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

Xiaoman Wei[1,2†], Yicong Chen[1,2†], Guangqian Duan[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

Abstract

Endogenous retroviruses (ERVs) can be thought of as 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*), an endangered reptile species endemic to New Zealand. Phylogenetic analyses revealed that SpuEFV has likely co-diverged with its host over a period of many millions of years. 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; tuatara

Introduction

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 (Hayward, et al. 2013; Cui, et al. 2014; Hayward, et al. 2015; Xu, et al. 2018) and constitute 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 pattern with their hosts; and they have also been reported in primates and the Cape golden mole (Han and Worobey 2012b, 2014). The discovery of a EFV in the coelacanth genome indicated that foamy viruses could have an ancient evolutionary history (Han and Worobey 2012a), likely co-diverging with their vertebrate hosts over time-scales of hundreds of million years (Aiewsakun and Katzourakis 2017). Although EFVs or foamy-like elements have been reported in fish, amphibians and mammals, they have currently not been reported in genomes of two other major classes of vertebrates - reptiles and birds (Tristem, et al. 1995; Hayward, et al. 2015; Xu, et al. 2018).

Materials and Methods

Genomic mining and consensus genome construction

To identify foamy viruses in reptiles, the TBLASTN program (Altschul, et al. 1990) was used to screen relevant taxa from 28 reptile genomes (Supplementary Table S1) and 130 bird genomes (Supplementary Table S2) (as of October 2018) downloaded from GenBank (www.ncbi.nlm.nih.gov/genbank). In each case, the amino acid sequences of the Pol genes of representative EFVs (endogenous foamy viruses), endogenous foamy-like viruses, and exogenous 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% region, with an e-value set to 0.00001. Genomes that contained only single hits for EFVs were excluded as likely false-positives. 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). Sequences highly similar to foamy virus proteins found in tuatara were aligned to generate a SpuEFV consensus genome (Supplementary Table S5). Conserved domains were identified using CD-Search service in NCBI (Marchler-Bauer and Bryant 2004).

Phylogenetic analysis

To determine the evolutionary relationship of EFVs and retroviruses, the 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. The best-fit models of amino acid substitution were determined by ProtTest 3.4.2 (Abascal, et al. 2005): RtREV+Γ+I for Pol, LG+Γ+I+F for concatenated gag, pol and env. All alignments used in the phylogenetic analyses can be found in Data set S1-S2.

Results and Discussion

Discovery of foamy viral elements in reptile genomes

To search for potential foamy (-like) viral elements in reptiles and birds, we collated 28 reptilian genomes (Supplementary Table S1) and 130 bird genomes (Supplementary Table S2) and performed *in silico* TBLASTN with full-length Pol protein sequences of various foamy viruses, including EFVs, as screening probes (Supplementary Table S3). We only considered viral hits within long genomic scaffold (>20 kilobases in length) to be *bona fide* ERVs. This genomic mining identified 117 ERV hits in tuatara (*Sphenodon punctatus*) and none in bird genomes. Hence, a total of 117 ERV hits in the tuatara genome were extracted and subjected to evolutionary analysis (Supplementary Table S4). We named this new ERV as SpuEFV (***S****phenodon* ***pu****nctatus* **e**ndogenous **f**oamy **v**irus).

Genomic organization

We extracted all significant SpuEFV viral elements and constructed a consensus genomic sequence of SpuEFV (Supplementary Fig. S1, Table S5). The consensus genome harbored a pairwise long terminal repeats (LTRs) and exhibits a typical spuma virus structure, encoding three mainly open reading frames (ORF) – *gag*, *pol* and *env* – and one putative additional accessory genes, ORF1 (Fig. 1). Interestingly, this accessory ORF 1 exhibit no sequence similarity to known foamy accessory genes. Notably, by searching the Conserved Domains Database (www.ncbi.nlm.nih.gov/Structure/cdd), we identified three typical foamy conserved domain for both consensus and one full-length original SpuEFV (Accession no. QEPC01003194.1): (i) Spuma virus Gag domain (pfam03276) (Winkler, et al. 1997), (ii) Spuma aspartic protease (A9) domain (pfam03539) which exists in all mammalian foamy virus pol protein (Aiewsakun and Katzourakis 2017), and (iii) foamy virus envelope protein

103 domain (pfam03408) (Han and Worobey 2012a) (Supplementary Fig. S2, Fig. S3),
104 confirming that SpuEFV is indeed of foamy virus origin.

105 **Phylogenetic analysis**

106 The Pol (490 amino acids) of SpuEFVs were used for phylogenetic analysis. Our maximum
107 likelihood (ML) phylogenetic trees revealed that the EFVs present in that tuatara genome
108 formed a close monophyletic group within the foamy clade, indicative of a single origin, and
109 with high bootstrap support (Fig. 2). The divergent phylogenetic position of SpuEFV is
110 compatible with virus-host co-divergence for the entire history of the vertebrates. However, it
111 is possible that this pattern will change with a larger sampling of taxa such that the EFV
112 phylogeny expands. Failure to detect any SpuEFV related elements in the remaining reptilian
113 genome screening suggests that the virus was not vertically transmitted among reptiles,
114 although this will clearly need to be reassessed with a larger sample size.
115
116 Previous studies provided strong evidence for the co-divergence of foamy viruses and their
117 vertebrate hosts over extended time-periods (Katzourakis, et al. 2009). That the reptilian
118 SpuEFV newly described here seemingly follows the same pattern (Fig. 3) thereby implies
119 that it could diverge from the other mammalian foamy viruses with its tuatara host more than
120 320 million years ago (http://www.timetree.org/). As such, the discovery of SpuEFV fills a
121 major gap in our knowledge of the evolutionary history of the foamy viruses and provides
122 important insights into the early evolution of retroviruses.
123

124 **Acknowledgments**

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

Data availability

All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Conflict of interest: None declare

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.

Cui J, Zhao W, Huang Z, Jarvis ED, Gilbert MT, Walker PJ, Holmes EC, Zhang G. 2014. Low frequency of paleoviral infiltration across the avian phylogeny. Genome Biol 15:539.

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

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

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

151 Han GZ, Worobey M. 2012b. An endogenous foamy virus in the aye-aye (Daubentonia
152 madagascariensis). J Virol 86:7696-7698.

153 Han GZ, Worobey M. 2014. Endogenous viral sequences from the Cape golden mole
154 (Chrysochloris asiatica) reveal the presence of foamy viruses in all major placental mammal
155 clades. PLoS One 9:e97931.

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 Hayward A, Grabherr M, Jern P. 2013. Broad-scale phylogenomics provides insights into
159 retrovirus-host evolution. Proc Natl Acad Sci U S A 110:20146-20151.

160 Johnson WE. 2015. Endogenous Retroviruses in the Genomics Era. Annu Rev Virol
161 2:135-159.

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

168 Marchler-Bauer A, Bryant SH. 2004. CD-Search: protein domain annotations on the fly.
169 Nucleic Acids Res 32:W327-331.

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

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

175 Tristem M, Myles T, Hill F. 1995. A highly divergent retroviral sequence in the tuatara
176 (Sphenodon). Virology 210:206-211.

177 Winkler I, Bodem J, Haas L, Zemba M, Delius H, Flower R, Flugel RM, Lochelt M. 1997.
178 Characterization of the genome of feline foamy virus and its proteins shows distinct features
179 different from those of primate spumaviruses. J Virol 71:6727-6741.

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

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

184

185

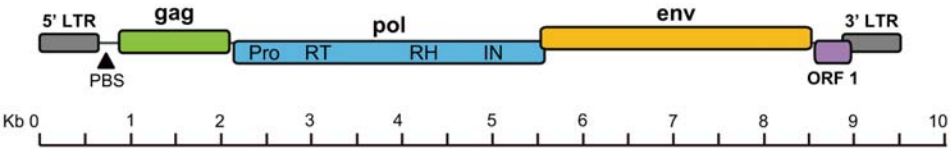


Figure 1. Genomic organizations of SpuEFV. LTR, long-terminal repeat; PBS, primer-binding site; Pro, aspartic protease; RT, reverse transcriptase; RH, ribonuclease H; IN, integrase.

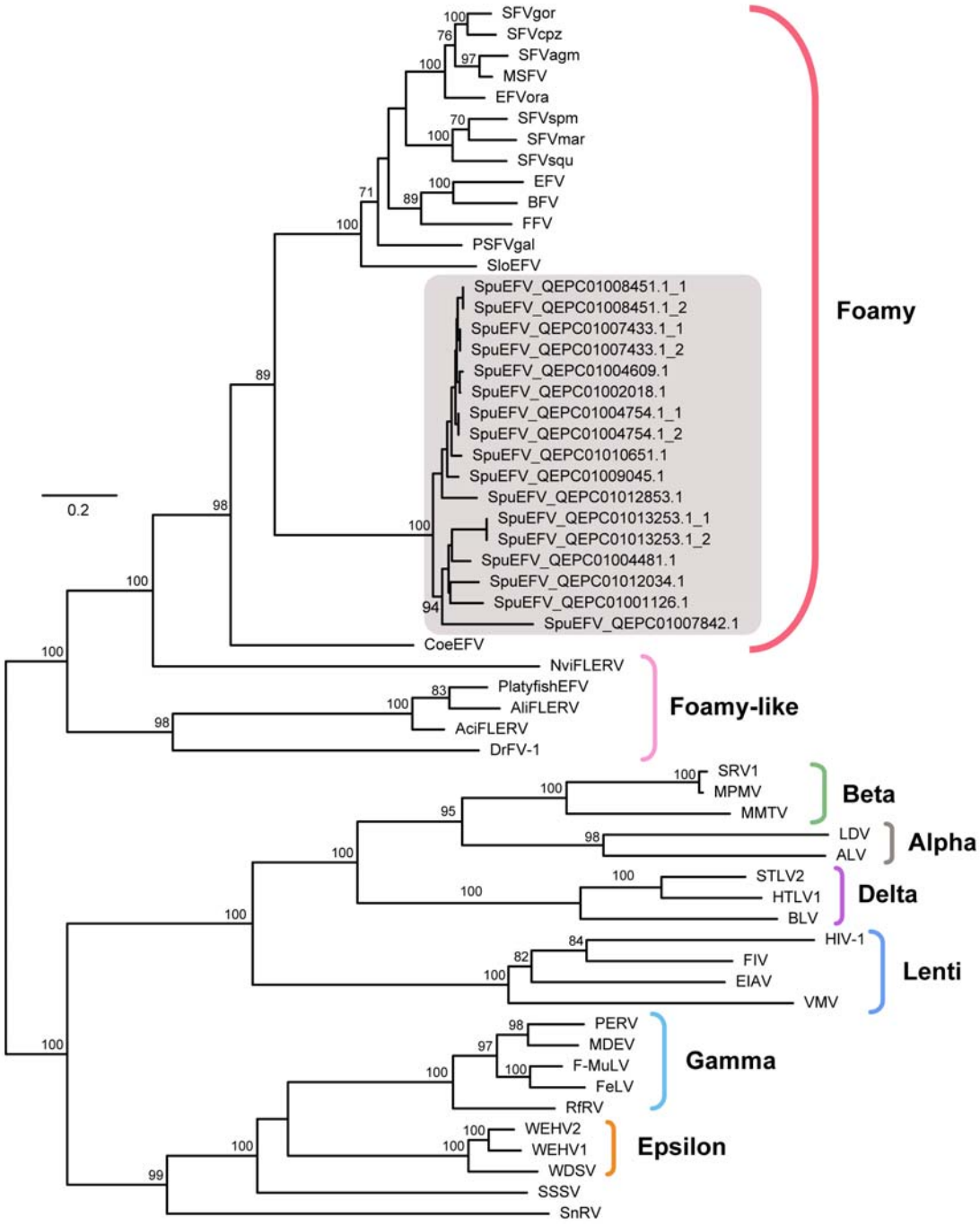
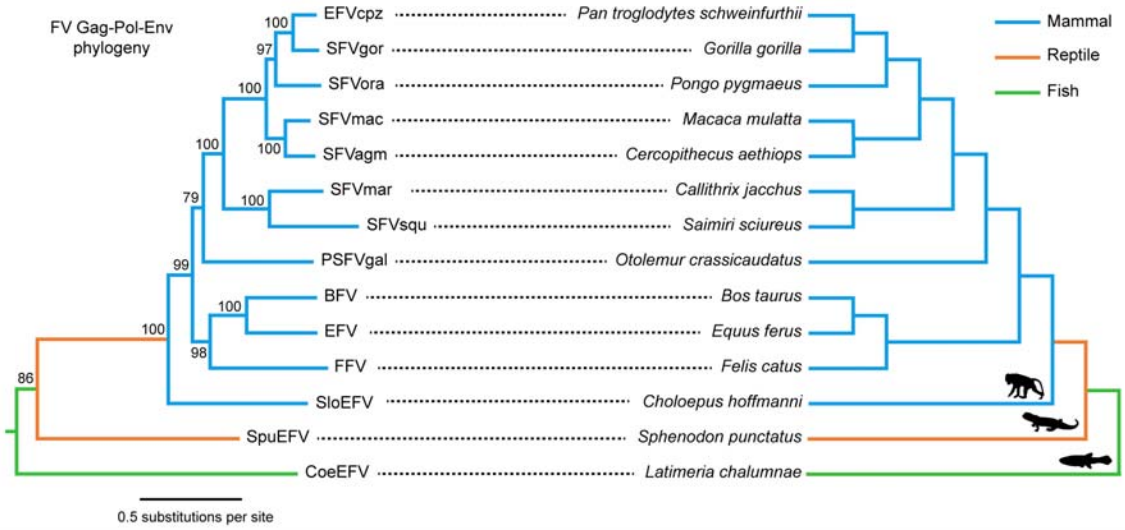


Figure 2. Phylogenetic tree of retroviruses, including SpuEFVs, inferred using amino acid sequences of the Pol gene (490aa). The tree is midpoint rooted for clarity only. The newly identified SpuEFVs are labelled using a grey-shaded box with their accession numbers (different pol sequences in same contig are numbered in the suffix). The scale bar indicates

197 the number of amino acid changes per site. Bootstrap values <70% are not shown. The
198 alignment of pol amino acid sequences is provided in Data set S1.
199
200
201



202
203 **Figure 3.** A simplified evolutionary relationship between foamy viruses (left) and their
204 vertebrate hosts (right). The scale bar on the virus phylogeny indicates number of amino acid
205 changes per site with bootstrap support values provided at each node. The alignment of FV
206 gag-pol-env amino acid sequences is provided in Data set S2.
207