C., Elias, O. T., Achonduh-Atijegbe, O., Daodu, R. O., Adedokun, O. A., Adeyemo, A., Ogundana, K. E., Lawal,

O. Z., Parker, E., Nosamiefan, I., Okolie, J., Parker, Z. F., McCauley, M. D., … Happi, C. T. (2023). Lassa Virus in Novel Hosts: Insights into the Epidemiology of Lassa Virus Infections in Southern Nigeria. *Emerging Microbes & Infections*, *0*(ja), 2294859. https://doi.org/10.1080/22221751.2023.2294859

Huddleston, J., Hadfield, J., Sibley, T. R., Lee, J., Fay, K., Ilcisin, M., Harkins, E., Bedford, T., Neher, R. A., & Hodcroft, E. B. (2021). Augur: A bioinformatics toolkit for phylogenetic analyses of human pathogens. *Journal of Open Source Software*, *6*(57), 2906. https://doi.org/10.21105/joss.02906

Ibukun, F. I. (2020). Inter-Lineage Variation of Lassa Virus Glycoprotein Epitopes: A Challenge to Lassa Virus Vaccine Development. *Viruses*, *12*(4), Article 4. https://doi.org/10.3390/v12040386

Inusa, B. P. D., Hsu, L. L., Kohli, N., Patel, A., Ominu-Evbota, K., Anie, K. A., & Atoyebi, W. (2019). Sickle Cell Disease—Genetics, Pathophysiology, Clinical Presentation and Treatment. *International Journal of Neonatal Screening*, *5*(2), 20. https://doi.org/10.3390/ijns5020020

Jae, L. T., Raaben, M., Herbert, A. S., Kuehne, A. I., Wirchnianski, A. S., Soh, T., Stubbs, S. H., Janssen, H., Damme, M., Saftig, P., Whelan, S. P., Dye, J. M., & Brummelkamp, T. R. (2014). Lassa virus entry requires a trigger-induced receptor switch. *Science (New York, N.Y.)*, *344*(6191), 1506–1510. https://doi.org/10.1126/science.1252480

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066.

Katz, M., Weinstein, J., Eilon-Ashkenazy, M., Gehring, K., Cohen-Dvashi, H., Elad, N., Fleishman, S. J., & Diskin, R. (2022). Structure and receptor recognition by the Lassa virus spike complex. *Nature*, *603*(7899), 174–179. https://doi.org/10.1038/s41586-022-04429-2

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., & Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Research*, *21*(3), 487–493. https://doi.org/10.1101/gr.113985.110

Kim, J., Lee, K., Rupasinghe, R., Rezaei, S., Martínez-López, B., & Liu, X. (2021). Applications of Machine

Learning for the Classification of Porcine Reproductive and Respiratory Syndrome Virus Sublineages Using Amino Acid Scores of ORF5 Gene. *Frontiers in Veterinary Science*, *8*. https://doi.org/10.3389/fvets.2021.683134

Köster, J., & Rahmann, S. (2012). Snakemake—A scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Larsson, A. (2014). AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, *30*(22), 3276–3278. https://doi.org/10.1093/bioinformatics/btu531

Leifer, E., Gocke, D. J., & Bourne, H. (1970). Lassa fever, a new virus disease of man from West Africa. II. Report of a laboratory-acquired infection treated with plasma from a person recently recovered from the disease. *The American Journal of Tropical Medicine and Hygiene*, *19*(4), 677–679. https://doi.org/10.4269/ajtmh.1970.19.677

Macher, A. M., & Wolfe, M. S. (2006). Historical Lassa Fever Reports and 30-year Clinical Update. *Emerging Infectious Diseases*, *12*(5), 835–837. https://doi.org/10.3201/eid1205.050052

Mahdy, M. S., Chiang, W., McLaughlin, B., Derksen, K., Truxton, B. H., & Neg, K. (1989). Lassa fever: The first confirmed case imported into Canada. *Canada Diseases Weekly Report = Rapport Hebdomadaire Des Maladies Au Canada*, *15*(39), 193–198.

Manning, J. T., Forrester, N., & Paessler, S. (2015). Lassa virus isolates from Mali and the Ivory Coast represent an emerging fifth lineage. *Frontiers in Microbiology*, *6*, 1037. https://doi.org/10.3389/fmicb.2015.01037

McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Mertens, P. E., Patton, R., Baum, J. J., & Monath, T. P. (1973). Clinical presentation of Lassa fever cases during the hospital epidemic at Zorzor, Liberia, March-April 1972. *The American Journal of Tropical Medicine and Hygiene*, *22*(6), 780–784. https://doi.org/10.4269/ajtmh.1973.22.780

Mire, C. E., Cross, R. W., Geisbert, J. B., Borisevich, V., Agans, K. N., Deer, D. J., Heinrich, M. L., Rowland,

M. M., Goba, A., Momoh, M., Boisen, M. L., Grant, D. S., Fullah, M., Khan, S. H., Fenton, K. A., Robinson, J. E., Branco, L. M., Garry, R. F., & Geisbert, T. W. (2017). Human-monoclonal-antibody therapy protects nonhuman primates against advanced Lassa fever. *Nature Medicine*, *23*(10), 1146–1149. https://doi.org/10.1038/nm.4396

Monath, T. P., Mertens, P. E., Patton, R., Moser, C. R., Baum, J. J., Pinneo, L., Gary, G. W., & Kissling, R. E. (1973). A hospital epidemic of Lassa fever in Zorzor, Liberia, March-April 1972. *The American Journal of Tropical Medicine and Hygiene*, *22*(6), 773–779. https://doi.org/10.4269/ajtmh.1973.22.773

Nembaware, V., Bennett, D., Chimusa, E. R., Chikowore, T., Daodu, R., Bitoungui, V. N., Williams, S. M., Fatumo, S., Healy, S., Seoighe, C., Wonkam, A., Landouré, G., Ndiaye, R., Dandara, C., Mutesa, L., Ramsay, M., El-Kamah, G., Sirugo, G., Makani, J., … Coughlan, S. C. (2023). The African Society of Human Genetics successfully launches global data science workshops. *Trends in Genetics*, *39*(11), 803–807. https://doi.org/10.1016/j.tig.2023.06.004

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Olayemi, A., Cadar, D., Magassouba, N., Obadare, A., Kourouma, F., Oyeyiola, A., Fasogbon, S., Igbokwe, J., Rieger, T., Bockholt, S., Jérôme, H., Schmidt-Chanasit, J., Garigliany, M., Lorenzen, S., Igbahenah, F., Fichet, J.-N., Ortsega, D., Omilabu, S., Günther, S., & Fichet-Calvet, E. (2016). New Hosts of The Lassa Virus. *Scientific Reports*, *6*, 25280. https://doi.org/10.1038/srep25280

Olono, A., Mitesser, V., Happi, A., & Happi, C. (2024). Building genomic capacity for precision health in Africa. *Nature Medicine*, 1–9. https://doi.org/10.1038/s41591-024-03081-9

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Cournapeau, D. (2011). Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*.

Perrett, H. R., Brouwer, P. J. M., Hurtado, J., Newby, M. L., Liu, L., Müller-Kräuter, H., Müller Aguirre, S., Burger, J. A., Bouhuijs, J. H., Gibson, G., Messmer, T., Schieffelin, J. S., Antanasijevic, A., Boons, G.-J., Strecker, T., Crispin, M., Sanders, R. W., Briney, B., & Ward, A. B. (2023). Structural conservation of Lassa virus glycoproteins and recognition by neutralizing antibodies. *Cell Reports*, *42*(5), 112524. https://doi.org/10.1016/j.celrep.2023.112524

Richmond, J. K., & Baglole, D. J. (2003). Lassa fever: Epidemiology, clinical features, and social consequences. *BMJ*, *327*(7426), 1271–1275. https://doi.org/10.1136/bmj.327.7426.1271

Robinson, J. E., Hastie, K. M., Cross, R. W., Yenni, R. E., Elliott, D. H., Rouelle, J. A., Kannadka, C. B., Smira, A. A., Garry, C. E., Bradley, B. T., Yu, H., Shaffer, J. G., Boisen, M. L., Hartnett, J. N., Zandonatti, M. A., Rowland, M. M., Heinrich, M. L., Martínez-Sobrido, L., Cheng, B., … Garry, R. F. (2016). Most neutralizing human monoclonal antibodies target novel epitopes requiring both Lassa virus glycoprotein subunits. *Nature Communications*, *7*(1), 11544. https://doi.org/10.1038/ncomms11544

Sagulenko, P., Puller, V., & Neher, R. A. (2018). TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution*, *4*(1), vex042. https://doi.org/10.1093/ve/vex042

Salam, A. P., Duvignaud, A., Jaspard, M., Malvy, D., Carroll, M., Tarning, J., Olliaro, P. L., & Horby, P. W. (2022). Ribavirin for treating Lassa fever: A systematic review of pre-clinical studies and implications for human dosing. *PLoS Neglected Tropical Diseases*, *16*(3), e0010289. https://doi.org/10.1371/journal.pntd.0010289

Salami, K. A., Mandi, H. E., Imbault, N., & Tornieporth, N. G. (2022). The promise, problems, and pitfalls of including pregnant women in clinical trials of Lassa fever vaccine: A qualitative assessment of sub-Sahara Africa investigators' perception. *The Pan African Medical Journal*, *41*, 242. https://doi.org/10.11604/pamj.2022.41.242.33863

Sharaf, A., Ndiribe, C. C., Omotoriogun, T. C., Abueg, L., Badaoui, B., Badiane Markey, F. J., Beedessee, G., Diouf, D., Duru, V. C., Ebuzome, C., Eziuzor, S. C., Jaufeerally Fakim, Y., Formenti, G., Ghanmi, N.,

Guerfali, F. Z., Houaga, I., Ideozu, J. E., Katee, S. M., Khayi, S., … Ebenezer, T. E. (2023). Bridging the gap in African biodiversity genomics and bioinformatics. *Nature Biotechnology*, *41*(9), 1348–1354. https://doi.org/10.1038/s41587-023-01933-2

Shlaeffer, F., Sikuler, E., & Keynan, A. (1988). [Lassa fever—First case diagnosed in Israel]. *Harefuah*, *114*(1), 12–14.

Simons, D. (2023). Lassa fever cases suffer from severe underreporting based on reported fatalities. *International Health*, *15*(5), 608–610. https://doi.org/10.1093/inthealth/ihac076

Whitmer, S. L. M., Strecker, T., Cadar, D., Dienes, H.-P., Faber, K., Patel, K., Brown, S. M., Davis, W. G., Klena, J. D., Rollin, P. E., Schmidt-Chanasit, J., Fichet-Calvet, E., Noack, B., Emmerich, P., Rieger, T., Wolff, S., Fehling, S. K., Eickmann, M., Mengel, J. P., … Günther, S. (2018). New Lineage of Lassa Virus, Togo, 2016. *Emerging Infectious Diseases*, *24*(3), 599–602. https://doi.org/10.3201/eid2403.171905

WHO. (n.d.). *Prioritizing diseases for research and development in emergency contexts*. Retrieved January 20, 2024, from https://www.who.int/activities/prioritizing-diseases-for-research-and-development-in-emergency-contexts

Wolf, T., Ellwanger, R., Goetsch, U., Wetzstein, N., & Gottschalk, R. (2020). Fifty years of imported Lassa fever: A systematic review of primary and secondary cases. *Journal of Travel Medicine*, *27*(4), taaa035. https://doi.org/10.1093/jtm/taaa035

Wolff, S., Schultze, T., Fehling, S. K., Mengel, J. P., Kann, G., Wolf, T., Eickmann, M., Becker, S., Hain, T., & Strecker, T. (2016). Genome Sequence of Lassa Virus Isolated from the First Domestically Acquired Case in Germany. *Genome Announcements*, *4*(5), e00938-16. https://doi.org/10.1128/genomeA.00938-16

Woodruff, A. W. (1975). Handling patients with suspected Lassa fever entering Great Britain. *Bulletin of the World Health Organization*, *52*(4–6), 717–721.

Woodruff, A. W., Monath, T. P., Mahmoud, A. A., Pain, A. K., & Morris, C. A. (1973). Lassa fever in Britain: An imported case. *British Medical Journal*, *3*(5881), 616–617. https://doi.org/10.1136/bmj.3.5881.616

Woolsey, C., Cross, R. W., Prasad, A. N., Agans, K. N., Borisevich, V., Deer, D. J., Dobias, N. S., Fears, A. C., Harrison, M. B., Heinrich, M. L., Fenton, K. A., Garry, R. F., Branco, L. M., & Geisbert, T. W. (2024). Monoclonal antibody therapy demonstrates increased virulence of a lineage VII strain of Lassa virus in nonhuman primates. *Emerging Microbes & Infections*, *13*(1), 2301061. https://doi.org/10.1080/22221751.2023.2301061

Yadouleton, A., Picard, C., Rieger, T., Loko, F., Cadar, D., Kouthon, E. C., Job, E. O., Bankolé, H., Oestereich, L., Gbaguidi, F., Pahlman, M., Becker-Ziaja, B., Journeaux, A., Pannetier, D., Mély, S., Mundweiler, S., Thomas, D., Kohossi, L., Saizonou, R., … Fichet-Calvet, E. (2020). Lassa fever in Benin: Description of the 2014 and 2016 epidemics and genetic characterization of a new Lassa virus. *Emerging Microbes & Infections*, *9*(1), 1761–1770. https://doi.org/10.1080/22221751.2020.1796528

Yanase, O., Motomiya, T., Watanabe, K., Tokuyasu, Y., Sakurada, H., Tejima, T., Hiyoshi, Y., Sugiura, M., Yabata, Y., & Kitazumi, H. (1989). [Lassa fever associated with effusive constrictive pericarditis and bilateral atrioventricular annular constriction: A case report]. *Journal of Cardiology*, *19*(4), 1147–1156.

Zweighaft, R. M., Fraser, D. W., Hattwick, M. A., Winkler, W. G., Jordan, W. C., Alter, M., Wolfe, M., Wulff, H., & Johnson, K. M. (1977). Lassa fever: Response to an imported case. *The New England Journal of Medicine*, *297*(15), 803–807. https://doi.org/10.1056/NEJM197710132971504