bioRxiv preprint doi: https://doi.org/10.1101/426585; this version posted September 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A Reptilian Endogenous Foamy Virus Sheds Light on the Early Evolution of Retroviruses

Xiaoman Wei[1,2†], Yicong Chen[1,2†], Edward C. Holmes[3], Jie Cui[1*]

[1]Key Laboratory of Special Pathogens and Biosafety, Center for Emerging Infectious Diseases, Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan 430071, China.

[2]University of Chinese Academy of Sciences, Beijing 100049, China.

[3]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and Faculty of Medicine and Health, University of Sydney, Sydney, NSW 2006, Australia.

[†]These authors contributed equally to this work.

[*]Corresponding author: E-mail: jiecui@wh.iov.cn

bioRxiv preprint doi: https://doi.org/10.1101/426585; this version posted September 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Endogenous retroviruses (ERVs) represent host genomic fossils of ancient viruses. Foamy viruses, including those that form endogenous copies, provide strong evidence for virus-host co-divergence across the vertebrate phylogeny. Endogenous foamy viruses (EFV) have previously been discovered in mammals, amphibians and fish. Here we report a novel endogenous foamy virus, named SpuEFV, in genome of the tuatara (*Sphenodon punctatus*), a reptile species endemic to New Zealand. Surprisingly, SpuEFV robustly grouped with the coelacanth EFV on virus phylogenies, rather than with the mammalian foamy viruses as expected with virus-host co-divergence, and indicative of a major cross-species transmission event in the early evolution of the foamy viruses. In sum, the discovery of SpuEFV fills a major gap in the fossil record of foamy viruses and provides important insights into the early evolution of retroviruses.

Key words: endogenous retroviruses; foamy virus; reptiles; evolution; cross-species transmission

bioRxiv preprint doi: https://doi.org/10.1101/426585; this version posted September 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Retroviruses (family *Retroviridae*) are viruses of major medical significance as some are associated with severe infectious disease or are oncogenic (Hayward, et al. 2015; Aiewsakun and Katzourakis 2017; Xu, et al. 2018). Retroviruses are also of note because of their ability to integrate into the host germ-line, generating endogenous retroviruses (ERVs) that then exhibit Mendelian inheritance (Stoye 2012; Johnson 2015). ERVs are widely distributed in vertebrates and provide important molecular "fossils" for the study of retrovirus evolution. ERVs related to all seven major retroviral genera have been described, although some of the more complex retroviruses, such as lenti-, delta- and foamy viruses, rarely appear as endogenous copies.

As well as being agents of disease, foamy viruses are of importance because of their long-term virus-host co-divergence (Switzer, et al. 2005). Endogenous foamy viruses (EFVs), first discovered in sloths (class Mammalia) (Katzourakis, et al. 2009) also exhibit co-divergence. The later discovery of a fish EFV in the coelacanth genome indicated that foamy viruses have an ancient evolutionary history (Han and Worobey 2012) and hence have likely co-diverged with their vertebrate hosts for hundreds of million years (Aiewsakun and Katzourakis 2017). However, although EFVs or foamy-like elements have been reported in fish, amphibians and mammals, they have currently not been reported in genomes of reptiles.

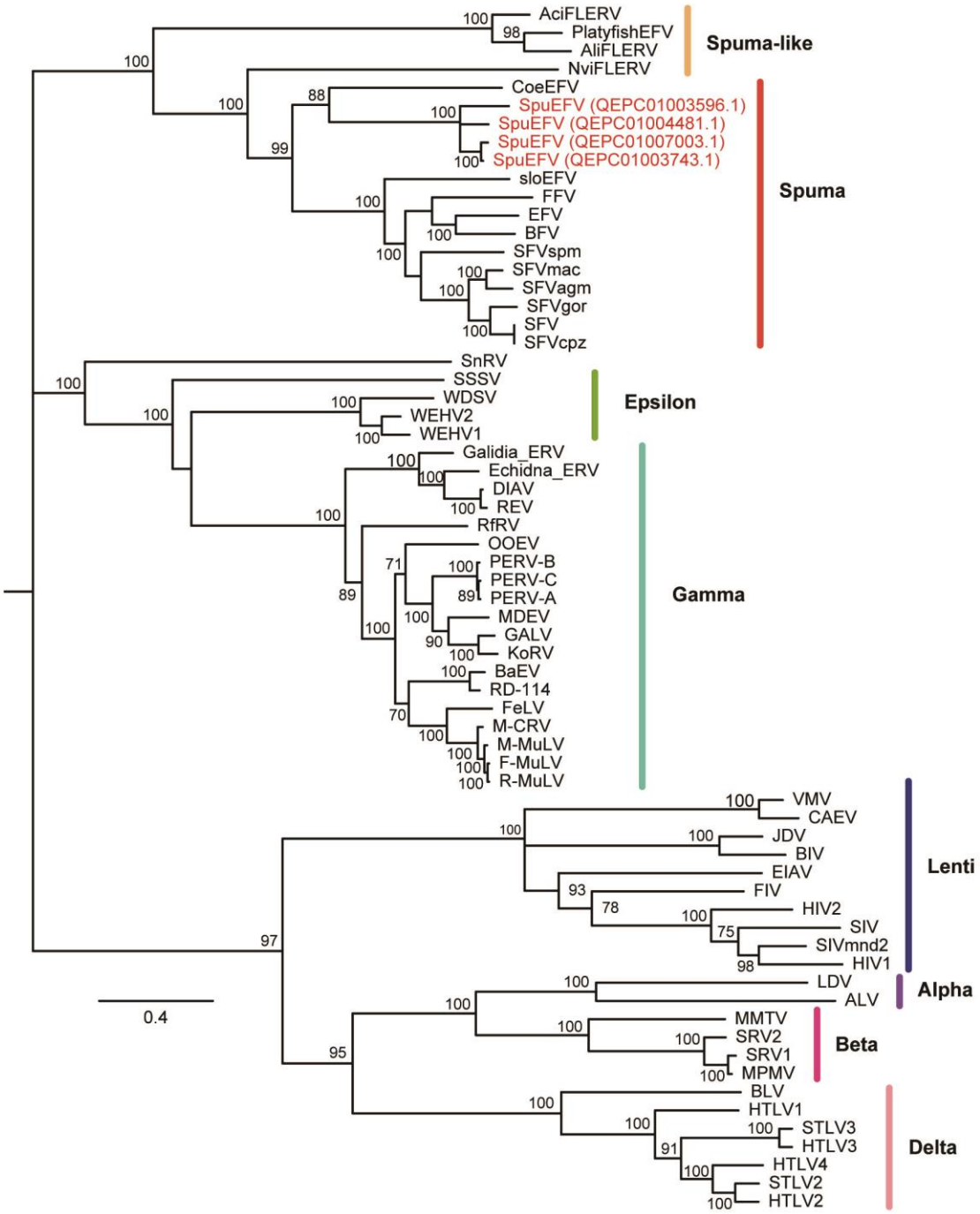
To search for potential foamy (-like) viral elements in reptiles, we collated 28 reptilian genomes (Supplementary Table S1) and performed *in silico* TBLASTN with full-length Pol sequences of various foamy viruses, including EFVs, as screening probes (Supplementary Table S2). We only considered viral hits within long genomic scaffold (>20 kilobases in length) to be *bona fide* ERVs. This genomic mining identified 175 ERV hits in three species: tuatara (*Sphenodon punctatus*), Schlegel's Japanese Gecko (*Gekko japonicas*) and

bioRxiv preprint doi: https://doi.org/10.1101/426585; this version posted September 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Madagascar ground gecko (*Paroedura picta*). However, because only one viral hit of each was found in the Schlegel's Japanese Gecko and Madagascar ground gecko (accession number: LNDG01066615.1 and BDOT01000314.1), which could represent false-positives, they were excluded. Hence, a total of 173 ERV hits in the tuatara genome were extracted and subjected to evolutionary analysis (Supplementary Table S3).

The long Pol (>700 amino acids) and Env (>350) sequences of these ERVs were then selected for phylogenetic analysis. Our maximum likelihood (ML) phylogenetic tree revealed that the ERVs discovered in tuatara genome formed a close monophyletic group within the foamy clade, indicative of a single origin, and with high bootstrap supportsin both phylogenies (Fig. 1; Fig. S1). We named this new ERV as SpuEFV (***S****phenodon* ***pu****nctatus* **e**ndogenous **f**oamy **v**irus). To our surprise, SpuEFV was consistently and robustly related to the fish EFV – CoeEFV – derived from the coelacanth genome (Han and Worobey 2012), and hence in conflict with the known host phylogeny. Although this phylogenetic pattern is compatible with cross-class virus transmission from fish to reptiles, it is possible that this pattern will change with a larger sampling of taxa such that the EFV phylogeny expands. Failure to detect any SpuEFV related elements in the remaining reptilian genome screening suggests that the virus was not vertically transmitted among reptiles, although this will clearly need to be reassessed with a larger sample size.

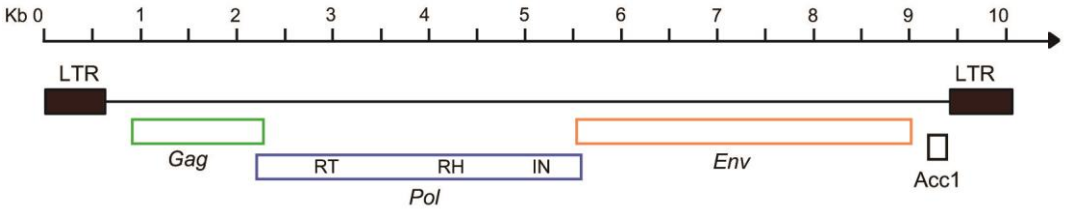
bioRxiv preprint doi: https://doi.org/10.1101/426585; this version posted September 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



We successfully retrieved two full-length SpuEFV viral genomes and annotated one in detail (Fig. S2). The annotated sequence exhibits a typical spuma virus structure, encoding three mainly open reading frames (ORF) – *gag*, *pol* and *env* – and one additional accessory genes, Acc1 (Fig. 2). Interestingly, this accessory ORF (Acc1) exhibit no sequence similarity to known genes. Notably, by searching the Conserved Domains Database

bioRxiv preprint doi: https://doi.org/10.1101/426585; this version posted September 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

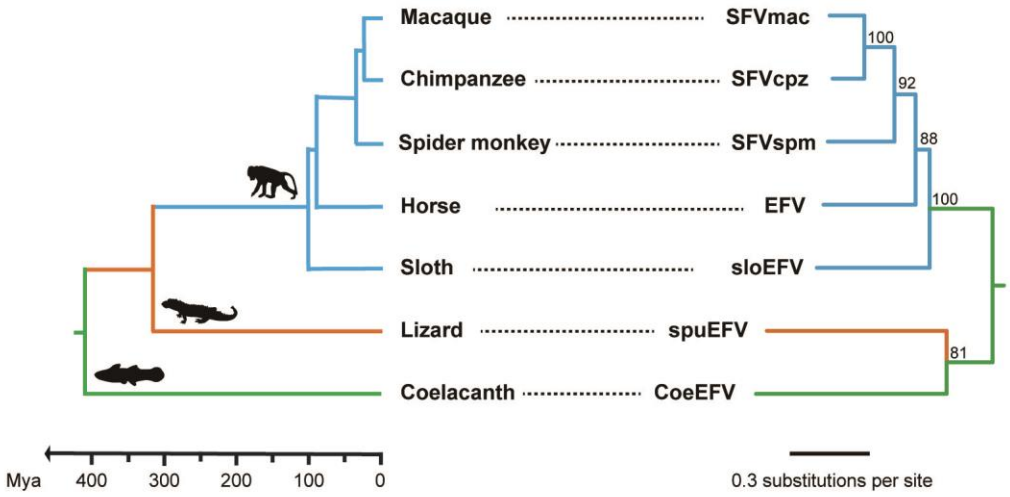
(www.ncbi.nlm.nih.gov/Structure/cdd), we identified a typical conserved foamy virus envelope protein domain (pfam034308) (Han and Worobey 2012), they further confirming that SpuEFV is of foamy virus origin.



To broadly estimate the integration time of SpuEFVs, we employed the LTR (long terminal repeat)-divergence method, which analyzes the degree of divergence between 5' and 3'LTRs assuming a known rate of nucleotide substitution (Johnson and Coffin 1999). In total, five pairwise LTRs flanking SpuEFV elements were used for date estimation (Supplementary Table S4), from which we estimated an integration time of SpuEFV ranging from 1.3 to 35.47 MYA (million years ago). Although these dates are young relative to the age of reptiles, LTR dating may severely underestimate ERV ages (Kijima and Innan 2010; Aiewsakun and Katzourakis 2017), such that all estimates of integration time should be treated with caution.

Previous studies provided strong evidence for the co-divergence of foamy viruses and their vertebrate hosts over extended time-periods (Katzourakis, et al. 2009). That the reptilian SpuEFV newly described here was most closely related to fish EFVs than those found in mammalian genomes (Fig. 3) indicates that cross-species virus transmission on a back-bone of long-term virus-host co-divergence may also play a major role in shaping the early evolution of retroviruses.

bioRxiv preprint doi: https://doi.org/10.1101/426585; this version posted September 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Materials and Methods

Genomic mining

To identify foamy viruses in reptiles, the TBLASTN program (Altschul, et al. 1990) was used to screen relevant taxa from 28 reptile genomes downloaded from GenBank (www.ncbi.nlm.nih.gov/genbank) (Supplementary Table S1). In each case amino acid sequences of the Pol and Env genes of representative EFVs (endogenous foamy viruses), foamy-like sequences, and foamy viruses were chosen as queries. As filters to identify significant and meaningful hits, we chose sequences with more than 30% amino acid identity over a 30% genomic region, with an e-value set to 0.00001. We extended viral flanking sequences of the hits to identify the 5'- and 3'-LTRs using LTR finder (Xu and Wang 2007) and LTR harvest (Ellinghaus, et al. 2008).

Phylogenetic analysis

To determine the evolutionary relationship of EFVs and retroviruses, Pol and Env protein sequences were aligned in MAFFT 7.222 (Katoh and Standley 2013) and confirmed manually in MEGA7 (Kumar, et al. 2016). The phylogenetic relationships among these sequences were then determined using the maximum-likelihood (ML) method in PhyML 3.1 (Guindon, et al. 2010), incorporating 100 bootstrap replicates to determine node robustness.

bioRxiv preprint doi: https://doi.org/10.1101/426585; this version posted September 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The best-fit models of amino acid substitution were determined by ProtTest 3.4.2 (Abascal, et al. 2005): RtREV+Γ+I for Pol, and WAG+Γ for Env. All alignments used in the phylogenetic analyses can be found in Data set S1.

Molecular dating

The ERV integration time can be calculated using the following simple relation: T = (D/R)/2, in which T is the integration time (million years, MY), D is the number of nucleotide differences per site between the two LTRs, and R is the genomic substitution rate (i.e. number of nucleotide substitutions per site, per year). We used the previously estimated neutral substitution rate for squamate reptiles (7.6×10^{-10} nucleotide substitutions per site, per year) (Perry, et al. 2018). LTRs less than 300 bp in length were not included in this analysis.

Acknowledgments

J.C. is supported by National Natural Science Foundation of China (31671324) and CAS Pioneer Hundred Talents Program. ECH is supported by an ARC Australian Laureate Fellowship (FL170100022).

References

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104-2105.

Aiewsakun P, Katzourakis A. 2017. Marine origin of retroviruses in the early Palaeozoic Era. Nat Commun 8:13954.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403-410.

bioRxiv preprint doi: https://doi.org/10.1101/426585; this version posted September 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

149 Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for
150 de novo detection of LTR retrotransposons. BMC Bioinformatics.

151 Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New
152 algorithms and methods to estimate maximum-likelihood phylogenies: assessing the
153 performance of PhyML 3.0. Syst Biol 59:307-321.

154 Han GZ, Worobey M. 2012. An endogenous foamy-like viral element in the coelacanth
155 genome. PLoS Pathog 8:e1002790.

156 Hayward A, Cornwallis CK, Jern P. 2015. Pan-vertebrate comparative genomics unmasks
157 retrovirus macroevolution. Proc Natl Acad Sci U S A 112:464-469.

158 Johnson WE. 2015. Endogenous Retroviruses in the Genomics Era. Annu Rev Virol 2:135-
159 159.

160 Johnson WE, Coffin JM. 1999. Constructing primate phylogenies from ancient retrovirus
161 sequences. Proc Natl Acad Sci U S A 96:10254-10260.

162 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
163 improvements in performance and usability. Mol Biol Evol 30:772-780.

164 Katzourakis A, Gifford RJ, Tristem M, Gilbert MT, Pybus OG. 2009. Macroevolution of
165 complex retroviruses. Science 325:1512.

166 Kijima TE, Innan H. 2010. On the estimation of the insertion time of LTR retrotransposable
167 elements. Mol Biol Evol 27:896-904.

168 Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis
169 Version 7.0 for Bigger Datasets. Mol Biol Evol 33:1870-1874.

170 Perry BW, Card DC, McGlothlin JW, Pasquesi GIM, Adams RH, Schield DR, Hales NR,
171 Corbin AB, Demuth JP, Hoffmann FG, et al. 2018. Molecular Adaptations for Sensing and

bioRxiv preprint doi: https://doi.org/10.1101/426585; this version posted September 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Securing Prey and Insight into Amniote Genome Diversity from the Garter Snake Genome. Genome Biol Evol 10:2110-2129.

Stoye JP. 2012. Studies of endogenous retroviruses reveal a continuing evolutionary saga. Nat Rev Microbiol 10:395-406.

Switzer WM, Salemi M, Shanmugam V, Gao F, Cong ME, Kuiken C, Bhullar V, Beer BE, Vallet D, Gautier-Hion A, et al. 2005. Ancient co-speciation of simian foamy viruses and primates. Nature 434:376-380.

Xu X, Zhao H, Gong Z, Han GZ. 2018. Endogenous retroviruses of non-avian/mammalian vertebrates illuminate diversity and deep history of retroviruses. PLoS Pathog 14:e1007072.

Xu Z, Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35:W265-268.

Figure Legends

Figure 1. Phylogenetic tree of retroviruses, including SpuEFV, using amino acid sequences of the Pol gene. The phylogenetic tree was rooted using *Caenorhabditis elegans* retrotransposon Cer1 (GenBank accession no. U15406). The newly identified SpuEFVs are labelled in red along with their accession numbers. The scale bar indicates the number of amino acid changes per site. Bootstrap values <70% are not shown.

Figure 2. Genomic organizations of SpuEFV. LTR, long-terminal repeat; RT, reverse transcriptase; RH, ribonuclease H; IN, integrase.

Figure 3. A simplified evolutionary relationship between foamy viruses and their hosts. Phylogenies representing mammals, reptile and fish and their associated viruses are shown.

bioRxiv preprint doi: https://doi.org/10.1101/426585; this version posted September 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

197 The scale bar indicates host speciation time (million years ago, MYA) or the number of

198 amino acid changes per site in the viral genomes.