Identification and evolution of avian endogenous foamy viruses

Yicong Chen[1,2,†], Xiaoman Wei[1,2,†], Guojie Zhang[3], Edward C. Holmes[4], Jie Cui[1,*]

[1]Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan, China; Institut Pasteur of Shanghai, Chinese Academy of Sciences, Shanghai, China.

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

[3]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China; China National GeneBank, BGI-Shenzhen, Shenzhen, China; Section for Ecology and Evolution, Department of Biology, University of Copenhagen, Copenhagen, Denmark; Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China.

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

[*]Corresponding author: E-mail: jcui@ips.ac.cn

[†]These authors contributed equally to this work.

Abstract

A history of long-term co-divergence means that foamy viruses (family *Retroviridae*) provide an ideal framework to understanding virus-host evolution over extended time-periods. Endogenous foamy viruses (EFVs) are rare, and to date have only been found in a limited number of genomes from mammals, amphibians, reptiles and fish. By screening 510 avian genomes we identified endogenous foamy viruses in avian species - from the Maguari Stork (*Ciconia maguari*) and Oriental Stork (*C. boyciana*). Phylogenetic analysis and analysis of the genome structures and flanking sequences indicated a single origin of EFVs into *Ciconia* during evolutionary history, and the marked incongruence with the host phylogeny suggested that this integration event occurred independently in birds. In sum, this study provides evidence that birds can be infected with foamy viruses, filling the last major gap in the taxonomic distribution of foamy viruses and their animal hosts.

Key words: endogenous foamy viruses; birds; incongruence; vertical transmission

Introduction

Retroviruses (family *Retroviridae*) are viruses of substantial medical and economic 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 evolutionary importance as they have occasionally invaded the host germ line, leading to the generation of endogenous retroviruses (ERVs) and hence genomic 'fossils' (Stoye 2012; Johnson 2015, 2019). ERVs are widely distributed in vertebrates (Hayward, et al. 2013; Cui, et al. 2014; Hayward, et al. 2015; Xu, et al. 2018) and provide important insights into the origin and long-term evolution of retroviruses. However, some complex retroviruses such as lenti-, delta- and spuma viruses, only relatively rarely appear as endogenized forms.

As well leaving a litany of endogenous copies in host genomes, foamy viruses are of particular importance because they exhibit a history of long-term co-divergence with their vertebrate hosts (Switzer, et al. 2005). Endogenous foamy viruses (EFVs) were first discovered in sloths (Katzourakis, et al. 2009), and then found in several primate genomes (Han and Worobey 2012b, 2014; Katzourakis, et al. 2014). The subsequent discovery of EFV and EFV-like copies in fish genomes indicated that foamy viruses may have a deep evolutionary history within the vertebrates in general (Han and Worobey 2012a; Ruboyianes and Worobey 2016; Aiewsakun and Katzourakis 2017). Recently, a novel endogenous virus - denoted ERV-Spuma-Spu - was identified in genome of the reptile *Sphenodon punctatus* (the tuatara), suggesting that foamy viruses have co-diverged with their vertebrate hosts for hundreds of million years (Aiewsakun, et al. 2019; Wei, et al. 2019). However, although

57 endogenous foamy viruses have been found in the genomes of fish, reptiles, amphibians and
58 mammals, to date they have not been identified in avian genomes.
59
60 **Materials and Methods**
61 **Genome screening and viral genome structure identification**
62 All 147 avian genomes available in GenBank as of June 2019 (Supplementary Table S1) and
63 363 genomes from the 'Bird 10K' program were screened for endogenous foamy viruses
64 using the TBLASTN algorithm (Altschul, et al. 1990) and the protein sequences of
65 representative exogenous foamy viruses, EFVs and endogenous foamy-like viruses
66 (Supplementary Table S2). A 35% sequence identity over a 30% region with an e-value set to
67 0.0001 was used to filter significant hits (Supplementary Table S3). Viral hits within large
68 scaffold (>20 kb) were assumed to represent *bona fide* ERVs. We then extended the flanking
69 sequence of these hits to identify the viral long terminal repeats (LTRs) using BLASTN
70 (Altschul, et al. 1990), LTR Finder (Xu and Wang 2007) and LTRharvest (Ellinghaus, et al.
71 2008). In accordance with the nomenclature proposed for ERVs (Gifford, et al. 2018),
72 endogenous foamy viruses were identified in the genomes of the Maguari Stork (*Ciconia*
73 *maguari*) and the Oriental stork (*C. boyciana*), were termed 'ERV-Spuma.n-Cma'
74 and 'ERV-Spuma.n-Cbo', respectively (in which n represents the number of the viral
75 sequences extracted from host genome) (Supplementary Table S4). Putative genome
76 structures and conserved EFV domains were identified using BLASTP, CD-search
77 (Marchler-Bauer and Bryant 2004; Marchler-Bauer, et al. 2017) and ORFfinder
78 (https://www.ncbi.nlm.nih.gov/orffinder/) in NCBI.

Molecular dating

ERV integration time can be calculated using the 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 pairwise LTRs, and R is the genomic substitution rate (nucleotide substitutions per site, per year). We used the previously estimated neutral nucleotide substitution rate for birds (1.9×10^{-9} nucleotide substitutions per site, per year) (Zhang, et al. 2014). Two full-length ERVs-Spuma-Cma containing a pairwise intact LTRs were used to estimate integration time in this manner (Supplementary Table S5). We excluded ERV-Spum.1-Cbo from this dating exercise due to its defective 5' LTR.

Phylogenetic analysis

To describe the evolutionary relationship of EFVs in relation to other representative retroviruses, sequences of the Pol (Supplementary data set S1) and concatenated Gag-Pol-Env proteins (Supplementary data set S2) were aligned using MAFFT 7.222 (Katoh and Standley 2013) and confirmed manually in MEGA X (Kumar, et al. 2018). A phylogenetic tree of these data was then inferred using the maximum likelihood (ML) method in PhyML 3.1 (Guindon, et al. 2010), incorporating 100 bootstrap replicates to assess node robustness. The best-fit LG+Γ+I+F was selected for both Pol and concatenated Gag-Pol-Env protein data sets using ProtTest (Abascal, et al. 2005).

Results and Discussion

Discovery and characterization of endogenous foamy viral elements in avian genomes

To identify potential foamy (-like) viral elements in birds, we collected all available bird genomes from GenBank (Supplementary Table 1) and the 'Bird 10K' project (Zhang, et al. 2015) and performed *in silico* TBLASTN, using the amino acid sequences of representative retroviruses (Supplementary Table 2). This genomic mining identified 16 significant hits in the Maguari Stork and the 12 in Oriental Stork (Supplementary Table 3). We designated these ERV-Spuma.n-Cma and Spuma.n-Cbo, respectively (Gifford, et al. 2018) (Supplementary Table S3, Table S4). We considered hits within large scaffolds (>20 kilobases in length) to represent *bona fide* ERVs. We then extended the flanking sequences of these EFVs on both sides to search for LTRs as these define the boundary of the viral elements. Through this analysis we were able to discover two full-length EFVs in the Maguari stork genome and one in the Oriental stork genome. The low copy number of EFVs found in both two bird species accords with the observation that avian genomes generally harbor small numbers of endogenous viruses (Cui, et al. 2014).

To further elucidate the relationship between these novel avian EFVs and other retroviruses, the long Pol (>800 amino acid) gene sequences were used in a phylogenetic analysis (Fig. 1). Accordingly, our maximum likelihood phylogenetic tree revealed that the EFVs discovered in birds formed a close and well supported monophyletic group within the foamy virus clade compatible with the idea that these avian EFVs might have a single origin. Notably, however, because they were most closely related to the endogenous foamy viruses found in mammals rather than to those found in reptiles, the phylogenetic position of the avian EFVs described

here was incongruent with that of the host phylogeny (although the node associated with the tuatara EFV was relatively poorly supported) (Fig. 2). This, and the overall rarity of EFVs in birds, suggests that these avian EFVs have an independent origin in birds and were not acquired through virus-host co-divergence.

Genomic structure characterization

The genomes of the new EFVs documented here contained a pair of LTRs, although with no sequence similarity to other EFV LTRs, and exhibited a typical spuma virus structure, with three main open reading frames (ORF) - gag, pol, and env - and one putative additional accessory gene, ORF 1 (Fig. 3, Supplementary Fig. S1). Similar to the LTRs, this accessory ORF 1 exhibits no sequence similarity to any known accessory genes from foamy viruses. In addition, by searching for conserved domains against the Conserved Domains Database (www.ncbi.nlm.nih.gov/Structure/cdd), we identified three typical foamy conserved domains in the three full-length avian EFVs: (1) the Spuma virus Gag domain (pfam03276) (Winkler, et al. 1997), (2) the Spuma aspartic protease (A9) domain (pfam03539) that is present in all mammalian foamy virus Pol proteins (Aiewsakun and Katzourakis 2017), and the (3) foamy virus envelope protein domain (pfam03408) (Han and Worobey 2012a; Wei, et al. 2019) (Supplementary Fig. S2).

Vertical transmission of bird EFVs

Surprisingly, ERV-Spuma.2-Cma and ERV-Spuma.1-Cbo shared 98% nucleotide sequence identity and contained the same deletion in the pol gene (Supplementary Fig. S3). By

145 comparing the flanking sequences of these two EFVs using BLASTN, we discovered that two
146 scaffolds containing EFV (BDFF02011124.1 in the Oriental stork and scaffold3222 in the
147 Maguari Stork) shared 99% coverage with 98.66% sequence identity (e-value = 0.0).
148 Furthermore, upon the ERV insertion, the target DNA fragment is also duplicated, resulting
149 in target site duplication (TSDs) that differs among ERV insertions (Hughes and Coffin
150 2001). We were able to identify the same TSD flanking these two avian EFVs
151 (Supplementary Fig. S3), confirming that they have been vertically inherited. However,
152 neither EFVs nor any solo LTR were present in other bird species from same order
153 (Pelecaniformes), including little egret (*Egretta garzetta)*, crested ibis (*Nipponia nippon*) and
154 Yellow-throated sandgrouse (*Pterocles gutturalis)*. Clearly, the study of additional genomes
155 sampled across the avian phylogeny is merited.
156
157 **Estimation of ERV insertion times**
158 To approximately estimate the insertion time of endogenous foamy viruses in two birds, we
159 utilized the LTR-divergence based on the degree of sequence divergence between the 5' and
160 3' LTRs with a known host nucleotide substitution rate (Dangel, et al. 1995; Johnson and
161 Coffin 1999). Only two intact pairwise LTRs of ERVs-Spuma-Cma were selected for time
162 estimation (Supplementary Table S5). This analysis revealed insertion times of 3.15 and
163 13.95 MYA (million years ago). Although, the insertion time is surprisingly young compared
164 to the age of birds, the presence of multiple premature stop codon suggests the invasion was
165 ancient, and all estimates of integration times should be treated with caution (Kijima and
166 Innan 2010).

In sum, we describe the presence and evolution of two novel avian endogenous foamy viruses. This discovery fills the last major gap in our understanding of the taxonomic distribution of the foamy viruses and will help us to understand the evolutionary interactions between retroviruses and their hosts over extended time-periods

Acknowledgements

J.C. is supported by National Natural Science Foundation of China (31671324) and CAS Pioneer Hundred Talents Program. E.C.H. 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.

Aiewsakun P, Simmonds P, Katzourakis A. 2019. The First Co-Opted Endogenous Foamy Viruses and the Evolutionary History of Reptilian Foamy Viruses. Viruses 11:641.

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.

189 Dangel AW, Baker BJ, Mendoza AR, Yu CY. 1995. Complement component C4 gene intron
190 9 as a phylogenetic marker for primates: long terminal repeats of the endogenous retrovirus
191 ERV-K(C4) are a molecular clock of evolution. Immunogenetics 42:41-52.
192 Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for
193 de novo detection of LTR retrotransposons. BMC Bioinformatics.
194 Gifford RJ, Blomberg J, Coffin JM, Fan H, Heidmann T, Mayer J, Stoye J, Tristem M,
195 Johnson WE. 2018. Nomenclature for endogenous retrovirus (ERV) loci. Retrovirology
196 15:59.
197 Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New
198 algorithms and methods to estimate maximum-likelihood phylogenies: assessing the
199 performance of PhyML 3.0. Syst Biol 59:307-321.
200 Han GZ, Worobey M. 2012a. An endogenous foamy-like viral element in the coelacanth
201 genome. PLoS Pathog 8:e1002790.
202 Han GZ, Worobey M. 2012b. An endogenous foamy virus in the aye-aye (Daubentonia
203 madagascariensis). J Virol 86:7696-7698.
204 Han GZ, Worobey M. 2014. Endogenous viral sequences from the Cape golden mole
205 (Chrysochloris asiatica) reveal the presence of foamy viruses in all major placental mammal
206 clades. PLoS One 9:e97931.
207 Hayward A, Cornwallis CK, Jern P. 2015. Pan-vertebrate comparative genomics unmasks
208 retrovirus macroevolution. Proc Natl Acad Sci U S A 112:464-469.
209 Hayward A, Grabherr M, Jern P. 2013. Broad-scale phylogenomics provides insights into
210 retrovirus-host evolution. Proc Natl Acad Sci U S A 110:20146-20151.

211 Hughes JF, Coffin JM. 2001. Evidence for genomic rearrangements mediated by human
212 endogenous retroviruses during primate evolution. Nat Genet 29:487-489.
213 Johnson WE. 2015. Endogenous Retroviruses in the Genomics Era. Annu Rev Virol
214 2:135-159.
215 Johnson WE. 2019. Origins and evolutionary consequences of ancient endogenous
216 retroviruses. Nat Rev Microbiol 17:355-370.
217 Johnson WE, Coffin JM. 1999. Constructing primate phylogenies from ancient retrovirus
218 sequences. Proc Natl Acad Sci U S A 96:10254-10260.
219 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
220 improvements in performance and usability. Mol Biol Evol 30:772-780.
221 Katzourakis A, Aiewsakun P, Jia H, Wolfe ND, LeBreton M, Yoder AD, Switzer WM. 2014.
222 Discovery of prosimian and afrotherian foamy viruses and potential cross species
223 transmissions amidst stable and ancient mammalian co-evolution. Retrovirology 11:61.
224 Katzourakis A, Gifford RJ, Tristem M, Gilbert MT, Pybus OG. 2009. Macroevolution of
225 complex retroviruses. Science 325:1512.
226 Kijima TE, Innan H. 2010. On the estimation of the insertion time of LTR retrotransposable
227 elements. Mol Biol Evol 27:896-904.
228 Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary
229 Genetics Analysis across Computing Platforms. Mol Biol Evol 35:1547-1549.
230 Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, Chitsaz F, Derbyshire MK, Geer
231 RC, Gonzales NR, et al. 2017. CDD/SPARCLE: functional classification of proteins via
232 subfamily domain architectures. Nucleic Acids Res 45:D200-d203.

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

Ruboyianes R, Worobey M. 2016. Foamy-like endogenous retroviruses are extensive and abundant in teleosts. Virus Evol 2:vew032.

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.

Wei X, Chen Y, Duan G, Holmes EC, Cui J. 2019. A reptilian endogenous foamy virus sheds light on the early evolution of retroviruses. Virus Evol 5:vez001.

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

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.

Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW, et al. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. Science 346:1311-1320.

254 Zhang G, Rahbek C, Graves GR, Lei F, Jarvis ED, Gilbert MT. 2015. Genomics: Bird

255 sequencing project takes off. Nature 522:34.

256

257

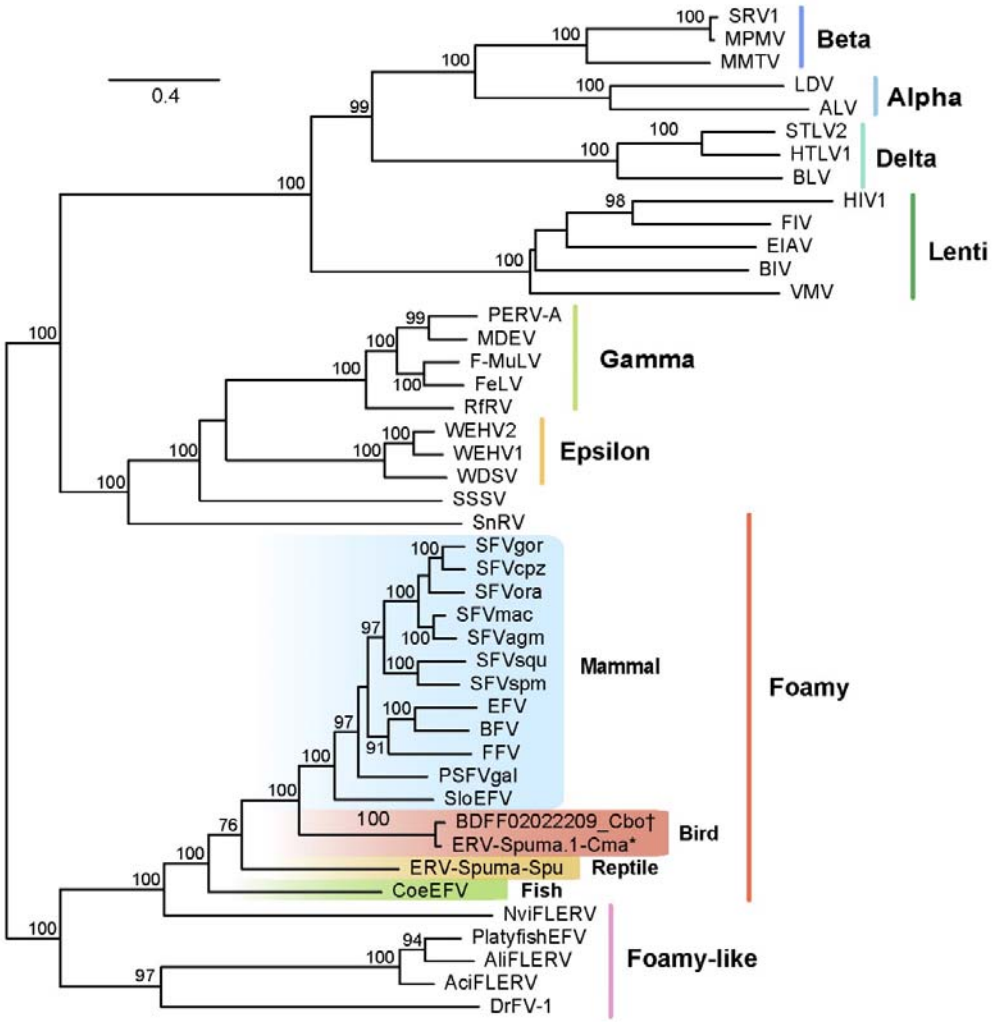


Fig. 1. Phylogenetic tree of retroviruses and endogenous retroviruses, including the endogenous foamy viruses found in avian genomes. The tree was inferred using amino acid sequences of the Pol gene, and is midpoint rooted for clarity only. The newly identified viral elements are marked with a red-shaded box. * indicates the EFV found in the Maguari Stork genome, while † indicates the EFV found in the Oriental Stork genome. The scale bar indicates the number of amino acid changes per site. Bootstrap values >70 % are shown.

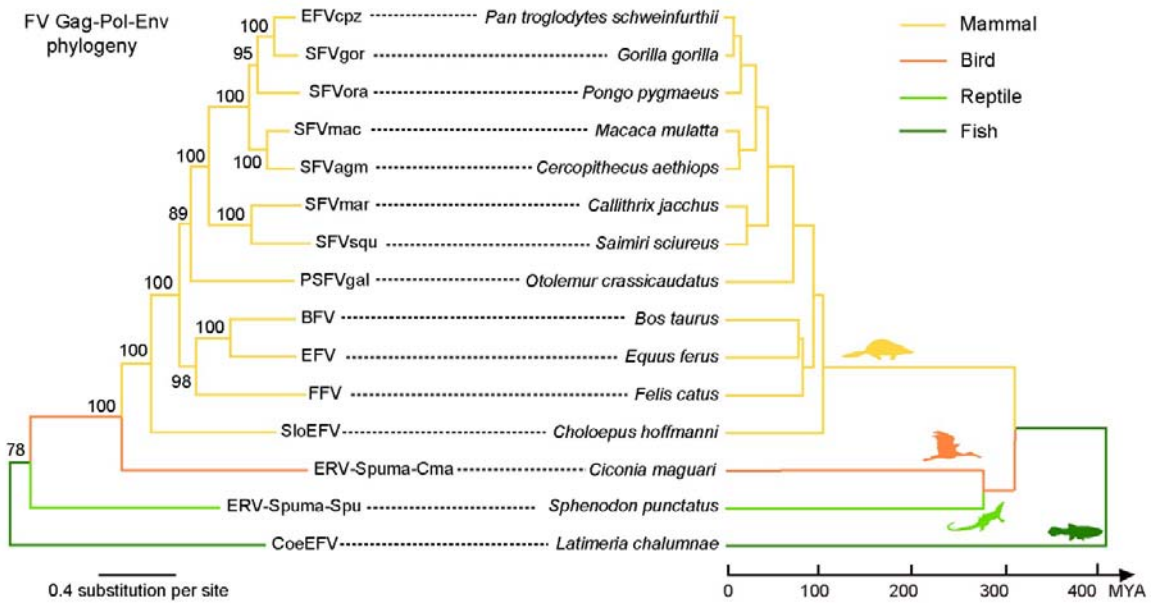


Fig. 2. Evolutionary history of foamy viruses (left) and their vertebrate hosts (right). Associations between foamy viruses and their hosts are indicated by connecting lines, although note the avian EFVs are more closely related to mammalian foamy viruses than the reptilian EFV. The scale bars indicate the number of amino acid changes per site in the viruses and the host divergence times (million years ago, MYA).

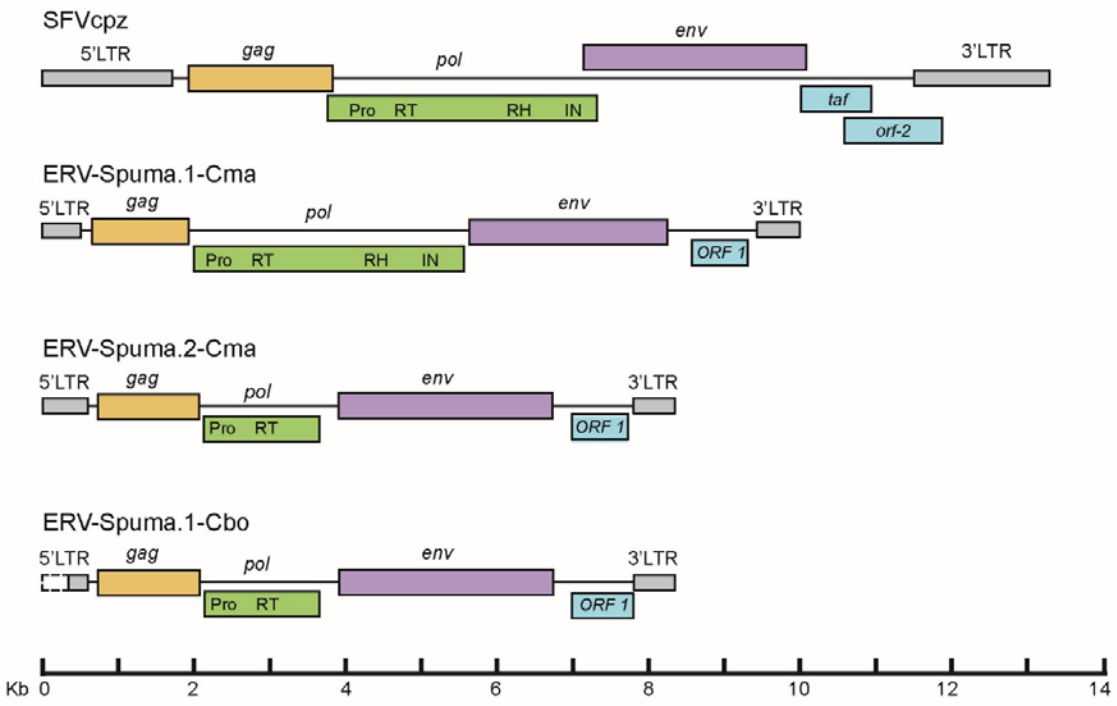
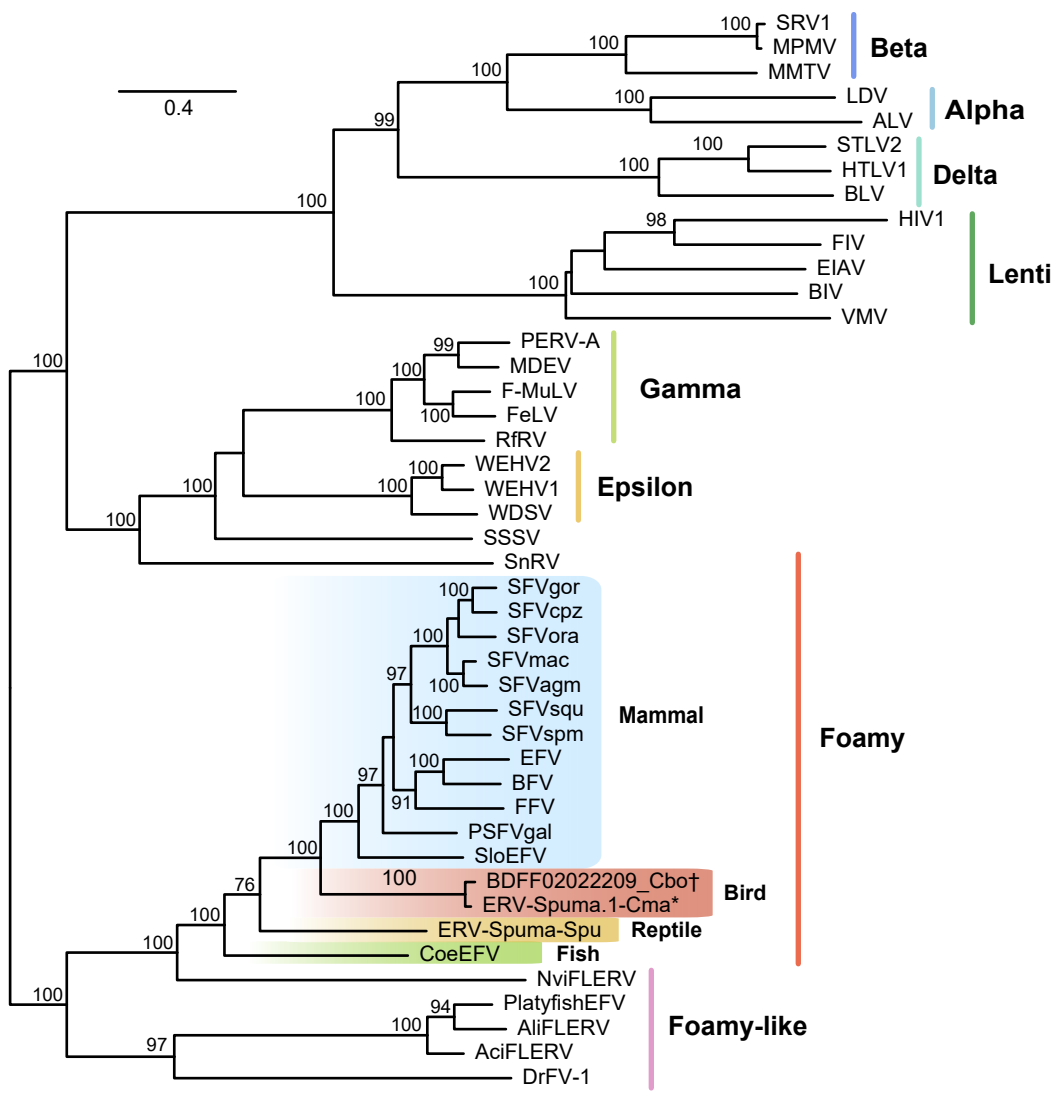
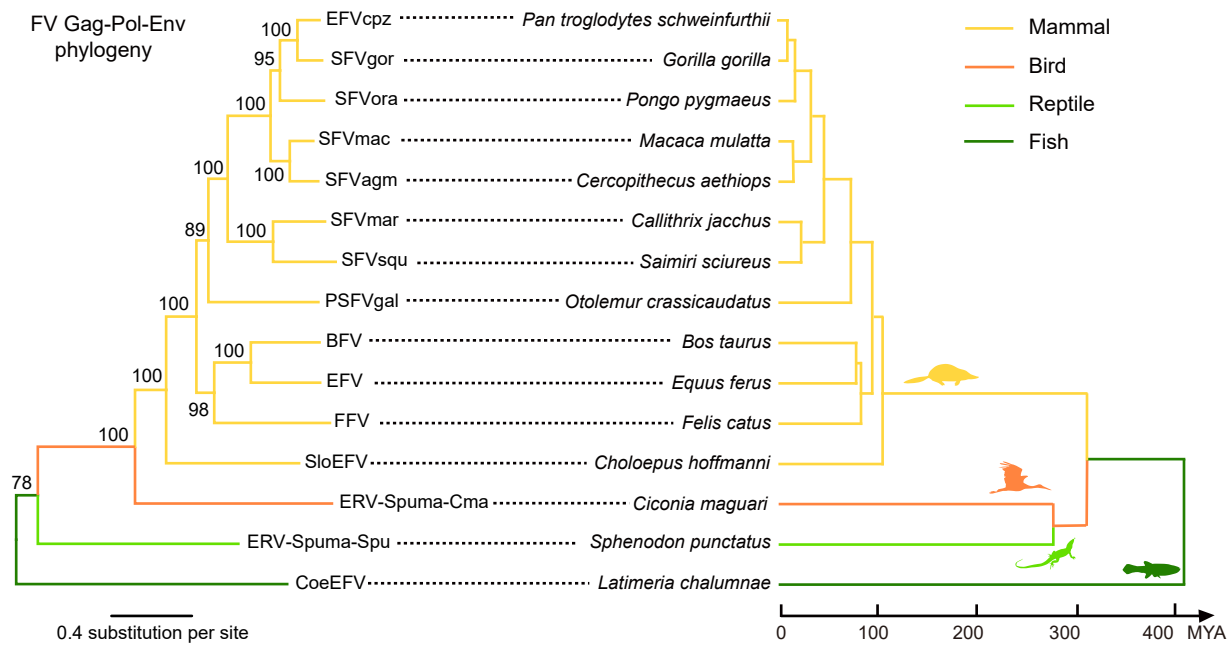
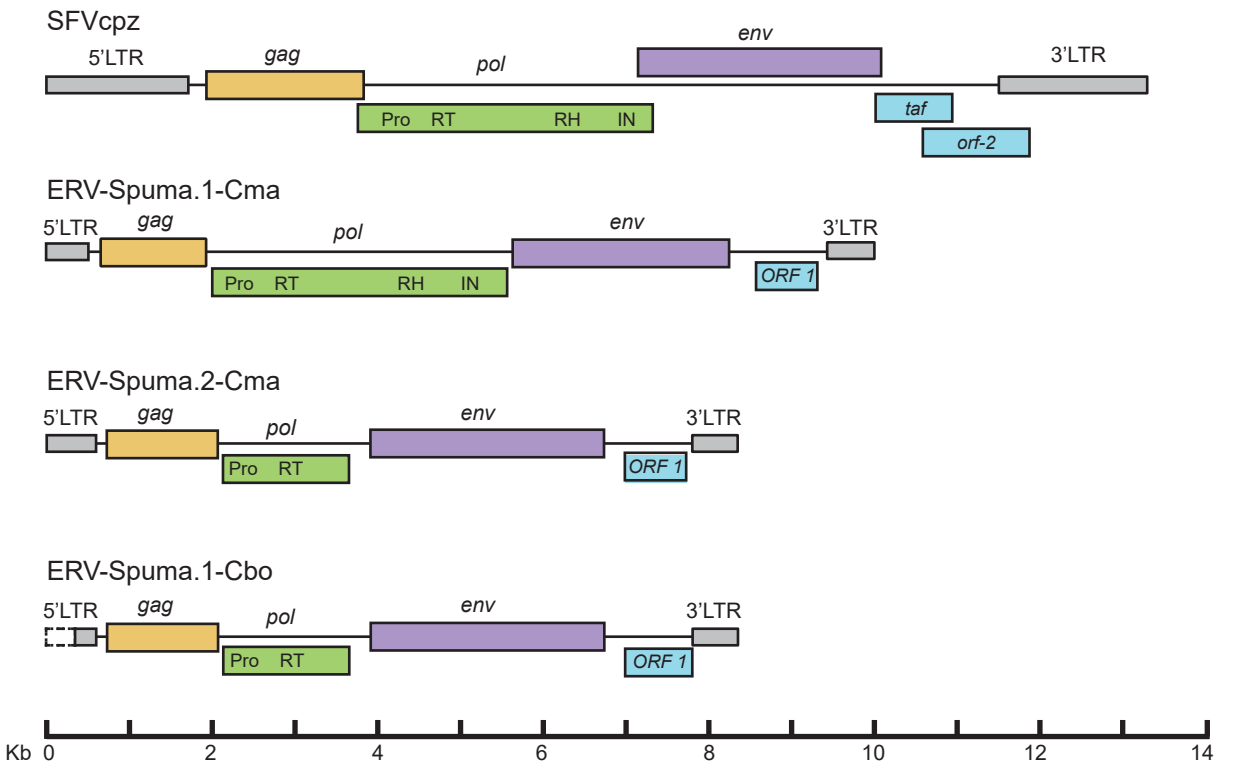


Fig. 3. Genomic organization of exogenous foamy viruses (acc: NC_001364) and avian endogenous foamy viruses. The dotted box marks the LTR deletion. LTR, long terminal repeat; Pro, protease; RT, reverse transcriptase; RH, ribonuclease H; IN, integrase.


0.4
SRV1
MPMV
MMTV
Beta
LDV
ALV
Alpha
STLV2
HTLV1
BLV
Delta
HIV1
FIV
EIAV
BIV
VMV
Lenti
PERV-A
MDEV
F-MuLV
FeLV
RfRV
Gamma
WEHV2
WEHV1
WDSV
Epsilon
SSSV
SnRV
SFVgor
SFVcpz
SFVora
SFVmac
SFVagm
SFVsqu
SFVspm
Mammal
EFV
BFV
FFV
PSFVgal
SloEFV
BDFF02022209_Cbo†
ERV-Spuma.1-Cma*
Bird
ERV-Spuma-Spu
Reptile
CoeEFV
Fish
Foamy
NviFLERV
PlatyfishEFV
AliFLERV
AciFLERV
DrFV-1
Foamy-like


FV Gag-Pol-Env phylogeny
EFVcpz
Pan troglodytes schweinfurthii
SFVgor
Gorilla gorilla
SFVora
Pongo pygmaeus
SFVmac
Macaca mulatta
SFVagm
Cercopithecus aethiops
SFVmar
Callithrix jacchus
SFVsqu
Saimiri sciureus
PSFVgal
Otolemur crassicaudatus
BFV
Bos taurus
EFV
Equus ferus
FFV
Felis catus
SloEFV
Choloepus hoffmanni
ERV-Spuma-Cma
Ciconia maguari
ERV-Spuma-Spu
Sphenodon punctatus
CoeEFV
Latimeria chalumnae
100
95
100
100
100
89
100
100
100
98
100
78
Mammal
Bird
Reptile
Fish
0.4 substitution per site
0
100
200
300
400
MYA


SFVcpz
5'LTR
gag
pol
env
3'LTR
Pro
RT
RH
IN
taf
orf-2
ERV-Spuma.1-Cma
5'LTR
gag
pol
env
3'LTR
Pro
RT
RH
IN
ORF 1
ERV-Spuma.2-Cma
5'LTR
gag
pol
env
3'LTR
Pro
RT
ORF 1
ERV-Spuma.1-Cbo
5'LTR
gag
pol
env
3'LTR
Pro
RT
ORF 1
Kb
0
2
4
6
8
10
12
14