Both Epistasis and Diversifying Selection Drive the Structural Evolution of the Ebola Virus Glycoprotein Mucin-Like Domain

Neke Ibeh,^a Jean Claude Nshogozabahizi,^a Stéphane Aris-Brosou^{a,b}

Department of Biology, Faculty of Science, University of Ottawa, Ottawa, Ontario, Canada^a; Department of Mathematics and Statistics, Faculty of Science, University of Ottawa, Ottawa, Ontario, Canada^b

ABSTRACT

Throughout the last 3 decades, Ebola virus (EBOV) outbreaks have been confined to isolated areas within Central Africa; however, the 2014 variant reached unprecedented transmission and mortality rates. While the outbreak was still under way, it was reported that the variant leading up to this outbreak evolved faster than previous EBOV variants, but evidence for diversifying selection was undetermined. Here, we test this selection hypothesis and show that while previous EBOV outbreaks were preceded by bursts of diversification, evidence for site-specific diversifying selection during the emergence of the 2014 EBOV clade is weak. However, we show strong evidence supporting an interplay between selection and correlated evolution (epistasis), particularly in the mucin-like domain (MLD) of the EBOV glycoprotein. By reconstructing ancestral structures of the MLD, we further propose a structural mechanism explaining how the substitutions that accumulated between 1918 and 1969 distorted the MLD, while more recent epistatic substitutions restored part of the structure, with the most recent substitution being adaptive. We suggest that it is this complex interplay between weak selection, epistasis, and structural constraints that has shaped the evolution of the 2014 EBOV variant.

IMPORTANCE

The role that selection plays in the emergence of viral epidemics remains debated, particularly in the context of the 2014 EBOV outbreak. Most critically, should such evidence exist, it is generally unclear how this relates to function and increased virulence. Here, we show that the viral lineage leading up to the 2014 outbreak underwent a complex interplay between selection and correlated evolution (epistasis) in a protein region that is critical for immune evasion. We then reconstructed the three-dimensional structure of this domain and showed that the initial mutations in this lineage deformed the structure, while subsequent mutations restored part of the structure. Along this mutational path, the first and last mutations were adaptive, while the intervening ones were epistatic. Altogether, we provide a mechanistic model that explains how selection and epistasis acted on the structural constraints that materialized during the 2014 EBOV outbreak.

The five viruses that constitute the genus *Ebolavirus* (*Zaire ebolavirus* [EBOV], *Sudan ebolavirus*, *Bundibugyo ebolavirus*, *Reston ebolavirus*, and *Tai Forest ebolavirus*) have been the cause of a major public health concern in sub-Saharan Africa for over 3 decades (1). Historically, outbreaks have been confined to isolated areas within Central Africa; however, the 2014 outbreak reached an unprecedented level, making this the largest outbreak since the discovery of the virus in 1976 (2).

The primary analysis of EBOV isolates conducted by Gire and colleagues found a large number of nonsynonymous mutations between the 2014 EBOV sequences and all previously published EBOV sequences (3). In particular, 50 fixed nonsynonymous changes were observed, all of which were distinct to the 2014 EBOV variant. While these fixed mutations spanned all seven viral proteins (NP, VP35, VP40, GP, VP30, VP24, and L Polymerase), both epidemiological (4) and molecular (5) evidence suggest that these mutations did not affect the virulence or the transmissibility of the 2014 EBOV strain. Nonetheless, the authors of the primary study fell short of testing for the role of selection in the emergence of this outbreak.

A subsequent study further investigated the rate of nonsynonymous substitutions in the 2014 EBOV strain, focusing on the glycoprotein (GP) (6). They discovered a strikingly high ratio of nonsynonymous-to-synonymous mutations concentrated in the heavily glycosylated mucin-like domain (MLD) of GP, a region

that spans residues 308 to 501 (7, 8). While it has been proposed that the MLD plays a key role in immune evasion (9), the flexible structure and dispensability of this domain (10) might explain its propensity to maintain high substitution rates throughout the outbreak. Alternatively, as GP plays a critical role in host cell attachment and membrane fusion (6), selective pressure exerted by the host's immune system may contribute to the high incidence of nonsynonymous mutations in this protein. Park and colleagues found a significantly higher rate of nonsynonymous mutations in experimentally validated GP epitopes than expected by chance, supporting the selection hypothesis (6).

While selection favors the strains (or phenotypes) best adapted to the changing host environment, epistatic interactions among

Received 19 February 2016 Accepted 18 March 2016

Accepted manuscript posted online 23 March 2016

Citation Ibeh N, Nshogozabahizi JC, Aris-Brosou S. 2016. Both epistasis and diversifying selection drive the structural evolution of the Ebola virus glycoprotein mucin-like domain. *J Virol* 90:5475–5484. doi:10.1128/JVI.00322-16.

Editor: A. García-Sastre, Icahn School of Medicine at Mount Sinai

Address correspondence to Stéphane Aris-Brosou, sarisbro@uottawa.ca.

Supplemental material for this article may be found at <http://dx.doi.org/10.1128/JVI.00322-16>.

Copyright © 2016, American Society for Microbiology. All Rights Reserved.

these mutations are known, at least from a theoretical point of view, to increase the number and magnitude of diversification events (11). Epistasis within and among genes is defined as the effect of one mutation being dependent on (i.e., correlated to) the presence or absence of one or more different mutations (12). The existence of such nonadditive interactions within (and among) genes can be experimentally validated in a variety of viral genomes (13). However, experimental validation is time-intensive and cumbersome, and it does not always capture the complex patterns of epistatic interactions (12). Some of these complex patterns are manifested structurally through the stabilizing interactions among mutations. The presence of permissive (that expand the margin of stability) and compensatory (that restore the stability of a compromised structure) mutations dictates the physical conformation of proteins and may therefore offer critical insights into the evolutionary trajectory of the Ebola virus (14).

Here, for the very first time, we aim to disentangle how episodic diversifying selection and epistasis have shaped the evolution of the 2014 EBOV strain. Using MEME (mixed-effects model of evolution) (15) as a tool to test for site-specific evidence for selection, we identify in the EBOV genome codons that experienced diversifying selection in restricted lineages. Using AEGIS (analysis of epistasis and genomic interacting sites) (J. C. Nshogozababizi, J. Dench, and S. Aris-Brosou, submitted for publication), we identify sites that evolved under epistasis. We show that these two categories of sites overlap along the lineage leading up to the 2014 sequences in the MLD. By means of ancestral reconstructions of the MLD, we further develop a mechanistic model explaining how selection and epistasis brought about conformational changes that led to the MLD found in the 2014 sequences.

MATERIALS AND METHODS

Sequence data. We retrieved from GenBank a total of 124 complete Ebola virus genome sequences that were previously published (3). Of these, 99 EBOV genomes were from 78 confirmed 2014 Ebola virus disease patients, and two EBOV genomes (Guinea) were from the start of the outbreak in February 2014. Twenty-three additional genomes from earlier outbreaks were also retrieved in order to allow for a complete phylogenetic reconstruction of the *Ebolavirus* genus. These 23 genomes consisted of 10 Sudan sequences, seven Reston sequences, five Bundibugyo sequences, and one Tai Forest sequence, all spanning 1976 to 2012.

Phylogenetic analyses. We aligned the nucleotide sequences (124) for each gene independently using MUSCLE (16). Poorly aligned positions were eliminated from our alignment using default parameters in GBLOCKS (17). Model selection was conducted next based on the Akaike information criterion (AIC) as implemented in jModelTest (18). The nucleotide alignments for VP40, VP24, and VP30 were analyzed with PhyML (19) using a general time-reversible model with among-site rate variation modeled by a discrete gamma distribution (GTR+ Γ) while a transversion model (TVM+ Γ) was used for the proteins VP35, GP, NP, and L polymerase. The resultant phylogenetic trees for each gene were visualized using FigTree (<http://tree.bio.ed.ac.uk/software/figtree>).

Selection analysis. We tested for evidence of episodic diversifying selection within the Ebola virus genome using MEME (15). Available methods for identifying sites which undergo pervasive selection (constant through time) are not always able to detect episodic (sporadic) selection (20). However, MEME, a tool for detecting site-specific diversifying selection, circumvents this issue of low sensitivity by identifying episodes of diversifying selection affecting a subset of branches at individual sites (21). In particular, this analysis allowed us to identify codons under diversifying selection along the branch leading up to the 2014 EBOV strain. This data analysis was conducted using Datamonkey, a web server of the HyPhy v2.12 package (22).

Timeline of amino acid substitutions. In order to map the timeline of accumulation of substitutions under selection in the Ebola virus genome, we analyzed genome data using a relaxed molecular clock. To obtain a tree of concatenated coding sequences, we first examined our full alignment for evidence of recombination using SBP (single breakpoint recombination) of the HyPhy v2.12 package (22). Finding no recombinant sequences, an alignment of concatenated coding sequences was derived from our original seven protein alignments using Catfasta2phym (23). This new alignment was analyzed in BEAST using a GTR+ Γ_4 model of nucleotide substitution, assuming an uncorrelated lognormal relaxed clock and a constant-population coalescent prior. We ran two separate Markov chain Monte Carlo samplers to check for convergence, visually assessed these with Tracer (tree.bio.ed.ac.uk/software/tracer), and removed a conservative 10% of the 1×10^{-10} states as burn-in. The detected adaptive substitutions were then mapped onto the dated phylogenetic tree.

Detection of epistasis. To test for a link between sites under diversifying selection and those under epistasis, we ran an algorithm designed to detect correlated substitutions, as implemented in AEGIS (Nshogozababizi et al., submitted). Taking inspiration from a statistical approach originally developed to assess phenotypic correlation at pairs of traits (24), our approach determines the degree of correlation between pairs of amino acid positions. As the original model was developed to analyze binary traits, amino acid positions were recoded as ancestral/derived characters based on a majority-rule consensus. More specifically, using the 2014 EBOV sequences as a reference, amino acids (at each site) which are present in at least 90% of the EBOV 2014 clade are denoted “derived,” while those that differ from this reference are denoted “ancestral.” The appropriate majority-rule percentage was chosen based on the consensus amino acids observed across all 124 sequences.

Epistasis and selection as interacting agents in outbreak onsets. To elucidate how the interplay between epistasis and selection shaped the epidemiology of the virus, we reconstructed a lineage-through-time (LTT) plot derived from the BEAST tree. We used the `ltt.plot()` function from the R package `ape` (25) to do this.

Ancestral MLD structure predictions. To gain insights into the biological significance of our results, we mapped the substitutions that are both under selection and involved in epistatic interactions on the three-dimensional (3D) structure of the EBOV glycoprotein (PDB entry 3CSY) using Swiss PDB Viewer v4.1.0 (26). The corresponding chains (A to H) of the Fab KZ52 human antibody complex were removed to facilitate visualization of the prefusion conformation of the GP trimer. However, all of the sites of interest (identified as under selection and/or epistasis) are within a region, the GP mucin-like domain (MLD), that was excised from protein 3CSY to aid its crystallization (27). As such, homology modeling (28) could not be used to predict the structure of this region, and we had to resort to a more sophisticated approach. For this, we resorted to Robetta (29), a full-chain protein structure prediction server that uses the Rosetta software package (30). Domain parsing and *ab initio* structure prediction of the GP MLD were carried out using the Ginzu (29) protocol method implemented in Robetta.

In order to understand the structural changes within this MLD at different time points along the lineage leading up to the 2014 sequences, we performed marginal ancestral reconstructions. Using our GP sequence alignment, we first mapped the branches along which each mutation of interest (identified as under selection and/or epistasis) occurred. This led us to selecting all internal nodes between the first and the last occurrences. At each of these internal nodes, we inferred the GP MLD from the GP gene alignment and its corresponding phylogenetic tree using FastML (31), a web server for maximum-likelihood reconstruction of amino acid sequences. To reconstruct a representative 2014 GP MLD, a sixth sequence was chosen at random from the 2014 clade (GenBank accession number [KM233078](https://www.ncbi.nlm.nih.gov/nuccore/KM233078)). We were able to select a random representative sequence without compromising our analysis due to the high global similarity and low pairwise distance among the GP MLD regions of the 2014 EBOV sequences. Using the BLOSUM62 scoring matrix (32) implemented in

FIG 1 Lineage-through-time (LTT) plot derived from dated BEAST tree. (A) Number of lineages (log transformed) as a function of time (years). The five major Ebola virus outbreaks are marked in red dashed lines (1976, 1995, 2000, 2007, and 2014), with the most recent one being the 2014 outbreak. (B) The glycoprotein (GP) network of epistatic sites is mapped onto the BEAST tree, with the emergence of the adaptive substitutions (red circles) indicated along the respective branches (solid red line). The X represents a group of sites with identical site patterns (i.e., identical recoding pattern based on majority-rule consensus, as described in the text). This group of sites is listed in full in Table S1 in the supplemental material. Branches are color coded (yellow, turquoise, and purple) according to the emergence of each mutation. The elliptic arrow represents the timing of emergence for substitutions, starting from sites that mutated along the yellow branch to the substitution at position 308; a full chronology for each site could not be inferred from the alignment.

CLUSTAL W (33), we obtained an average global similarity score of 0.996 (minimum, 0.89; maximum, 1.00) comparing the MLD region among the 81 2014 sequences. To further test the level of similarity within the MLD region of our 81 2014 sequences, we obtained an average pairwise distance using ConSurf 2010 (34), a server for calculating evolutionary conservation in protein sequences. With an average pairwise distance of 1.02×10^{-7} (compared to 0.67 for the entire alignment of 124 sequences), our

2014 sequences exhibit a high level of conservation within the MLD region of the glycoprotein. Given these findings, any one of our 2014 sequences would yield accurate MLD reconstructions representing the 2014 clade.

MLD structure evolution. To understand how selection and epistasis affected the structure of the MLD, we compared each predicted structure to the oldest reconstructed structure. To do this, we first performed structural alignments with Swiss PDB Viewer v4.1.0 (26). From these, we de-

TABLE 1 MEME analysis results for codon-specific episodic diversifying selection in EBOV genes

EBOV gene	No. of codons	Model of evolution ^a	Detected codon ^b (<i>P</i> value; <i>q</i> value)
L Polymerase	2,212	GTR	226 (0.0066; 1), 790 (0.0039; 1), 1,060 (0.0099; 1), 1,364 (0.0038; 1), 1,694 (0.0096; 1)
GP	676	GTR	44* (0.0041; 1), 215 (0.0062; 1), 308 (0.0061; 1), 428* (0.0044; 1)
VP40	326	GTR	NA
NP	739	GTR	35* (0.0059; 1), 111* (0.0004; 0.3176), 398 (0.0026; 0.9617), 721 (0.0036; 0.8912)
VP35	341	GTR	NA
VP30	288	GTR	NA
VP24	251	GTR	NA

^a The GTR (general time-reversible) model was selected as the best fitting model for all genes.

^b NA, genes which had no signal of episodic diversifying selection. Codons marked by an asterisk were detected along the 2014 EBOV lineage. Sites which are also involved in epistatic interactions (see Tables S1 to S7 in the supplemental material) are in boldface.

rived root mean squared deviations (RMSD) and plotted these as a function of time (as derived from the BEAST analysis).

Because of the complexity of the MLD structural changes, we further computed difference distance matrices among pairs of residues for each structural alignment with SuperPose, a server specializing in structure alignments and distance statistics (35). Difference distance matrices were obtained by first generating a distance matrix for each molecule (pairwise

distances between all pairs of aligned backbone alpha carbon atoms) and then subtracting these matrices from one another. Corresponding heat maps were then plotted to reveal where and when structural divergence/convergence occurred and hence explain the interplay between epistasis and selection on the evolving shape of the MLD. A brief illustration of how these maps were obtained is available at https://en.wikipedia.org/wiki/Distance_matrix.

FIG 2 Timeline of adaptive amino acid substitutions using a relaxed-molecular-clock phylogenetic tree. Sites detected as evolving under episodic diversifying selection (MEME analysis) for three out of seven Ebola virus genes (L Polymerase, glycoprotein, and nucleoprotein) are mapped onto the dated BEAST phylogenetic tree. All substitutions along the boldfaced backbone of the tree lead up to the 2014 EBOV clade. Numbers in parentheses denote the five internal nodes and one 2014 EBOV sequence (GenBank accession number [KM233078](#)) used for the ancestral reconstruction of the GP mucin-like domain. The GP sites (308 and 428) around which these nodes were chosen are in red boldface type.

FIG 3 Detected epistatic interactions within the seven EBOV genes. (A to G) Epistatic interactions detected in all seven genes at a significance level of 4×10^{-4} (after FDR). Each circle represents an amino acid site, and connecting lines denote epistatic interactions detected by AEGIS. Sites in red are those which were also detected as adaptive by the MEME analysis for episodic diversification. Overlapping site labels were omitted for ease of visualization; all interacting pairs are listed in Tables S1 to S7 in the supplemental material. An X represents a group of two or more sites with identical site patterns (i.e., identical recoding pattern based on majority-rule consensus, as described in the text). These groups of sites with identical site patterns are listed in full in Tables S1 to S7.

RESULTS AND DISCUSSION

Major Ebola outbreaks are preceded by diversification events.

First, using a lineage-through-time (LTT) analysis, we reconstructed the diversification of the Ebola virus through time and showed that all known major outbreaks are immediately preceded by a distinct diversification event (Fig. 1A). Such a pattern prompts the question of whether these diversification events are the result of diversifying selection acting episodically just prior to each outbreak.

Evidence for selection is weak. To test this hypothesis, we ran a MEME analysis that aims at detecting evidence for diversifying selection acting episodically, i.e., along particular lineages of the tree, without specifying these lineages *a priori* in order not to bias our analysis toward outbreak lineages. This analysis revealed that of the seven proteins, three (GP, L Polymerase, and NP) were evolving under episodic diversifying selection, with *P* values of ≤ 0.01 (Table 1). GP, L Polymerase, and NP play a major role in viral replication and host cell attachment (36), which suggests that these three proteins are the most vulnerable to selective pressures exerted by the host cell. However, these adaptive substitutions do not show the phylogenetic clustering (Fig. 2) that we expected if they were, on their own, responsible for viral diversification bursts identified in the LTT analysis.

This disagreement is not caused by improper model selection: this MEME analysis was based on AIC model selection, which showed that the best model to use was GTR. To ensure that the MEME results were robust for this model choice, we ran the analysis again under a range of models of evolution and found that the location of inferred sites undergoing diversifying selection did not

change depending on the model used: these substitutions still do not cluster on lineages leading up to outbreaks.

On the other hand, because the MEME analysis involves multiple rounds of testing of sites and branches, it may be critical to correct the test procedure by computing the false discovery rate (FDR) of sites undergoing episodic diversifying selection (15). However, after correcting for FDR in our analysis, no statistically significant sites remained.

Just like ours, a previous study also found strong evidence for episodic diversifying selection within the GP and L Polymerase proteins (37) at a 5% significance level and without correcting for FDR. Increasing our significance level from 1 to 5%, we found that only one residue, amino acid position 1492 of the L Polymerase protein, was detected in both our study and the previous analysis, with no other common sites. One possible explanation for this discrepancy is that the previous study used only 78 EBOV sequences (and 13 additional unnamed sequences) instead of the 124 full-genome sequences used here.

Evidence for epistasis links to selection. In the absence of strong statistical evidence supporting selection at these positions, we assessed their potential biological significance. Based on the Bayesian relaxed-molecular-clock analysis conducted for the LTT analysis, we reconstructed a timeline of the emergence of these substitutions putatively under selection (Fig. 2). The sequential and sustained mode of emergence of these adaptive substitutions suggests that epistatic interactions have played a role in the evolution of the 2014 EBOV clade.

To test this hypothesis, we analyzed each protein alignment with AEGIS. At a significance level of 4×10^{-4} (after FDR), we

FIG 4 Molecular representation of EBOV GP trimer crystal structure. (A) Top view of the three GP_{1,2} monomers in prefusion conformation (PDB entry 3CSY). The human antibody Fab KZ52 was removed to facilitate visualization. The Robetta structure prediction of the mucin-like domain (gray) is superimposed onto one of the GP_{1,2} monomers. In theory, each GP_{1,2} monomer has an identical mucin-like domain (MLD). The putative location of the MLD was determined using the MagicFit function in Swiss PDB viewer. (B) A closeup view of the GP MLD showing distribution of epistatic sites, each labeled in blue. (C) A closeup view of the GP MLD showing the position of the epistatic sites 308 and 428 that also were detected as adaptive by the MEME analysis. All graphic representations were produced using Swiss PDB Viewer.

detected epistatic interactions in all seven Ebola virus genes (see Tables S1 to S7 in the supplemental material). As previously uncovered (Nshogozabahizi et al., submitted), we found some overlap among pair members of the epistatic interactions, so that long chains or networks of epistatic sites could be reconstructed by transitivity (Fig. 3A to G).

Critically, three of these networks of epistatic sites also contained sites potentially under selection (Table 1): GP, L Polymerase, and NP. This kind of overlap suggests a link between epistasis and selection. Such a link was theoretically shown to be possible in prokaryotes (11), but to the best of our knowledge it has never been shown to exist in actual data, let alone in viruses such as Ebola.

Selection and epistasis are linked to conformation and function in the MLD. To investigate the nature of this link between selection and epistasis, we focused on the GP gene for three rea-

sons. First, GP has become the primary target of drug design and human treatments. The newly developed recombinant, replication-competent, vesicular stomatitis virus-based vaccine (rVSV-ZEBOV) expresses EBOV glycoproteins in order to elicit an immune response against the complete virus (38). Second, only in this gene are the detected epistatic sites clustered in a small region of the protein. Indeed, all positions are in the tight range of amino acids 308 to 428 (Fig. 3) and are physically close to each other on a 3D model of the protein structure (Fig. 4). This clustering of epistatic sites could be the product of a relatively high mutation rate in this region, reflecting its dispensability for cell entry (10), but this possibility has, to date, not been fully investigated. Third, and most critically, this small cluster is right in the mucin-like domain (MLD), a region that spans residues 308 to 501 (7, 8) and that has been hypothesized to shield the glycoprotein from antibodies (9, 39).

FIG 5 All 6 MLD ancestral reconstructions with labeled sites of interest. (A to F) MLD reconstructions for ancestral nodes at time points (years) 1918, 1963, 1969, 1995, 2004, and 2014. The locations of GP linear epitopes within the MLD are labeled for each reconstruction, with amino acid numbers in parentheses (panel A only). GP epitopes 401-417 (orange) and 389-405 (teal) overlap at sequence ATQVE. Antibody 13F6-1-2 is directed against residues 401 to 417, antibody 6D8-1-2 is directed against residues 389 to 405, and antibody 12B5-1-1 is directed against residues 477 to 493. Known N-linked glycosylation sites throughout the MLD are marked in blue (stick representation is used). The locations of the detected adaptive sites (from the MEME analysis) are shown in red.

Several lines of evidence support the conclusion that episodic diversifying selection coupled with epistasis is operating on the GP gene. Most notably, the MLD is critical for evading the humoral immune response through the inhibition of surface protein recognition (9). This region also contains multiple binding sites of early cocktail antibodies that are, however, nonneutralizing, because the MLD is excised by host cathepsins in the endosome (40, 41). It also has been speculated that in the absence of cleavage by furin, the host protein necessary for GP activation, it is the flexible MLD that provides the necessary conformational freedom within the internal fusion loop to activate GP (10). As such, substitutions within this domain may be necessary to facilitate a metastable conformation of GP prior to fusion. The possibility of such compensatory interactions is made evident by the tight physical clustering of our detected sites, which suggests an interaction network driven by steric hindrance (42). In such a network, any conformational change in one of the amino acids may necessitate a compensatory change in another neighboring amino acid. As such, this would lead to an intriguing hypothesis: a series of epistatic changes might have been required at some point in time to compensate for a previous substitution. Such an epistatic constraint, present for functional reasons, might have favored the emergence of an initial adaptive substitution (at position 428) (Fig. 1B). While conferring a

selective advantage, this substitution might have distorted the 3D structure of the MLD, thereby requiring some epistatic changes for conformational reasons. These changes, in turn, permitted the emergence of a second adaptive substitution (at position 308), which appeared in the 2014 Ebola strain (Fig. 2).

Adaptive substitutions as endpoints of epistatic changes. In order to test the plausibility of this scenario, we reconstructed the 3D structure of the MLD at key time points, from just before the emergence of the first adaptive substitution (at position 428) all the way to the emergence of the second adaptive substitutions (at position 308). This led us to five reconstructions at nodes placed on the lineage leading up to the 2014 sequences (Fig. 2, nodes 1 to 5) and one reconstruction representing a 2014 EBOV sequence (Fig. 5); all six models are available as PDB files in the supplemental material. The pairwise superimpositions of the five latest MLD structures to the earliest structure reveal an approximate trajectory of structural divergence through time (Fig. 6).

Notably, the difference distance plot of MLD1 versus MLD5 reveals a distinct square area of low amino acid distance bounded by the two adaptive substitutions at sites 308 and 428 (Fig. 6, 1 versus 5). These sites, therefore, seem to act as beginning and ending points for the structural reestablishment of the mucin-like domain, both in time and 3D space. We suggest that adaptive

FIG 6 Root mean squared deviation (RMSD) plot of MLD structural alignments. The line graph shows RMSD as a function of time for six reconstructed 3D structures of the GP MLD at key time points (before and after adaptive substitutions). The oldest reconstructed structure (year 1918) was used as a reference upon which each of the remaining structures were superimposed. The six plotted points (1918, 1963, 1969, 1995, 2004, and 2014) correspond to the following structure comparisons: MLD1 versus MLD2, MLD1 versus MLD3, MLD1 versus MLD4, MLD1 versus MLD5, and MLD1 versus MLD6. Corresponding heat maps of amino acid distance are plotted for each pairwise comparison (except for MLD1 versus itself). The number of residues compared for the first three pairs is 195, and it is 177 for the last two (due to deleted sequence gaps for Robetta prediction). 2D locations of epistatic GP amino acid sites are represented with vertical and horizontal red lines. Adaptive sites 308 and 428 are annotated in red on the MLD1 versus MLD5 heat map.

mutation P428L arose as a permissive substitution that first allowed for a wide array of substitutions to accrue, changing the MLD structure considerably up to year 1966. The new conformation of the MLD then might have required compensatory changes that (i) drastically reduced the extent of divergence from the original (1918) MLD structure, potentially creating a much more adapted glycoprotein secondary structure, and (ii) facilitated the emergence of the adaptive mutation at site 308, which appeared in the 2014 strain.

It can be noted that the structural similarities between the six reconstructions (Fig. 5) do not seem to coincide with the global RMSD values obtained for the pairwise comparisons (Fig. 6). For instance, MLD 1 and MLD 5 appear more diverged than MLD 1 versus MLD 4 (Fig. 5), while our global RMSD values show that the MLD 1 reference structure is the most similar to the MLD 5 structure. In cases where the aligned structures do not have identical or very similar sequences (as seen with the MLD), RMSD calculations only consider backbone alpha carbons and disregard

side chain residues (43). Hence, RMSDs do not reflect the rotameric states of all side chain residues shown in Fig. 5.

Insight into the GP MLD structure. The computational nature of our ancestral reconstructions of the MLD may make them somehow approximative, yet they are based on one of the best-performing algorithms for *ab initio* structure prediction (29, 44). As such, these reconstructions provide some structural insight that has not been previously available in the literature. In particular, we can better understand the spatial configuration of the GP MLD in the absence of a crystallized structure to work from. Our structural prediction and superimposition (Fig. 4) onto the glycoprotein (PDB entry 3CSY) show that the MLD is situated at the top-most vertex (spike) of each GP1 monomer and protrudes at the sides of the GP structure. These results agree with previous predictions and findings about the MLD (7). Each GP1 monomer contains a moderately glycosylated glycan cap, which is located atop the trimeric GP spike (27). The spatial orientation of the predicted MLD (Fig. 4) suggests that there is no physical contact

among all three MLDs on the GP structure. Our reconstruction therefore supports the proposition that the MLD extends outwards from this glycan cap, forming a heavily glycosylated blanket across the top of the GP (27, 45).

Although the MLD has been predicted to be a generally unstructured region (7), it contains three polypeptide sequences that have been experimentally identified as linear epitopes (46). For the Zaire species in particular, antibodies 13F6-1-2, 6D8-1-2, and 12B5-1-1 target residues 389 to 405, 401 to 417, and 477 to 493, respectively (47, 48). We can now highlight these epitopes on our ancestral reconstructions to gain functional insight through time. Our models suggest that both relative position and structure of these three epitopes vary widely from the first reference time point (1918) to the last (2014) (Fig. 5). Notably, the first four MLD reconstructions (Fig. 5A to D) predict the GP epitopes 389 to 405 and 477 to 493 as alpha helices, which then become highly unstructured regions in the last two reconstructions (Fig. 5E and F). The loss of the alpha helix secondary structure at these epitopes between 1995 and 2004 is intriguing and may warrant experimental data to understand its implications. Nonetheless, our reconstructions support previous hypotheses that regions involved in critical interactions with the host proteins are often these unique alpha helices (49, 50). Note that one of the experimentally validated GP-bound linear epitopes (401-417) for 1976 and 1995 Ebola Zaire isolates (46) contains two of our detected epistatic sites: amino acids 410 and 414 (see Table S1 in the supplemental material). Although this overlap on its own does not substantiate that epistasis is a major factor in epitope-antibody interaction, it opens up potential avenues for future epitope investigation. It is likely that these apical surface epitopes are more exposed and more flexible for antibody interaction than other GP epitopes situated deeper inside the GP spike, allowing the MLD to tolerate a greater number of substitutions (and consequent epistatic interactions).

The MLD is also populated with N- and O-linked glycosylation sites (10, 51), some of which border the above-mentioned MLD linear epitopes. While our reconstructions (Fig. 5) provide the structural positions of the validated N-linked glycans in the MLD, very little can be gathered about the potential structural consequences of these glycans on MLD function. It has been previously demonstrated that N-linked glycans are not critical for GP folding and that their removal from the MLD does not affect antiserum sensitivity (51). It is therefore possible that the N-linked glycans which flank the GP epitope residues 389 to 405 and 401 to 417 influence the stability of this region but are unlikely to have any effect on antibody sensitivity, as previously investigated (51).

Linking selection, epistasis, and outbreak onsets. While our initial MEME analyses failed to show evidence for selection in most other major outbreaks, our epistasis analyses show widespread evidence for correlated evolution throughout the Ebola genome (see Tables S1 to S7 in the supplemental material). This begs the question as to whether epistasis can act alone to accelerate divergence. Based on the above-described findings, we argue that selection requires epistatic interactions, which are in turn driven by a tight physical linkage among amino acids.

Although experimental evidence of our argument would lead to the usual dual-use conundrum (52) and hence would most likely be unfeasible, we presented here tantalizing evidence supporting a key role for epistasis, in conjunction with selection, in the diversification process leading up to the 2014 Ebola outbreak.

We further support our claim with structural evidence, linking all processes together.

FUNDING INFORMATION

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canada Foundation for Innovation (S.A.-B.) and by the University of Ottawa (N.I. and J.C.N.).

REFERENCES

- Baize S, Pannetier D, Oestereich L, Rieger T, Koivogui L, Magassouba N, Soropogui B, Sow MS, Keita S, De Clerck H, Tiffany A, Dominguez G, Loua M, Traoré A, Kolié M, Malano ER, Heleze E, Bocquin A, Mély S, Raoul H, Caro V, Cadar D, Gabriel M, Pahlmann M, Tappe D, Schmidt-Chanasit J, Impouma B, Diallo AK, Formenty P, Van Herp M, Günther S. 2014. Emergence of Zaire Ebola virus disease in Guinea. *N Engl J Med* 371:1418–1425. <http://dx.doi.org/10.1056/NEJMoa1404505>.
- House T. 2014. Epidemiological dynamics of Ebola outbreaks. *eLife* 3:e03908.
- Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G, Wohl S, Moses LM, Yozwiak NL, Winnicki S, Matranga CB, Malboeuf CM, Qu J, Gladden AD, Schaffner SF, Yang X, Jiang P-P, Nekoui M, Colubri A, Coomber MR, Fonnies M, Moigboi A, Gbakie M, Kamara FK, Tucker V, Konuwa E, Saffa S, Sellu J, Jalloh AA, Kovoma A, Koninga J, Mustapha I, Kargbo K, Foday M, Yillah M, Kanneh F, Robert W, Massally JLB, Chapman SB, Bochicchio J, Murphy C, Nusbaum C, Young S, Birren BW, Grant DS, Scheffelin JS, Lander ES, Hapci C, Gevaio SM, Gnirke A, Rambaut A, Garry RF, Khan SH, Sabeti PC. 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* 345:1369–1372. <http://dx.doi.org/10.1126/science.1259657>.
- WHO Ebola Response Team. 2014. Ebola virus disease in West Africa—the first 9 months of the epidemic and forward projections. *N Engl J Med* 371:1481–1495. <http://dx.doi.org/10.1056/NEJMoa1411100>.
- Hoenen T, Safronetz D, Groseth A, Wollenberg KR, Koita OA, Diarra B, Fall IS, Haidara FC, Diallo F, Sanogo M, Sarro YS, Kone A, Togo ACG, Traore A, Kodio M, Dosseh A, Rosenke K, de Wit E, Feldmann F, Ebihara H, Munster VJ, Zoon KC, Feldmann H, Sow S. 2015. Mutation rate and genotype variation of Ebola virus from Mali case sequences. *Science* 348:117–119. <http://dx.doi.org/10.1126/science.aaa5646>.
- Park DJ, Dudas G, Wohl S, Goba A, Whitmer SL, Andersen KG, Sealfon RS, Ladner JT, Kugelman JR, Matranga CB, Winnicki SM, Qu J, Gire SK, Gladden-Young A, Jalloh S, Nosamiefan D, Yozwiak NL, Moses LM, Jiang PP, Lin AE, Schaffner SF, Bird B, Towner J, Mamoh M, Gbakie M, Kanneh L, Kargbo D, Massally JL, Kamara FK, Konuwa E, Sellu J, Jalloh AA, Mustapha I, Foday M, Yillah M, Erickson BR, Sealy T, Blau D, Paddock C, Brault A, Amman B, Basile J, Bearden S, Belser J, Bergeron E, Campbell S, et al. 2015. Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. *Cell* 161:1516–1526. <http://dx.doi.org/10.1016/j.cell.2015.06.007>.
- Tran EE, Simmons JA, Bartschaghi A, Shoemaker CJ, Nelson E, White JM, Subramaniam S. 2014. Spatial localization of the Ebola virus glycoprotein mucin-like domain determined by cryo-electron tomography. *J Virol* 88:10958–10962. <http://dx.doi.org/10.1128/JVI.00870-14>.
- Martinez O, Tantral L, Mulherkar N, Chandran K, Basler CF. 2011. Impact of Ebola mucin-like domain on antigenic protein antibody responses induced by Ebola virus-like particles. *J Infect Dis* 204:S825–S832. <http://dx.doi.org/10.1093/infdis/jir295>.
- Martinez O, Ndungo E, Tantral L, Miller EH, Leung LW, Chandran K, Basler CF. 2013. A mutation in the Ebola virus envelope glycoprotein restricts viral entry in a host species- and cell-type-specific manner. *J Virol* 87:3324–3334. <http://dx.doi.org/10.1128/JVI.01598-12>.
- Lee JE, Saphire EO. 2009. Ebolavirus glycoprotein structure and mechanism of entry. *Future Virol* 4:621–635. <http://dx.doi.org/10.2217/fvl.09.56>.
- Griswold CK. 2015. Epistasis can accelerate adaptive diversification in haploid asexual populations. *Proc Biol Sci* 282:20142648. <http://dx.doi.org/10.1098/rspb.2014.2648>.
- Gong LI, Suchard MA, Bloom JD. 2013. Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife* 2:e00631. <http://dx.doi.org/10.7554/eLife.00631>.
- Elena SF, Solé RV, Sardanyés J. 2010. Simple genomes, complex inter-

- actions: epistasis in RNA virus. *Chaos* 20:026106. <http://dx.doi.org/10.1063/1.3449300>.
14. Bloom JD, Gong LI, Baltimore D. 2010. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science* 328:1272–1275. <http://dx.doi.org/10.1126/science.1187816>.
 15. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. 2012. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet* 8:e1002764. <http://dx.doi.org/10.1371/journal.pgen.1002764>.
 16. Edgar RC. 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797. <http://dx.doi.org/10.1093/nar/gkh340>.
 17. Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* 56:564–577. <http://dx.doi.org/10.1080/10635150701472164>.
 18. Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* 9:772. <http://dx.doi.org/10.1038/nmeth.2109>.
 19. Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704. <http://dx.doi.org/10.1080/10635150390235520>.
 20. Grueber CE, Wallis GP, Jamieson IG. 2014. Episodic positive selection in the evolution of avian toll-like receptor innate immunity genes. *PLoS One* 9:e89632. <http://dx.doi.org/10.1371/journal.pone.0089632>.
 21. Arunachalam R. 2013. Detection of site-specific positive Darwinian selection on pandemic influenza A/H1N1 virus genome: integrative approaches. *Genetica* 141:143–155. <http://dx.doi.org/10.1007/s10709-013-9713-x>.
 22. Delpont W, Poon AFY, Frost SDW, Kosakovsky Pond SL. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* 26:2455–2457. <http://dx.doi.org/10.1093/bioinformatics/btq429>.
 23. Nylander J. 2013. Catusa2phym. Program distributed by the author. Evolutionary Biology Center, Uppsala University, Uppsala, Sweden.
 24. Pagel M, Meade A. 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *Am Nat* 167:808–825. <http://dx.doi.org/10.1086/503444>.
 25. Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290. <http://dx.doi.org/10.1093/bioinformatics/btg412>.
 26. Guex N, Peitsch MC. 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* 18:2714–2723. <http://dx.doi.org/10.1002/elps.1150181505>.
 27. Lee JE, Fusco ML, Hessel AJ, Oswald WB, Burton DR, Saphire EO. 2008. Structure of the Ebola virus glycoprotein bound to an antibody from a human survivor. *Nature* 454:177–182. <http://dx.doi.org/10.1038/nature07082>.
 28. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L, Schwede T. 2014. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res* 42:W252–W258. <http://dx.doi.org/10.1093/nar/gku340>.
 29. Kim DE, Chivian D, Baker D. 2004. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res* 32:W526–W531. <http://dx.doi.org/10.1093/nar/gkh468>.
 30. Simons KT, Bonneau R, Ruczinski I, Baker D. 1999. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* 37:171–176.
 31. Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, Pupko T. 2012. FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res* 40:W580–W584. <http://dx.doi.org/10.1093/nar/gks498>.
 32. Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 89:10915–10919. <http://dx.doi.org/10.1073/pnas.89.22.10915>.
 33. Larkin MA, Blackshields G, Brown N, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947–2948. <http://dx.doi.org/10.1093/bioinformatics/btm404>.
 34. Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. 2010. Con-Surf 2010: calculating evolutionary conservation in sequence and structure of proteins and Nucleic acids. *Nucleic Acids Res* 38:W529–W533. <http://dx.doi.org/10.1093/nar/gkq399>.
 35. Maiti R, Van Domselaar GH, Zhang H, Wishart DS. 2004. SuperPose: a simple server for sophisticated structural superposition. *Nucleic Acids Res* 32:W590–W594. <http://dx.doi.org/10.1093/nar/gkh477>.
 36. de Wit E, Feldmann H, Munster VJ. 2011. Tackling Ebola: new insights into prophylactic and therapeutic intervention strategies. *Genome Med* 3:5. <http://dx.doi.org/10.1186/gm219>.
 37. Volz E, Pond S. 2014. Phylodynamic analysis of Ebola virus in the 2014 Sierra Leone epidemic. *PLoS Curr* 6:ecurrents.outbreaks.6f7025f1271821d4c815385b08f5f80e. <http://dx.doi.org/10.1371/currents.outbreaks.6f7025f1271821d4c815385b08f5f80e>.
 38. Henao-Restrepo AM, Longini IM, Egger M, Dean NE, Edmunds WJ, Camacho A, Carroll MW, Dombia M, Draguez B, Duraffour S, Enwere G, Grais R, Gunther S, Hossmann S, Kondé MK, Kone S, Kuisma E, Levine MM, Mandal S, Norheim G, Riveros X, Soumah A, Trelle S, Vicari AS, Watson CH, Kéita S, Kiény MP, Röttingen JA. 2015. Efficacy and effectiveness of an rVSV-vectored vaccine expressing Ebola surface glycoprotein: interim results from the Guinea ring vaccination cluster-randomised trial. *Lancet* 386:857–866. [http://dx.doi.org/10.1016/S0140-6736\(15\)61117-5](http://dx.doi.org/10.1016/S0140-6736(15)61117-5).
 39. Simon-Lorière E, Faye O, Faye O, Koivogui L, Magassouba N, Keita S, Thiberge J-M, Diancourt L, Bouchier C, Vandenbogaert M, Caro V, Fall G, Buchmann JP, Matranga CB, Sabeti PC, Manuguerra J-C, Holmes EC, Sall AA. 2015. Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic. *Nature* 524:102–104. <http://dx.doi.org/10.1038/nature14612>.
 40. Murin CD, Fusco ML, Bornholdt ZA, Qiu X, Olinger GG, Zeitlin L, Kobinger GP, Ward AB, Saphire EO. 2014. Structures of protective antibodies reveal sites of vulnerability on Ebola virus. *Proc Natl Acad Sci U S A* 111:17182–17187. <http://dx.doi.org/10.1073/pnas.1414164111>.
 41. Davidson E, Bryan C, Fong RH, Barnes T, Pfaff JM, Mabila M, Rucker JB, Doranz BJ. 2015. Mechanism of binding to Ebola virus glycoprotein by the ZMapp, ZMab, and MB-003 cocktail antibodies. *J Virol* 89:10982–10992. <http://dx.doi.org/10.1128/JVI.01490-15>.
 42. Ortlund EA, Bridgman JT, Redinbo MR, Thornton JW. 2007. Crystal structure of an ancient protein: evolution by conformational epistasis. *Science* 317:1544–1548. <http://dx.doi.org/10.1126/science.1142819>.
 43. Godzik A. 1996. The structural alignment between two proteins: is there a unique answer? *Protein Sci* 5:1325–1338. <http://dx.doi.org/10.1002/pro.5560050711>.
 44. Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, DiMaio F, Lange O, Kinch L, Sheffler W, Kim B-H, Das R, Grishin NV, Baker D. 2009. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* 77(Suppl 9):S89–S99. <http://dx.doi.org/10.1002/prot.22540>.
 45. Dube D, Brecher MB, Delos SE, Rose SC, Park EW, Schornberg KL, Kuhn JH, White JM. 2009. The primed ebolavirus glycoprotein (19-kilodalton GP1, 2): sequence and residues critical for host cell binding. *J Virol* 83:2883–2891. <http://dx.doi.org/10.1128/JVI.01956-08>.
 46. Wilson JA, Hevey M, Bakken R, Guest S, Bray M, Schmajohn AL, Hart MK. 2000. Epitopes involved in antibody-mediated protection from Ebola virus. *Science* 287:1664–1666. <http://dx.doi.org/10.1126/science.287.5458.1664>.
 47. Lee JE, Kuehne A, Abelson DM, Fusco ML, Hart MK, Saphire EO. 2008. Complex of a protective antibody with its Ebola virus GP peptide epitope: unusual features of a V λ x light chain. *J Mol Biol* 375:202–216. <http://dx.doi.org/10.1016/j.jmb.2007.10.017>.
 48. Lee JE, Saphire EO. 2009. Neutralizing ebolavirus: structural insights into the envelope glycoprotein and antibodies targeted against it. *Curr Opin Struct Biol* 19:408–417. <http://dx.doi.org/10.1016/j.sbi.2009.05.004>.
 49. Chakraborty S, Rao BJ, Asgerisson B, Dandekar A. 2014. Characterizing alpha helical properties of Ebola viral proteins as potential targets for inhibition of alpha-helix mediated protein-protein interactions. *F1000Res* 3:251. <http://dx.doi.org/10.12688/f1000research.5573.1>.
 50. Azzarito V, Long K, Murphy NS, Wilson AJ. 2013. Inhibition of [alpha]-helix-mediated protein-protein interactions using designed molecules. *Nat Chem* 5:161–173. <http://dx.doi.org/10.1038/nchem.1568>.
 51. Lennemann NJ, Rhein BA, Ndungo E, Chandran K, Qiu X, Maury W. 2014. Comprehensive functional analysis of N-linked glycans on Ebola virus GP1. *mBio* 5:e00862-13. <http://dx.doi.org/10.1128/mBio.00862-13>.
 52. Berg P. 2012. The dual-use conundrum. *Science* 337:1273. <http://dx.doi.org/10.1126/science.1229789>.