PLOS PATHOGENS

RESEARCH ARTICLE

Genomic transfers help to decipher the ancient evolution of filoviruses and interactions with vertebrate hosts

Derek J. Taylor*, Max H. Barnhart¤

Department of Biological Sciences, University at Buffalo, Buffalo, New York, United States of America

¤ Current address: Department of Plant Biology, University of Georgia, Athens, Georgia, United States of America
* djtaylor@buffalo.edu

Citation: Taylor DJ, Barnhart MH (2024) Genomic transfers help to decipher the ancient evolution of filoviruses and interactions with vertebrate hosts. PLoS Pathog 20(9): e1011864. https://doi.org/10.1371/journal.ppat.1011864

Editor: Jens H. Kuhn, Division of Clinical Research, UNITED STATES OF AMERICA

Received: November 27, 2023

Accepted: August 21, 2024

Published: September 3, 2024

Data Availability Statement: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper and its Supporting Information files.

Funding: The author(s) received no specific funding for this work.

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

Abstract

Although several filoviruses are dangerous human pathogens, there is conflicting evidence regarding their origins and interactions with animal hosts. Here we attempt to improve this understanding using the paleoviral record over a geological time scale, protein structure predictions, tests for evolutionary maintenance, and phylogenetic methods that alleviate sources of bias and error. We found evidence for long branch attraction bias in the L gene tree for filoviruses, and that using codon-specific models and protein structural comparisons of paleoviruses ameliorated conflict and bias. We found evidence for four ancient filoviral groups, each with extant viruses and paleoviruses with open reading frames. Furthermore, we found evidence of repeated transfers of filovirus-like elements to mouse-like rodents. A filovirus-like nucleoprotein ortholog with an open reading frame was detected in three subfamilies of spalacid rodents (present since the Miocene). We provide evidence that purifying selection is acting to maintain amino acids, protein structure and open reading frames in these elements. Our finding of extant viruses nested within phylogenetic clades of paleoviruses informs virus discovery methods and reveals the existence of Lazarus taxa among RNA viruses. Our results resolve a deep conflict in the evolutionary framework for filoviruses and reveal that genomic transfers to vertebrate hosts with potentially functional co-options have been more widespread than previously appreciated.

Author summary

Filoviruses are a family of RNA viruses discovered in 1967 and notorious for spillover of the dangerous pathogens, Ebola virus and Marburg virus. However, their origins, deeper relations, diversity, and interactions with animal hosts remain controversial. Part of the confusion may be that differing rates of evolution among divergent viral lineages can create a bias termed long branch attraction (LBA). We tested for this scenario in the L protein gene sequence of filoviruses and found evidence that LBA is occurring leading to a false pairing of filovirus lineages associated with a fish and a snake. We found that using nucleotides instead of amino acids when inferring trees, paleoviral sequences with open reading

frames, additional conserved genes, and comparisons of predicted protein structures can resolve the LBA. We found four major groups of filoviruses, with the paleoviral record and trees being consistent with an ancient fish origin for the family. Moreover, we found evidence of a filovirus-like element in spalacid rodents that has been evolutionarily maintained at the open-reading frame, amino acid sequence and structural level for over 20 million years. We conclude that genomic interactions of filoviruses with vertebrates likely predates the separation of lobe-finned and ray-finned fish and that filoviral elements (including potential co-opted genes) are more widespread than previously appreciated.

Introduction

Despite the ongoing importance of RNA viruses as zoonotic human pathogens and key components of ecosystems, we know little about their pre-historic evolution and host interactions. However, recent advances in paleovirology, evolutionary genomics, structural biology and artificial intelligence are enabling rapid insights [1–5]. There is now evidence that many families of RNA viruses have long evolutionary histories of interactions with vertebrate hosts (including gene co-option) and branching orders that recapitulate those from ancient host phylogenies [6]. Still, host jumps have occurred (often from prey to predator) and the propensity to harbor viruses that jump hosts varies among host taxa [2].

The transfer of non-retroviral RNA viral elements to the host is a rare macromutation. While the majority of these elements appear to be pseudogenes, some of these elements have expression products. For example, a bornavirus-like nucleoprotein element in humans expresses a protein that interacts with mitochondrial proteins and affects cell viability [7]. A tandem gene family in yeast is derived from a capsid gene of totiviruses and expresses protein products that appear to have antiviral function [8,9]. Another totivirus-derived gene in insects, also expressed as a protein, affects fecundity and development [10].There are cases where non-coding RNA products derived from viral genes may also have functional significance. A non-coding bornavirus-like element in humans has been shown to reduce the expression of a neighboring gene, *COMMD3 (COMM Domain Containing 3)*, thereby enhancing the NF-κB (nuclear factor kappa B) pathway for pathogen defense [11].

One group that still has mysterious origins and host interactions is the filoviruses–a family of negative stranded unsegmented RNA genomes whose best-known members are dangerous human pathogens, Ebola virus (EBOV) and Marburg virus (MARV). Filovirus genomes contain from 6 to 10 genes, with the nucleoprotein (NP) and RNA-dependent RNA polymerase (L) genes being the only apparent common homologs. In all known filoviruses, save the thamnoviruses, the first gene (3') is NP, the second gene is the polymerase cofactor (viral protein 35, VP35) and the last gene is L [12]. For decades, filoviruses were thought to be closely related African viruses that diverged less than a few thousand years ago [13,14]. However, the discovery of filovirus-like elements (paleoviruses) and filoviruses from several continents suggested that the group was much more diverse and ancient than previously proposed [15–18]. Indeed, fossil-calibrated orthologs from mammalian genomes indicated that the family itself is older than tens of millions of years and that host infections have occurred in marsupials and eutherians [16,19–21]. Some of these paleoviruses have the potential to be co-opted elements that function in the host [22]. For example, an open reading frame (ORF) of a VP35-like element has been maintained by purifying selection throughout a radiation of mouse-eared bats and relatives and has a similar protein structure to that found in the extant human pathogen EBOV [23–25]. Another ORF for VP35-like elements has been identified in African spiny

mice [26]. Putative filoviral co-options in marsupials lack an ORF but show tissue-specific RNA expression [16,21]. Nevertheless, there are no known filovirus-like elements with extended ORFs, evidence for purifying selection, and evidence for expression.

There is regional evidence that rodents harbor RNA viruses with the greatest potential for host jumps and shrews harbor the greatest richness of viruses [2]. While rodents are important non-human study systems for filovirus infection [27], natural filovirus-host interactions with rodents and shrews remain poorly studied. Although the immune responses differ among host taxa, adult rodents fail to show significant disease with wild-type filoviral infections [28], suggesting evolved immune adaptions to filoviral infections. Indeed, the genomes of rodents and shrews have filovirus-like paleoviral elements indicating prior infections with filoviruses [16]. Taylor et al. [19] reported filovirus-like orthologs in three genomes of hamsters and voles that were phylogenetically nested inside the clade that contains human pathogens, (orthomarburgviruses and orthoebolaviruses). The genomic location of this VP35-like element was the last intron of *Tax1bp1*. One function of TAX1BP1 is the suppression of the innate immune response by down regulating NF-kB and IRF3 (interferon regulatory factor 3) [29,30]. TAX1BP1 also plays key roles in $CD4^+$ T cell-dependent antiviral responses [30]. Interestingly, mononegaviruses related to filoviruses (respiratory syncytial virus and measles virus), directly interact with and co-opt TAX1BP1 to suppress host immune response [31,32]. Filoviruses have evolved multiple approaches to disrupt the interferon response including inhibition of IF-1 by the VP35 gene and sequestering IRF3 in inclusion bodies [33,34]. While rodents have evolved immunoprotection to filoviruses, the nature of the protection appears to differ among species. For example, CD4+ cells appear to play a more important role than CD8+ cells in the antibody response to EBOV for hamsters but not for mice [35,36]. Presently, it is unknown if the VP35-like elements of hamster *Tax1bp1* are maintained by evolution or if the expression of TAX1BP1 is affected by ebola virus disease.

Recently, partial viral genome sequences related to MARV and to EBOV were obtained from bats in China [18,37]. In addition, seven genomes of filoviruses with marked sequence divergences from MARV were assembled from non-mammalian vertebrate transcriptomes (from percomorph fishes and a snake [6,12,38]). Shi et al. [6] found that one of the fish filoviruses (*Striavirus antennarii*, Xīlǎng virus, XILV) was basal to mammal-associated viruses even when fossil calibrated mouse/rat paleoviral elements [16] are included on the tree. This suggests that divergent fish filoviruses are far older than the tens of millions of years from known mammal calibrations. Geoghan et al. [39] later detected fragments (68–69 aa) of filovirus-like L-protein sequence in a percomorph (blue spotted goatfish) and a zeiform (John Dory fish). These grouped with HUJV (*Thamnovirus thamnaconi*, Huángjiāo virus) and are consistent with a fish origin for filoviruses. Presently, all of the known piscine filovirus-like sequences are from the acanthomorph clade which includes percomorphs and Zeiformes [40]. Many groups of fishes beyond the acanthomorph clade remain poorly sampled. The result is several deep phylogenetic gaps in potential vertebrate hosts of filoviruses. However, a divergent filovirus genome (Tapajós virus, TAPV, *Tapjovirus bothropis*) was recently assembled from the transcriptome of a common lancehead snake (*Bothrops atrox* (Linnaeus, 1758) [12]. TAPV forms a well-supported phylogenetic sister group (based on the conserved L or RDRP protein sequences) with the fish-associated virus, XILV [12,41]. The L gene protein sequences are the standard for comparing divergent RNA viral genomes because this gene is often the longest and most conserved genomic region [42]. Moreover, significant BLAST similarity for divergent viruses and elements is often present for protein sequences but absent for nucleotide sequences. At face value, the L protein tree suggests a host jump between fish and reptiles. However, for TAPV, the gene order, nucleotide sequence identity, and nucleotide L tree showed similarity to mammal-associated filoviruses [12,26,43]. Moreover, gene trees (NP and

VP35, but without fish-associated sequences or an outgroup) support the basal position of TAPV to one of the mammal-associated clades of paleoviruses [26]. A retrovirus-like domain in the glycoprotein gene of TAPV groups with lizards, while a similar element in mammal-associated filoviruses groups with cartilaginous fish [44]. Presently, these conflicts hinder our understanding of the deeper relations of filoviruses and the importance of host co-option of filovirus-like elements.

In this study, we examine over 500 filovirus-like paleoviruses within vertebrate genomes, seeking a deeper understanding of the ancient evolution and interactions of filoviruses with their hosts. We mitigated potential biases and artifacts, by employing simulations, codon-partitioned substitution models, taxon additions with paleoviral sequences, and comparisons of predicted protein structure distances between viruses and paleoviruses. Additionally, we assessed the evolutionary maintenance of putative co-opted elements in vertebrate genomes. This approach aims to inform about filovirus-vertebrate interactions over a geological time scale.

Results

Long branch attraction in filoviruses and its resolution

The phylogenetic analysis based on unfiltered L-protein amino acid sequences from filoviruses revealed a distinctive TAPV-XILV grouping in the phylogram, characterized by two extended branches connected by a short intermediary branch (Fig 1A). This tree shape is characteristic of potential long branch artifacts (LBA). Kapli et al. [45] reported that partitioned nucleotide models under simulated LBA (with no or weak compositional heterogeneity) have a higher probability of recovering the correct tree than amino acid models. When the phylogram from the present analysis was based on nucleotides with a partitioned codon model including all or only first and second positions, TAPV grouped with the sequences from genomes with a shared architecture, the mammalian MARV-like clade (albeit with weak support), instead of with the fish-associated virus, XILV (Figs 1B, S1 and S2). Significant differences in amino acid composition were also present between sequences linked to fish and those associated with tetrapods (S1 Table). However, our use of a method that can reduce LBA due to amino acid site compositional heterogeneity, PMSF (Posterior Mean Site Frequency profiles analysis [46]), found the same tree as the putative LBA phylogram (S3 Fig). We then carried out simulations to assess if the branch lengths and amino acid compositions are sufficient to form a long branch attraction involving TAPV and XILV. In the first simulation, the clade observed from the L-protein amino acid tree (TAPV/XILV) is enforced in the constraint parameter. As expected, most of the trees (95%) estimated from simulated alignments were consistent with the constrained TAPV/XILV clade (Fig 2A). However, when simulations were constrained to favor the non-LBA grouping (TAPV/MARV-like grouping, Fig 2B), the putative LBA grouping of TAPV/XILV remained predominant, being recovered in 61% of the simulations. A further simulation with the same TAPV/MARV-like clade constraint but with terminal branches leading to TAPV and XILV shortened by approximately half, reduced the putative LBA grouping of TAPV/XILV to 6% (Fig 2C). Another approach that alleviated presumptive LBA was the addition of outgroup sequences from related families of RNA viruses (paramyxoviruses, lispiviruses, and rhabdoviruses). This approach also led to strong support for the monophyly of filoviruses (Fig 3), and L-like paleoviral elements from Neotropical opossums being placed within the MARV-like clade.

Affiliations of a snake-associated filoviruses with paleoviruses

We then estimated phylograms from the NP and VP35 gene regions of the filovirus genome. Unlike the L-protein analyses, we initially added paleoviruses from vertebrate genomes with

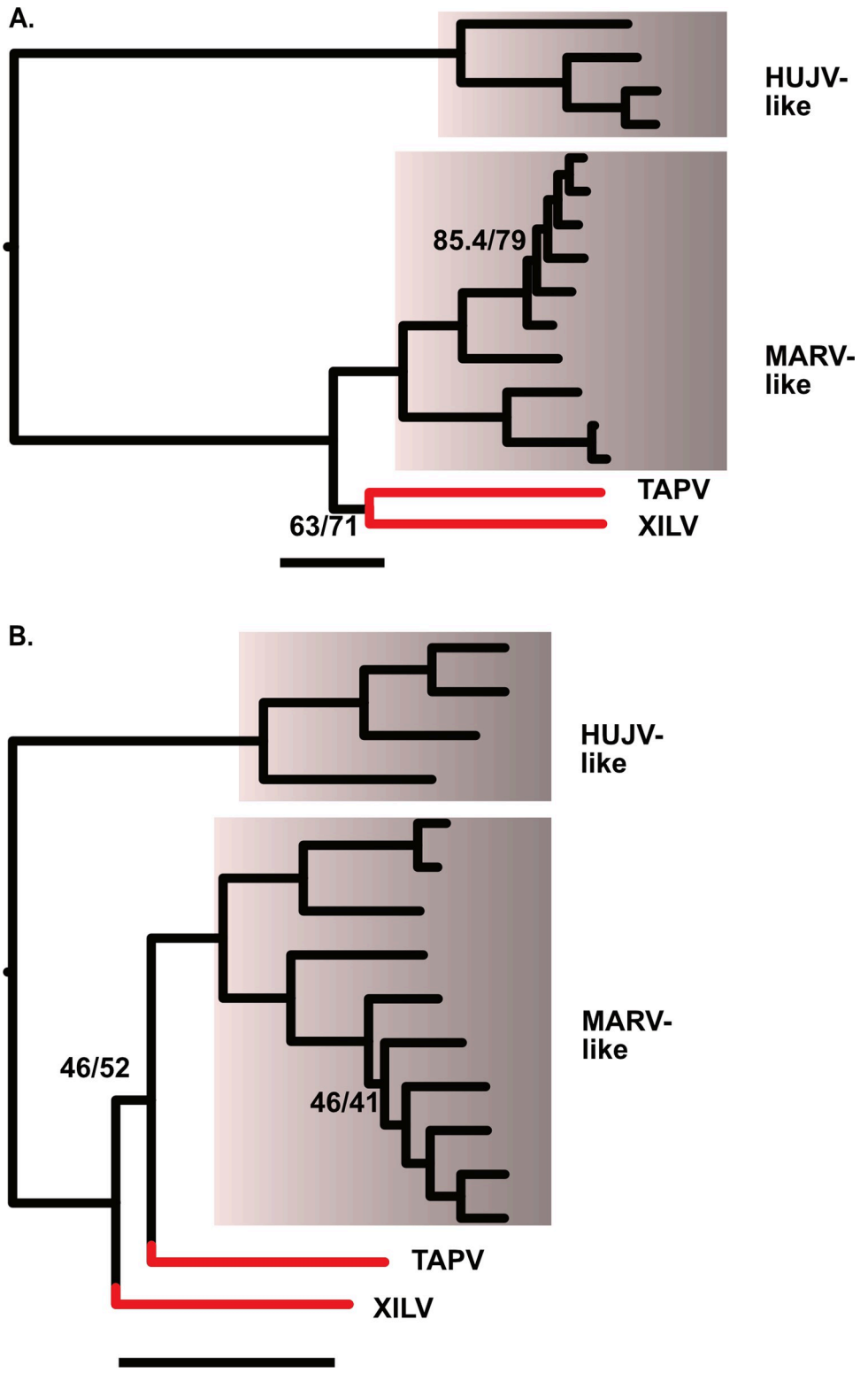

Fig 1. Phylogenies showing relationships of four deep lineages/clades of known filoviruses. The x-axis for each graph is proportional in length to genetic distances (the scale bar is 1 substitution/site). Acronyms are Huángjiāo virus (HUJV), Marburg virus (MARV), Tapajós virus (TAPV), and Xīlǎng virus (XILV). Note that TAPV and XILV form a long branch pair with the amino acid data. Numbers represent branches with the lowest approximate likelihood ratio tests and bootstrap values. The remaining branches had support greater than 96. A. ML tree based on the L Protein

amino acid sequence alignment B. ML tree based on partitioned codon model (nucleotides) for the same alignment as in A.

https://doi.org/10.1371/journal.ppat.1011864.g001

open reading frames that might "break up" long branches. With the NP gene, four deep clades were apparent that we termed HUJV-like, XILV-like, TAPV-like, and MARV-like after the first described filovirus in each clade. Notably, the putative LBA is absent in the NP data as the TAPV-like clade groups with MARV-like sequences associated with mammals with strong support values (Fig 4). Indeed, TAPV groups with open reading frame paleoviruses from three spalacid rodents (bamboo rats, mole-rats, and zokors). The XILV branch is paired with paleoviruses from the genome of the freshwater fish, *Paedocypris* (Cyprinoformes). Switching the data and model for the NP gene to nucleotides with a partitioned codon models increased the support values of the TAPV/spalacid rodent grouping from 59/78 to above 96 (S4 and S5 Figs). The main supported topological difference between AA and codon-based analyses was the movement of the paleoviruses of *Borostomias* from the base of the HUJV-like clade (AA tree) to within the HUJV-like clade (partitioned codon model tree). When the analysis of the NP-like sequences is expanded to include paleoviruses with disrupted reading frames, TAPV groups strongly with mammalian sequences (Figs 5 and S6). Indeed, TAPV is nested within a clade of paleoviruses from shrews (*Sorex* sp.) with strong support values. NP-like paleoviruses from *Acomys* (African spiny mice) are nested within the cricetid rodent clade that is more closely related to EBOV than MARV is (S6 Fig). Indeed, while several groups of mammals are represented in the paleoviruses by single paleoviruses or small monophyletic clades of paleoviruses (vespertilionid bats, a tenrec, anteaters, shrews, opossums, diprotodontid marsupials and tarsiers), filovirus-like sequences are common in every rodent suborder (save Sciuromorpha) and dispersed throughout the NP-like phylogeny (with some groups such as cricetid rodents present in several divergent phylogenetic groups).

For the VP35 gene, the sequences from fish lack significant similarity with the VP35 of the rest of the known filoviruses. Thus, LBA involving fish-associated sequences with other filoviral sequences such as TAPV cannot be assessed. However, the TAPV VP35 does have significant similarity to mammal-associated viruses and groups with paleoviral ORF's from vespertilionid bats (S7 Fig). In the analysis without fish- associated viruses, there was no significant amino acid composition heterogeneity found. The VP35-like elements had an ORF in African spiny mice that groups more closely with EBOV than MARV does. Applying an alignment filter (clipKIT with dynamic determination of gaps) removed 8.79% of sites, increased many support values, but did not significantly alter the topology (S8 Fig). Using a model that accounted for heterotachy (within site variation), altered the branch lengths but did not affect the major groupings or topological placements of paleoviruses (S9 Fig). Using nt with a partitioned model had little change on the topology of the phylogeny, with the exception of the movement of the paleovirus from the bat, *Murina*, moving to a position within the paleoviruses from *Myotis* (S10 Fig). When paleoviruses with disrupted reading frames are included, the topologies are similar to the ORF tree (Fig 6). As with the NP gene tree, *Acomys* VP35-like sequences are nested within cricetid rodents that are in turn more closely related to EBOV-like sequences than to MARV-like sequences. The *Acomys* sequences group with the cricetid ortholog found in the last intron of the TAX1BP1 gene. As with the NP gene, VP35-like paleoviruses in the MARV-like clade are widespread in the genomes of diprotodont marsupials.

Protein structure recapitulates phylogeny

Analyses of predicted protein structures revealed conservation of some NP protein structures from divergent filovirus sequences. For example, the predicted structure of NP from TAPV

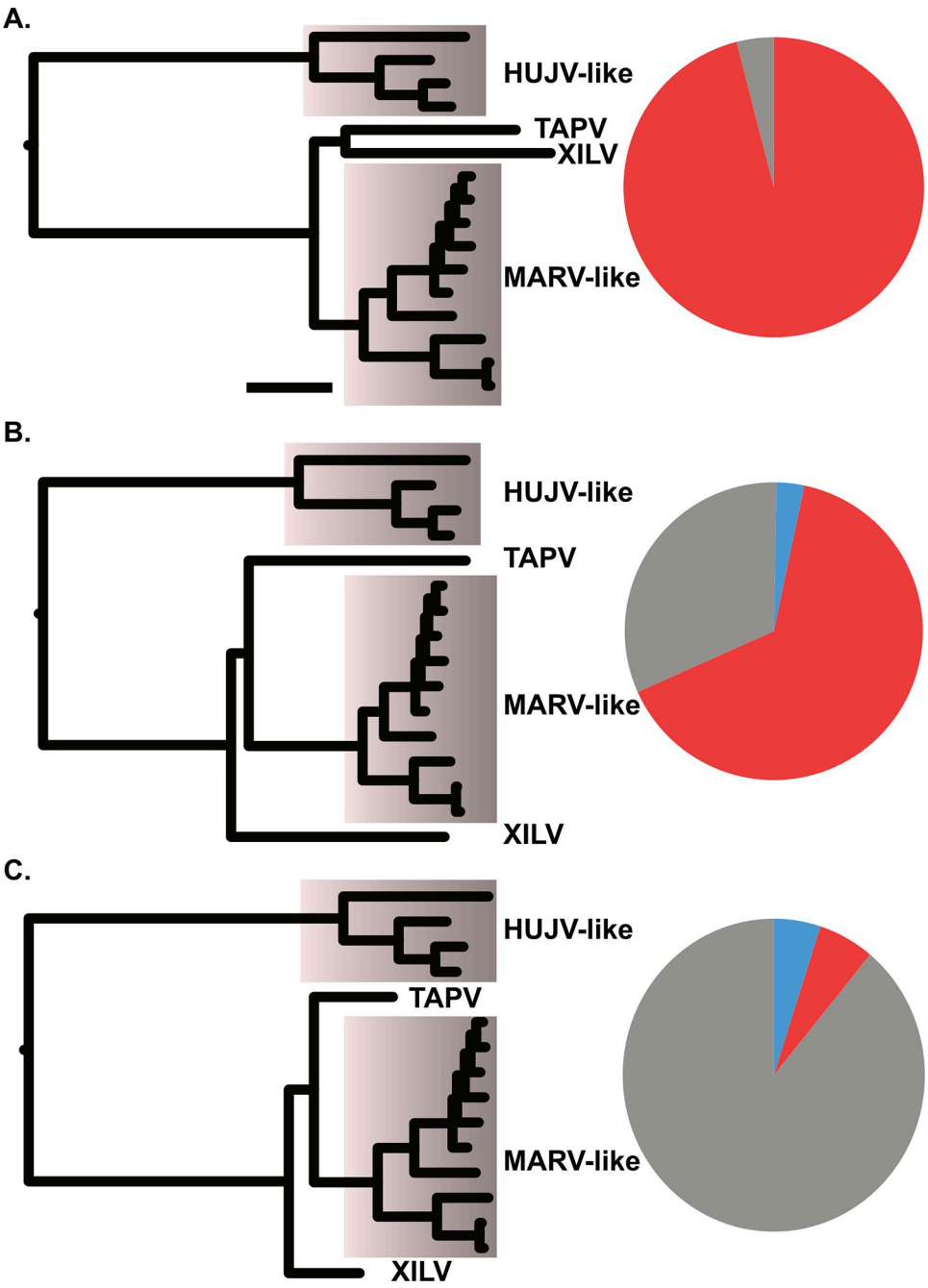

Fig 2. Proportions of topologies observed from maximum likelihood analysis of alignments from parametric simulations that included tree parameters (shown on left side cartoon). Acronyms are Huángjião virus (HUJV), Marburg virus (MARV), Tapajós virus (TAPV), and Xīlǎng virus (XILV). A) a TAPV/XILV sister group, B) a TAPV basal to MARV-like taxa with observed branch lengths, or C) TAPV basal to MARV-like taxa where the branches leading to TAPV and XILV are shortened to 0.5 in length. Red fill on the pie graphs indicates proportion of simulations with a TAPV/XILV group (putative long branch attraction), gray indicates proportion of simulations with a TAPV basal to MARV-like taxa, and blue indicates proportion of topologies observed that differ from red or gray.

https://doi.org/10.1371/journal.ppat.1011864.g002

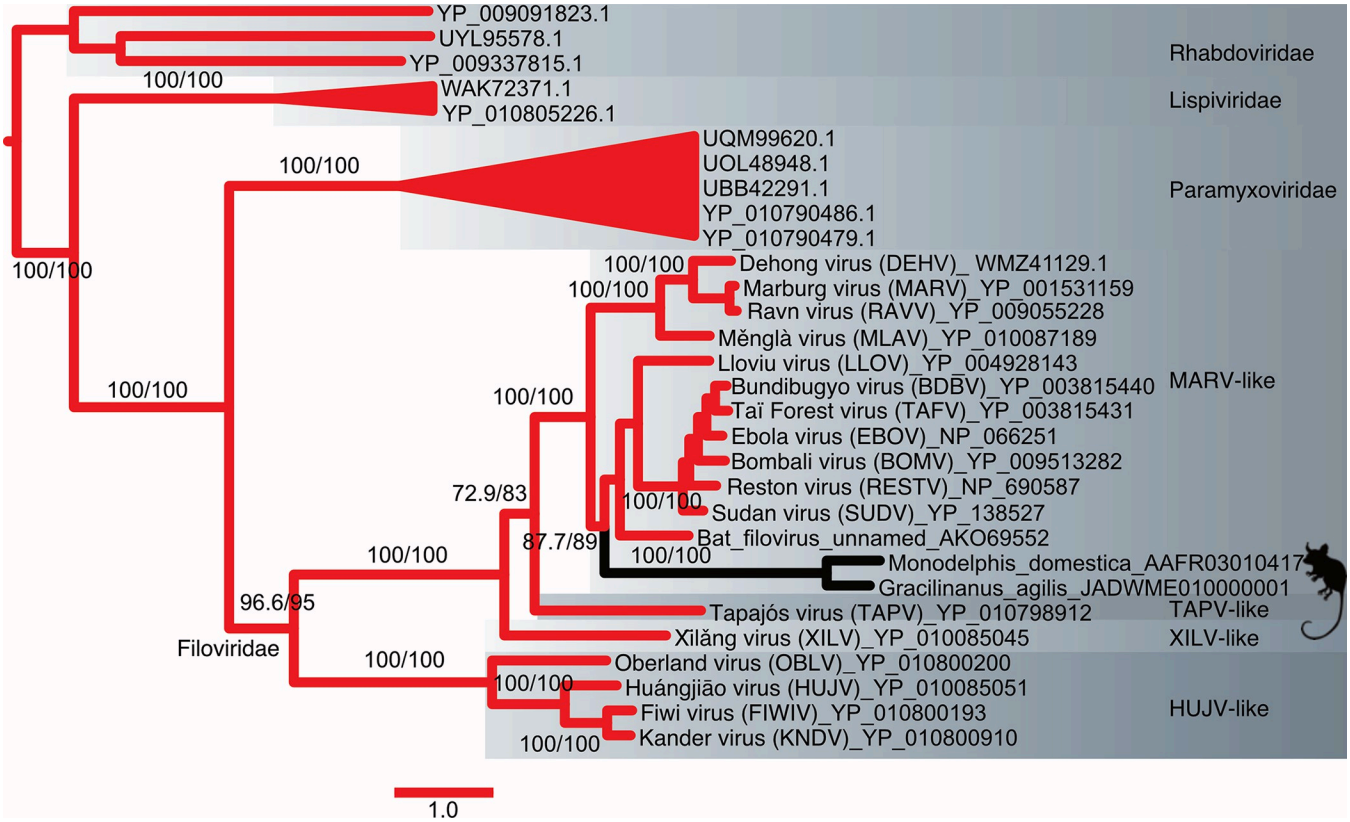

Fig 3. Maximum Likelihood phylogram of L-protein amino acid sequences from filoviruses and filovirus-like paleoviruses with outgroup rooting using sequences of rhabdoviruses. Numbers represent support values from approximate likelihood ratios and ultrafast bootstraps. Major clades of filoviruses are named after the original sequence of each group.

https://doi.org/10.1371/journal.ppat.1011864.g003

was more similar to that of *Nannospalax* (root mean squared deviations, RMSD = 0.916; Fig 7A), than to the predicted NP structure of XILV (RMSD = 1.587; Fig 7B). Indeed, the multidimensional scaling (MDS) analysis of pairwise distances from estimated protein structures largely recapitulated the major groupings found in the phylogenies (Fig 7C). For the NP gene, XILV grouped closely with the paleovirus from the fish *Paedocypris*, while TAPV from a snake again grouped with mammalian sequences from bats (*Myotis*) and spalacid rodents (e.g. *Nannospalax*). Predicted structures from *Acomys* and cricetid rodents grouped with structures of the MARV-like clade. The MDS plot for VP35 structures (S11 Fig) was less resolved than the MDS from the NP structures. Still, in agreement with the other analyses, TAPV was most closely grouped with bat sequences, while *Acomys* grouped closely to MARV-like sequences. The predicted structure based on a paleoviral pseudogene from the hamster (*Phodopus sungorus*) was an outlier and did not group within the MARV-like clade based on sequences alone. We do note that when flexibility was permitted (FATCAT), the structure of the hamster pseudogene had significant similarity to the structure predicted from EBOV VP35(P = 2.56e-13; RMSD = 1.78 with 3 twists).

Candidates for co-opted filovirus-like elements

Several new potential co-options were detected in the present study with filovirus-like elements containing open reading frames present in each major clade. For the XILV-like clade, the fish *Paedocypris*, contained open reading frames of an NP-like element. One near full

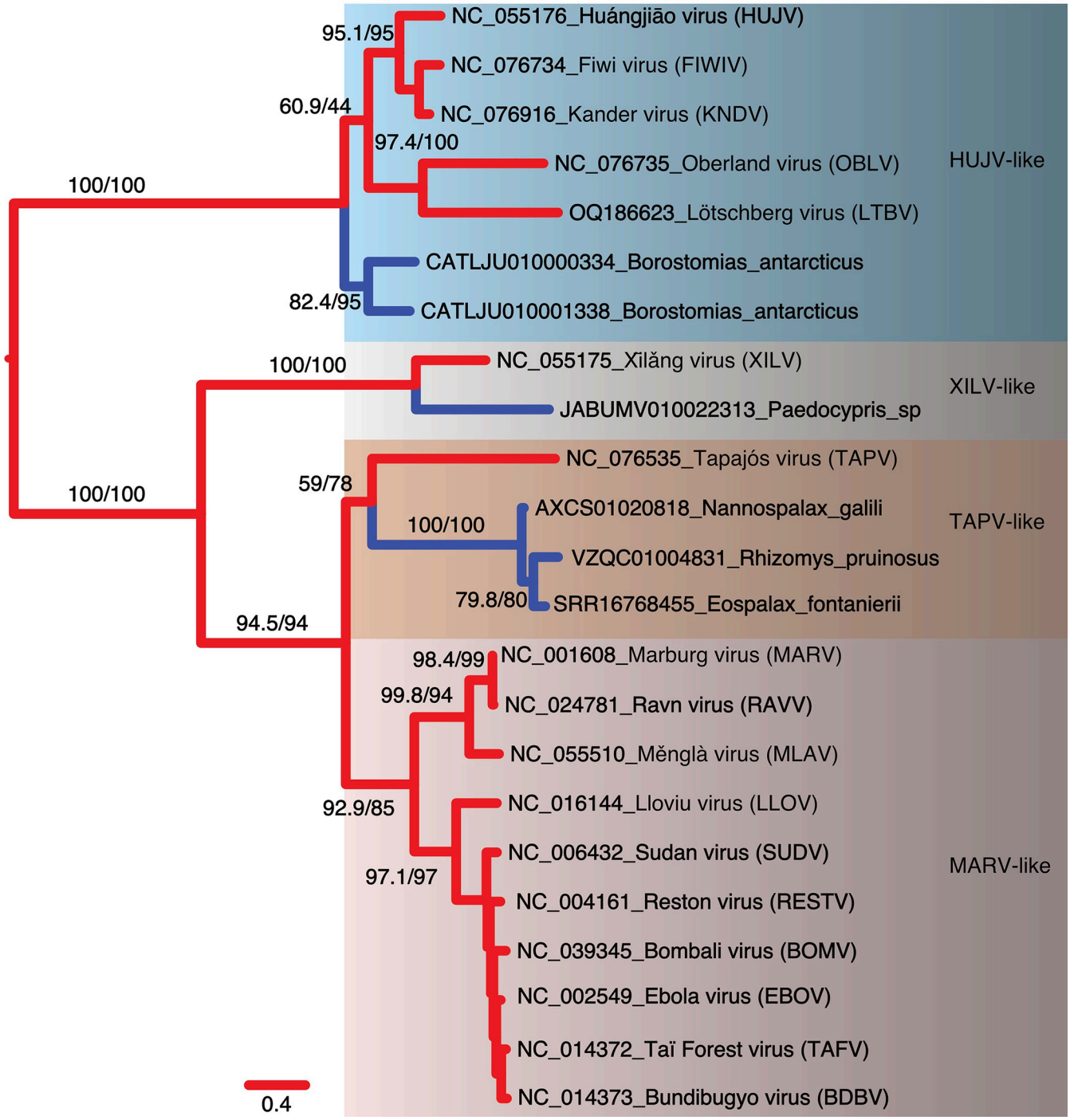

Fig 4. Maximum likelihood phylogram based on amino acid sequences of the nucleoprotein gene for filoviruses and filovirus-like paleoviruses (from vertebrate genomes) with open reading frames (blue lines). Four major clades are identified. Numbers represent approximate likelihood ratio test values and bootstrap values.

https://doi.org/10.1371/journal.ppat.1011864.g004

length XILV NP-like ORF was detected (JABUMV010022313.1) in the genome of a specimen of *Paedocypris* sp. from Singkep Island, Indonesia. As the contig was short (2197 bp), we could not assess the genomic context of the element. However, there are no further gene matches to XILV-like sequences on this contig or in this genome assembly, suggesting that this is ORF is

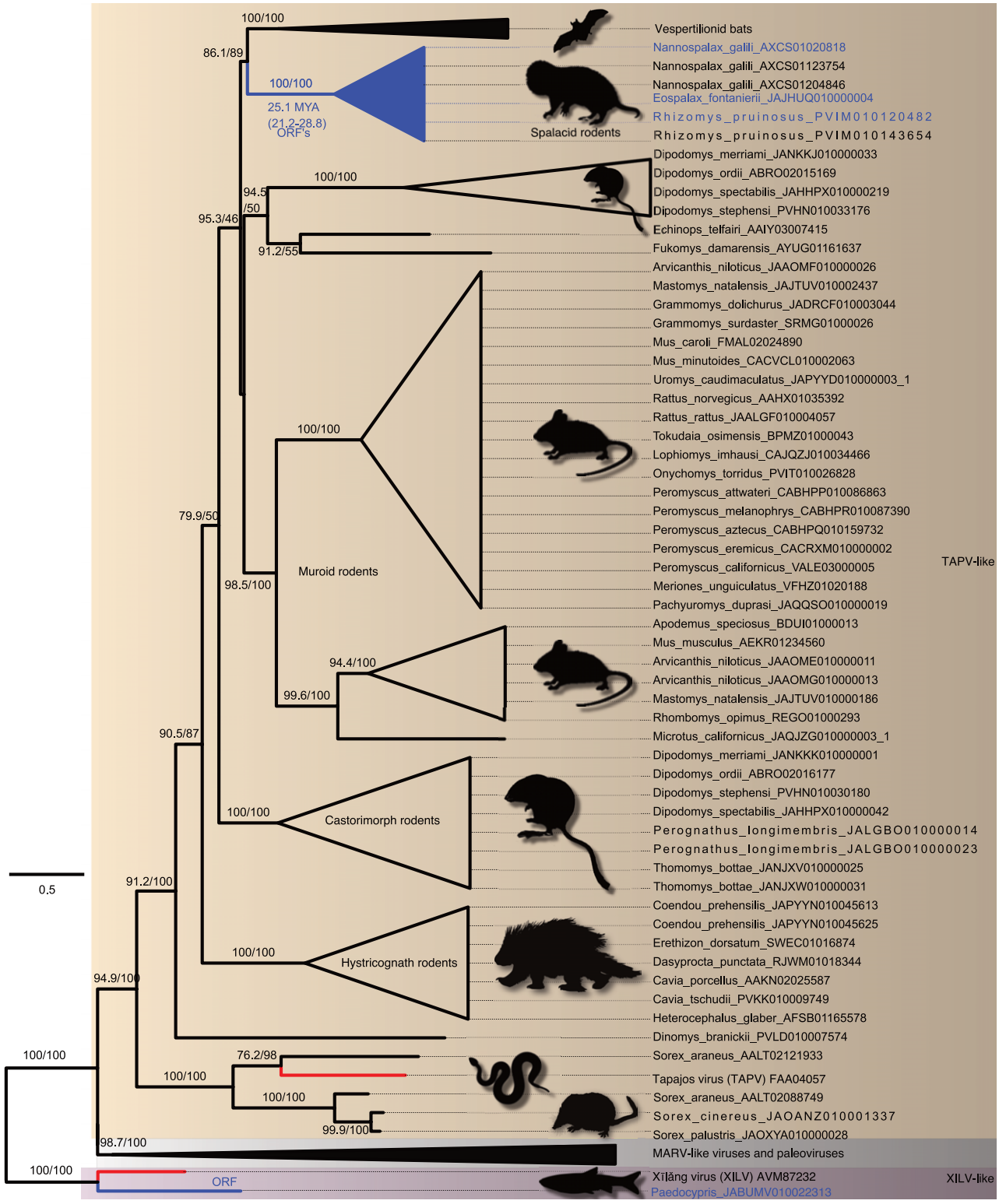

Fig 5. Maximum likelihood phylogram based on the amino acid sequences of the NP or nucleoprotein gene of filoviruses and the NP-like paleoviruses from vertebrates. Black shaded triangles are large clades that were collapsed to save space. Red lines indicate viral lineages, blue lines indicate vertebrate sequences with open reading frames and black lines indicate vertebrate paleoviral sequences that are pseudogenes. Numbers represent approximate likelihood ratio test values and bootstrap values. The scale bar is present. Tree is rooted by XILV and the full expanded tree is presented in S6 Fig.

https://doi.org/10.1371/journal.ppat.1011864.g005

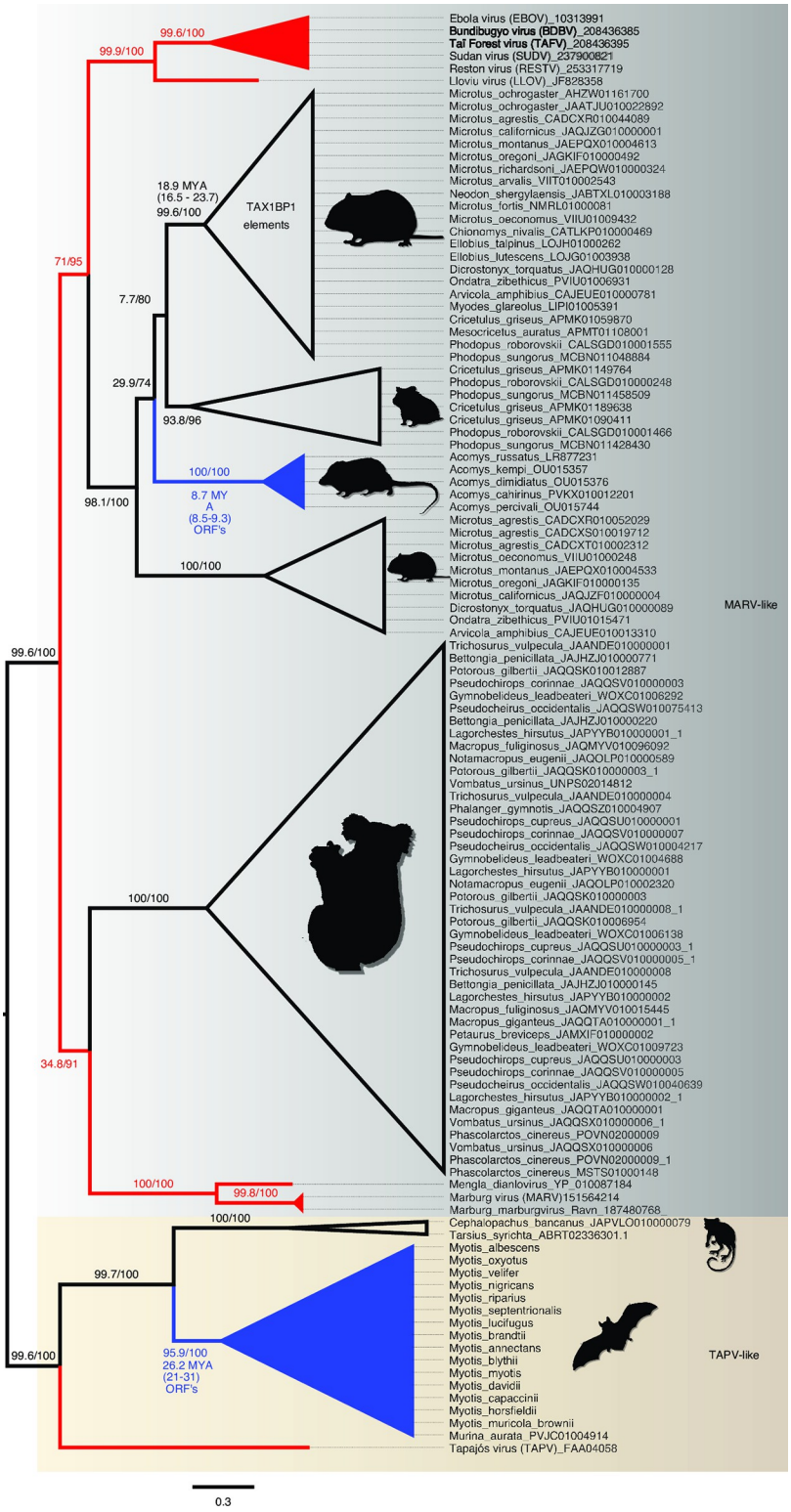

Fig 6. Maximum Likelihood phylogram based on the amino acid sequences of VP35 of filoviruses and the VP35-like paleoviruses from vertebrates. Red lines indicate viral lineages, blue lines indicate vertebrate sequences with open reading frames and black lines indicate vertebrate paleoviral sequences that are pseudogenes. Numbers represent approximate likelihood ratio test values and bootstrap values. Tree is midpoint rooted and two major clades are shown in shaded rectangles. Additional accession numbers for *Myotis* sp. are MH43104.1-MH31036.1, ALWT01033109.1, and ANKR01158691.1.

https://doi.org/10.1371/journal.ppat.1011864.g006

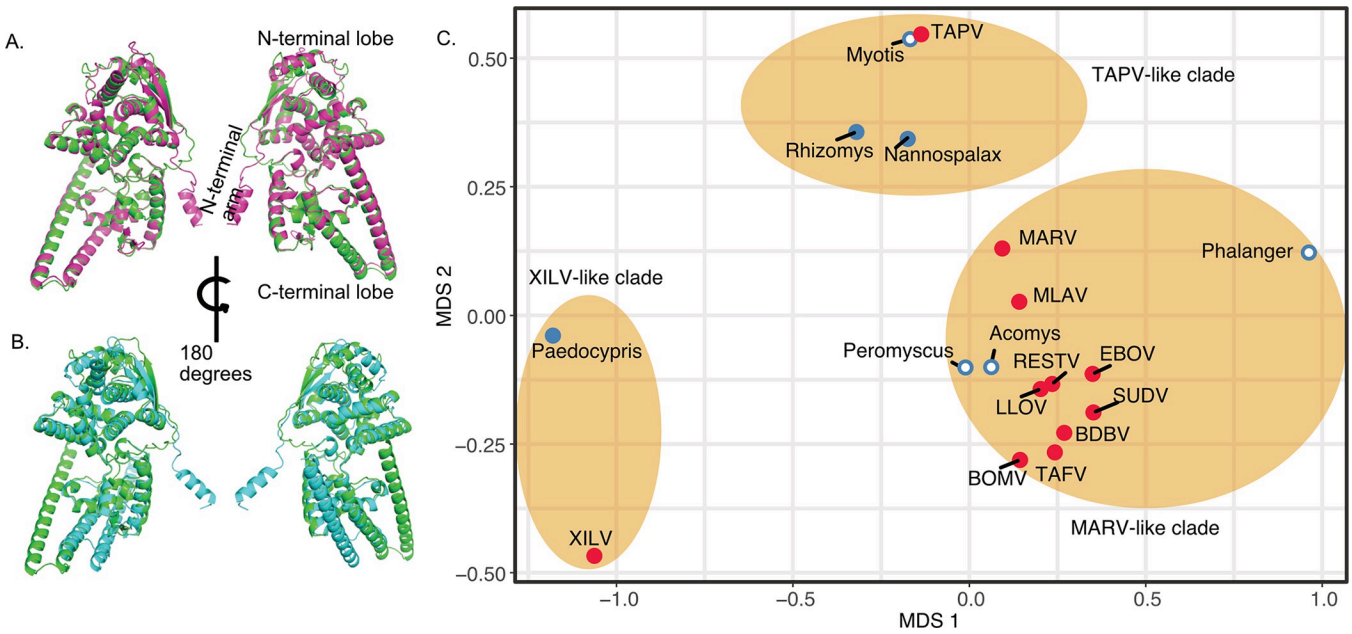

Fig 7. Protein Structure model divergence predicted by alphafold2 and aligned and visualized in PyMol 2.55 for the nucleoprotein gene of filoviruses and filovirus-like sequences in vertebrate genomes. A. Purple cartoon indicates the predicted the structure from the NP-like open reading frame sequence of *Nannospalax* (African mole rat; predicted template modeling score, pTM = 0.83) and green cartoon represents the predicted structure of NP from TAPV (assembled from a lancehead snake, pTM = 0.87) B. Blue cartoon indicates the predicted the structure from the NP-like open reading frame sequence of XILV (a fish-associated filovirus, pTM = 0.77) and the green cartoon represents the predicted structure of NP from TAPV (assembled from a lancehead snake). C. Multidimensional Scaling plot of root mean squared deviations (RMSD) of predicted protein structure of the Nucleoprotein NP from filoviruses and filovirus-like sequences in vertebrates after alignment in PyMol. Ovals indicate major clades found in the phylogenetic analyses presented here. Red shaded stimuli are based on viral structures while blue stimuli are predicted from vertebrate genome sequences. Solid shading indicates extended open reading frames are present.

https://doi.org/10.1371/journal.ppat.1011864.g007

from a genomic element rather than an RNA virus. Chromosome 6 of another species, *Paedocypris micromegethes* (QKNR01000012), had a significant blast match (E = 4e-16, 46% coverage, 29% identity) to the second gene of the XILV genome, (Xilang_striavirus, YP_010085037, presumptive VP35). This match was part of a longer open reading frame (QKNR01000012:17238562–17239635). RNA sequences from the sequence read archive (SRA) of two specimens of *Paedocypris* from Sarawak, Malaysia (SRX8475542-SRX8475543) had 1037 (female) and 685 (male) matched reads to the NP element from Singkep.

We also found paleoviruses with ORFs in the HUJV-clade from the Antarctic snaggletooth fish (*Borostomias antarcticus*). At least fifteen NP-like elements from fifteen different contigs are present in the genome assembly (CATLJU000000000.1), with nine having extended open reading frames and a maximum length of 352 codons (S12 Fig). There are two clades with one group having about 57–66% identity to sequences of the other clade. The only other HUJV-like gene detected was hypothetical protein QKR09_gp3 (**YP_010800189**, Fiwi virus) which matched *Borostomias antarcticus* (CATLJU010001711: 251857–249174, 1e-20, 40.8% identity), but had a disrupted reading frame. No reads from these clades were detected in the transcriptome (SRA: ERX10375716).

In the TAPV-like clade, novel NP-like ORF elements were detected in spalacid rodents (mole rats, bamboo rats and zokors) and the previously described bat VP35-like orthologs with an ORF were present in assemblies from *Myotis* and *Murina* (bats). Along the branches from the common ancestor of *Murina* and *Myotis*, FEL detected 26 sites under significant pervasive purifying selection and 16 sites under diversifying positive selection. In comparison, FEL detected 39 sites under significant pervasive purifying selection and 11 sites under positive

selection in an NP-like with disrupted ORFs along branches from the same common ancestor (*Murina*/*Myotis*). Both the pseudogenic elements (NP-like in bats) and the VP35 ORFs showed site-specific dN/dS distributions consistent with relaxed purifying selection and less diversifying selection. That is, a large peak of sites occurred where dN/dS < 0.5 (S13A and S13B Fig), with a gradual decrease in represented sites between 0.5 and 1 and beyond. The open reading frame was preserved in every bat VP35-like sequence. Only one copy was detected per genome and the genomic context appeared to be the same between inserts in *Myotis* and *Murina*. That is, the inserts appeared in the same local colinear block in the MAUVE alignment of sequences from the two genera (S14 Fig). In the reference genome of *Myotis lucifugus*, the upstream neighboring gene of the VP35-like insert is *TRIM36*, a negative regulator of the interferon response in humans [47].

Spalacid rodents also had ORF elements (NP-like). Here, FEL analysis returned 56 sites under significant purifying selection and 10 sites under positive selection. The distribution of site-specific dN/dS sites for these ORF's was concentrated below a dN/dS of 0.5 (S13C Fig). The aligned ORF region for the three species (*Nannospalax galili*, *Rhizomys pruinosus*, and *Eospalax fontanierii*) was 1320 nt in length with seven apparent indels (each a multiple of three nucleotides in length). The filovirus-like inserts were orthologous based on the nucleotide similarities of flanking regions of rodent contigs with elements: *N. galili* (AXCS01020818.1, 22694 nt) had 81% nt identity to a larger contig with 64% coverage from *R. pruinosus* (VZQC01004831.1). Likewise *N. galili* (AXCS01020818.1) had 83.4% nt identity with 73% coverage using a contig (SRA:SRR16768455.3388524.1) from *E. fontanierii*. On a chromosomal scale MAUVE alignment (Fig 8), the inserts are present in the same local colinear blocks in the three spalacid genomes. The *Kif2a* gene is the upstream neighbor to the NP-like insert (S15 Fig). 1000 parametric simulations of evolution (assuming no purifying selection to maintain codon structure, starting from an ancestral reconstructed sequence of the spalacid NP-like element, and using observed branch length parameters) yielded no cases of alignments that lacked stop codons (Fig 9). Indeed, the mean number of codons detected was 9.2 per alignment. The probability of observing no ORF disruptions for the spalacid paleoviruses under a scenario of pseudogenization is thus very low ($P<0.001$).

We also detected reads for the spalacid filovirus-like elements. For example, searching a transcriptome project in bamboo rats (*R. pruinosus*: 270 G bases, 36 runs, SRP367919 [48]) that separated RNA into three tissue sources (liver, colon, duodenum) using BLASTn, found 60 positive read matches to the NP-like element with 42 (70%) of positive reads being from the

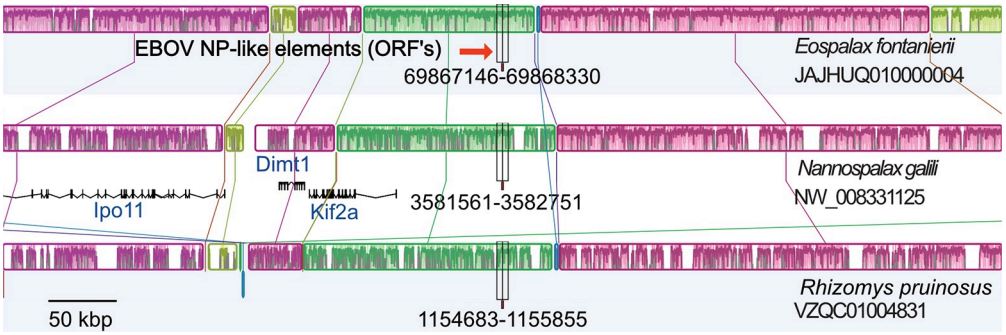

Fig 8. Mauve progressive alignment of chromosomal segments of spalacid rodents showing the positionally homologous region of the filovirus NP-like insertion. Colored boxes are local colinear blocks (aligned regions that lack internal rearrangements). Internal nucleotide similarity plots are shown inside the boxes (higher peaks are more similar). The graphs are positioned at the NP-like insertions (location in each Accession is shown below the insert box) for comparison. Gene tracks are presented for the reference genome of *Nannospalax galili*. Scale bar is 50 kbp.

https://doi.org/10.1371/journal.ppat.1011864.g008

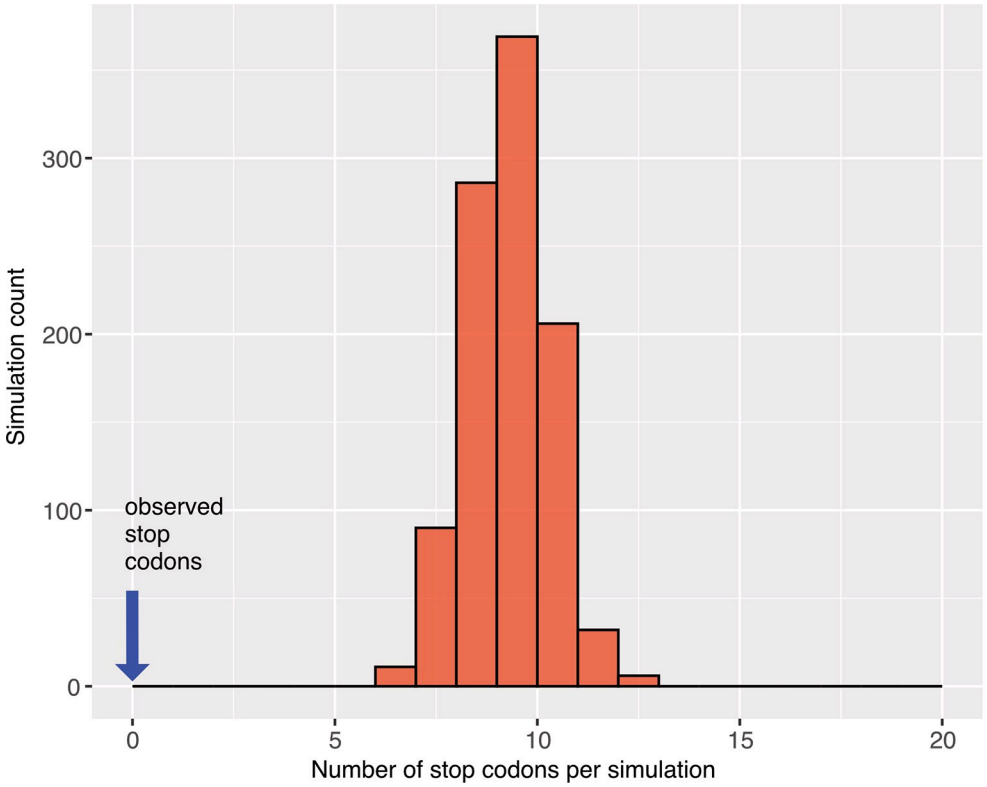

Fig 9. Histogram of stop codon counts from 1000 simulated alignments based on orthologous filovirus NP-like sequences of spalacid rodents. The parametric simulations used an ancestral reconstructed sequence as the starting sequence and the observed tree and substitution model values.

https://doi.org/10.1371/journal.ppat.1011864.g009

liver (complete sequence coverage of the element was achieved). Another study with tissue specific RNA-seq experiments (*Eospalax fontanierii*: SRS3962618, nine experiments for each of four tissue types, 271 Gbases total) had 36 matches (maximum read depth of 11) to the NP-like element from *E. fontanierii* with liver experiments, 4 with brain tissue, 2 with heart and 0 with skeletal tissue. RNA-seq experiments with *Nannospalax galili* also yielded significant matches to its NP-like element. 338 matches (SRP331054, 8 tissue types) were found with top matches being skin (133) and lung (70). RNA extracted from a fourth spalacid taxon, *Tachyorcytes splendens* (SRR2141217), had 10 significant matches to the NP-like element from the related *R. pruinosus* (VZQC01004831.1), resulting in incomplete coverage.

The VP35-like element in the genomes of African spiny mice (*Acomys*) also formed a clade of ORF's. Although the VP35-like ORF's in *Acomys* group phylogenetically with cricetid elements found in the *Tax1bp1* gene, the genomic locations differ. In *Acomys*, the insert is located downstream (142106 bp) of *Casein kinase 1* as part of what was previously annotated as a long noncoding RNA (ENSRACG00000015834 lncRNA; S16 Fig). An RNA-seq study of six tissue types (SRP350516) had the most matches to the VP35-like element with lung (42 reads) and brain (54 reads) tissue.

Unlike with *Acomys*, the related VP35-like elements from cricetid rodents lacked an open reading frame and were present within the *Tax1bp1* gene region (S17 Fig). More specifically, these elements appeared at the same genomic location in the genomes of all examined hamsters, voles, lemmings, and musk rats (S17 Fig), suggesting long term presence in the *Tax1bp1* gene (last intron).

Tax1bp1 expression in EBOV infected mice

The mean expression of TAX1BP1(as measured by exonic RPM) in MA-EBOV infected mouse livers (susceptible CC strains 5 days post infection) was 1.6 times greater than in the mock treatment (single-end reads mapped; S2 Table and S18 Fig). This shift in RPM was greater than expected by chance (Wilcoxon rank-sum test, W = 146, p-value = 5.346e-05).

Discussion

Our results improve our understanding of the deeper evolutionary relationships and host interactions of a group of RNA viruses with dangerous human pathogens. The finding of filovirus-like paleoviruses in fish genomes in each of the major lineages of viruses proposed to be associated with fishes (XILV and HUJV), is consistent with these viruses being piscine rather than the result of more recent jumps from terrestrial hosts. Indeed, the phylogenetic and structural associations between XILV (isolated from a marine frogfish) and the paleoviruses from *Paedocypris* sp. (a freshwater cyprinoform) suggest an ancient association between fish and the XILV-like lineage itself (frogfish and *Paedocypris* shared a common ancestor about 224 MYA; S19 Fig [49]). However, a lack of paleoviral orthologs across genomes precludes estimates of a specific timescale for filoviral elements in teleost fish. The sequences from *Paedocypris* differ from other piscine filovirus-like sequences in being associated with fish from one of the early branching clades of teleosts (Otocephala- herrings, catfishes, cyprinids and others; S19 Fig). The results also suggest that XILV-like viruses infect a broader taxonomic range of bony fish than previously thought. The paleoviruses with open reading frames in both the XILV-like and HUJV-like clades for NP, suggest possible co-options of viral proteins by the host. Indeed, the existence of two divergent clades of elements in the same genome of the Antarctic snaggletooth fish (*Borostomias antarcticus*), both containing extended ORFs, is unique for filovirus-like elements.

Our results support the hypothesis of at least four ancient major clades of filoviruses—each with extant viruses and vertebrate paleoviruses. The results are consistent with the hypotheses that filoviruses have been interacting with vertebrates since the divergence of ray-finned and lobe-finned fishes (more than 400 MYA; S19 Fig). For the NP-like tree, the grouping of a sequence from South American *Dromiciops gliroides* with Australian marsupials agrees with the marsupial species tree [50]. This further supports the hypothesis that a MARV-like filovirus lineage was associated with Neotropical marsupials before the Australian radiation [16,26]. However, the monophyly of marsupial NP-like sequences is prevented by paleoviral sequences from several geographically divergent rodent groups (cricetids, gundis, jerboas) and anteaters. The presence of sequences from murid rodents (e.g. cricetids such as *Peromyscus* sp.) in several disparate positions of the NP-like tree, suggests multiple independent endogenizations and host transfer events.

Our results also suggest that the standard of using amino acid sequence from the RDRP (L-protein), for comparison of divergent viruses can be susceptible to long branch attraction artifacts. Our simulations reveal that XILV and TAPV are prone to LBA for the amino acid data. Using a codon-partitioned nt substitution model appeared to alleviate the LBA. We recommend that evolutionary analyses of divergent RNA viruses include codon-partitioned models. Such data uses the same alignment hypothesis as for amino acids but have three times as much characters and tends to have a higher probability of recovering the correct tree under simulated LBA [45]. Adding divergent outgroups can reduce ingroup LBA (as occurred here with the L protein), but also introduce new sources of bias and LBA [51]. A site-specific model failed to affect the topology (S3 Fig), suggesting that such models may not account for some LBA scenarios (such as lineage specific rate accelerations).

Other lines of evidence also support TAPV grouping with MARV-like sequences (with mammal hosts) and not as sister group to XILV. For example, NP phylogenies (AA and partitioned nt) with outgroup taxa strongly supported a placement of TAPV with the MARV-like sequences. The grouping of TAPV (from a snake) *within* a clade of filovirus-like sequences from shrews (*Sorex* sp.) with strong support, suggests at least one relatively recent host transfer involving snakes and shrews. The nesting of TAPV within a clade of elements from shrews and related to bat elements (i.e., in a clade of elements that reflects the mammal species tree) suggests a transfer from shrew to snake. The opportunity for prey-predator transfer of viruses between snakes and small mammals has been persistent in evolutionary time. Presently, shrews (Soricidae) overlap in geographic distribution with lancehead snakes in the northern Neotropical zone. TAPV structures are most similar to those predicted for paleoviruses of the same phylogenetic clade, bats and spalacid rodents. Taken together, our paleoviral evidence and simulations to address bias, suggest that TAPV is nested within mammal-associated filoviruses, perhaps a result of prey-predator transfers.

TAPV appears to be a viral Lazarus taxon—a recently discovered virus nested within a known ancient clade composed entirely of fossil-like viral elements. The result is consistent with the notion that the study of major paleoviral groupings can inform on the time scale of virus-host interactions, virus discovery and host range. However, the formation of stabilized nonretroviral elements are almost certainly affected by biased processes (as with the taphonomic biases of fossil analogs). Factors such as host effective population size, the activity level of retroelements, viral persistence, and tissue tropisms are but a few potential biases. Host related biases such as effective population size and retroelement activity, however, fail to explain why non-retroviral RNA virus families have very different distributions of paleoviruses in vertebrates. For example, bornavirus-like elements have a broad distribution in the genomes of large eutherian mammals, reptiles and birds where filovirus-like genomic elements are presently unknown [1,52]. Despite the vagaries of formation and maintenance, paleoviruses have, on several occasions, enabled the discoveries of divergent RNA viruses in the expected host taxa. For example, the ancient and widespread association of phasmaviruses with insect hosts, and the finding of totiviruses that infect hosts with a modified genetic code were predicted by paleoviral evidence [8,53].

In general, estimates of divergence based on predicted protein structures recapitulated the phylogenetic associations. This includes the phylogenetic position of paleoviruses from rodents being nested inside the clade containing human filoviral pathogens. Although evolutionary analyses of predicted proteins and phylogenies are based on the same sequences, the two approaches differ in models, sources of error and potential biases. Moreover, viral protein structure is more conserved over evolutionary time than the primary sequences [42].

Our results show that protein structure has likely been highly conserved since fish origins (between XILV, TAPV, and MARV-like clades). Note that the structural predictions shown here from alphafold 2 have limitations as we didn't include the disordered region for the NP sequences or consider RNA binding and interactions among viral proteins. As we used a conservative approach, we also didn't consider structural flexibility for most comparisons. However, the close agreement of the major phylogroups with the structural divergences from viruses and paleoviruses supports the existence of evolutionary signal and overall structural conservation over geological time. Note that the crystal structure of a bat VP35-like paleovirus had strong identity to the structure of VP35 from EBOV [24], indicating the finding of general deep structural conservation of filovirus protein structures is consistent with structural biological evidence. As such, structural paleovirology has the potential to add important information for understanding ancient RNA virus evolution and detecting paleoviruses.

Our results suggest that ebola virus disease can increase expression of TAX1BP1 in the mouse MA-EBOV system (Day five post infection). It is unknown if direct interactions with viral proteins are occurring (as happens with other mononegaviruses). Our results also expand the presence of the *Tax1bp1* VP35-like paleoviruses in rodents from 3 species to 21 species. While this ortholog, found in the last intron of *Tax1bp1*, is non-coding, it is expressed (in the primary transcript and excised introns) and has remained present since at least the Miocene (i.e., since the of the common ancestor of hamsters and voles). It is striking that a rare expressed EBOV-element is present in the *Tax1bp1*gene of these rodent hosts. Their immunity to ebola virus disease is imparted by an antibody response that is dependent on CD4+ T cells [36], which, in turn, is regulated by TAX1BP1 [30]. It is plausible, then, that these elements have an antiviral role, either by affecting the expression of TAX1BP1 or by non-canonical interference [22]. Manipulations or comparisons of *Tax1bp1* inserts in rodents and EVD treatments are needed to address a role these VP35-like inserts. Understanding the details of rodent immune protection to EBOV has direct implications for human medicine and for studies that use these rodents as models for filovirus research.

While viruses of the clade that contains human pathogens are most often associated with bat hosts, isolation of infectious virus from bat hosts has occurred for only MARV and LLOV. So, there is still much unknown about natural host reservoirs for filoviruses. The presence of independently inserted EBOV-like elements suggests that rodents have had significant evolutionary interactions with EBOV-like viruses since the common ancestor of MARV and EBOV. Although we detected reads of VP35-like elements in the transcriptomes of *Acomys*, it is unknown if proteins are produced. The close association of *Acomys* sequences to sequences of LLOV/EBOV did not appear to be due to a bias such as lineage-specific rate differences between mammal and viral sequences. First, an analysis that accounted for heterotachy did not change the phylogenetic position of the filovirus-like elements in *Acomys*. Also, the estimated protein structure distance analysis, which is unaffected by the fit of substitution models, showed a close relationship of the *Acomys* sequences with LLOV/EBOV-like sequences. Under a scenario of heterotachy, we might expect the mammalian sequences to be the outlier and basal to the most recent extant viral clades rather than nested within an extant viral clade (MARV/LLOV/EBOV) as seen here. Finally, the phylogenetic association of *Acomys* filovirus-like elements with LLOV/EBOV is found for both VP35-like genes and NP-like genes. The finding of independent paleoviruses related to EBOV in African and cricetid mouse-like rodents suggests that the clade of filoviruses containing human pathogens is more diverse (even in Africa) than presently known.

The NP-like ORF's present in genomes of spalacid rodents are a unique case of a potential filoviral gene co-option with strong evidence of evolutionary maintenance of ORF's, amino acids, predicted protein structure and detectible RNA-seq reads. As the virus-like elements are flanked by significantly similar genomics contexts, we infer that the integration was present in the common ancestor of spalacids (about 28 MYA [54]). The parametric simulations and site tests for selection suggest that purifying selection is acting to maintain ORF's and many amino acids in the spalacid NP-like orthologs. Our test of ORF maintenance was conservative as our simulations did not include indels—a common ORF disruptor for pseudogenes. While we did detect a low to modest number of reads in the existing RNA-seq data, their biological significance is difficult to assess (especially when the use of paired reads and the scale of RNA-seq experiments is considered). For filoviral elements that have strong evidence for maintenance of protein coding sequence, low counts may reflect pronounced tissue-specific expression. For example, a protein-coding bornaviral element (*hsEBLN-1*) is actively silenced in nearly every tissue type but strongly expressed in testes [11]. Specific experiments with viral infection,

comprehensive tissue sampling, and evidence of paleoviral proteins are required to further assess functional antiviral hypotheses of paleoviruses.

The deeper evolutionary context of RNA viral pathogens is critical for understanding viral diversity, virus-host coevolution and the biology of spillover. Our evolutionary analyses generated hypotheses about gene co-option of filoviral genes with strong evidence of multiple levels of purifying selection and evidence for scenarios of prey-host transfer. We expect that paleoviral elements of pathogenic RNA viruses, such as filoviruses, will continue to inform the biology of virus-host interactions.

Materials and methods

Filovirus and vertebrate genomic sequences were obtained from NCBI and EMBL-EBI. Filoviral taxonomy followed Biedenkopf et al. [55]. Local databases were made of genomic contigs using EBOV protein sequences (from NCBI reference sequence, NC_002549.1) as queries for tBLASTn of the WGS and reference databases with default settings except for "significant" expect scores of E<1.0e-10 and a max hits value of 1000. Taxonomic IDS were used to reduce the size of the searches. Additional queries used protein sequences of Tapajós virus (TAPV, NCBI reference sequence, NC_076535.1), NP-like sequences from contigs (*Myotis myotis* PVIZ010081685.1, *Myodes glareolus* LIPI01005391.1:385698–386592, and *Mesocricetulus auratus* APMT01047719.1), and VP35-like sequences (*Acomys cahirensis*, PVKX01001220.1). Positive contigs were then searched using a tFASTy-based translation search, with the VT200 substitution matrix and the same query sequences. FASTA approaches to pairwise alignment help to reduce paleoviral fractioning [19]. High-scoring segment pairs (HSP's) with expect scores (E<1.0e-10) and greater than 200 amino acids were parsed from the output. We used ORF finder (https://www.ncbi.nlm.nih.gov/orffinder/) to verify the completeness of the ORF and obtain the nucleotide sequences (display ORF as nt option) when a contig had an HSP with an open reading frame. Outgroup sequences for the L protein were chosen based on sequences from three families (excluding filoviruses) with the highest expect values from Blastp comparisons with the L sequence of Xīlǎng virus (XILV, YP_010085045). To examine RNA-seq reads that map to filovirus-like elements with extended ORF's, we exported available experimental runs from the sequence read archive (SRA) to Blast. Queries were conspecific nucleotide sequences of filovirus-like elements. The algorithm parameters were Megablast (highly similar sequence matches) and a maximum target sequences set to 5000.

Alignment

For translation alignment we used MAFFT 7 EINSI as a plugin in Aliview [56,57]. Amino Acid alignments with >200 sequences were carried out by MAFFT using JT200 as a substitution model. Alignment trimming was carried out with either GBLOCKS [58] (least stringent parameters), ClipKIT 2.01 and the smart-gap option [59], or simply removing the disordered region of NP. The flanking sequences of filovirus VP-35-like elements associated with *Tax1bp1* were obtained by BLAST searches using the hamster *Tax1bp1* 3' and Jazf1 region from the reference genome (*Mesocricetus auratus*: NW_024429266.1, 116226539–116233724) as a query for rodent WGS genomes. Chromosomal sequences containing spalacid filovirus NP-like elements (NW_008331125, VZQC01004831, JAJHUQ010000004) and bat VP35-like elements (each with open reading frames) were aligned and visualized using Mauve progressive alignment [60]. Reference genomes were used to identify putative gene regions. For paleoviral ORF's and the rodent *Tax1bp1* insert, orthology was inferred from phylogenetic groupings that matched mammalian taxonomy, and from synteny based on flanking region similarity.

Simulations

Parametric simulations to assess long branch attraction were carried out using the Ali-sim module in IQ-TREE [61]. Three conditions were simulated for the L-protein sequences of filoviruses. A Gblocks alignment filter was used to reduce indels. ML trees and substitution parameters were estimated for the observed topology of the L-protein gene tree where TAPV and XILV are sister taxa. The three simulations had the following constraint trees: TAPV grouping with XILV as in the observed tree (presumptive LBA), TAPV grouping with MARV-like sequences (LBA disrupted) and TAPV grouping with MARV-like sequences (LBA disrupted but branch lengths leading to TAPV and XILV shortened by approximately half). 100 simulations were carried out for each set of parameters. Four branch-specific amino acid compositions (XILV, TAPV, HUJV-like clade and MARV-like clade) were included in the tree parameters and based on empirical values to simulate unequal amino acid compositions across the alignments (see S1 Text for parameters and commands). ML analyses of the simulated alignments for each condition were carried out in IQ-TREE using the "-S" function and a file containing boundaries of simulated alignments. Topologies were tallied and summarized in pie graphs.

Parametric simulations to test for evolutionary retention of the open reading frames were carried out using an ancestral sequence reconstruction of the NP-like elements in spalacid rodents (with a complete open reading frame) estimated using the IQ-TREE -asr option (the TAPV NP sequence was used as an outgroup). Simulations used a tree and nucleotide substitution parameters estimated for the paleoviruses in IQ-TREE. 1000 simulated alignments of 1320 nucleotides were estimated using Seq-gen 1.34 [62] with the specific values: Seq-Gen-1.3.4/source/seq-gen -mGTR -r3.7022 11.3352 1.7544 3.7222 11.0280 1.0 -f0.2884,0.2339,0.2387,0.239 -l1320 -k4 -n1000 -op spalacid_np_only_for_seqgen_tree.phy. The resulting alignments in Phylip format (-op) were concatenated, translated in Aliview. We tallied stop codons and taxon-specific stop codons per simulated alignment. A histogram of the stop codons was created using ggplot2 in R.

Phylogenetic methods

Phylogenies and substitution models were estimated using Maximum Likelihood and IQ-TREE 2.2.2.6 [63]. Models partitioned by codon position used a partition file and the -p command. Trees were visualized in Figtree 1.44. Branch support was estimated by approximate likelihood ratios (aLRTs) and ultrafast bootstraps. Site-specific frequency models (PMSF using C60, the ML version of Bayesian CAT models) in IQ-TREE were used with a guide tree parameter to examine the role of site-specific frequency effects on topology. The effect of heterotachy (within-site variation) on tree topology was examined using the GHOST (General Heterogeneous evolution On a Single Topology) model implemented in IQ-TREE [64]. Divergence times of filovirus-associated vertebrates were estimated from a dated phylogeny based on the output of Time tree [49].

We used the Fixed Effects Likelihood (FEL) routine in HyPhy to assess site-specific patterns in selection with the default threshold of significance of $P <= 0.1$ [65]. Additional sequences from previous sequencing of bat VP35-like and NP-like elements [23,24] were added to those assembled here. Test clades were selected, translation aligned in Aliview and calculations of dN/dS were made for partial NP sequences (several bat sequences were available for partial NP). NP and VP35 bat sequences were chosen to compare paleoviruses with and without an open reading frame over the same evolutionary time scale. Because FEL requires an open reading frame, we corrected stop codons and replaced disrupting indels with consensus nucleotides for paleoviruses that were pseudogenes. Kernal density estimate plots of dN/dS distributions for three clades of NP-like sequences were made in Datamonkey.

Structural comparisons

Protein structures were predicted with Colabfold:AlphaFold2 using MMseqs2 and a PDB100 template mode [66]. Structure models with the highest per-residue estimate of confidence (pLDDT) were aligned in PyMol 2.55 [67] and the average distance between atoms was calculated using pairwise root mean square deviation (RMSD). Divergent RMSD's were tested for significant similarity using a flexible protein structure alignment algorithm in FATCAT [68]. Representative paleoviruses for major clades were used for structural analyses if they had open reading frames or, if pseudogenized, had the highest FASTA expect values compared to EBOV. A pairwise matrix of the RMSD's was exposed to multidimensional scaling in R and the resulting values were plotted using ggplot [69] and compared with phylogenetic clade.

Tax1bp1 expression in mice

To assess if Tax1bp1 shows differential expression in rodents with ebolavirus disease we downloaded publicly available reads (S2 Table [70]) from the sequence read archive (SRA) and exposed them to RNA seq analysis (CLC genomics workbench 24, Qiagen). We chose day 5 post infection livers to examine as EBOV targets liver and is present in high copy number at day 5 in mice [70]. Reads were mapped onto the exonic regions of the Tax1bp1 nucleotide sequence (NM_001355596.1) using strains of mice (*Mus musculus*) that are susceptible to ebolavirus disease. Publicly available single-end reads mapped per million (RPM) from liver tissue experiments (day 5 post infection with mouse-adapted EBOV, an evolved virus that can produce ebolavirus disease in adult rodents; S2 Table) were compared to RPM values from mock treatments [70]. A nonparametric Wilcoxon rank-sum test was used to test the null hypothesis of no shift in the distribution of expressions from mock ebolavirus treatments.

Supporting information

S1 Fig. Maximum likelihood phylogram based on nucleotide sequences of the L gene (RDRP) for filoviruses. The substitution model was partitioned by the three codon positions. Genbank accession numbers are part of tip names. Numbers on branches represent approximate likelihood ratio test values and bootstrap values.
(PDF)

S2 Fig. Maximum likelihood phylogram based on partitioned nucleotide sequences of the L gene (RDRP) for filoviruses. The substitution model was partitioned by the first two codon positions, while the third codon position was omitted. Genbank accession numbers are part of tip names. Numbers on branches approximate likelihood ratio test values and bootstrap values.
(PDF)

S3 Fig. Maximum likelihood phylogram based on Posterior Mean Site Frequency Profiles (PMSF) for amino acids of filoviruses. Genbank accession numbers are part of tip names. Numbers on branches approximate likelihood ratio test values and bootstrap values.
(PDF)

S4 Fig. Maximum likelihood phylogram based on nucleotide sequences of the nucleoprotein gene (NP) for filoviruses. The substitution model was partitioned by the first two codon positions with third codon positions being omitted. Genbank accession numbers are part of tip names. Numbers on branches approximate likelihood ratio test values and bootstrap values. Blue lines indicate branches leading to paleoviruses from vertebrate genomes with extended

open reading frames.
(PDF)

S5 Fig. Maximum likelihood phylogram based on nucleotide sequences (all codon positions) of the nucleoprotein gene (NP) for filoviruses. The substitution model was partitioned by three codon positions. Genbank accession numbers are part of tip names. Numbers on branches approximate likelihood ratio test values and bootstrap values. Blue lines indicate branches leading to paleoviruses from vertebrate genomes with extended open reading frames.
(PDF)

S6 Fig. Maximum Likelihood phylogram based on the amino acid sequences of the nucleoprotein gene (NP) of filoviruses and the NP-like paleoviruses from vertebrates. Red lines indicate viral lineages, blue lines indicate vertebrate sequences with open reading frames and black lines indicate vertebrate paleoviral sequences that have disrupted open reading frames. Numbers on branches represent approximate likelihood ratio test values and bootstrap values. Three major clades are shown in shaded rectangles.
(PDF)

S7 Fig. Maximum likelihood phylogram based on amino acid sequences of the VP35 gene for filoviruses and filovirus-like paleoviruses (from vertebrate genomes) with open reading frames (blue lines). Two major clades (MARV-like and TAPV-like) are identified and shaded. Numbers represent approximate likelihood ratio test values and bootstrap values. Genbank accession numbers are part of tip names. Additional accession numbers for Myotis sp. are MH431024.1-MH431036.1, ALWT01033109.1, and ANKR01158691.1.
(PDF)

S8 Fig. Maximum likelihood phylogram based on filtered amino acid sequences of the VP35 gene for filoviruses and filovirus-like paleoviruses (from vertebrate genomes) with open reading frames (blue lines). The alignment was filtered using clipKIT. Two major clades (MARV-like and TAPV-like) are identified and shaded. Numbers represent approximate likelihood ratio test values and bootstrap values. Additional accession numbers for Myotis sp. are MH431024.1-MH431036.1, ALWT01033109.1, and ANKR01158691.1.
(PDF)

S9 Fig. Maximum likelihood phylogram based on amino acid sequences of the VP35 gene for filoviruses and filovirus-like paleoviruses (from vertebrate genomes) with open reading frames. The tree was midpoint rooted and based on a substitution model that specifically accounts for heterotachy (within site rate variation). Additional accession numbers for Myotis sp. are MH431024.1-MH431036.1, ALWT01033109.1, and ANKR01158691.1.
(PDF)

S10 Fig. Maximum likelihood phylogram based on nucleotide sequences of the VP35 gene for filoviruses and filovirus-like paleoviruses (from vertebrate genomes) with open reading frames (blue lines). The substitution model was partitioned by codon position. Two major clades are identified. Numbers represent approximate likelihood ratio test values and bootstrap values. Additional accession numbers for Myotis sp. are MH431024.1-MH431036.1, ALWT01033109.1, and ANKR01158691.1.
(PDF)

S11 Fig. Multidimensional Scaling plot of root mean squared deviations (RMSD) of predicted protein structure of the VP35 from filoviruses and filovirus-like sequences in vertebrates after alignment in PyMol. Ovals indicate major clades found in phylogenetic analyses.

Red shaded stimuli are based on viral structures while blue stimuli are predicted from vertebrate genome sequences. Solid shading indicates open reading frames are present.
(PDF)

S12 Fig. Multiple sequence alignment (amino acids) of NP-like elements from the genome of the fish, large-eye snaggletooth (*Borostomias antarcticus*). Stop codons are depicted by an asterisk. Nine of the fifteen sequences are open reading frames.
(PDF)

S13 Fig. Kernal density estimates of dN/dS per site (calculated by the fixed effects likelihood method or FEL) in the alignments of filovirus-like elements of vertebrate genomes. The black vertical bar indicates a neutral ratio. A) estimates from NP-like elements with reading frame disruptions of bats from *Murina* and *Myotis*; B) estimates from filovirus VP35-like elements (extended ORFs) in genomes of bats (*Murina* and *Myotis*); C) estimates from the filovirus NP-like elements in spalacid rodents with an open reading frame and expression products.
(PDF)

S14 Fig. Mauve progressive alignment of assembly segments of bat genomes showing the positionally homologous region of the filovirus VP35-like insertion. Colored boxes are local colinear blocks (aligned regions that lack internal rearrangements). Internal nucleotide similarity plots are shown inside the boxes (higher peaks are more similar). The graphs are positioned at the VP35-like insertions (location in each Accession is shown below the insert box) for comparison. Gene tracks are presented for the reference genome of *Myotis lucifugus*. Scale bar is 50 kbp.
(PDF)

S15 Fig. Genomic context of the filovirus NP-like insertion in the reference genome of the spalacid rodent, *Nannospalax galili*. Gene tracks are presented below in green. Scale bar is above the green bar.
(PDF)

S16 Fig. Genomic context of the filovirus VP35-like insertion in the reference genome (NC_067156) of the spiny mouse, *Acomys russatus*. Gene tracks from NCBI Refseq and Ensembl are presented below in green and gray. Scale bar is above the dark gray bar.
(PDF)

S17 Fig. Cartoon of the 3' end of the *Tax1bp1* gene showing microsynteny of filovirus VP35-like elements (taxon names and accession numbers highlighted by pink rectangle) in the genomes of cricetine (hamsters) and arvicoline (voles, lemmings, muskrats) rodents. Dark shaded bars indicate exonic regions and gray shaded bars indicate the 3' intron of Tax1bp1. The red bar below the alignment indicates region that has significant similarity to VP35 protein sequences of pathogenic filoviruses. Vertical bars above the alignment indicate sequence similarity (including differences in sequences that lack the insert). Genomic regions of four muroid rodents that lack the insert are shown for comparison.
(PDF)

S18 Fig. Box and whiskers plot of *Tax1bp1* exonic read maps under two treatments in ebola virus susceptible Collaborative Cross strains of mice. MA-EBOV are reads from livers of mice (5 days post infection with mouse adapted ebolavirus). Mock results are reads from livers of the same strains (5 days post mock infection). Reads are from SRA Project PRJNA540840 and the mapping results for each mouse are presented in S2 Table. The Y-axis

is a normalized read map score, reads assigned per million reads (RPM).
(PDF)

S19 Fig. Time tree of Filovirus-associated vertebrates. Boxes indicate major groups of vertebrates (teleost fish, reptiles, marsupials and eutherians). Vertical gray lines indicate 100 million-year intervals. Red branches lead to hosts with extant viral lineages, blue lines indicate vertebrates with paleoviral lineages that have extended open reading frames and black lines indicate vertebrates with paleoviral lineages that have only disrupted open reading frames.
(PDF)

S1 Text. Parameters and commands for simulations to test for long branch attraction for the L protein gene of filoviruses using IQtree2.
(TXT)

S1 Table. Base compositional heterogeneity test of the L protein amino acid sequences from Filoviruses.
(PDF)

S2 Table. Summary of *Tax1bp1* exonic read maps under two infection treatments in ebola virus susceptible Collaborative Cross (CC) strains of mice. MA_EBOV are reads from livers of mice (5 days post infection with mouse adapted ebolavirus). Mock results are reads from livers of the same strains (5 days post mock infection). Reads are from SRA Project PRJNA540840. RPM is a normalized read map score, reads assigned per million reads (RPM).
(PDF)

S1 Data. Pairwise root-mean-square deviations of atomic positions (RMSDs) of NP and NP-like sequences (without the terminal disordered region) from filoviruses and vertebrate genomes. The distance matrix was used for S12 Fig.
(PDF)

S2 Data. Pairwise root-mean-square deviations of atomic positions (RMSDs) of VP35 and VP35-like sequences (without the terminal disordered region) from filoviruses and vertebrate genomes. The distance matrix was used for Fig 6C.
(PDF)

S3 Data. Multiple sequence alignment of L protein and L protein-like sequences from filoviruses and filovirus-like elements in vertebrate genomes. The alignment was trimmed using Clipkit 2. The data are used for the analysis in Fig 3.
(TXT)

S4 Data. Multiple sequence alignment of NP and NP-like sequences (nucleotides) from filoviruses and open reading frame (ORF) filovirus-like elements in vertebrate genomes. The data were used for the analysis in S5 Fig and related to Fig 4 (amino acids).
(TXT)

S5 Data. Multiple sequence alignment of VP35 and VP35-like sequences (nucleotides) from filoviruses and open reading frame (ORF) filovirus-like elements in vertebrate genomes. Additional accession numbers for Myotis sp. are MH43104.1-MH31036.1, ALWT01033109.1, and ANKR01158691.1.The data were used for the analysis in S7 Fig and related analyses.
(TXT)

S6 Data. NP and NP-like sequences (amino acids) from filoviruses and filovirus-like elements in vertebrate genomes (from tFasty results). The data were used for the analysis in Fig

5 and related analyses.
(TXT)

S7 Data. Genomic locations and Accession numbers of vertebrate filovirus NP-like elements detected by FASTA.
(TXT)

S8 Data. Genomic locations and Accession numbers of vertebrate filovirus VP35-like elements detected by FASTA. See genome context figures for locations of species with ORFs.
(TXT)

Author Contributions

Conceptualization: Derek J. Taylor, Max H. Barnhart.

Formal analysis: Derek J. Taylor, Max H. Barnhart.

Investigation: Derek J. Taylor.

Software: Derek J. Taylor, Max H. Barnhart.

Visualization: Derek J. Taylor.

Writing – original draft: Derek J. Taylor, Max H. Barnhart.

Writing – review & editing: Derek J. Taylor, Max H. Barnhart.

References

1. Kawasaki J, Kojima S, Mukai Y, Tomonaga K, Horie M. 100-My history of bornavirus infections hidden in vertebrate genomes. Proc Natl Acad Sci U S A. 2021; 118(20). Epub 2021/05/16. https://doi.org/10.1073/pnas.2026235118 PMID: 33990470; PubMed Central PMCID: PMC8157955.
2. Chen YM, Hu SJ, Lin XD, Tian JH, Lv JX, Wang MR, et al. Host traits shape virome composition and virus transmission in wild small mammals. Cell. 2023. Epub 20230914. https://doi.org/10.1016/j.cell.2023.08.029 PMID: 37734372.
3. Cui X, Fan K, Liang X, Gong W, Chen W, He B, et al. Virus diversity, wildlife-domestic animal circulation and potential zoonotic viruses of small mammals, pangolins and zoo animals. Nat Commun. 2023; 14 (1):2488. Epub 20230429. https://doi.org/10.1038/s41467-023-38202-4 PMID: 37120646; PubMed Central PMCID: PMC10148632.
4. Hou X, He Y, Fang P, Mei S-Q, Xu Z, Wu W-C, et al. Artificial intelligence redefines RNA virus discovery. bioRxiv. 2023; https://doi.org/10.1101/2023.04.18.537342
5. Hu S.NT. Filovirus helical nucleocapsid structures. Microscopy (Oxf). 2023; 72(3):178–90. Epub 2023/06/08. https://doi.org/10.1093/jmicro/dfac049 PMID: 36242583.
6. Shi M, Lin XD, Chen X, Tian JH, Chen LJ, Li K, et al. The evolutionary history of vertebrate RNA viruses. Nature. 2018; 556(7700):197–202. Epub 20180404. https://doi.org/10.1038/s41586-018-0012-7 PMID: 29618816.
7. Fujino K, Horie M, Kojima S, Shimizu S, Nabekura A, Kobayashi H, et al. A Human Endogenous Bornavirus-Like Nucleoprotein Encodes a Mitochondrial Protein Associated with Cell Viability. J Virol. 2021; 95(14):e0203020. https://doi.org/10.1128/JVI.02030-20 PMID: 33952640
8. Taylor DJ, Ballinger MJ, Bowman SM, Bruenn JA. Virus-host co-evolution under a modified nuclear genetic code. PeerJ. 2013; 1:e50. https://doi.org/10.7717/peerj.50 PMID: 23638388
9. Warner BE, Ballinger MJ, Yerramsetty P, Reed J, Taylor DJ, Smith TJ, et al. Cellular production of a counterfeit viral protein confers immunity to infection by a related virus. PeerJ. 2018; 6:e5679. Epub 20180928. https://doi.org/10.7717/peerj.5679 PMID: 30280045; PubMed Central PMCID: PMC6166632.
10. Huang HJ, Li YY, Ye ZX, Li LL, Hu QL, He YJ, et al. Co-option of a non-retroviral endogenous viral element in planthoppers. Nat Commun. 2023; 14(1):7264. Epub 20231109. https://doi.org/10.1038/s41467-023-43186-2 PMID: 37945658; PubMed Central PMCID: PMC10636211.

11. Sofuku K, Parrish NF, Honda T, Tomonaga K. Transcription Profiling Demonstrates Epigenetic Control of Non-retroviral RNA Virus-Derived Elements in the Human Genome. Cell Rep. 2015; 12(10):1548–54. Epub 20150828. https://doi.org/10.1016/j.celrep.2015.08.007 PMID: 26321645.
12. Horie M. Identification of a novel filovirus in a common lancehead (Bothrops atrox (Linnaeus, 1758)). J Vet Med Sci. 2021; 83(9):1485–8. Epub 20210719. https://doi.org/10.1292/jvms.21-0285 PMID: 34275961; PubMed Central PMCID: PMC8498845.
13. Suzuki Y, Gojobori T. The origin and evolution of Ebola and Marburg viruses. Mol Biol Evol. 1997; 14 (8):800–6. https://doi.org/10.1093/oxfordjournals.molbev.a025820 PMID: 9254917
14. Carroll SA, Towner JS, Sealy TK, McMullan LK, Khristova ML, Burt FJ, et al. Molecular Evolution of Viruses of the Family Filoviridae Based on 97 Whole-Genome Sequences. J Virol. 2013; 87(5):2608–16. https://doi.org/10.1128/JVI.03118-12 PMID: 23255795
15. Belyi VA, Levine AJ, Skalka AM. Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes. PLoS pathogens. 2010; 6(7): e1001030. https://doi.org/10.1371/journal.ppat.1001030 PMID: 20686665
16. Taylor DJ, Leach RW, Bruenn J. Filoviruses are ancient and integrated into mammalian genomes. BMC Evol Biol. 2010; 10(1):193. https://doi.org/10.1186/1471-2148-10-193 PMID: 20569424
17. Negredo A, Palacios G, Vázquez-Morón S, González F, Dopazo H, Molero F, et al. Discovery of an ebolavirus-like filovirus in europe. PLoS pathogens. 2011; 7(10):e1002304. https://doi.org/10.1371/journal.ppat.1002304 PMID: 22039362
18. He B, Feng Y, Zhang H, Xu L, Yang W, Zhang Y, et al. Filovirus RNA in Fruit Bats, China. Emerg Infect Dis. 2015; 21(9):1675–7. https://doi.org/10.3201/eid2109.150260 PMID: 26291173; PubMed Central PMCID: PMC4550138.
19. Taylor DJ, Ballinger MJ, Zhan JJ, Hanzly LE, Bruenn JA. Evidence that ebolaviruses and cuevaviruses have been diverging from marburgviruses since the Miocene. PeerJ. 2014; 2:e556. https://doi.org/10.7717/peerj.556 PMID: 25237605
20. Ng M, Ndungo E, Kaczmarek ME, Herbert AS, Binger T, Kuehne AI, et al. Filovirus receptor NPC1 contributes to species-specific patterns of ebolavirus susceptibility in bats. Elife. 2015;4. Epub 20151223. https://doi.org/10.7554/eLife.11785 PMID: 26698106; PubMed Central PMCID: PMC4709267.
21. Harding EF, Russo AG, Yan GJH, Waters PD, White PA. Ancient viral integrations in marsupials: a potential antiviral defence. Virus Evol. 2021; 7(2):veab076. Epub 2021/09/23. https://doi.org/10.1093/ve/veab076 PMID: 34548931; PubMed Central PMCID: PMC8449507.
22. Ogawa H, Honda TViral Sequences Are Repurposed for Controlling Antiviral Responses as Non-Retroviral Endogenous Viral Elements. Acta Medica Okayama. 2022; 76(5):503–10. https://doi.org/10.18926/AMO/64025 PMID: 36352796
23. Taylor DJ, Dittmar K, Ballinger MJ, Bruenn JA. Evolutionary maintenance of filovirus-like genes in bat genomes. BMC Evol Biol. 2011; 11(1):336. https://doi.org/10.1186/1471-2148-11-336 PMID: 22093762
24. Edwards MR, Liu H, Shabman RS, Ginell GM, Luthra P, Ramanan P, et al. Conservation of Structure and Immune Antagonist Functions of Filoviral VP35 Homologs Present in Microbat Genomes. Cell Rep. 2018; 24(4):861–72 e6. https://doi.org/10.1016/j.celrep.2018.06.045 PMID: 30044983; PubMed Central PMCID: PMC6474348.
25. Levantis I. A phylogenomic study characterising the co-option and evolutionary history of endogenous viral elements in bats (order Chiroptera): Queen Mary University of London; 2021.
26. Blanco-Melo D, Campbell MA, Zhu H, Dennis TPW, Modha S, Lytras S, et al. A novel approach to exploring the dark genome and its application to mapping of the vertebrate virus 'fossil record'. bioRxiv. 2023:2023.10.17.562709. https://doi.org/10.1101/2023.10.17.562709
27. Kuhn JH, Schmaljohn CS. Of mice and Mike—An underappreciated Ebola virus disease model may have paved the road for future filovirology. Antiviral Res. 2023; 210:105522. https://doi.org/10.1016/j.antiviral.2022.105522 PMID: 36592667
28. Bennett RS, Huzella LM, Jahrling PB, Bollinger L, Olinger GG, Jr., Hensley LE. Nonhuman Primate Models of Ebola Virus Disease. Curr Top Microbiol Immunol. 2017; 411:171–93. https://doi.org/10.1007/82_2017_20 PMID: 28643203.
29. Whang MI, Tavares RM, Benjamin DI, Kattah MG, Advincula R, Nomura DK, et al. The Ubiquitin Binding Protein TAX1BP1 Mediates Autophagasome Induction and the Metabolic Transition of Activated T Cells. Immunity. 2017; 46(3):405–20. Epub 20170314. https://doi.org/10.1016/j.immuni.2017.02.018 PMID: 28314591; PubMed Central PMCID: PMC5400745.
30. Sarango G, Manoury B, Moris A. TAX1BP1 a novel player in antigen presentation. Autophagy. 2023; 19 (7):2153–5. Epub 20221206. https://doi.org/10.1080/15548627.2022.2153570 PMID: 36448736; PubMed Central PMCID: PMC10283426.

31. Petkova DS, Verlhac P, Rozieres A, Baguet J, Claviere M, Kretz-Remy C, et al. Distinct Contributions of Autophagy Receptors in Measles Virus Replication. Viruses. 2017; 9(5). Epub 20170522. https://doi.org/10.3390/v9050123 PMID: 28531150; PubMed Central PMCID: PMC5454435.

32. Descamps D, Peres de Oliveira A, Gonnin L, Madrieres S, Fix J, Drajac C, et al. Depletion of TAX1BP1 Amplifies Innate Immune Responses during Respiratory Syncytial Virus Infection. J Virol. 2021; 95(22): e0091221. Epub 20210825. https://doi.org/10.1128/JVI.00912-21 PMID: 34431698; PubMed Central PMCID: PMC8549506.

33. Basler CF, Wang X, Mühlberger E, Volchkov V, Paragas J, Klenk H-D, et al. The Ebola virus VP35 protein functions as a type I IFN antagonist. Proc Nat Acad Sci USA. 2000; 97(22):12289–94. https://doi.org/10.1073/pnas.220398297 PMID: 11027311

34. Zhu L, Jin J, Wang T, Hu Y, Liu H, Gao T, et al. Ebola Virus Sequesters IRF3 in Viral Inclusion Bodies to Evade Host Antiviral Immunity. eLife. 2024; 12:RP88122. https://doi.org/10.7554/eLife.88122 PMID: 38285487

35. Bradfute SB, Warfield KL, Bavari S. Functional CD8+ T Cell Responses in Lethal Ebola Virus Infection. J Immunol. 2008; 180(6):4058–66. https://doi.org/10.4049/jimmunol.180.6.4058 PMID: 18322215

36. Prescott J, Falzarano D, Feldmann H. Natural Immunity to Ebola Virus in the Syrian Hamster Requires Antibody Responses. J Infect Dis. 2015; 212 Suppl 2(Suppl 2):S271–6. Epub 20150505. https://doi.org/10.1093/infdis/jiv203 PMID: 25948862; PubMed Central PMCID: PMC4564546.

37. Yang XL, Zhang YZ, Jiang RD, Guo H, Zhang W, Li B, et al. Genetically Diverse Filoviruses in Rousettus and Eonycteris spp. Bats, China, 2009 and 2015. Emerg Infect Dis. 2017; 23(3):482–6. https://doi.org/10.3201/eid2303.161119 PMID: 28221123; PubMed Central PMCID: PMC5382765.

38. Hierweger MM, Koch MC, Rupp M, Maes P, Di Paola N, Bruggmann R, et al. Novel Filoviruses, Hantavirus, and Rhabdovirus in Freshwater Fish, Switzerland, 2017. Emerg Infect Dis. 2021; 27(12):3082–91. https://doi.org/10.3201/eid2712.210491 PMID: 34808081; PubMed Central PMCID: PMC8632185.

39. Geoghegan JL, Di Giallonardo F, Wille M, Ortiz-Baez AS, Costa VA, Ghaly T, et al. Virome composition in marine fish revealed by meta-transcriptomics. Virus Evol. 2021; 7(1):veab005. Epub 20210204. https://doi.org/10.1093/ve/veab005 PMID: 33623709; PubMed Central PMCID: PMC7887440.

40. Chen W-J, Santini F, Carnevale G, Chen J-N, Liu S-H, Lavoue S, et al. New insights on early evolution of spiny-rayed fishes (Teleostei: Acanthomorpha). Frontiers in Marine Science. 2014;1. https://doi.org/10.3389/fmars.2014.00053

41. Jingkai J, Cixiu L, Tao H, Zhongshuai T, Juan L, Lin X, et al. Diverse RNA viruses in the venom-related microenvironment of different animal phyla. Virus Evol. 2024. https://doi.org/10.1093/ve/veae024/7619247

42. Charon J, Buchmann JP, Sadiq S, Holmes EC. RdRp-scan: A bioinformatic resource to identify and annotate divergent RNA viruses in metagenomic sequence data. Virus Evol. 2022; 8(2):veac082. Epub 20220901. https://doi.org/10.1093/ve/veac082 PMID: 36533143; PubMed Central PMCID: PMC9752661.

43. Biedenkopf N, Bukreyev A, Chandran K, Di Paola N, Formenty PB, Griffiths A, et al. ICTV Virus Taxonomy Profile: Filoviridae 2024. J Gen Virol. 2024; 105(2):001955. https://doi.org/10.1099/jgv.0.001955 PMID: 38305775

44. Nino Barreat JG, Katzourakis A. Deep-mining of vertebrate genomes reveals an unexpected diversity of endogenous viral elements. bioRxiv. 2023:2023.10. 26.564176.

45. Kapli P, Kotari I, Telford MJ, Goldman N, Yang Z. DNA Sequences Are as Useful as Protein Sequences for Inferring Deep Phylogenies. Syst Biol. 2023. Epub 20230627. https://doi.org/10.1093/sysbio/syad036 PMID: 37366056.

46. Wang HC, Minh BQ, Susko E, Roger AJ. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. Syst Biol. 2018; 67(2):216–35. https://doi.org/10.1093/sysbio/syx068 PMID: 28950365.

47. Maarifi G, Smith N, Maillet S, Moncorgé O, Chamontin C, Edouard J, et al. TRIM8 is required for virus-induced IFN response in human plasmacytoid dendritic cells. Sci Adv. 2019; 5(11):eaax3511. https://doi.org/10.1126/sciadv.aax3511 PMID: 31799391

48. Xiao K, Liang X, Lu H, Li X, Zhang Z, Lu X, et al. Adaptation of gut microbiome and host metabolic systems to lignocellulosic degradation in bamboo rats. ISME J. 2022; 16(8):1980–92. Epub 20220514. https://doi.org/10.1038/s41396-022-01247-2 PMID: 35568757; PubMed Central PMCID: PMC9107070.

49. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol. 2017; 34(7):1812–9. https://doi.org/10.1093/molbev/msx116 PMID: 28387841.

50. Duchêne DA, Bragg JG, Duchêne S, Neaves LE, Potter S, Moritz C, et al. Analysis of Phylogenomic Tree Space Resolves Relationships Among Marsupial Families. Syst Biol. 2017; 67(3):400–12. https://doi.org/10.1093/sysbio/syx076 PMID: 29029231

51. Zou H, Jakovlić I, Zhang D, Hua C-J, Chen R, Li W-X, et al. Architectural instability, inverted skews and mitochondrial phylogenomics of Isopoda: outgroup choice affects the long-branch attraction artefacts. R Soc Open Sci. 2020; 7(2):191887. https://doi.org/10.1098/rsos.191887 PMID: 32257344

52. Blanco-Melo D, Campbell MA, Zhu H, Dennis TPW, Modha S, Lytras S, et al. A novel approach to exploring the dark genome and its application to mapping of the vertebrate virus fossil record. Genome Biol. 2024; 25(1):120. Epub 20240513. https://doi.org/10.1186/s13059-024-03258-y PMID: 38741126; PubMed Central PMCID: PMC11089739.

53. Ballinger MJ, Bruenn JA, Hay J, Czechowski D, Taylor DJ. Discovery and evolution of bunyavirids in arctic phantom midges and ancient bunyavirid-like sequences in insect genomes. J Virol. 2014:JVI-00531. https://doi.org/10.1128/JVI.00531-14 PMID: 24850747

54. Guo YT, Zhang J, Xu DM, Tang LZ, Liu Z. Phylogenomic relationships and molecular convergences to subterranean life in rodent family Spalacidae. Zool Res. 2021; 42(5):671–4. https://doi.org/10.24272/j.issn.2095-8137.2021.240 PMID: 34490760; PubMed Central PMCID: PMC8455469.

55. Biedenkopf N, Bukreyev A, Chandran K, Di Paola N, Formenty PBH, Griffiths A, et al. Renaming of genera Ebolavirus and Marburgvirus to Orthoebolavirus and Orthomarburgvirus, respectively, and introduction of binomial species names within family Filoviridae. Arch Virol. 2023; 168(8):220. Epub 20230803. https://doi.org/10.1007/s00705-023-05834-2 PMID: 37537381.

56. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics. 2014; 30(22):3276–8. Epub 20140805. https://doi.org/10.1093/bioinformatics/btu531 PMID: 25095880; PubMed Central PMCID: PMC4221126.

57. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform. 2019; 20(4):1160–6. https://doi.org/10.1093/bib/bbx108 PMID: 28968734; PubMed Central PMCID: PMC6781576.

58. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 2007; 56(4):564–77. https://doi.org/10.1080/10635150701472164 PMID: 17654362.

59. Steenwyk JL, Buida TJ 3rd, Li Y, Shen XXRokas AClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. PLoS Biol. 2020; 18(12):e3001007. Epub 20201202. https://doi.org/10.1371/journal.pbio.3001007 PMID: 33264284; PubMed Central PMCID: PMC7735675.

60. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PloS one. 2010; 5(6):e11147. https://doi.org/10.1371/journal.pone.0011147 PMID: 20593022

61. Ly-Trong N, Naser-Khdour S, Lanfear R, Minh BQ. AliSim: A Fast and Versatile Phylogenetic Sequence Simulator for the Genomic Era. Mol Biol Evol. 2022;39(5). https://doi.org/10.1093/molbev/msac092 PMID: 35511713; PubMed Central PMCID: PMC9113491.

62. Rambaut A, Grass NC. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics. 1997; 13(3):235–8. https://doi.org/10.1093/bioinformatics/13.3.235 PMID: 9183526

63. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020; 37(5):1530–4. https://doi.org/10.1093/molbev/msaa015 PMID: 32011700; PubMed Central PMCID: PMC7182206.

64. Crotty SM, Minh BQ, Bean NG, Holland BR, Tuke J, Jermiin LS, et al. GHOST: recovering historical signal from heterotachously evolved sequence alignments. Syst Biol. 2020; 69(2):249–64. https://doi.org/10.1093/sysbio/syz051 PMID: 31364711.

65. Kosakovsky Pond SL, Frost SD. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 2005; 22(5):1208–22. Epub 20050209. https://doi.org/10.1093/molbev/msi105 PMID: 15703242.

66. Mirdita M, Schutze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022; 19(6):679–82. Epub 20220530. https://doi.org/10.1038/s41592-022-01488-1 PMID: 35637307; PubMed Central PMCID: PMC9184281.

67. Schrodinger LLC. The PyMOL Molecular Graphics System, Version 1.8. 2015.

68. Li Z, Jaroszewski L, Iyer M, Sedova M, Godzik A. FATCAT 2.0: towards a better understanding of the structural diversity of proteins. Nucleic Acids Res. 2020; 48(W1):W60–W4. https://doi.org/10.1093/nar/gkaa443 PMID: 32469061; PubMed Central PMCID: PMC7319568.

69. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag; 2016.

70. Price A, Okumura A, Haddock E, Feldmann F, Meade-White K, Sharma P, et al. Transcriptional Correlates of Tolerance and Lethality in Mice Predict Ebola Virus Disease Patient Outcomes. Cell Rep. 2020; 30(6):1702–13 e6. https://doi.org/10.1016/j.celrep.2020.01.026 PMID: 32049004.