Clinical Sequencing Uncovers Origins and Evolution of Lassa Virus

Graphical Abstract

Highlights

- Lassa virus is a life-threatening pathogen that is endemic in West Africa
- Lassa virus has diverse and ancient origins in Nigeria
- Viral strains from Nigeria and Sierra Leone differ in their translation efficiency
- The virus evolves within hosts to evade immune-determined selection pressures

Authors
Kristian G. Andersen, B. Jesse Shapiro, Christian B. Matranga, ..., Christian T. Happi, Robert F. Garry, Pardis C. Sabeti

Correspondence
andersen@scripps.edu (K.G.A.), happic@run.edu.ng (C.T.H.), psabeti@oeb.harvard.edu (P.C.S.)

In Brief
Sequencing analysis of ~200 Lassa virus genomes reveals its ancient origins and distinct evolutionary features compared to the Ebola virus.
Clinical Sequencing Uncovers Origins and Evolution of Lassa Virus

Kristian G. Andersen,1,2,4,21 B. Jesse Shapiro,1,2,4,21 Christian B. Matranga,2,4,21 Rachel Sealfon,2,5 Aaron E. Lin,1,2 Lina M. Moses,6 Onikepe A. Folakin,7,8 Augustine Goba,9 Ikpomwomosa Odia,7 Philomena E. Ehiane,7 Mambu Momoh,9,10 Eleina M. England,2 Sarah Winnicki,1,2 Luis M. Branco,11 Stephen K. Gire,1,2 Eric Phelan,2 Ridhi Tariyal,2 Ryan Tewhey,1,2 Omowummi Omoniwa,7 Mohammed Fullah,8,10,23 Richard Fonnie,9,23 Mbalu Fonnie,9,23 Lansana Kanneh,9 Simbirie Jalloh,9 Michael Gbakie,9 Sidiki Saffa,9,23 Kandeh Karbo,9 Adrianne D. Gladden,2 James Qu,2 Matthew Stremlau,1,2 Mahan Nekoui,1,2 Hilary K. Finucane,2 Shervin Tabrizi,1,2 Joseph J. Vitti,1 Bruce Birren,2 Michael Fitzgerald,2 Caryn McCowan,2 Andrea Ireland,2 Aaron M. Berlin,2 James Bochicchio,2 Barbara Tazon-Vega,2 Niall J. Lennon,2 Elizabeth M. Ryan,2 Zach Bjornson,12 Danny A. Milner, Jr.,13 Amanda K. Lukens,13 Nisha Broodie,14 Megan Rowland,11 Megan Heinrich,11 Marjan Akdag,11 John S. Schieffelin,11 Danielle Levy,15 Henry Akpan,15 Daniel G. Bausch,6 Kathleen Rubins,16 Joseph B. McCormick,17 Eric S. Lander,2 Stephan Günter,18 Lisa Hensley,19 Sylvanus Okogbenin,19 Viral Hemorrhagic Fever Consortium,20 Stephen F. Schaffner,2 Peter O. Okokhere,2 S. Humarr Khan,23 Donald S. Grant,23 George O. Akpede,7 Danny A. Asogun,7 Andreas Gnirke,7 Joshua Z. Levin,2,22 Christian T. Happi,7,8,22, Robert F. Garry,8,22 and Pardis C. Sabeti1,2,4,22,23

1FAS Center for Systems Biology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
2Broad Institute, Cambridge, MA 02142, USA
3The Scripps Research Institute, Scripps Translational Science Institute, La Jolla, CA 92037, USA
4Department of Biological Sciences, University of Montréal, Montréal, QC H2V 2S9, Canada
5Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
6Tulane Health Sciences Center, Tulane University, New Orleans, LA 70118, USA
7Institute of Lassa Fever Research and Control, Irrua Specialist Teaching Hospital, Irrua, Edo State, Nigeria
8Department of Biological Sciences, College of Natural Sciences, Redeemer’s University, Redemption City, Osun State, Nigeria
9Lassa Fever Laboratory, Kenema Government Hospital, Kenema, Eastern Province, Sierra Leone
10Eastern Polytechnic College, Kenema, Eastern Province, Sierra Leone
11Zaibling Labs, Germantown, MD 20876, USA
12Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA 94304, USA
13Department of Immunology and Infectious Disease, Harvard School of Public Health, Boston, MA 02115, USA
14College of Medicine, Columbia University, New York, NY 10032, USA
15Nigerian Federal Ministry of Health, Abuja, Federal Capital Territory, Nigeria
16The National Aeronautics and Space Administration, Johnson Space Center, Houston, TX 77058, USA
17The University of Texas School of Public Health, Brownsville, TX 77030, USA
18Department of Virology, Bernhard-Nocht-Institute for Tropical Medicine, 20259 Hamburg, Germany
19NIAID Integrated Research Facility, Frederick, MD 21702, USA
20Tulane University, New Orleans, LA 70118, USA
21Co-first author
22Co-senior author
23Deceased

*Correspondence: andersen@scripps.edu (K.G.A.), happic@run.edu.ng (C.T.H.), psabeti@oeb.harvard.edu (P.C.S.)
http://dx.doi.org/10.1016/j.cell.2015.07.020

SUMMARY

The 2013–2015 West African epidemic of Ebola virus disease (EVD) reminds us of how little is known about biosafety level 4 viruses. Like Ebola virus, Lassa virus (LASV) can cause hemorrhagic fever with high case fatality rates. We generated a genomic catalog of almost 200 LASV sequences from clinical and rodent reservoir samples. We show that whereas the 2013–2015 EVD epidemic is fueled by human-to-human transmissions, LASV infections mainly result from reservoir-to-human infections. We elucidated the spread of LASV across West Africa and show that this migration was accompanied by changes in LASV genome abundance, fatality rates, codon adaptation, and translational efficiency. By investigating intrahost evolution, we found that mutations accumulate in epitopes of viral surface proteins, suggesting selection for immune escape. This catalog will serve as a foundation for the development of vaccines and diagnostics.

INTRODUCTION

Viruses that cause human hemorrhagic fevers, such as Ebola, Marburg, and Lassa, are classified as BL-4 agents due to their high fatality rates and lack of effective treatment (Paessler and Walker, 2013). With increasing globalization, changing climatic conditions, and an ever-expanding human population, our interactions with these pathogens are likely to increase (Gire et al.,
The 2013–2015 Ebola virus disease (EVD) epidemic (Baize et al., 2014) is a stark reminder that a better understanding of these viruses is required to develop effective therapeutics and vaccines, as standard containment and isolation can be insufficient to prevent large-scale outbreaks (Pandey et al., 2014).

Lassa virus (LASV) is unique among BL-4 agents in being a common human pathogen, causing endemic disease in much of West Africa, primarily in Sierra Leone, Guinea, Liberia, and Nigeria (Figure 1A). Infection with LASV can lead to acute Lassa fever (LF) with symptoms similar to EVD. LASV is estimated to hospitalize tens of thousands and cause several thousand deaths each year. Case fatality rates (CFRs) among hospitalized LF patients can exceed 50%, although numerous sub-clinical infections are believed to occur (Troup et al., 1970; McCormick and Fisher-Hoch, 2002). Most patients are infected by exposure to excreta from the rodent *Mastomys natalensis*, which functions as a reservoir and maintains persistent infections (Lecompte et al., 2006); human-to-human transmissions have also been reported, however, primarily in hospital settings (McCormick and Fisher-Hoch, 2002; Lo Iacono et al., 2015).

LASV is a single-stranded RNA virus in the family Arenaviridae. Its genome consists of two segments, L (7.3 kb) and S (3.4 kb), which encode four proteins: Z (matrix), L (polymerase), NP (nucleoprotein), and GPC, which is post-translationally cleaved into two peptides, GP1 and GP2, that form the transmembrane glycoprotein (Figure 1B). EBOV (*Zaire ebolavirus*) is a single-stranded RNA virus in the family Filoviridae with a 19-kb genome encoding seven proteins. While the prevalence of LASV makes it a rare model for studying the evolution of a BL-4 pathogen, only 12 whole-genome LASV sequences were available prior to this study (Djavani et al., 1997; Vieth et al., 2004).

**RESULTS**

**Generation of a Large Dataset of Lassa Virus Genomes**

We established partnerships with Kenema Government Hospital (KGH), Sierra Leone, and Irrua Specialist Teaching Hospital (ISTH), Nigeria, and collected samples from LF patients between 2008 and 2013. We implemented diagnostics, training, and infrastructure to ensure high-quality and safe sample collection from patients hospitalized with LF (Shaffer et al., 2014).

We sequenced 183 LASV genomes from these clinical samples, 11 LASV genomes from *M. natalensis* field samples, and two genomes from viral laboratory isolates (Figure 1C; Table S1); we deposited all sequence data at NCBI (BioProject PRJNA254017) before publication. Most samples contained >50% human material and yielded <1% LASV reads (Figures S1A and S1B; Table S1). Genome coverage was fairly uniform, with higher coverage of the S than the L segment (Figure 1D), consistent with a greater copy number of S (Southern, 1996).

Since we used an unbiased sequencing approach, we were also able to assemble 7,028 unique open reading frames from the transcriptome of *M. natalensis*, a species not previously sequenced (Figures S1C–S1E; Data S1).

---

Figure 1. Lassa Fever Is a Viral Hemorrhagic Fever Endemic in West Africa, where Ebola Virus Disease Broke Out in 2014

(A) Overview of the LF endemic zone. Study sites are marked. The LF risk zone was defined according to Fichet-Calvet and Rogers (2009).

(B) Schematic of LASV virions.

(C) Summary of LASV sequence data (% ORF coverage = average coverage of open reading frames; x Coverage = median base pair (bp) coverage; % bp > Q32 = fraction of bp with a phred-score > 32.

(D) Plot of the combined normalized (to the sample average) genome coverages (matched dataset, n = 167).

See also Figure S1, Tables S1 and S2, and Data S1.
Lassa Virus Strains Are Genetically Diverse and Cluster Based on Geographic Location

We first examined patterns of variation and phylogenetic relationships. We found high levels of LASV nucleotide diversity, with strain variation up to 32% and 25% for the L and S segments (Figures 2A, S1F, and S1G). This is substantially higher than previous findings based on LASV fragments (Bowen et al., 2000), and much higher than EBOV, which is more than 97% conserved across all sequenced strains (Figures 2B and S1H). We confirmed previous findings (Bowen et al., 2000) that LASV clusters into four major clades: three in Nigeria and one from the Mano River Union countries (MRU) of Sierra Leone, Guinea, and Liberia (Figure 2A; Data S1, S2, and S3). We found no evidence for host-specific clades of LASV lineages; rather, samples from humans and *M. natalensis* clustered together (Figure 2A; Data S1, S2, and S3). We did not identify any recombination events within segments, but did find evidence for reassortment between segments in three samples (Figures S2A–S2G). This could be explained by infections of individual hosts with multiple LASV lineages, followed by shuffling of segments, a process previously observed in vitro with LASV (Lukashevich, 1992) and in vivo with other arenaviruses (Stenglein et al., 2015).

**Figure 2. LASV Is More Diverse than EBOV and Has Ancient Origins in Nigeria**

(A) Phylogenetic tree of LASV S segments (n = 211) (outer ring: gray, previously sequenced; orange, sequenced from *M. natalensis*; scale bar, nucleotide substitutions/site; I–IV, lineages as defined by Bowen et al. (2000). (B) Scaled trees of LASV L and S segments, as well as EBOV. Trees are shown with the same scale of genetic distance (0.1 nucleotide substitutions/site), except for EBOV, which was magnified 10^3 (0.01 nucleotide substitutions/site). LASV lineages are shown (Nig., Nigeria; MRU, Mano River Union). (G and D) Root-to-tip distance versus collection date for (C) EBOV from the West African EVD epidemic (2014; n = 131) or (D) LASV from Sierra Leone (2012; n = 21). Confidence intervals (95%) for linear regression fits are shown in blue. (E) The % pairwise differences (log scale) in EBOV lineages from the 2014 EVD epidemic (March–October 2014; n = 116) and LASV lineages from Sierra Leone (SL: 2009–2013; n = 60) and Nigeria (NG; 2009–2012; n = 83). The % divergence was calculated within the countries for each year separately and pooled. Error bars represent SD. (F–H) Bayesian coalescent analysis of LASV samples (matched dataset, n = 179). (F) Substitution rates. (G) LASV L segment tMRCA for each country (median values; ya, years ago). Gray arrows depict the likely spread of LASV. An asterisk indicates a tMRCA that was dependent on only one sequence (AV) from outside Nigeria and MRU. (H) Probability distributions for the estimated tMRCA with median marked.

See also Figures S1F–S1H, S2, S3, S4, and S5, Table S2, and Data S1, S2, and S3.
Lassa Virus Infections Are the Result of Multiple Independent Reservoir-to-Human Transmissions

Recent studies suggest that the 2013–2015 EVD epidemic is maintained by sustained human-to-human transmission (Gire et al., 2014) after an initial “spillover” event from a likely animal reservoir (Baize et al., 2014). Similarly, it has been suggested that up to 20% of LF cases also arise from human-to-human transmissions (Lo Iacono et al., 2015). Sustained human-to-human transmission should result in a “ladder-like” structure of the phylogenetic tree along with a strong correlation between a sample’s collection date and its genetic distance from the root of the tree over a short time period. Based on data from the 2013–2015 EVD epidemic (Team, 2014), we defined that time period as one year. While collection date is strongly correlated with root-to-tip distance for EBOV from the 2013–2015 EVD epidemic (R² = 0.64; Figure 2C; Table S2), the same correlation is absent for LASV sampled over a similar time period (R² = 0.0001; Figure 2D; Table S2).

Human-to-human transmission should also result in clustering of contemporaneous viral sequences on the tree. While this is pervasive across the 2013–2015 EVD epidemic samples (Gire et al., 2014) (Data S1), we found that only 5 out of 169 (3%) LASV sequences from patients resulted in such clusters (Data S1 and S3). As M. natalensis serves as the reservoir host for LASV—and presumably maintain LASV diversity via sustained rodent-to-rodent transmission chains—we would expect rodent samples to group into more defined clusters. Indeed, 5 out of 10 (50%) LASV sequences from M. natalensis formed clusters consistent with rodent-to-rodent transmissions (Data S1 and S3). Finally, we also found that the average pairwise divergence for EBOV lineages in Sierra Leone from the 2013–2015 EVD epidemic was much lower than that observed for LASV lineages with individual years from Sierra Leone (Figure 2E), despite similar observed substitution rates (Figure 2F). These three lines of evidence suggest that, while EBOV during the 2013–2015 EVD epidemic was transmitted through human-to-human contact, of evidence suggest that, while EBOV during the 2013–2015 EVD epidemic was transmitted through human-to-human contact, most human LASV infections represent independent transmissions from a genetically diverse reservoir.

Lassa Virus Has Ancient Origins in Modern-Day Nigeria and Has Recently Spread Across West Africa

While EBOV and LASV were both discovered in the latter part of the 20th century—1976 and 1969, respectively—their origins likely vary greatly (Commission, 1978; Frame et al., 1970), Reports suggest that all EVD outbreaks share a common ancestor within the last fifty years (Carroll et al., 2013; Dudas and Rambaut, 2014; Calvignac-Spencer et al., 2014; Gire et al., 2014). In contrast, the widespread persistence of LASV in M. natalensis, and evidence in the human genome of natural selection linked to LASV resistance (Andersen et al., 2012), suggest that LASV might be a long-standing human pathogen. Using molecular dating, we found that extant LASV strains likely originated in modern-day Nigeria more than a thousand years ago and spread into neighboring West African countries within the last several hundred years (Figures 2G and 2H). We first examined evidence for a molecular clock by comparing sample collection dates and root-to-tip distances across the entire LASV tree. In contrast to the shorter timescales analyzed above (Figure 2D), here we found significant evidence for a molecular clock (R² = 0.38, p value < 0.0001). This allowed us to calculate the time to the most recent common ancestor (tMRCA) using Bayesian coalescent analysis (Drummond et al., 2012). We estimated the tMRCA of sampled extant LASV strains to be a little over one thousand years for the L segment (Figure S2H; Table S2; median = 1,057 years ago [ya]; 915 ya–1,218 ya; 95% highest posterior density [HPD]) and 650 years for the S segment (Figure S2I; Table S2; median = 631 ya; 519 ya–748 ya, 95% HPD). While LASV strains in Nigeria have the same tMRCA as all extant strains, those in Sierra Leone have an estimated tMRCA of only 150 years (Figures 2G and 2H; median = 153 ya; 137 ya–171 ya; 95% HPD).

We tested the sensitivity of our results to key analysis parameters that could severely affect our tMRCA estimates (Wertheim and Kosakovsky Pond, 2011; Wertheim et al., 2013). We found that our estimates were robust to the choice of all tested parameters, including evolutionary model, geographical separations, and exclusion or inclusion of older “anchoring” sequences, e.g., the 1969 Pinneo strain (Figures S3 and S4; Table S2). In linear regression of root-to-tip distance of samples on the date of collection, the sequences from the MRU showed the strongest evidence of temporal structure, suggesting that the dating is most reliable within that region (Figure S4).

Non-Nigerian Lassa Virus Strains Have Higher Codon Adaptation to Mammalian Hosts

Previous studies have shown that viruses can adapt their codon usage to that of their hosts for translational efficiency (Sharp and Li, 1987; Bahir et al., 2009; Butt et al., 2014; Hershberg and Petrov, 2008). We examined the codon adaptation index (CAI) of LASV and EBOV to different hosts. CAI quantifies how well synonymous codon choice in the viral genome matches that of a potential host genome.

We found that LASV had a higher mean CAI than EBOV, and a similar CAI distribution across different potential mammalian hosts (Figure 3A). There was a strong linear correlation between the CAI of LASV to human and to M. natalensis, regardless of which organism LASV was sequenced from (Figure SS5A). In agreement with previous studies (Bahir et al., 2009), this suggests that codon adaptation to one mammal also leads to adaptation to another.

We also compared LASV sequences from patients in Sierra Leone to those from Nigeria and found that the former had significantly higher CAI (p value < 0.001, permutation test) (Figure 3B). This apparent “burst” of codon adaptation as LASV spread into Sierra Leone began on the branch leading out of Nigeria and remained high in most non-Nigerian strains (Figures 3C, 3D, and SSB–SSE), with an even distribution across the LASV genome (Figure S5F).

As it has been suggested that dinucleotide usage play a role in determining translational efficiency of RNA viruses (Tulloch et al., 2014), we investigated whether there was a difference between Nigerian and Sierra Leonian strains, but did not observe any significant skew (Figures S5G and S5H).

Lassa Virus Genome Abundance and Case Fatality Rates Differ between Nigeria and Sierra Leone

Increased codon optimization might lead to increased viral output (Plotkin and Kudla, 2011) and therefore higher viral titers...
(Lauring et al., 2012) for non-Nigerian strains. With standardized inclusion criteria at our field sites (Supplemental Experimental Procedures), we tested this hypothesis by using qPCR to quantify LASV genome abundance. We found significantly more LASV genomes in patients from Sierra Leone than in those from Nigeria (Figure 4A). LASV genome abundance in Sierra Leone was similar to that observed in EBOV patients from the same hospital (KGH; Figure 4A) and decreased over the course of the infection (Figure S5I), likely due to treatment with the antiviral drug ribavirin (McCormick et al., 1986).

Next, we binned the LASV samples into those in the top or bottom 50% CAI from within each country, and compared LASV genome abundance between bins. In Sierra Leone, individual LASV sequences with high CAI tended to have higher genome copy numbers (p value < 0.05, Mann-Whitney test) but no trend was visible in Nigeria (Figures 4B and 4C). This suggests that CAI may affect LASV replication rate and abundance.

Since increased viremia of LASV in LF patients is correlated with higher fatality rates (McCormick and Fisher-Hoch, 2002), we might also expect CFRs to be higher in patients from Sierra Leone than in Nigeria. Again using strict criteria for inclusion, we found a significantly higher CFR (p value = 0.01; Fisher’s exact test) in Sierra Leonean patients than in their Nigerian counterparts (81% versus 60%; Figure 4D). While the treatment options for LF patients are similar in the two countries, other factors could also affect genome abundances and CFRs. In particular, delay in clinical care could bias our estimates; however, self-reported times from onset of symptoms to hospital admission are the same in the two countries (average = 9.3 days; Figure 4E).

Nigerian Lassa Virus Strains Have Higher Protein Output than Do Sierra Leonean Strains

Although we observed a correlation among CAI, viral genome abundance, and CFR, it remained unclear whether this is driven by differences in protein translation efficiency between Nigerian and Sierra Leonean LASV strains. We designed an experimental system to estimate translational activity for a single LASV gene with different CAI values. We randomly selected 20 LASV sequences from Nigeria and Sierra Leone and fused the first 699 bp of their NP genes (NP1–699) to luciferase, before cloning into expression vectors for transfection or in vitro translation experiments (Figure 4F). Readout of luciferase activity allowed us to detect differences in translational activity of the chimeric transcripts. As controls, we codon-optimized one LASV sequence from Nigeria and one from Sierra Leone, for an upper bound on NP1–699-luciferase translational efficiency.

For both transfection and in vitro translation experiments, we observed a significant difference in translational output of the tested NP1–699-luciferase genes, with Nigerian versions having higher outputs (Figures 4G and 4H). This was the opposite of the expectation based on CAI because the Sierra Leonean sequences had higher CAI (Table S2). Nigerian versions also had higher outputs for the codon-optimized forms of NP1–699 (Figure 4H), suggesting that Nigerian sequences are intrinsically more efficient or stable.

To test whether these observations were specific to NP, we repeated the in vitro translation experiment using the first 736 bp of ten LASV GPC genes (Figure 4F). Once again, we found that Nigerian genes had significantly higher translational output (Figure 4I). These results suggest that there is a difference in the translational output between LASV strains from Nigeria and Sierra Leone that is independent of the variation in CAI.
Lassa Virus Is a More Diverse Intrahost than Is Ebola Virus

The long-term evolution of viruses ultimately depends on mutation and selection within individual hosts (Parameswaran et al., 2012). Our deep sequencing allowed us to examine LASV intrahost single-nucleotide variants (iSNVs) within individual human and rodent hosts (Figure 5A). We called variants at a minimum minor allele frequency (minMAF) of 5% and applied stringent filtering (Supplemental Experimental Procedures). We validated subsets of iSNVs using different sequencing technologies and found that our results were consistent across platforms, experimental replicates, library preparations, and variant calling methods (Figures S5J–S5L and S6).

We found that *M. natalensis* generally harbors more LASV iSNVs than humans (median iSNVs/kb = 1.5 versus 0.1; p value < 0.0001; Mann-Whitney test), consistent with longer, more chronic infections (Figures 5B and S7A–S7D). LASV is a significantly more diverse intrahost than is EBOV (accounting for differences in sequence coverage between the two; median bp coverage \(2,000\) for EBOV [Gire et al., 2014] and \(250\) for LASV; Figure 1C; p value \(= 0.0005\); Mann-Whitney test), with an average number of iSNVs per covered site of \(2.1 \times 10^{-3}\) in...
LF patients, but only $1.3 \times 10^{-4}$ in EVD patients (Figure 5C). This difference is primarily driven by a subset of LASV-infected individuals that have $>15$ iSNVs—diversity similar to that observed in *M. natalensis* (Figure 5C). Such high diversity—with iSNV frequencies that appear stable over the course of infection (Figure S7E)—was never observed in EVD patients (Figures 5B and 5C).

**Natural Selection Is Acting on the Lassa Virus Glycoprotein**

Next, we investigated the role of natural selection in shaping intrahost variation. In LASV, we observed a significantly higher dN/dS (p value = 0.0013; permuted McDonald-Kreitman test)—a measure of selective constraint at the protein level—within hosts than between hosts (Figure 5D). For EBOV, the trend was in the same direction but was not statistically significant (Figure 5D). Assuming that dN represents mostly deleterious mutations (Sha-...
we used a machine-learning method (El-Manzalawy et al., 2008) to predict B cell epitopes in each LASV protein.

Nonsynonymous iSNVs in LASV GPC occurred in predicted B cell epitopes significantly more than expected by chance (Figures 6A and 6B; p value < 0.01; binomial test). This was true for LASV samples from patients and M. natalensis independently, although the signal was stronger in patients (Figure 6A). In contrast, synonymous iSNVs were randomly distributed across GPC, consistent with their lack of impact on epitope structure (Figures 6A and 6B). We observed a similar but weaker trend for NP, although this difference only reached statistical significance in M. natalensis (Figure 6A).

To test if nonsynonymous iSNVs interfere with B cell epitope recognition, we reran the B cell epitope predictions, changing single amino acids within the epitopes from the consensus call to the iSNV variant. For 14 of the 18 predicted B cell epitopes, changing the iSNV from the consensus to the variant allele significantly reduced the epitope score (Table S2; p value = 0.015, Sign test).

To test if nonsynonymous iSNVs also appear to fall within T cell epitopes, we predicted T cell epitopes in each LASV protein (Supplemental Experimental Procedures). We found that nonsynonymous iSNVs accumulated to some extent in LASV GPC, although the results did not reach statistical significance (Figure 6C; p value = 0.07; binomial test).

Intrahost Variants Interfere with Antibody Binding

To investigate the functional effects of a subset of LASV iSNVs, we created iSNV mutations in predicted B cell epitopes in GP1 (Supplemental Experimental Procedures), expressed them in HEK293 cells, and tested their binding to a panel of GPC-specific monoclonal antibodies (mAbs) using flow cytometry. These mutations led to a significant drop in the average mean fluorescence intensity (MFI) for GP1-specific mAbs (Figure 6D), consistent

Figure 6. Nonsynonymous iSNVs Are Overrepresented within Predicted B Cell Epitopes in LASV GPC

(A) The fraction of iSNVs within predicted B cell epitopes. The observed fraction is compared to the expected fraction (*p < 0.01, *p < 0.05, binomial test).

(B) Overlap between GPC epitopes and iSNVs. Epitopes were predicted separately in each sample (y axis) and overlaid with iSNVs from that sample.

(C) Fraction of iSNVs falling within predicted T cell epitopes (p value = binomial test).

(D and E) Binding of monoclonal antibodies (mAbs) to iSNV mutants in predicted B cell epitopes was tested in HEK293 cells. (D) Each circle corresponds to the normalized average mean fluorescence intensity (MFI) measured by flow cytometry of each LASV GPC construct carrying either wild-type or iSNV mutations (Supplemental Experimental Procedures). Each tested mAb is shown on the x axis. The MFI was normalized to the MFI of the empty vector control for each experiment. (E) Binding to the GP1-specific mAbs 12.1F and 19.7E using constructs carrying either the major or minor population-wide allele at positions 89 and 114. For comparison, binding to mAb 36.1F, which requires GP2, is also shown. All MFI values are normalized to the MFI of binding to the GP2-specific mAb 37.2D. Error bars show the SD from four independent experiments; *p < 0.05, Mann-Whitney test.

See also Table S2 and Data S1.
with diminished mAb binding. Similarly, when we investigated the effects of single-point mutations within GP1 epitopes, we found that minor alleles in the LASV population displayed significantly reduced binding to GP1-specific mAbs (Figure 6E).

These observations suggest that the host adaptive immune system imposes selective pressures on the intrahost viral population, driving an accumulation of nonsynonymous iSNVs in LASV GPC.

**Nonsynonymous Lassa Virus Intrahost Variants Tend Not to Become Fixed in Other Hosts**

To further explore the evolution of LASV within and between hosts, we investigated how often iSNVs become fixed in other consensus sequences. We defined an iSNV as “fixed” if its minor allelic variant was observed in one or more LASV consensus sequences. We observed a significantly higher nonsynonymous-to-synonymous ratio (N/S) for unfixed compared to fixed iSNVs (Figure 7A), suggesting a selective bias against the fixation of nonsynonymous iSNVs. LASV and EBOV both have similar numbers of unfixed iSNVs, but LASV has many more fixed iSNVs, likely due to higher rates of iSNV fixation (or transmission) in LASV than in EBOV (Figure 7B). However, the putative transmitted (fixed) iSNVs tend to be biased toward synonymous mutations. This bias is much stronger in LASV (Figure 7C, top) but still detectable in EBOV (Figure 7C, middle). The bias cannot be attributed to differences in minor allele frequencies between nonsynonymous and synonymous iSNVs (p value > 0.1; Kolmogorov-Smirnov test) or to a correlation between MAF and prevalence in consensus sequences (p value > 0.1 for both N and S; Pearson’s correlation); therefore, it is best attributed to selection against transmission and/or fixation of nonsynonymous iSNVs.

A single suspected transmission event, between a pair of *M. natalensis* captured on the same day from the same household, provided an opportunity to observe iSNV fixation dynamics on short timescales. The two samples, Z0947 and Z0948, are nearest-neighbors on the LASV phylogeny (Data S1, S2, and S3), suggesting recent (but not necessarily direct) transmission (Supplemental Experimental Procedures). Assuming that transmission occurred from Z0948 to Z0947, we observed that derived alleles reaching high frequency (DAF > 0.5) in Z0947 tended to be non-synonymous, while derived alleles remaining at lower frequency (DAF < 0.5) were always synonymous (Figure 7C, bottom). Other transmission scenarios (Figure S7H; Supplemental Experimental Procedures) also confirm that nonsynonymous iSNVs reach high frequency within a host, but fail to be transmitted to the next host. Along with the dN/dS and epitope analyses, this supports a model in which nonsynonymous iSNVs rise to high frequency within an individual due to positive selection, but are less likely to become fixed in other hosts due to purifying selection.

**DISCUSSION**

**Comparing Ebola Virus and Lassa Virus Evolutionary Dynamics**

EBOV and LASV are RNA viruses that can lead to illnesses with similar clinical symptoms, yet they differ markedly in their epidemiology and evolutionary dynamics. LASV is more than an order of magnitude more diverse than EBOV (Figure 2B), and molecular dating suggests that it has been circulating in Nigeria for over a thousand years, followed by a more recent spread across West Africa (Figure 2G). In contrast, it has been suggested that the Makona variant of EBOV responsible for the EVD epidemic in West Africa was introduced over the last decade (Dudas and Rambaut, 2014; Calvignac-Spencer et al., 2014; Gire et al., 2014).

These analyses, however, provide lower-bound tMRCA estimates of sampled (extant) viral lineages; the true ages of all LASV and EBOV lineages are likely much older (Taylor et al., 2010). Because of limited sampling from Guinea and Liberia,
our LASV dating analysis is likely most accurate within Sierra Leone, although we achieved comparable results when considering the entire dataset or the individual regions alone (Figure S3E). Furthermore, our 400-year-old “out of Nigeria” estimate relies on a single sequence from the Ivory Coast; additional sampling outside the MRU and Nigeria could push back this date.

Because of the high heterogeneity among LASV lineages, continuous monitoring of its mutational spectrum and evolutionary change will be critical for the development of effective vaccines and diagnostics. Since LASV strains cluster by geography, it is more conserved within individual countries. For example, average sequence identity among lineages from Sierra Leone is 90% at the nucleotide level and 95% at the amino acid level (Figure S1F). A useful strategy might therefore be to develop diagnostics, vaccines, and sequence-specific therapeutics that are country specific or that target the most conserved features of the viral genome.

The 2013–2015 West African EVD epidemic likely originated from a single zoonotic transmission event (Baize et al., 2014), followed by sustained human-to-human transmission and clock-like, linear accumulation of mutations (Figure 2C). In contrast, LASV has a clock-like signature on the timescale of decades (Figure S4B), but not on shorter timescales (Figure 2D). Combined with the intermingling of human and M. natalensis samples on the phylogenetic tree (Figure 2A), this is consistent with a genetically diverse pool of LASV being maintained in its rodent reservoir, with most human infections caused by genetically distinct viruses. A recent study suggested that human-to-human transmission of LASV may account for up to 20% of all cases (Lo Iacono et al., 2015), but we found little support for this in our dataset. This does not rule out human-to-human transmission entirely, but it suggests that human transmission chains are the exception rather than the rule.

LASV is more polymorphic within hosts than EBOV, and M. natalensis hosts harbor more polymorphic LASV populations than humans (Figures 5B and 5C). Since most LF and EVD patients have 0 or 1 iSNVs, the difference between LASV and EBOV is mostly driven by a subset of LF patients with many LASV iSNVs (Figure 5B and 5C). LASV iSNV frequencies tend to remain stable over the relatively short period of hospitalization (Figure S7E), suggesting that intrahost de novo mutations and frequency changes may take longer to develop or may occur early in the infection. These observations suggest that—at least in some patients—LASV infections could last longer than EBOV infections, allowing more time for the generation of polymorphism within hosts.

Longer infections also provide more time for natural selection to eliminate deleterious mutations from the viral population. Consistent with longer infection periods in LASV, dN/dS ratios are lower within LF patients than in EVD patients (Figure 5D).

While these findings are consistent with the existence of chronic LASV infections in humans, they do not constitute proof. Further studies are needed to verify the causes of high-diversity LASV infection and the prevalence of nonacute human infections. Compelling evidence could come from longitudinal sampling asymptomatic carriers of LASV, for example.

**Codon Adaptation, Translational Efficiency, and Genome Abundance**

We uncovered significant differences in genome abundance, CFR, CAI, and translational efficiency of LASV strains from Sierra Leone and from Nigeria. The increase in CAI of non-Nigerian strains (Figures 3B–3D), along with higher viral copy numbers and CFRs in human patients (Figures 4A–4D), would seem to suggest that the virus evolved toward greater human virulence. There are indeed many examples suggesting that pathogens with natural reservoirs may evolve toward greater human virulence, so long as they remain avirulent in the reservoir host (Ewald, 1994). We have not, however, been able to establish causality among these observations. While clinical care for LF patients is similar at ISTH and KGH (Supplemental Experimental Procedures)—and time from symptom onset to hospitalization appears to be the same (Figure 4E)—several additional parameters are beyond our control, including variations in socioeconomic, clinical, and human genetic factors between Sierra Leone and Nigeria. These prevent us from determining whether the difference in CFRs has a viral genetic basis and whether variations in CAI, viral genome abundance, and CFRs are causally linked. In addition, while we observed significantly higher LASV genome abundance in LF patients from Sierra Leone, we could not determine whether this difference translates into higher infectious titers. Controlled animal studies in BL-4 laboratories comparing strains from the two countries would help to resolve this important question.

We expected that the increased CAI of Sierra Leonan LASV lineages would lead to increased translational output in an experimental system. Surprisingly, the opposite was true. Nigeri an LASV lineages had significantly higher in vitro protein outputs than their Sierra Leonan counterparts (Figures 4G–4I). Translation-independent mechanisms, such as post-translational modifications and protein stability, could explain these observations, regardless of CAI differences.

If the progenitor to Sierra Leonan LASV strains indeed had lower translational output, the emergence of LASV strains with increased CAI could have been driven by positive selection to compensate. The higher CAI could then have led to higher viral titers in human patients. Alternatively, the increase in CAI could have been caused by genetic drift. Under this scenario, mutational biases (Supplemental Experimental Procedures) and drift—combined with insufficient time for mutations to be exposed to purifying selection in Sierra Leone—led to an increase in CAI, independently of positive selection. The Sierra Leonan tMRCA of the LASV population is indeed relatively recent (Figure 2G), having possibly undergone a recent population bottleneck (Lalis et al., 2012). Further work will be required to disentangle the adaptive and neutral contributions to the evolution of CAI and determine whether changes in CAI affect viral fitness and CFRs.

**Immune Selection on Lassa Virus within Hosts**

The observation that LASV GPC has the highest intrahost dN/dS (Figure 5E) and the most nonsynonymous iSNVs in predicted epitopes (Figure 6) suggests that it is a target of immune-driven diversifying selection within hosts, both in humans and M. natalensis. This effect was most strongly supported when
we investigated the involvement of B cells (Figure 6A), although we also found weaker evidence for the role of T cells (Figure 6C).

Given that human LASV infections are typically thought to be of short duration, it is perhaps surprising that there would be enough time for the host to mount an antibody response, and for viral diversity to be measurably shaped by this selective pressure. However, this is not entirely implausible, since it is known that LASV-specific antibodies appear relatively quickly, with experimentally infected nonhuman primates developing detectable IgM titers after 9 days and IgG titers after 12 days (Baize et al., 2009). Furthermore, it has also been shown that 25% of LF patients in Sierra Leone are IgM-positive upon admission and 10% are IgG-positive (Shaffer et al., 2014). Combined with the observation that several LF patients have LASV ISNV numbers similar to those observed in M. natalensis (Figures 5B and 5C), sufficient time for antibody-mediated responses has likely elapsed in a number of LF patients when they present at the hospital. We were unable to test seropositivity in our cohort, but future studies could assess whether there is a correlation between the presence or absence of LASV-specific antibodies and the number of ISNVs. Our prediction would be that LF patients with high numbers of LASV ISNVs should also tend to be IgM- and/or IgG-positive upon admission.

Conclusions

Our dataset of full-length LASV genomes yields insights into the biogeography, evolution, and spread of a hemorrhagic fever-causing virus. Further improvements in collecting, sequencing, and analyzing LASV and EBOV samples combined with concurrent experiments in BL-4 laboratories will allow us to pursue open areas of inquiry. The catalog of LASV genetic diversity presented here is thus an essential foundation for the development of vaccines and diagnostics for a highly diverse, continuously evolving, and unusually prevalent BL-4 agent.

EXPERIMENTAL PROCEDURES

Ethics Statement

Subjects were recruited for this study using protocols approved by relevant human subjects committees. All patients were treated with a similar standard of care.

Samples

All samples were acquired on the day of admission before any treatment regimen had begun. 10 ml of whole blood was collected, and plasma or serum was prepared. Diagnostic tests for the presence of LASV were performed on-site. Rodents (all from Sierra Leone) were trapped in case households and humanely sacrificed, and samples were collected from serum and/or spleen. All samples were inactivated in either Buffer AVL (QIAGEN) or TRIzol (Life Technologies), following the manufacturer’s standard protocols and stored in solar-driven –20°C freezers.

Case Fatality Rates

CFRs were calculated from patients that had all of the following characteristics: (1) known outcome, (2) positive for LASV in the field, (3) confirmed positive upon retesting in the United States, and (4) sequencing confirmed the presence of LASV.

Sequencing

cDNA synthesis and Illumina library construction were performed, libraries were pooled, and sequenced on the illumina platform. All the data were deposited at NCBI (BioProject PRJNA254017). EBOV data are available under PRJNA257197.

Assembly of LASV Genomes

LASV genomes were de novo assembled, followed by read mapping and duplicate removal.

Trees

Maximum likelihood phylogenies were made with RAxML (v.7.3.0) (Stamatakis et al., 2006) and rooted using the 1969 Pineino strain.

Molecular Dating

BEAST (v.1.7.4) (Drummond and Rambaut, 2007) was used with a model incorporating a log normal relaxed clock, exponential growth, HKY+I with four categories, and codon partitioning (sr96).

Intrahost Variants

iSNVs were called and filtered taking into consideration minimum coverage, minimum frequency (5%), strand-bias, and base quality.

Intrahost Selection

A modified version of the McDonald-Kreitman test (McDonald and Kreitman, 1991) was applied to assess evidence for natural selection.

Epitope Prediction

B cell epitopes were predicted using BCPRED (El-Manzalawy et al., 2008), T cell epitopes were predicted using NetCTL (Larsen et al., 2007).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, two tables, and three data files and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.07.020.

AUTHOR CONTRIBUTIONS


ACKNOWLEDGMENTS

The authors thank C. Edwards, X. Yang, M. Zody, B. Han, A. Andersen, I. Shlyakhter, D. Park, M. Henn, M. Busby, M. Kellis, C. Bishop, A. Haislip, L. Burchfield, A. Matthews, and F. Viloria for their help and advice. We also thank the Ministry of Health and Sanitation Government of Sierra Leone, Oyo State Ministry of Health, and the Federal Ministry of Health, Nigeria. A full list of members of the Viral Hemorrhagic Fever Consortium can be found at www.vhfc.org. This project was funded with federal funds from the NIH, Department of Health and Human Services (1DP2OD006514-01) and NIAID Contracts (HHSN272200900049C, HHSN272201000022C, HHSN27220090018C, and U19AI11081). The authors received additional support from USAMRRA (W81XWH-10-1-0008), a Packard Foundation Fellowship for Science and Engineering, a Broad Institute SPARC award, and the German Research Foundation (GI 883/1-1). K.G.A. was supported by a fellowship from the Karlsberg Foundation; B.J.S. was supported by a Harvard MIDAS CCD postdoctoral fellowship and Canada Research Chair; and R.S. and A.E.L. received NSF GRFP fellowships (DGE 1122374 and 1144152). Tulane University and industry...
partners have filed patent applications for several Lassa-related technologies and may receive royalties or profits if commercial products are developed.

Received: February 12, 2015
Revised: April 26, 2015
Accepted: June 12, 2015
Published: August 13, 2015

REFERENCES


Figure S1. LASV, Human, and Mastomys Sequences Extracted from Illumina cDNA Libraries, Related to Figures 1 and 2

(A) Illumina sequencing reads from the individual samples were aligned against LASV or the human genome. Each circle corresponds to a single library. (B) The amount of LASV genome copies in final cDNA libraries were quantified by qPCR, converted to copies/μl and plotted against the average depth of LASV coverage from the same library. Values were converted to base-10 logarithm. (C-E) Illumina sequence data sets from eight rodents were combined and used for de novo assembly of the *M. natalensis* transcriptome. (C) Table summarizing the analyzed data. (D) Contigs (n = 7,028) generated by de novo assembly were analyzed using BLASTn against the nt database. The fractions of contigs belonging to specific taxonomic families are shown. (E) All unique *M. natalensis* open reading frames (ORFs, n = 4,850; > 600 bp) were used to find orthologous ORFs in relevant rodents and primates. ORFs from the individual species were concatenated, aligned, and a maximum likelihood tree was generated using RAxML (Stamatakis et al., 2005) and midpoint rooted. Bootstrap values (500 pseudoreplicates) are shown (scale bar, nucleotide substitutions/site). Based on our phylogenetic analyses it is clear that *M. natalensis* is more closely related to mice than rats and that the analyzed rodents are genetically more diverse than the analyzed primates. (F-H) Alignment summary statistics for the LASV and EBOV datasets. (F) Average pairwise identity for the L and S segments in different geographical areas (nt = nucleotide identity; aa = amino acid identity). (G) Average pairwise identity for the individual LASV genes. (H) Average pairwise identity for EBOV from Central Africa and West Africa (2014 outbreak).
Figure S2. Evidence for Reassortment in LASV Strains, Related to Figure 2

(A–G) Multiple sequence alignments of L and S segments from each LASV lineage were constructed and screened for the presence of recombinant or reassorted strains. (A) Number of sequences screened from Sierra Leone and Nigeria (lineage II) together with the identified number of recombinants (0 total) and reassortants (3 total). (B–D) The RDP ‘BOOTSCAN’ bootstrap support (see the Supplemental Experimental Procedures) for the pairing between the parents and the identified reassortment events. The reassortant is shown in bold underlined text, the bootstrap support for the reassortant being nearest neighbor with the major parent is shown in blue, and the minor parent in red. (E–G) The identified reassortants, parents, and relevant in- and out-groups were selected and L and S segment alignments were made followed by RAxML (Stamatakis et al., 2005) maximum likelihood trees of the individual segments. L segment (top) and S segment (bottom) trees were rooted based on the out-groups identified in the entire LASV tree. The reassortant is shown in bold underlined text, the L segment parent in blue (same as ‘major’ above), and the S segment (same as ‘minor’ above) parent in red. Bootstrap values (500 pseudoreplicates) are shown for each node. (A–G) In support of these analyses, we also compared the full LASV L and S segment trees using a Shimodaira-Hasegawa test (Shimodaira and Hasegawa, 1999) and likewise found evidence for phylogenetic incongruence between the L and S segments (p value < 0.01).

(H and I) Time-aware phylogenetic analysis of LASV L and S segments. Bayesian coalescent analysis was performed on the full-length (H) LASV L segments and (I) S segments from the Matched dataset (n = 179) and phylogenetic trees were created using BEAST (Drummond and Rambaut, 2007). Branch-lengths are proportional to time before present (years). Orange bars correspond to the 95% HPD for key nodes.
Figure S3. Estimates for the Time to the Most Recent Common Ancestor Are Consistent across a Variety of Models and Data Subsets, Related to Figure 2

(A–I) Bayesian coalescent analyses were performed using several different parameters and models (Batch 1 dataset; n = 95). (A and B) The tMRCAs for the L segment (A) and the S segment (B) were compared under a variety of models. Each column represents a single model. The spectrum of points in each column represent the median times to each node in the tree under a specific model, with the topmost dot in each column representing the estimated median tMRCA under the given model. For each segment, eighteen models were compared, including a variety of clock models (lognormal, relaxed, or strict), population models (constant, exponential growth, or Bayesian Skyline Plot), and codon partitioning schemes (rate heterogeneity but no explicit knowledge of codon positions [HKY], separate partitions for the first and second codon positions combined and the third codon position [srd06], or separate partitions for the first, second, and third codon position [3pos]). The HKY model of nucleotide evolution was used in each model; altering the model of nucleotide evolution to e.g., GTR does not alter the estimated tMRCA. (C and D) To test whether the estimates are dependent on the exact set of samples, we performed molecular dating on random subsets of size n (for n = 10 to 90), with 20 random samples chosen for each subset size. The median estimated tMRCA across the 20 random samples is shown (red line). (E) The time to the most recent common ancestor to geographic subsets of nodes was computed on the full dataset and on the geographic subset only for the L and S segments (MRU = Mano River Union; Nig = Nigeria; Nig_II = Lineage II from Nigeria; oNig = Out of Nigeria; SL = Sierra Leone). The error bars indicate the 95% HPD interval. For parts C–E the full L and S segment alignments with a strict clock, an HKY model of nucleotide evolution, and a constant model of population growth were used. (F and G) Each of the 95 sequences in the Batch 1 dataset were removed from the alignment, and a molecular dating analysis was performed for (F) the L segment and (G) the S segment. For each leave-one-out replicate, the time to the most recent common ancestor as well as to all nodes in the tree was plotted. Each column represents the spectrum of median node times for a single leave-one-out replicate, with the tMRCA (time to the root of the tree) the top node in the spectrum. The leave-one-out replicates are sorted from left to right based on the age of the omitted sequence, with oldest sequence omitted on the leftmost side of the plot. (H and I) Randomization test of the data for (H) the L segment and (I) the S segment. We randomized the tip dates in 20 distinct ways and examined the clock rate in the resulting models. The error bars indicate the 95% HPD interval. (F–I) All runs were performed with the HKY model of nucleotide evolution, a strict clock, and a constant model of population growth. (A–I) For all parts of the figure, each run had a minimum estimated sample size (ESS) of at least 100 for all parameters.
Figure S4. Randomization Test within Geographically Grouped Subsets of Isolates and Root-to-Tip Linear Regression, Related to Figure 2
(A) We performed a randomization test on individual geographic taxa using L segments from the Batch 1 dataset (n = 95). We randomized the tip dates in 20 distinct ways and examined the clock rate in the resulting models. Molecular dating was performed on the full segment alignment with the HKY model of nucleotide evolution, a strict clock, and a constant model of population growth, and a minimum ESS of 100 for all parameters was obtained for each run. The error bars indicate the 95% HPD interval.

(B) Linear regression of root-to-tip distances versus date of isolate for the L segments in the Batch 1 dataset (n = 95). We performed a linear regression on the root-to-tip distances of samples versus the date of the isolate. We computed root-to-tip distances using the PATH-O-GEN program based on maximum-likelihood trees constructed using RAxML (Stamatakis et al., 2005).
Figure S5. Codon Adaptation, Nucleotide Content, and Mutation Spectra of LASV, Related to Figure 3

(A) CAI of LASV to human and M. natalensis are highly correlated. Normalized CAI was calculated for LASV to human and M. natalensis hosts. Each dot represents a single LASV sequence.

(B–F) LASV from outside Nigeria have higher CAI. (B and C) Gene trees for (B) NP and (C) GPC are shown separately (the Z gene was excluded because of its short sequence length). The CAI (to human codon usage) was calculated for each branch, according to the maximum-likelihood ancestral sequence reconstruction (PAML) and converted to a Z-score (by subtracting the mean CAI across branches and dividing by the SD). Because Z-scores can be less than one, the minimum Z-score in each tree was subtracted from the Z-score for each branch of the tree, such that the displayed branch have length zero or greater (scale bar, scaled (legend continued on next page)
Z-score). (D and E) Normalized CAI to (D) M. natalensis or (E) human of LASV sequences using the full gene-sets available for each host (** = p value < 0.0001, Mann-Whitney test). (F) The average CAI (to human codon usage) of all Sierra Leonean strains minus the average CAI of all the Nigerian strains plotted across the LASV genome (window size = 1).

(G and H) Dinucleotide frequencies in LASV strains from Nigeria and Sierra Leone. (G) Dinucleotide frequencies in the full-length open reading frames of LASV. (H) Dinucleotide frequencies of the first 699 bp of the open reading frames of LASV NPs from Nigeria and Sierra Leone.

(I) LASV genome copies decreases over the course of an infection. Relative abundance of LASV genome copies (log ratio of LASV copies/µl to 18S rRNA copies/µl) over the course of an infection was calculated using qPCR. Only a small subset of patients with multiple blood draws were included in these calculations. Each color correspond to an individual sample.

(J–L) Intrahost variation not likely due to systematic biases in variant-calling or sample degradation/damage. (J) The number of iSNVs called using two different methods is congruent (Batch 1 dataset; iSNVs were called at 5% minMAF and filtered as described in the Supplemental Experimental Procedures). (K) The number of LASV iSNVs called is not correlated with the number of human ‘iSNVs’, as would be expected if sample degradation inflated the observed intrahost mutation rate. Human ‘iSNVs’ were called by aligning all the sequencing reads back to the human genome. All LASV and human iSNVs were called at 5% minMAF with the same filters and settings using GATK. (L) Mutation spectra do not differ significantly within and between hosts. The legend shows exchanges (either iSNVs or fixed differences between consensus sequences) between each possible pair of bases. The mutation frequency distributions do not differ significantly between iSNVs and fixed (interhost) pairwise differences between consensus sequences (Pearson’s chi-square test, p value > 0.2), in both Sierra Leone (SL) or Nigeria (NIG). As expected due to the longer evolutionary history of LASV in Nigeria than Sierra Leone, there is a significant difference in the interhost mutation spectra between these two countries (Pearson’s chi-square test, p value < 2.2E-16), due to an excess of transversions in Nigerian strains relative to Sierra Leonean strains.
Figure S6. Confirmation of Illumina iSNV Calls, Related to Figure 5

(A and B) Independent cDNA library preps and Illumina sequencing runs were performed for two human samples containing high numbers of iSNVs and iSNV frequencies were called for each run. (A) Sample G733 and (B) Sample G1727.

(C) Unbiased Illumina sequencing and amplicon-based 454 sequencing were performed on *M. natalensis* sample Z0947. The iSNV frequencies were called and compared across the two sequencing technologies. No amplicons could be produced from the first ~1500 bp of the L segments so no 454 sequence data could be generated from this region. (A–C) The x axis numbers correspond to base pairs within the two segments. The height of each bar denotes the frequency of the iSNV (the maximum likelihood minor allele frequency estimate). Error bars span the 95% confidence interval of this estimate, assuming a binomial probability distribution.

(D and E) Amplicon-based (no cloning) Sanger sequencing was performed around iSNVs identified using Illumina data in (D) the L segment or (E) the S segment of sample G733. The identified iSNVs are represented in bold and minor curves corresponding to the identified minor allele variant calls can be seen in the individual chromatograms. In all cases, only iSNVs within the open reading frames are considered.
Figure S7. Additional iSNV Analyses, Related to Figures 5 and 7

(A–D) The number of iSNVs called reaches a plateau in most samples with increasing sequencing depth. Sequence data (Batch 1 dataset) was downsampled by randomly selecting subsets of the total sequence reads. The downsampled data were then used to call iSNVs. Each individual sequenced sample is shown in a different color, with (A and C) humans (minMAF = 5%) and (B and D) rodents (minMAF = 10%) plotted separately. (A and B) The fraction of iSNVs called (relative to the total called with full coverage) in (A) humans or (B) M. natalensis is plotted against the mean coverage per site (means across 20 replicate down samplings). Human samples often have > 1000x coverage; these were first downsampled to 500x, then further downsampled to fractions of 500x. (C and D) The same data, displaying the total mean number of iSNVs called in each downsampling and the fraction of total coverage from (C) humans or (D) M. natalensis (with total = 500x for humans).

(E) iSNV frequencies are relatively stable over the course of an infection and vary across sites within individual samples. iSNV frequencies (5% minMAF) were calculated in samples where more than one blood sample had been collected over the course of LASV infection. Each color correspond to an individual sample; each line correspond to an individual iSNV.

(legend continued on next page)
Limited evidence for multiple infections. iSNV frequencies were calculated in all samples with ≥ 10 iSNVs (5% minMAF) and plotted as the percent of total reads. Each circle corresponds to an iSNV within an individual sample.

(i) dN/dS ratios decrease with higher iSNV frequency. LASV and EBOV iSNVs were called at different minor allele frequencies and the dN/dS ratios were calculated. For iSNVs with minMAF < 5% only samples with independent replicate library preparations were considered and iSNVs had to be present at similar frequencies in all replicates to be called.

(h) Transmission scenarios between rodents Z0947 and Z0948. Three different transmission scenarios are illustrated: (1) outgroup to Z0948 and Z0947 independently, (2) outgroup to Z0948 to Z0947, and (3) outgroup to Z0947 to Z0948. Here, the outgroup is G1190 and G1727. Only sites at which G1190 and G1727 share the same fixed allele are included. Circles indicate hosts (Z0947, Z0948, or outgroup), with black and white indicating alternative alleles at iSNV sites or fixed differences between consensus sequences. Different observed patterns of iSNV/fixed differences (A–D) are shown. Patterns C and D are ambiguous because the Z0947 derived allele is sometimes the major and sometimes the minor allele, depending on the transmission scenario.

(i) Nonsynonymous iSNVs are overrepresented within predicted B cell epitopes in LASV GPC. B cell epitopes were predicted in LASV GPC genes from M. natalensis and humans hosts using FBCPred. The observed fraction is compared to the expected fraction, which is simply the fraction of amino acids in LASV GPC containing predicted epitopes (** = p value < 0.01, Binomial test).
Supplemental Information

Clinical Sequencing Uncovers Origins and Evolution of Lassa Virus

SUPPLEMENTAL INFORMATION

EXTENDED EXPERIMENTAL PROCEDURES

Ethics statement. Subjects were recruited for this study using protocols approved by human subjects committees at Tulane University, Harvard University, Broad Institute, Irrua Specialist Teaching Hospital (ISTH), Kenema Government Hospital (KGH), Oyo State Ministry of Health, Ibadan, Nigeria and Sierra Leone Ministry of Health. All patients were treated with a similar standard of care and were offered the drug Ribavirin, whether or not they decided to participate in the study. Treatment with Ribavirin followed the currently recommended guidelines (McCormick et al., 1986) and was generally offered as soon as Lassa fever (LF) was strongly suspected.

Sample collections, study subjects. Human samples were obtained from patients with LF; all samples were acquired on the day of admission before any treatment regimens had been started. Ten ml of whole blood was collected and plasma or serum was prepared by centrifugation at 2500 rpm for 15 minutes. Diagnostic tests for the presence of LASV were performed on-site using PCR (Olschlager et al., 2010) and/or ELISA antigen capture assays (Branco et al., 2011b). Both assays have comparable sensitivity (Branco et al., 2011a). All samples were re-tested by PCR upon receipt at Harvard University; in general ~60% of samples from both Nigeria and Sierra Leone that tested positive in the field also tested positive in the US. Only samples that tested positive for LASV in the field and by PCR at Harvard University were used in this study. Rodents (all from Sierra Leone) were trapped in case-households, humanely sacrificed and samples were collected from serum and/or spleen.

All samples were inactivated in either Buffer AVL (Qiagen) or TRIzol (Life Technologies) following the manufacturer’s standard protocols and stored in solar-driven −20° C freezers. In later samples, carrier RNA was omitted from Buffer AVL as we found that this significantly improved the efficiency of library construction (Matranga et al., 2014; Stremlau et al., 2015). Every three to six months inactivated samples were shipped on dry ice to Harvard University where samples were kept at −80° C until further processing.
Case-fatality rate calculations. Case-fatality rates (CFRs) were calculated from patients from Sierra Leone and Nigeria that had all of the following characteristics: (i) we knew the outcome of their illness, (ii) they tested positive for LASV in the field, (iii) their samples were confirmed positive by PCR upon retesting at Harvard University, and (iv) their samples were successfully sequenced at the Broad Institute. In Sierra Leone the CFRs were calculated from LASV positive patients that came to KGH from 2009 – 2013. In Nigeria they were calculated from LASV positive patients that came to ISTH from 2009 – 2012.

Ebola virus dataset. Consensus sequences and raw data for all the EBOV samples produced by our group from Sierra Leone are available via NCBI BioGroup PRJNA257197 (Gire et al., 2014). Sequences from Guinea (KJ660346, KJ660347, KJ660348), Liberia (KP178538), and Mali (KP260799, KP260800, KP260801, KP260802) were downloaded directly from NCBI. All accession numbers can be found in Table S2.

Viral RNA isolation and QC. AVL-inactivated RNA was isolated using the QIAamp Viral RNA Minikit (Qiagen) according to the manufacturer’s protocol, except that 0.1 M final concentration of β-mercaptoethanol was added to each sample to dissolve any agglutinated products (Matranga et al., 2014). TRIzol-inactivated RNA was isolated according to the manufacturer’s protocol with slight modifications. Briefly, 200 µl 1-bromo-2 chloropropane (BCP) was added for every 1 ml TRIzol used for inactivation. After phase separation, 20 µg of linear acrylamide was added to the aqueous portion. All extracted RNA was resuspended in water and treated with Turbo DNAse (Ambion) to ensure removal of contaminating DNA.

Quantification of LASV and EBOV genome load. We were unable to calculate traditional titers for LASV and EBOV due to their BL-4 status. Instead, LASV/EBOV RNA and host 18S rRNA were quantified using the Power SYBR Green RNA-to-Ct 1-Step qRT-PCR assay (Life Technologies). For LASV, primers were designed to produce short (~60-90 bp) amplicons within the LASV NP gene for both Sierra Leone-like strains (Table S1, primers tab, SL1, SL2) and Nigerian-like strains (Table S1, primers tab, NG1, NG2). Standard PCR amplicons encompassing qRT-PCR products were prepared to determine viral copy number in qRT-PCR assays. This was done by using synthetic oligonucleotides representing a portion of the LASV S
segment within the NP gene as a template for PCR (Table S1, primers tab, SL3-5, NG3-5). These amplicons were cleaned up using Agencourt AMPure XT beads (Beckman Coulter Genomics) and quantified by Quant-iT (Invitrogen). Amplicon concentrations were converted to LASV copies/µl for quantification purposes. Host RNA was quantified using human 18S rRNA primers using genomic DNA (Promega) as a standard control (Table S1, primers tab, 18S1, 18S2). For EBOV, we used data from the Gire et al. publication (Gire et al., 2014). All reactions were performed on the ABI 7900HT (Applied Biosystems). To control for differences in degradation, RNA extraction, and sample handling, the amount of LASV/EBOV material was normalized to that of host 18S rRNA material to give a final log ratio of viral copies/µl to 18S rRNA copies/µl.

Selective RNA depletion. Samples with ≥ 1 detectable copy of LASV RNA were used for selective depletion. Briefly, prior to library construction, poly(rA) carrier and host ribosomal RNA was depleted from LASV samples using an RNase H based selective depletion approach (Adiconis et al., 2013). Oligo(dT) and DNA probes complementary to human rRNA were hybridized to the sample RNA. The sample was then treated with RNase H (Hybridase, Epicentre). All synthetic DNA oligos were removed by treating with RNase-free DNase (Qiagen) according to the manufacturer’s protocol. Carrier- and rRNA-depleted samples were purified using 1.8x volumes AMPure RNA clean beads (Beckman Coulter Genomics) and eluted into water for cDNA synthesis.

Illumina library construction and sequencing. cDNA synthesis and Illumina paired-end library construction were done similarly to published RNA-Seq methods (Matranga et al., 2014; Levin et al., 2010). Briefly, controls were used to monitor our library construction process. We spiked in 1 pg of one, unique synthetic RNA (ERCC, gift from M. Salit, National Institute of Standards and Technology (Jiang et al., 2011)) using a different RNA for each individual LASV sample to aid in tracking our viral sequencing process and potential index cross-contamination. Also, we prepared libraries from 200 ng human K-562 total RNA (Ambion) with each batch as a LASV negative control. RNA samples were fragmented for 4 minutes at 85° C using NEBNext Fragmentation buffer (New England Biolabs). After fragmentation, samples were purified using 2.2x volume AMPure RNA clean beads (Beckman Coulter Genomics). The fragmentation step was removed for libraries prepared using Nextera XT tagmentation (Illumina, see below). Next,
double stranded cDNA was prepared using randomly-primed reverse transcription (first strand) and replacement synthesis (second strand). For adapter-ligated Illumina libraries (~50% of libraries) used up to 18 cycles of PCR to generate our libraries. For Nextera XT tagmentation libraries (remaining 50% of libraries), we used 16 cycles of PCR to generate our libraries. Finally, we used qPCR to calculate copies of LASV in the final libraries. Each individual sample was indexed with an 8 bp unique barcode (or dually-barcoded in Nextera XT libraries) and libraries were pooled equally and sequenced on the HiSeq2000 (101 bp paired-end reads; Illumina), the HiSeq2500 (101 or 150 bp paired-end reads; Illumina), or the MiSeq (150 bp paired-end reads; Illumina) platforms. All the data were deposited at NCBI under BioProject PRJNA254017.

Library preparation, sequencing, and assembly of EBOV samples have previously been described and performed (Gire et al., 2014; Matranga et al., 2014). Since EBOV samples experienced a much shorter time frame between collection and sequencing (Gire et al., 2014), we found that EBOV sequencing led to a higher percentage of viral reads (Gire et al., 2014; Matranga et al., 2014) than LASV sequencing (Figure S1A). However, the normalized genome abundance was very similar for EVD and LF patients from KGH (Figure 4A).

Demultiplexing of sequencing runs and QC. Raw sequencing reads were demultiplexed using the Picard v1.4 pipeline (Institute, 2012b) and saved as BAM files (Li et al., 2009). To avoid barcode cross-contamination between samples the default settings were changed to allow for no mismatches in the barcode and a minimum quality score of Q15 in the individual bases of the index. Sequencing quality metrics were collected using FastQC v0.10.0 (Institute, 2012a) and only high-quality sequencing libraries were used in subsequent analyses.

Assembly of *M. natalensis* transcriptome. Eleven sequencing runs from eight individual rodent samples were pooled. Low quality reads and bases were removed with Trimmomatic v0.15 (Lohse et al., 2012) using the following parameters: LEADING:20 TRAILING:20 SLIDINGWINDOW:4:25 MINLEN:70. BMTagger v3.101 (Agarwala and Rotmistrovsky) was then used to remove all ‘contaminating’ reads (primarily E. coli, spike-in RNAs and LASV) and duplicates were removed using PRINSEQ-lite v0.19.3 (Schmieder and Edwards, 2011). Paired-end de novo assembly of the cleaned reads was performed using Trinity r2011-11-26 with a minimum contig length of 1,000. Out of the resultant contigs > 99% were classified as being of
rodent origins using BLASTn v2.2.24+ (Figure S1D). Next, ORFs were predicted using EMBOSS getorf v6.2.0 (Rice et al., 2000) with a minimum size cutoff of 600. The ORFs were clustered using USEARCH v6.0.203 (Edgar, 2010) with an identity threshold of 0.95 and a representative sequence from each cluster was selected (the ‘centroid’). The centroids (n = 7,028) are available in File S1, and were used in all downstream analyses.

**Assembly of full-length LASV genomes.** BAM files were converted to Fastq format and LASV reads were extracted using Lastal r247 (CBRC, 2012) with a custom-made database containing full-length arenavirus genomes. The filtered reads were de novo assembled using Trinity r2011-11-26 with a minimum contig size of 300 (Grabherr et al., 2011). The assembly pipeline can be accessed on GitHub ([https://github.com/broadinstitute/viral-ngs](https://github.com/broadinstitute/viral-ngs)). If the sequenced sample contained too few reads for efficient de novo assembly (usually when average depth of coverage < 20x) a hybrid between de novo assembly with Trinity and a manual alignment-based approach implemented in Geneious was used instead (‘Assisted assembly’, Table S1). Once contigs had been generated, all sequencing reads from individual samples were aligned back to its own LASV consensus using Novoalign v2.08.02 (Novocraft, 2012) with the following stringent parameters -k -l 40 -g 40 -x 20 -t 100. Duplicates were removed using Picard v1.4 and BAM files were subjected to local realignment using GATK v2.1 (McKenna et al., 2010). If multiple sequencing runs had been performed for the same sample, BAM files were merged using Picard v1.4 before further analyses. Consensus sequences were called from the LASV-aligned reads using samtools v0.1.19 (Li et al., 2009) with the following parameters: samtools mpileup -Q 15 -uB -q 1 -d 10000 -f reference.fasta sample.bam. bcftools view -A -g vcf2fq > sample-consensus.fastq. All generated genomes were manually inspected, checked and corrected for accuracy, such as the presence of intact ORFs, using Geneious v6.1. Regions were depth of coverage was less < 3x were called as ‘N’. Samples that failed to generate high-quality consensus sequences were excluded from all further analyses. We found that consensus sequences for the individual LASV genomes were consistent across technical replicates, sequencing machines (MiSeq, HiSeq2000, and HiSeq2500), library preparation methods, and laboratories used for library generation (Harvard University or the Broad Institute).
**Multiple sequence alignments.** Consensus sequences were aligned using MAFFT v6.902b (Katoh et al., 2002) with the following parameters (L-INS-i): --localpair --maxiterate 1000 --reorder --ep 0.123 before being trimmed using trimAl v1.4 (Capella-Gutierrez et al., 2009) with the maximum likelihood specific parameter: -automated1. Typically this would remove ~ 1-2% of the positions in the alignments. Codon-based alignments - which were used for the majority of our analyses - were then generated using MAFFT as implemented in Geneious v6.1 (Drummond et al., 2010). Key alignments can be found in File S1.

**Recombination and reassortment analyses.** Alignments containing either all sequences, Nigerian sequences or Sierra Leonean sequences were screened for recombinant viral strains using the programs RDP, GENECONV, MAXCHI, CHIMAERA, 3SEQ, BOOTSCAN and SISCAN as implemented in the RDP3 software package (Martin et al., 2010). Defaults settings were used, except for the following changes: RDP, window size: 100 bp; BOOTSCAN, number of bootstrap replicates: 100, window size: 100 bp, step size: 20 bp, model options: Felsenstein, 1984; MAXCHI, variable window size was used, strip gap was selected; MAXCHI and CHIMAERA variable sites per window set to 50; SISCAN, window size: 100 bp with step size 20 bp. Potential recombinant sequences were identified when three or more methods were in agreement with $P$-values of $< 0.001$ (Bonferroni corrected). No recombinant sequences were identified in any of our screens as we found no evidence for phylogenetic incongruence in our datasets.

To detect reassortment between L and S segments two different approaches were used. Firstly, concatenated alignments of the L and S segments from Nigeria or Sierra Leone were scanned for breakpoints between the two segments using RDP3 as described above. A reassorted strain was identified when four or more methods identified the breakpoints with $P$-values of $< 0.001$ (Bonferroni corrected). Secondly, the reassortments were confirmed using the program GiRaF v1.02 with default settings (Nagarajan and Kingsford, 2011) and only samples for which GiRaF and RDP agreed on the identified reassortant were called.

**Phylogenetic tree reconstruction.** Maximum likelihood phylogenies were made with RAxML v7.3.0 (Stamatakis et al., 2005) using the GTR$\gamma$ nucleotide substitution model. Fifty consecutive runs were performed and the tree with the best likelihood score was bootstrapped with 500
pseudoreplicates. All trees were rooted using the 1969 Pinneo strain of LASV (Bowen et al., 2000). To confirm the topology of the trees, and to verify Pinneo as the correct root of the tree, we also created alignments and trees using Mopeia virus (GenBank ID: DQ328874.1) as an outgroup. In all cases did we find that the topology of the trees were in agreement (File S1).

**Estimation of human-to-human transmission.** Human LF patient samples from Sierra Leone collected in 2012 with detailed collection dates (day, month, year) were used for the transmission analysis of LASV. For EBOV, samples from Guinea, Sierra Leone, Liberia and Mali were used. All the included samples can be found in Table S2. Maximum likelihood phylogenetic trees were created for the LASV S segment (n = 21) and EBOV (n = 131). For LASV, a strain from Liberia (G1200) was initially included and the tree rooted on that strain; this sample was then removed from subsequent analyses, alignments recreated, and the tree was rooted on the nearest neighbor to G1200 (G2259). The EBOV tree was rooted on the three strains from Guinea (KJ660346, KJ660347, and KJ660348). Root-to-tip distances were calculated using the program PATH-O-GEN and plotted against the collection dates. $R^2$ and $P$-values were calculated based on linear regression.

**Molecular dating using BEAST.** Phylogenies incorporating time of sampling were estimated using Bayesian Markov Chain Monte Carlo (MCMC) as incorporated into the program BEAST v1.7.4 (Drummond and Rambaut, 2007). Several evolutionary models (HKY, GTR, SRD06), clock models (lognormal relaxed, exponential relaxed, strict), codon partitioning (no partitioning, 1,2,3 partitioning, [1+2],3 partitioning) and population sizes (constant, exponential, Bayesian skyline) (Drummond et al., 2005; Drummond et al., 2006) were tested, all with very similar results (Figure S3A, B and Table S2). Model likelihoods were estimated using an importance sampling estimator (Suchard et al., 2003) and Bayes factor analysis was used to select the best parameters (Li and Drummond, 2012). All these analyses were performed on the Batch 1 dataset (Figure 1C). For the final analyses - using the Matched dataset (Figure 1C) - a model incorporating a lognormal relaxed clock, exponential growth, HKYγ with four categories and codon partitioning (srdo6) was run for 100 million generations, sampled every 1000 generations using a 10% burn-in rate until all statistics had ESS values > 1,000 (File S1). Maximum-clade credibility trees summarizing all MCMC samples were generated using TreeAnnotator v1.7.4
with a burn-in rate of 10%. To examine the effect of positive selection on the molecular dating estimates, branch lengths were also estimated using the branch site random effects likelihood model in HyPhy, with three dN/dS classes fit for each branch (Kosakovsky Pond et al., 2011; Pond et al., 2005).

**Codon adaptation analysis.** The synonymous codon usage and sequence composition of each of the full-length LASV consensus sequences in the Matched dataset, or the EBOV sequences from the Gire *et al.* dataset (Gire *et al.*, 2014), were compared to that of human, *M. natalensis*, mouse, and chimpanzee. Each ORF identified in the *M. natalensis* transcriptome was queried against the human genome using BLASTn v2.2.24+ with default settings, and the best hit was then matched to orthologs in the other genomes using ENSEMBL (Birney *et al.*, 2004). In cases where multiple isoforms existed for an ENSEMBL ortholog in a given genome, only the longest isoform was retained. The codon sequences of genes with a 'complete set' of orthologs, represented in all four mammals, were then aligned with PRANK v10.08.02 with the -translate option and default settings (Loytynoja and Goldman, 2005) and all gaps were trimmed, such that every orthologous codon was represented in each mammalian sequence. Codon frequencies and codon usage tables for each mammal were calculated from the resulting alignment of 2,489 orthologs (692,522 codons). The codon adaptation index (CAI) of each full-length viral genome (coding sequences only) was computed separately relative to each host codon usage table using CAIcal v1.4 (Puigbo *et al.*, 2008a). CAI was normalized by the 'expected neutral CAI,' based on 500 randomized viral sequences, conserving the observed GC content and amino acid usage (Puigbo *et al.*, 2008b). The results are interpreted such that a value above ‘1’ is higher than neutral and for all LASV strains analyzed these values were statistically significant.

To account for potential biases in selecting an ortholog dataset based on *M. natalensis* samples sequenced from Sierra Leone, we also calculated CAI of the individual LASV sequences using full gene-sets from humans (total ORFs ~19,000) and *M. natalensis* (assembled ORFs = 7,028). These results were in full agreement, and in both cases we observed that Sierra Leone LASV strains had significantly higher CAIs (Figure S5D, E).

For the analyses in Figures 3C, D and S5B, C we calculated the CAI (to human codon usage) for each branch, according to the maximum-likelihood ancestral sequence reconstruction (PAML v4.7) and converted it to a Z-score (by subtracting the mean CAI across branches and
dividing by the standard deviation). Z-scores can be less than one, so the minimum Z-score in each tree was subtracted from the Z-score for each branch of the tree, such that the displayed branch have length zero or greater. These analyses were performed using the Batch 1 dataset, and both trees were rooted on Pinneo (not shown).

A Kolmogorov-Smirnov test comparing Nigerian and non-Nigerian branches shows a significant difference in CAI ($D = 0.48$, $P$-value = 0.0005), but the non-independence among branches violates the assumptions of the test. For this reason we used empirical $P$-values, based on permutations accounting for the tree structure. The significance of the difference in CAI between Nigerian and non-Nigerian sequences was tested by permuting the reconstructed sequence changes across the maximum-likelihood tree topology and testing whether the permuted sequences had significantly different distributions of CAI between Nigerian and non-Nigerian branches (Kolmogorov-Smirnov test, $P$-value < 0.05). The permutation was repeated 1000 times to obtain an empirical $P$-value equal to the fraction of permutations showing a significant difference between Nigerian and non-Nigerian CAI.

Intra-host variant calling. For each sequenced sample, reads were realigned to the consensus sequence using Novoalign as described above and iSNVs were called using mpileup with the following parameters: samtools mpileup -Q 0 -B -q 1 -d 10000 and VarScan v2.3 (Koboldt et al., 2012) with the following parameters: varscan.jar pileup2snp --min-reads2 5 --min-var-freq 0.01 -p-value 0.1 --min-coverage 5 --min-avg-qual 5. Stringent post-call filtering were applied to remove false calls using the following variables:

- Minimum overall coverage of 5x.
- Minimum depth for variant calls at each position of 5.
- Minimum iSNV frequency of 5%
- Maximum strand-bias difference of 10-fold between variant and reference alleles
- Minimum base-quality call of Q25
- Maximum base-quality difference between variant and consensus call of Q5.

For comparison analyses (Figure S5J) GATK v2.1 was used as well with the following parameters: GATK calling: -baq OFF --useOriginalQualities -dt NONE --
We experimented with different minMAFs and found that the 5% cutoff used in this study gave reproducible results across platforms and replicates (Figures S5J and S6). The filtered iSNV calls (5% minMAF) can be found in File S1. In contrast, we found that calling iSNVs at 1% or below — even given very high coverage — required multiple technical replicates and independent library preparations (Gire et al., 2014). This is likely due to the random incorporation of RT-PCR and PCR errors during library construction.

Validation of intra-host variants using 454 and Sanger sequencing. Two samples with high numbers of iSNVs were selected for variant validation using 454 sequencing (rodent Z0947) or Sanger sequencing (patient G733). For 454 sequencing the entire LASV genome was amplified using four sets of ‘DemiHemi’ primers (Table S1, primers tab, L1, L2, L3a, Sa) and sequenced on the 454 platform using previously described protocols (Lennon et al., 2010; Henn et al., 2012). For Sanger sequencing a panel of 17 M13-tagged primer sets (Table S1, primers tab) was created and amplicons were generated and sequenced using standard Sanger sequencing protocols (GeneWiz).

Comparison of iSNV rates. In order to compare levels of polymorphism between samples sequenced with varying coverage, we computed an iSNV rate for each sample: the number of iSNVs called at sites with sufficient coverage (N) divided by the total number of sites with sufficient coverage (N). The probability of calling an iSNV (power) was modeled with a binomial distribution to calculate the minimum sufficient coverage (N) at a site to call an iSNV with a given minor allele frequency (MAF), with the minor allele present on at least 5 reads. At 90% iSNV calling power, N ≥ 926 is sufficient for MAF ≥ 1%, N ≥ 184 for MAF ≥ 5%, N ≥ 91 for MAF ≥ 10%, N ≥ 70 for MAF ≥ 12.5%. Specifically, in Figure 5C we calculated the iSNV rate using only sites with sufficient coverage to call iSNVs with 80% calling power in each
minor allele frequency (MAF) bin and tested the difference between human and *M. natalensis* separately within each MAF bin.

**Rarefaction analysis.** All LASV reads were extracted from aligned BAM files with duplicates removed and multiple runs from the same sample were combined. Depending on the average coverage the various samples were then downsampled to an average of 500x, 200x, 100x or 50x for 20 replicates. Each replicate was then subsampled in fractions of 0.1 (0.1 to 1.0) and the subsampled reads were aligned to their own LASV consensus sequences using Novoalign v2.08.02. Intra-host variants were then called with VarScan v2.3 as described above.

**Assessment of transition/transversion ratios.** Higher transition/transversion ratio does not explain the high CAI in Sierra Leone. The Sierra Leonean LASV population is relatively young (Figure 2G), and has a significantly different mutation spectrum, enriched in transitions relative to Nigerian strains (Figure S5L). However, changes in synonymous codon preference between Sierra Leone and Nigeria are not enriched in transition mutations, as might be predicted if the change in CAI had been driven by mutational biases. For each four-fold degenerate amino acid, we computed an odds ratio statistic to assess the relative preference for each synonymous codon between Sierra Leone and Nigeria. We ranked the absolute value of the odds ratios for each of the 5 four-fold degenerate amino acids. If synonymous changes involving transitions were consistently highly ranked, this would suggest that they are responsible for the differences in codon preference between Sierra Leone and Nigeria. However, we observed that transitions (A/G or C/T) had the top-ranked odds ratio for only 1 out of 5 amino acids (expected (2/6) * 5 = 1.67) on the S segment and 3 out of 5 on the L segment. This suggests that a transition bias is not entirely responsible for the difference in codon preference between Sierra Leone and Nigeria. We therefore cannot exclude the role of selection in shaping CAI; however more work will be needed to determine whether high or low CAI is more adaptive for LASV.

**Intra-host selection analysis.** To investigate signatures of natural selection within infected hosts, we examined the patterns of LASV iSNV substitutions observed in patients and *M. natalensis* using the McDonald-Kreitman (MK) framework. This test has classically been used to test for selection at the protein level by calculating the ratio of the number of nonsynonymous
and synonymous substitutions (N/S) within species versus between species (McDonald and Kreitman, 1991). More recently, it has been extended to distinguish selective pressures within versus between populations of viruses (Renzette et al., 2011; Bhatt et al., 2010). Using this framework, we compared variation in LASV within and between hosts, using a measure called the Neutrality Index (NI) (Rand and Kann, 1996), defined as the N/S ratio between iSNVs divided by N/S between consensus sequences found in different hosts. Under neutral evolution, these ratios should be equal, yielding NI = 1. A significantly high NI is generally interpreted as evidence for purifying (negative) selection between hosts; however it can also indicate diversifying (positive) selection within hosts. NI is the reciprocal of the more commonly used Fixation Index (McDonald and Kreitman, 1991). Because most of our discussion focuses on evolution within hosts, NI is the more intuitive statistic to use here.

The MK framework has been previously applied to quantify selective pressures in viruses (Bhatt et al., 2010), but care must be taken to accurately estimate between-host N and S in rapidly evolving viral populations, where multiple substitutions may occur at the same site. Therefore, we only considered substitutions on leaf branches, which generally contain only a few substitutions, unlike long internal branches, which are often saturated with multiple synonymous substitutions. Between-host variants were inferred as changes between a consensus sequence (leaf branch) and all other sequences in the same phylogenetic cluster. Phylogenetic clusters were defined as monophyletic groups of the maximum likelihood phylogeny with a maximum of 0.01 substitutions/site from the root of the group to the tips. Only branches containing iSNVs were included in the between-host counts, although including all leaf branches did not significantly change the between-host N/S estimate (data not shown). The significance of the deviation of NI from the neutral expectation was evaluated by permuting the 2x2 contingency tables across genes and samples, keeping the row and column marginals constant, as previously described (Shapiro et al., 2007). The P-value of the observed NI was calculated as the fraction of 10,000 permutations that yielded an NI greater or equal to the observed NI (Figure 5D and E). The expected NI was defined as the mean permuted NI, and the normalized NI defined as observed divided by expected NI. In Figures 5, 6 and 7 we used a minMAF of 5%. In Figure 5, the N and S counts are normalized by the number of nonsynonymous or synonymous sites to obtain dN and dS, respectively, and calculate the dN/dS ratio.
**Epitope prediction.** B cell epitopes were predicted in protein-coding regions of LASV consensus sequence using BCPREDS with 75% specificity and an epitope length of 20 amino acids (El-Manzalawy et al., 2008b). We also predicted B cell epitopes in LASV GPC using FBCPred (EL-Manzalawy et al., 2008a), which allows for flexible length epitope prediction and obtained very similar results (Figure S7I). Comparing the predicted B cell epitopes to a smaller subset of experimentally determined epitopes (GPC only) available in the ViPR database (as of April 6th, 2015, http://www.viprbrc.org/), we found that five out of eight of the experimental B cell epitopes had overlap with the predicted epitopes. Epitopes were predicted separately on each consensus sequence, and overlaid with the iSNVs from the matched sample. The overlap between iSNVs and epitopes in each protein was assessed using a binomial test, with binomial probability $p$ equal to the fraction of amino acids in the protein covered by predicted epitopes. The epitope predictions were highly consistent across different consensus sequences, therefore a single average $p$ was used for each protein. The iSNVs falling within or outside predicted epitopes were summed across consensus sequences. This assumes independence of iSNVs, which is a reasonable assumption given that there is little evidence for strong linkage between adjacent iSNVs (Table S2). Moreover, the vast majority of samples contain only 1 iSNV; therefore any linkage effects are expected to be minimal.

T cell epitope predictions were performed using NetCTL (Larsen et al., 2007) as implemented on the ViPR website (http://www.viprbrc.org/). A cutoff score of 1.5 was used with Josiah as the reference and epitopes were calculated using all available MHC class I supertypes.

The predicted B cell and T cell epitopes can be found in File S1.

**iSNV fixation analysis.** Fixation of iSNVs was considered separately, first, across all sequenced samples, and second, between a pair of rodents captured on the same day from the same household, for which recent transmission was suspected (Z0947, Z0948). Their consensus sequences were nearest-neighbors on the inferred LASV phylogeny (Files S1-3), suggesting a recent transmission event. We further suspected that transmission occurred from Z0947 to Z0948 based on two observations. First, Z0947 contained a substantially more diverse LASV population (27 iSNVs at 10% MAF) compared to Z0948 (3 iSNVs at 10% MAF). This suggests a longer infection period in Z0947, and a more recent transmission to Z0948. Second, all 3 iSNVs in Z0948 are nonsynonymous, resulting in a high N/S ratio (compared to an N/S ratio of 8/19 =
0.42 in Z0947). This suggests that slightly deleterious nonsynonymous variants are circulating in Z0948, consistent with a recent population bottleneck (expected during transmission) and/or a relatively recently established viral population in Z0948. While these data support a "Z0947 to Z0948" transmission scenario, we cannot formally exclude other scenarios (Figure S7H). To ensure that iSNVs in Z0947 were not also present at low frequencies in Z0948 (and vice versa), we manually inspected the aligned sequence reads. Therefore, either the major or minor allele at each Z0947 iSNV must necessarily be fixed (within the resolution of our sequencing depth) in the Z0948 consensus sequence.

Across all sequenced samples, an iSNV was defined to be 'fixed' if the minor allelic variant was observed in one or more consensus sequences in the entire dataset. The same analysis was performed, focusing on iSNVs in Z0947, either with Z0948 as the only outgroup, or with Z0948 plus two additional, more distant outgroups (G1190 and G1727), in order to account for different possible transmission scenarios (Figure S7H). Specifically, we considered only polymorphic sites (iSNVs or fixed differences between consensus sequences) for which minor/major Z0947 iSNVs could be unambiguously assigned as derived/ancestral with reference to additional outgroups (G1190 and G1727). This resulted in the exclusion of 3 ambiguous sites.

The assumption that iSNVs can be counted independently is justified based on the general lack of linkage between adjacent iSNVs on the same sequenced read (Table S2). Nevertheless, as a precaution to guard against undetected linkage effects in the data, we repeated the analysis of minor iSNV alleles, only counting a maximum of one nonsynonymous (N) or synonymous (S) iSNV (chosen at random) per sample per segment, from Batch 1 sequences. In this way, each potentially linked haplotype is only counted once, yielding the following counts: N unfixed = 10, N fixed = 18, S unfixed = 0, S fixed = 40 (Fisher’s exact test: Odds ratio > 22, P-value = 4.5E-5).

**Linkage between iSNVs.** In a few cases, nearby iSNV sites were sequenced on the same sequencing read (or read pair), allowing us to assess linkage between them. We gathered all such iSNV pairs, requiring that each iSNV have a quality score ≥ 20 and minMAF of 10%. We then counted major-major, major-minor, minor-major and minor-minor allele combinations (haplotypes) and computed the $D_A^*$ linkage measure, which ranges from 0 (unlinked) to 1 (perfectly linked), taking allele frequencies into account (Hedrick, 1987; Kalinowski and
On average, linkage was relatively weak (mean $D_A' \approx 0.4$; Table S2) but the number of reads spanning two iSNVs was generally also quite small; therefore we hesitate to draw firm conclusions about the extent of linkage in our dataset.

Luciferase constructs containing GBlocks. We designed 750 bp GBlocks (Integrated DNA Technologies) containing 699 bp of the 5’ end of the LASV NP gene, each for 10 Nigerian strains and 10 Sierra Leonean strains, along with 21-25 bps of flanking sequence at either end homologous to the m6pgfp vector (Andersen et al., 2008). As positive controls, GBlocks were ordered with human codon-optimized versions of LASV NP based on ISTH0073 (Nigerian) and G1442 (Sierra Leonean). Sequences were codon-optimized using the IDT Codon Optimization Tool (Integrated DNA Technologies) and the optimized sequences were between 70-75% identical to the parent sequences at the nucleotide level (100% at the amino acid level). GBlocks were assembled into PCR linearized m6pgfp vector using Gibson Assembly (New England Biolabs) to create an NP-firefly Luciferase (fLuc) fusion driven by the MMLV LTR. A single sequence-verified assembled clone was propagated overnight and plasmid DNA extracted using the Plasmid Plus Midi kit (Qiagen). These plasmids were designated ‘pKGAC-fLuc.’

To generate pKGAC-gLuc series constructs, each pKGAC-fLuc construct was double digested with EcoRI and NotI (New England Biolabs), heat inactivated, and treated with Antarctic phosphatase (New England Biolabs). Vector backbones were subsequently gel purified with the QIAquick Gel Extraction Kit (Qiagen). A DNA insert containing Gaussia-Dura luciferase (henceforth referred to as 'gLuc') was generated by PCR (95°C 2 minutes; 95°C 20 seconds, 55°C 15 seconds, 70°C 12 seconds, 35 cycles; 70°C 2 minutes) of pTK-Gaussia-Dura Luc (Thermo Scientific) with the primers gLuc_FW and gLuc_RV (Table S1, primers tab), which contain regions of homology to the template as well as 5' overhangs generating EcoRI and NotI restriction sites. The PCR product was run on an agarose gel and purified with the QIAquick Gel Extraction Kit. Subsequently, the product was double digested with EcoRI and NotI to expose sticky ends, and purified with QIAquick Spin Columns (Qiagen). The gLuc insert was ligated in-frame to each vector backbone with T4 DNA ligase (New England Biolabs), and subsequent ligation products were used to transform chemically competent NEB 10β E. coli (New England Biolabs). Transformants were plated, and individual colonies were picked and
sequenced to verify correct, in-frame insertion of the gLuc insert. These plasmids were designated ‘pKGAC-gLuc.’

**fLuc Transfection, luminescence, and RT-qPCR.** 100 µl of HEK293 cells (Life Technologies) were seeded into three opaque white (for detecting luminescence) 96-well plates, grown to 60-80% confluence, and transfected with Lipofectamine LTX & PLUS Reagent using the manufacturer's guidelines (Life Technologies). Briefly, each pKGAC-fLuc plasmid was incubated for 5 minutes at room temperature with pGL4.74-hRluc (Promega), which encodes the Renilla luciferase (rLuc) as an internal control, and PLUS Reagent. Next, Lipofectamine LTX Reagent was added, and the mixture was incubated for another 30 minutes at room temperature. The mixture was aliquoted into cells in the three 96-well plates. Each replicate contained 101.5 ng pKGAC-fLuc DNA, 10 ng pGL4.74-hRluc, 0.1 µl PLUS Reagent, and 0.4 µl Lipofectamine LTX Reagent, administered to the cell media in a 20 µl mixture.

To detect NP-fLuc fusion protein expression, a luciferase assay was performed using the Dual-Glo Luciferase Assay System (Promega). 16-20 hours after transfection, 50 µl of media was removed from each well of the 3 opaque white 96-well plates. 50 µl Dual-Glo Luciferase Reagent with substrate was added directly to the culture media of each well, mixed vigorously to ensure full cell lysis, and incubated for 10 minutes at room temperature. Firefly luminescence was detected on a Spectramax L with 5 seconds integration and 470 nm calibration wave. Next, 50 µl Dual-Glo Stop & Glo Reagent with substrate (Promega) was added to quench firefly luminescence and initiate renilla luminescence. Following 10 minutes incubation at room temperature, renilla luminescence was detected on a Spectramax L with 5 seconds integration and 470 nm calibration wave. In data analysis, firefly luminescence was normalized against renilla luminescence (fLuc/rLuc) for each sample. The three technical replicate plates were then averaged.

The entire experiment was performed in biological triplicate to produce the final data. For each of the 3 biological replicates, the fLuc/rLuc of each sample was then normalized against the average fLuc/rLuc of all non-control samples on that plate.

**In vitro transcription.** pKGAC-gLuc series plasmids were used to generate NP₁₋₆₉₉-gLuc RNA for *in vitro* translation assays. First, DNA amplicons containing the T7 RNA polymerase
promoter (T7Rpo) sequence upstream of NP$_{1-699}$-gLuc were generated by PCR (94° C 2 minutes; 94° C 15 seconds, 40° C 15 seconds, 72° C 25 seconds, 5 cycles; 94° C 15 seconds, 55° C 15 seconds, 72° C 25 seconds, 30 cycles; 70° C 1 minute) of pKGAC-gLuc plasmids with the primers pKGAC-gLuc_FW and pKGAC-gLuc_RV (Table S1, primers tab), which contain regions of homology to the template as well as a 5' overhang to generate T7Rpo in the sense direction. PCR products were gel purified with the QIAquick Gel Extraction Kit (Qiagen).

In addition, to investigate the translation of LASV GPC, we designed 1,325 bp GBlocks (Integrated DNA Technologies) containing the T7Rpo sequence, 736 bp of the 5' end of the LASV GPC gene, each for 5 Nigerian strains and 5 Sierra Leonean strains, and the sequence encoding gLuc. DNA amplicons were generated by PCR (94° C 2 minutes; 94° C 15 seconds, 58° C 15 seconds, 72° C 30 seconds, 40 cycles; 72° C 1 minute) with the primers pKGAC-gLuc_FW and pKGAC-gLuc_RV (Table S1, primers tab), and gel purified.

Next, DNA amplicons were used as templates for in vitro transcription by T7 RNA polymerase using a reduced version of the MEGAscript T7 Transcription Kit (Life Technologies) per the manufacturer's protocol. Briefly, each 20 µl reaction contained ~250-400 ng DNA template, 1X reaction buffer, 50 nmol of each NTP, and 2/3 µl of enzyme mix, and transcription proceeded at 37° C for 4 hours. To remove template DNA, samples were treated with 2 U TURBO DNase (Life Technologies) at 37° C for 20-30 minutes. For each of the 20 NP and 10 GPC DNA amplicons, in vitro transcription was performed in biological triplicate, resulting in 60 NP and 30 GPC RNA samples, in addition to codon optimized/de-optimized controls. Each in vitro transcription product was purified with Agencourt RNAClean XP beads (Beckman Coulter).

In vitro translation and gLuc detection. Purified in vitro transcription products were used as templates for in vitro translation using a reduced version of the Retic Lysate IVT Kit (Life Technologies). Briefly, we set up a modified reaction per the manufacturer's protocol except using 4.25 µl of reticulocyte lysate per 25 µl reaction (25% of the recommended lysate amount). 2.5 µg of each in vitro transcribed RNA was subjected to in vitro translation at 30° C for 21 hours. Translation was also performed in biological triplicate. Immediately following translation, each sample was incubated with 25 µl of coelenterazine substrate from the Pierce Gaussia Luciferase Glow Assay Kit (Thermo Scientific) at room temperature for 10 minutes.
Luminescence was detected on a Spectramax L with 0.5 seconds integration and 395 nm calibration wave.

**LASV GPC cloning and expression.** LASV Josiah GPC was used as a control. Several LASV GPC iSNV mutants based on the LM395 and LM776 *M. natalensis* samples were created as GBlocks (Integrated DNA Technologies) and cloned into the pcDNA3.1+zeo_intA vector for high level expression in mammalian cells (Life Technologies). The following mutants were made: LM395\(^{WT}\), LM395\(^{D89N}\), LM395\(^{L113I}\), LM395\(^{N114D}\), LM395\(^{R161M}\), LM395\(^{R248K}\), LM776\(^{WT}\), and LM776\(^{E67Q}\). All iSNV mutants were located in the GP1 part of LASV GPC and all LM395 mutants overlapped with predicted or experimental B cell epitopes. The cloning strategy was identical to that used for generation of the LASV Josiah GPC construct and clones were verified by DNA Sanger sequencing.

Endotoxin-free plasmid DNAs were generated from transformed E. coli TOP10 clones, quantitated and used in transient transfection studies. HEK293 cells were seeded the day before transfection in Poly-D-Lysine treated 6-well polystyrene plates in DMEM, 10% qFBS, 2mM L-Gln. On the day of transfection, fresh medium was exchanged and cells were incubated with transfection mix contain pre-optimized ratios of DNA and PEI (PolyPlus), according to manufacturer’s recommendations. Twenty-four hours after transfection cells were gently washed twice with PBS and incubated with fresh PBS for 15 minutes at room temperature to allow for gentle dislodging without mechanical disruption or enzymatic detachment. Cells were centrifuged for 5 minutes at 250xg, washed twice in FACS buffer (1X PBS, pH7.4, 2% FBS, 0.05% NaN3), and used in binding experiments. Cells (10\(^5\)/assay) were incubated in polypropylene 96-well U-bottom plates with prediluted LASV human monoclonal or control antibodies in FACS buffer, for 20 minutes on ice. Cells were washed twice by centrifugation in FACS buffer and incubated with diluted Goat anti-human IgG(H+L)-Alexa488 secondary reagent in FACS buffer for an additional 20 minutes on ice. Following 2 additional centrifugation washes, cells were resuspended in PI buffer (FACS buffer with 1 \(\mu\)g/mL propidium iodide). Ten thousand live cell events (PI negative) per samples were collected and analyzed in a BD Accuri C6 cytomter, and mean channel fluorescence values were derived.

To verify expression of LASV GPC constructs in HEK293 cells, extent of GP1 and GP2 cleavage, and sGP1 secretion western blots were performed on cell extracts and supernatants.
Briefly, cell extracts were prepared with the Sigma Mammalian Cell Lysis Kit (Sigma), and an equivalent input of protein from each extract was resolved on Bis-Tris 4-20% SDS-PAGE gels (Life Technologies), transferred to nitrocellulose membranes, and probed with LASV glycoprotein-specific murine antibodies raised against irradiated LASV antigen. Blots were probed with a Goat F(ab’)2 anti-mouse IgG(H+L)-HRP secondary reagent and LumiGlo chemiluminescent substrate. All images were captured on a GE Image Quant LAS4000 gel docking station. Similarly, equal volumes of centrifugation cleared supernatants from each transfection were resolved by SDS-PAGE and probed in western blots with the same reagents.

**Human LASV-specific monoclonal antibodies.** The human monoclonal antibodies (12.1F, 19.7E, 25.6A, 36.1F, 37.7H, and 37.2D) used in these studies were derived from PBMC of LF convalescent patients. Briefly, PBMC were isolated from buffy coats obtained from Ficoll-Isopaque centrifugation gradients, prepared at KGH. Cells were cryopreserved by slow cooling to -80°C, and shipped to Tulane University in IATA-approved dry shippers. PBMC were subsequently thawed, enriched for Pan B cells and seeded at nearly clonal densities in flat bottom 96 well plates, under activation conditions that specifically induced B cell proliferation. All emerging clones were screened for the presence of human antibodies specific for the LASV Josiah GPC, by ELISA and direct neutralization assays using a LASV GPC pseudotyped lentiviral particle system.

Antibodies with demonstrated binding and/or neutralization were cloned from corresponding cells by PCR amplification of light and heavy chain cDNAs using universal human IgG oligonucleotides and high fidelity DNA polymerases. The heavy and light chains were cloned in expression vectors, verified by DNA sequencing, and expressed in transiently transfected HEK293 cells. Recombinant antibodies that exhibited binding and neutralization to LASV GPC expressed on the surface of HEK293 cells were subsequently cloned in dual expression mammalian vectors for generation of stable NS0 (null secreting) murine myeloma cell lines (CHOLCelect). These studies were performed with NS0 cell produced antibodies. All antibodies were purified from chemically defined serum free stable NS0 cultures by Protein A chromatography, dialyzed in a proprietary buffer for injection, and diluted in FACS buffer for cytometry assays.
Limited mapping studies have been performed and it is believed that 19.7E binds an epitope in LASV GP1 around position 114 and 12.1F is in the same complementation group as 19.7E, but off to one side where it binds a different epitope. All other antibodies are believed to bind epitopes that lie within GP2 but may cover quaternary epitopes that also require sites in GP1.
SUPPLEMENTAL REFERENCES


CBRC (2012). Lastal available online.


Novocraft (2012). Novoalign available online.


Institute, T. B. (2012a). FastQC available online.

Institute, T. B. (2012b). Picard available online.
