Harding *et al. Retrovirology* (2025) 22:13
https://doi.org/10.1186/s12977-025-00669-y

Retrovirology

RESEARCH **Open Access**

Transcriptome mining reveals diversity and evolution of circulating and endogenous amphibian retroviruses

Emma F. Harding[1,2], Banjo Webster[1], Tanu Sridhar[1], Lewis K. Mercer[1] and Peter A. White[1*]

Abstract

Background The evolutionary history of retroviruses and their impact on vertebrate evolution remains poorly understood, particularly in non-mammalian hosts. In this study, we explore retroviruses associated with Amphibia through analysis of 169 RNA sequencing datasets from 102 amphibian species. Using a BLAST-based approach, we identified retroviral transcripts from assembled transcriptomes and phylogenetically characterise both their *pol* and *env* regions to elucidate their evolutionary history.

Results We identified the transcription of 18 novel and two previously described retroviruses with closest relatives in *gammaretrovirus*, *epsilonretrovirus*, *betaretrovirus* and *spumaretrovirinae*. Despite their differing *pol* phylogenies, we found that all amphibian retroviruses belong to the gamma-type envelope group (GTE). This suggests a common selection pressure for amphibian retroviruses to retain GTEs. Within these GTEs we also observed a new clade of alpharetrovirus-like envelopes in amphibians which form a sister clade to avian alpharetrovirus envelopes. Furthermore, we observe correlations between amphibian taxonomical order and retroviral diversity, with *Gymnophiona* (caecilians) harbouring the widest diversity of retroviruses whilst *Anura* (frogs and toads) harbour the fewest. Through mapping these transcribed retroviruses to their respective genomes (seven available) supplemented with observing ORF intactness, we determined that 14 of the 20 retroviruses are likely endogenous in origin yet are still transcribed in many amphibian tissues. These amphibian endogenous retroviruses (ERVs) have high genomic copy numbers: most (5/7) ERVs investigated have > 100 copies, and one of which has 9,219 integrations within the *Ichthyophis bannanicus* caecilian genome. This high retroviral load in amphibian genomes may suggest that these retroviruses have low pathogenicity, or may reflect a lack of transposon control mechanisms in amphibian cells.

Conclusions Through the characterisation of metatranscriptomic and genomic data from retroviruses in this study, we provide insights into their evolution in amphibians and exemplify the diversity of *Retroviridae* in vertebrate genomes. The identification of novel retroviral clades, widespread transcription of endogenous retroviruses in amphibians and abundance of ERV copies suggests that *Retroviridae* have played a significant role in amphibian evolution.

---

*Correspondence:
Peter A. White
p.white@unsw.edu.au
[1] School of Biotechnology and Biomolecular Science, UNSW Sydney, Sydney, Australia
[2] Present address: Department of Biology, University of Oxford, Oxford, UK

Introduction

Retroviruses are vertebrate-specific viruses capable of integrating into the host genome as an obligate part of their replication strategy and constitute a broad family - the *Retroviridae* [1, 2]. Members of this viral family are estimated to have arisen over 400 million years ago (mya) in ancient aquatic vertebrates- and are now ubiquitous across the animal kingdom [3, 4]. Retroviruses utilise a unique replication cycle that involves reverse transcription of their RNA genome and subsequent integration into the host chromosomal DNA. If the retroviral infection occurs in germline-cells, the resulting provirus can also be vertically transmitted and eventually become fixed within a population as an endogenous retrovirus (ERV). Following both genomic somatic and germ-cell integration, the retroviral DNA (provirus) can continue to produce active virions capable of horizontal transmission as exogenous retroviruses (XRVs) [5]. XRVs are classified based on the *pol* into two subfamilies, *Orthoretrovirinae* and *Spumaretrovirinae*, comprising six and five genera respectively. *Orthoretrovirinae* encompasses the majority of known retroviruses that cause disease in vertebrates and frequently host switch. The six *Orthoretrovirinae* genera include: alpharetrovirus (Rous sarcoma virus), betaretrovirus (Mouse mammary tumor virus), gammaretrovirus (Koala retrovirus), epsilonretrovirus (Walleye dermal sarcoma virus), deltaretrovirus (Simian T-cell leukemia virus) and lentivirus (Human immunodeficiency viruses). *Spumaretrovirinae* encompasses foamy viruses which stereotypically cause less pathogenesis and infrequently host switch. The six *Spumaretrovirinae* genera include: bovispumavirus, equispumavirus, felispumavirus, prosimiispumavirus and simiispumavirus.

Genetically ERVs often fall outside the 11 established XRV genera, either representing remnants of ancient extinct lineages or yet unsampled modern genera, therefore ERVs are classified into three larger groups to better portray their overarching characteristics: Clade I including gamma and epsilon-like ERVs, Clade II including beta, delta and alpha-like ERVs and Clade III including spuma-like ERVs [6]. Furthermore, the *env* regions of Clade I and II retroviruses are classified based on their whether their subunits covalently associate, as in the case of gamma-type envelopes (GTEs) or not as in beta-type envelopes (BTE) [7]. Comprehensive ERV studies have revealed the richness of retroviral diversity and concurrently highlighted the gaps in our understanding of *Retroviridae* evolution [8–10].

Amphibians consist of three major taxonomic groups: caecilians (*Gymnophiona*), salamanders/newts (*Urodela*), and frogs/toads (*Anura*). Amphibians occupy the ecological interface between aquatic and terrestrial environments and are in a unique position to accumulate and transmit viruses between animals from both environments. Evolutionarily, they represent an important animal group through which many viral families transitioned before infecting endotherms. For example both influenza and papilloma- viruses can trace their origins through amphibians [11, 12]. The *Retroviridae* is an ancient viral family that originated in fish and transitioned into terrestrial vertebrates [13, 14], yet their evolutionary path and diversification within amphibian hosts is largely unknown.

Further studies on retroviruses in non-mammalian hosts are crucial for developing a comprehensive picture of the evolutionary history of *Retroviridae* and their broader impact on vertebrate evolution. Current knowledge of amphibian retroviruses is limited to eight well-characterised representatives and few partial ERV genomes [15–21]. To that end, we explored the retroviruses associated with amphibian hosts through detection and characterisation of transcribed retroviral genomes from 169 RNA-Seq samples spanning 102 amphibian species. Where available, we mapped these retroviruses to amphibian genomes to explore ERV abundance. Additionally, we conducted phylogenetic analysis to describe these retroviruses and explore historical recombination events to gain insights into the mechanisms driving diversity within *Retroviridae.* Through the combined analysis of both XRVs and transcriptionally active ERVs, we leveraged ten million years of retroviral evolution to understand the diversity of retroviruses that infect amphibians.

Methods

Selection and assembly of amphibian datasets

For retrovirus sequence discovery in 102 different amphibian species, raw RNA-sequencing (RNA-seq) data sets (n = 169) were downloaded from the National Centre for Biotechnology Information's (NCBI) sequence read archive (SRA) using SRA toolkit v.3.0.0 (Table S1) in March 2022. Datasets were inclusive of all three orders of the *Amphibia* family including: Anura (frogs and toads) (n = 109), Urodela (salamanders and newts) (n = 40) and Gymnophiona (caecilians) (n = 20). Transcriptomes were assembled as previously published; briefly, dataset quality was assessed by FastQC prior to trimming with Trimmomatic and assembly with Trinity v.2.14.0 [12, 22], all using default settings.

Identification and annotation of retrovirus transcripts

To identify retroviral transcripts, transcriptomes were annotated using DIAMOND BLASTx [23] (default settings) using the NCBI non-redundant protein database from October 2022. Transcripts that were annotated as retroviral and >5 kB were imported into Geneious Prime v.2022.1.1 for further analysis. Retroviral homology was confirmed using a reciprocal DIAMOND BLASTx search and the detection of open reading frames (ORFs) for the core retroviral regions (*gag*/*pol*/*env*). Long terminal repeats (LTRs) on transcripts (where present) were identified using the 'Find Repeats' function in Geneious Prime with default parameters.

Phylogenetic analysis

Phylogenetic analysis was conducted to understand the evolutionary context of novel retroviruses by comparing them to previously characterised retroviruses. Trees of both the *pol* and *env* regions were generated to detect phylogenetic signals of recombination events. The polymerase and envelope from each amphibian retrovirus were translated in silico and aligned using MAFFT [24] with the localpair setting and 1,000 maximum iterations. Maximum likelihood trees were generated in iqTree2 [25] with 1,000 ultra-fast bootstrap replicates.

Host genomic mapping and ERV characterisation

To assess whether the retroviral transcripts arose from ERVs or XRVs, each amphibian retrovirus was mapped to the host genome where available ($n=7/102$) (Table S1). A Megablast search (E-value cut-off = 0.05) with retroviral transcript queries against the amphibian genome was used to identify the corresponding ERVs. ERVs were retained if they shared > 90% nt identity over > 6 kB of the corresponding retroviral transcript. The single ERV with the highest identity to the retroviral transcript query was designated as the representative ERV and used for further analysis.

LTRs were detected through a BLASTn search of the ERV to itself, with high nucleotide identity (>90%) regions (200–700 nt) flanking the *gag/pol/env* ORFs designated as LTRs. Integration dating was undertaken by aligning the LTRs using MUSCLE and calculating the nucleotide divergence using the Kimura 2-parameter model in MEGA 11 [26]. The time since integration was then calculated using the formula $T = (k/2)/2R$ where k is the nucleotide divergence and R is the rate of neutral evolution, estimated as 1.03×10^{-9} [27].

In the case where no host genome was available, retroviruses were categorised as endogenous if they contained frameshifts or stop codons within their *gag*, *pol* or *env* ORF. Additionally, if any motifs vital to retroviral function, namely the reverse-transcriptase motif C (YXDD) of *pol* and furin cleavage site (RXKR) of *env*, were mutated, it was also deemed endogenous. If a retrovirus was not predicted as endogenous by the above criteria, it was categorised as possibly exogenous.

Results

To discover novel active retroviruses associated with amphibian hosts, 169 raw RNA-sequencing datasets (Table S4) were screened for the presence of retroviral transcripts originating from ERVs and XRVs. We identified 20 transcribed amphibian retroviruses in this study, 18 of which are novel and two of which have previously been reported (Table 1) [19]. These viruses were discovered in 18% of the species investigated (n = 20/102), primarily in *Anura* species (n = 11) followed by *Gymnophiona* species (n = 5) and *Urodela* species (n = 4) (Fig. 1). Most retroviruses (n = 17/20) were identified within internal sampling tissue sites such as the liver, muscle, and mixed organs (Table S1). The remaining three retroviruses; dyeing poison dart frog retrovirus, rough skinned newt retrovirus and cayenne caecilian retrovirus were identified from skin tissue (Table S1).

Retroviruses from all three ERV clades were identified, nine of which were full length proviruses and 11 were partial, only missing the flanking LTR sequences. The retroviral genomes ranged in length from 5,557 to 8,933 nt with an average length of 7,129 ± 889 nt, as reported from the start of the *gag* ORF to the end of the *env* ORF (Table 1; Fig. 2). Clade I (*Gamma/Epsilon*) retroviruses were detected in all three amphibian orders, demonstrating a wide host range and long evolutionary association with *Amphibia* (Fig. 1; Table 1). Clade II (*alpha*/*beta*) retroviruses were detected only in the *Gymnophiona* order, infecting two caecilians from central America (Fig. 1; Table 1). Clade III (*Spuma*) retroviruses were detected solely in the *Urodela* class, specifically two newts and a salamander from Europe, the Middle East and North America [29] (Fig. 1; Table 1).

Clade II – *Betaretrovirus*

Clade II viruses are primarily associated with endothermic hosts, with only three previously discovered in ectotherms: two in pythons and one in a caecilian, all of which are endogenous [15, 30]. We identified two transcripts with *pol* identity to betaretroviruses in caecilian (*Gymnophiona*) hosts: the cayenne caecilian (*Microcaecilia unicolor*) and the Angoulême microcaecilia (*Microcaecilia dermatophaga*) (Fig. 1; Table 1), both native to central America [29, 31].

Both transcripts share high identity (96.7–99.5%) to ERVs-AB-Mun, a previously characterised endogenous retrovirus in the cayenne caecilian genome [15] (Table S1) and have the canonical *betaretrovirus* genome organisation of distinct *gag*, *pro*, *pol* and *env* ORFs (Fig. 2) and possess a YMDD RT-C motif, characteristic of Clade II viruses [32]. Both betaretrovirus transcripts contained full-length ORFs and could be XRVs, or transcriptionally active ERVs.

To determine if these betaretroviruses were also endogenised within the caecilians, we mapped the ERVs-AB-Mun transcript to the cayenne caecilian genome (aMicUni1.2). We identified 202 full-length ERVs with

Table 1 Retroviruses identified in amphibian hosts

Retrovirus	Endogenised	Intact ORFs	Host	Host order	Classification (pol)	Env classification	Length (nt)	Estimated age (MY)[a]	Genome available	Genomic ERV copies
ERV-AB-Mun	Yes	Yes	*Microcaecilia unicolor*	Gymnophiona	*Betaretrovirus*	*Alpharetrovirus*	7295	0	Yes	202
ERV-Epsilon-Dti	Yes	Yes	*Dendrobates tinctorius*	Anura	*Epsilonretrovirus*	*Epsilonretrovirus*	10,370	2.02	Yes	147
ERV-Epsilon-Pwa	Yes	Yes	*Pleurodeles waltl*	Urodela	*Epsilonretrovirus*	*Unclassified Clade I*	9906	9.27	Yes	953
ERV-Epsilon-Lpi	Yes	Yes	*Lithobates pipiens*	Anura	*Epsilonretrovirus*	*Unclassified Clade I*	8376	0	Yes	142
ERV-Gamma-Rbi	Yes	Yes	*Rhinatrema bivittatum*	Gymnophiona	*Gammaretrovirus*	*Alpharetrovirus*	9223	0.96	Yes	2
ERV-Gamma-Iba	Yes	No	*Ichthyophis bannanicus*	Gymnophiona	*Gammaretrovirus*	*Unclassified Clade I*	10,355	4.43	Yes	9,219
ERV-Epsilon-Rma	Yes	Yes	*Rhinella marina*	Anura	*Epsilonretrovirus*	*Epsilonretrovirus*	8592	2.47	Yes	20
Microcaecilia dermatophaga betaretrovirus	Unknown	Yes	*Microcaecilia dermatophaga*	Gymnophiona	*Betaretrovirus*	*Alpharetrovirus*	6444*	-	No	-
Brachycephalus rotenbergae retrovirus	Unknown	Yes	*Brachycephalus rotenbergae*	Anura	*Epsilonretrovirus*	*Epsilonretrovirus*	7231*	-	No	-
ERV-Epsilon-Bsu	Yes	No	*Brachycephalus sulfuratus*	Anura	*Epsilonretrovirus*	*Epsilonretrovirus*	6456*	-	No	-
ERV-Epsilon-Bci	Yes	No	*Boana cinerascens*	Anura	*Epsilonretrovirus*	*Epsilonretrovirus*	6220*	-	No	-
Emerald glass frog retrovirus	Unknown	Yes	*Espadarana prosoblepon*	Anura	*Epsilonretrovirus*	*Epsilonretrovirus*	7060*	-	No	-
Hong Kong paddy frog retrovirus	Unknown	Yes	*Fejervarya multistriata*	Anura	*Epsilonretrovirus*	*Epsilonretrovirus*	7354*	-	No	-
ERV-Epsilon-Npo	Yes	No	*Nymphargus posadae*	Anura	*Epsilonretrovirus*	*Epsilonretrovirus*	7067*	-	No	-
Red pumpkin toadlet retrovirus	Unknown	Yes	*Brachycephalus pitanga*	Anura	*Epsilonretrovirus*	*Epsilonretrovirus*	8933*	-	No	-
ERV-Epsilon-Hph	Yes	No	*Hyloscirtus phyllognathus*	Anura	*Epsilonretrovirus*	*Epsilonretrovirus*	5910*	-	No	-
Typhlonectes compressicauda retrovirus	Unknown	Yes	*Typhlonectes compressicauda*	Gymnophiona	*Gammaretrovirus*	*Gammaretrovirus*	7762*	-	No	-
ERV-Spuma-Clu	Yes	No	*Chioglossa lusitanica*	Urodela	*Spumaretrovirinae*	*Spumaretrovirinae*	11,977	-	No	-
ERV-Spuma-Nde	Yes	No	*Neurergus derjugini*	Urodela	*Spumaretrovirinae*	*Spumaretrovirinae*	10,890	-	No	-
ERV-Spuma-Tgr	Yes	No	*Taricha granulosa*	Urodela	*Spumaretrovirinae*	*Spumaretrovirinae*	6006*	-	No	-

*Nucleotide length is measured from the beginning of the *gag* ORF to the end of the *env* ORF, excluding leader regions and LTRs

[a] Age was only estimated when 2 LTRs were detected

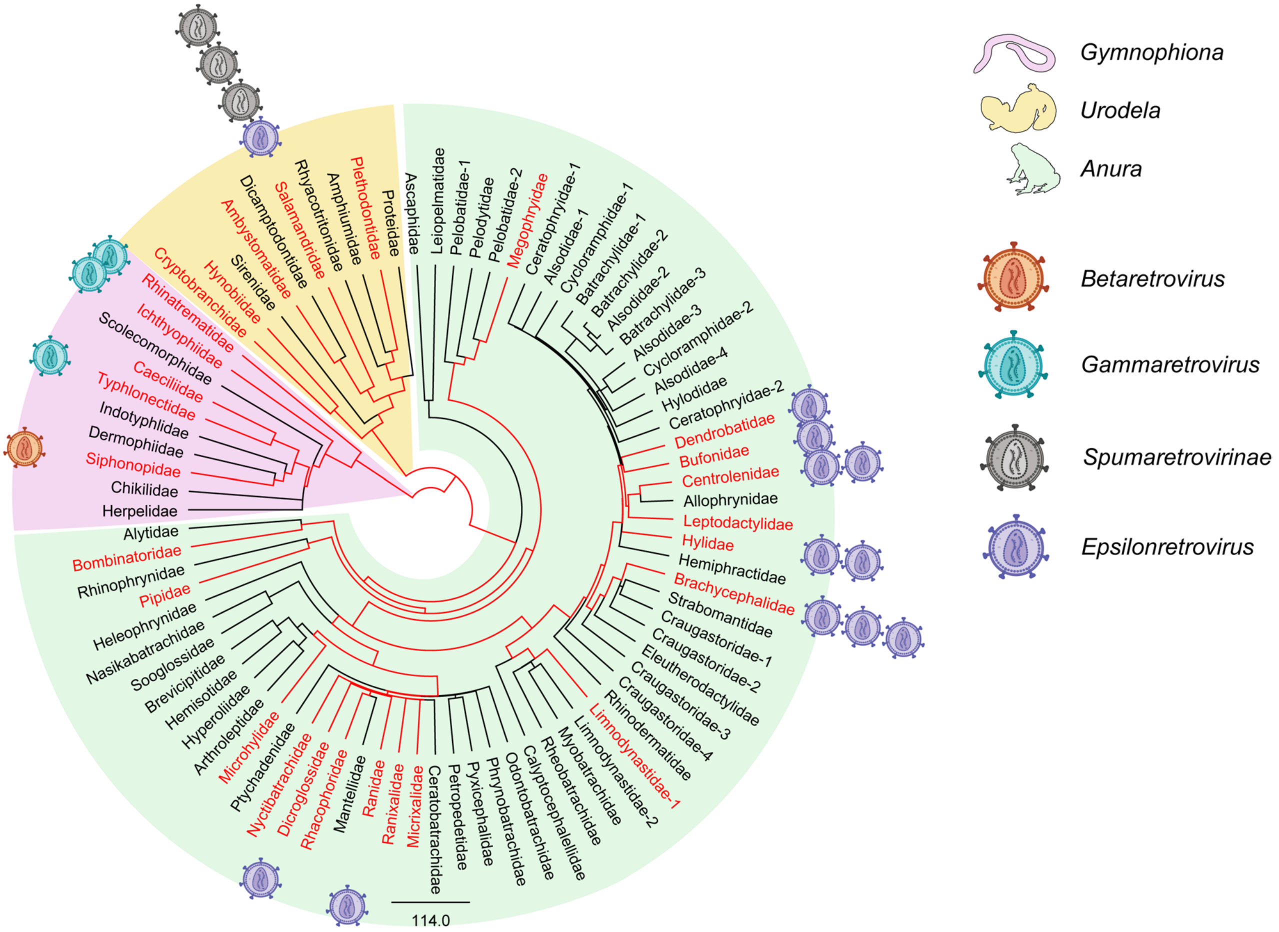

Fig. 1 Overview of 20 retroviruses discovered in amphibians. Amphibians with full-length retroviruses detected in their transcriptome are shown in red. The retroviruses are coloured depending on their polymerase phylogeny: orange denotes betaretrovirus, blue denotes gammaretrovirus, purple denotes epsilonretrovirus and grey denotes spumaretrovirus. The amphibian host phylogeny was generated with TimeTree.org [28] and coloured based on their class with pink denoting Gymnophiona (caecilians), yellow denoting Urodela (salamanders and newts) and green denoting Anura (frogs and toads)

>99% identity to the transcript, indicating either frequent retroviral infection of the germline or extensive retrotransposition within the genome to generate the abundance of ERV copies (Tables 1 and S2). The 325 bp LTRs flanking the representative ERV-AB-Mun provirus (NC_044044.1:50487792–50494734) are 100% identical, indicating the ERV was a very recent genomic integration (Tables S1 and S3). No genome was available for *Microcaecilia dermatophaga*.

To characterise the evolutionary history and context of these betaretroviruses we phylogenetically analysed both their *pol* and *env* regions (Fig. 3). The *pol* and *env* form a clade related to characterised alpharetroviruses of birds and show similar signs of recombination: they both contain a Clade II *pol* and a GTE (Fig. 3). Although phylogenetically they cluster closest to *alpharetrovirus*, their genome organisation contains a distinct and separate *pro* ORF, more similar to *betaretrovirus* than a*lpharetrovirus* (Fig. 2).

Clade I – *Gammaretrovirus* and epsilonretrovirus

Clade I retroviruses span *gammaretrovirus* and *epsilonretrovirus* genera and have been previously identified in a range of amphibian hosts [16–18, 33]. We identified 15 new amphibian Clade I retroviruses: three with identity to gammaretroviruses and twelve with identity to epsilonretroviruses (Fig. 1; Table 1).

The twelve epsilonretroviruses identified in this study were almost exclusively (n = 11/12) in Anuran hosts (Fig. 1; Table 1). The single non-anuran epsilonretrovirus was detected in the Iberian newt from northern Africa, whilst the 11 anuran epsilonretroviruses were detected in frogs and toads from across North, Central and South America and south-east Asia [29]. The three gammaretroviruses were all from caecilian (*Gymnophionan*) hosts: the two-lined caecilian (*Rhinatrema bivittatum*), the Banna caecilian (*Ichthyophis bannanicus*) and *Typhlonectes compressicauda* (Fig. 1; Table 1), all of which are native to central and South America [29].

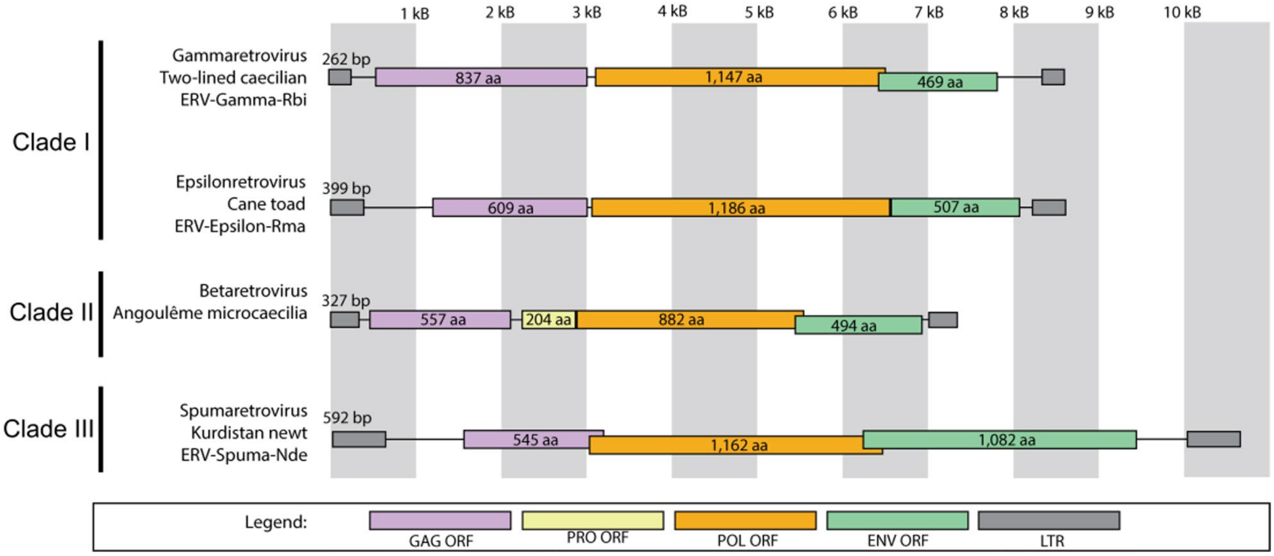

Fig. 2 Genome organisation of four representative amphibian retroviruses. Representative genomes from each retrovirus genera in this study are presented and annotated. Open reading frames and long terminal repeats are denoted by boxes and coloured as follows: purple: gag, yellow: pro, orange: pol, green: env, grey: LTR, white: unknown. Thin black lines denote untranslated regions of the retroviral genomes. The scale bar is shown along the top and drawn as background stripes

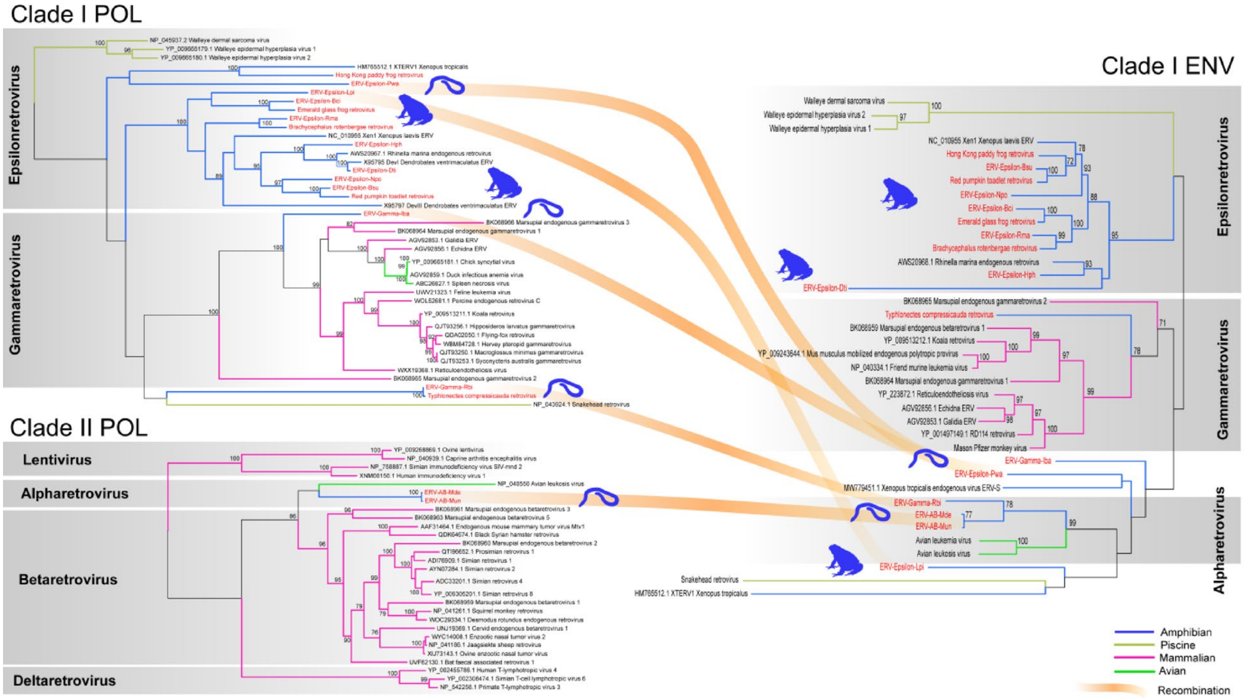

Fig. 3 Phylogeny of Clade I and Clade II amphibian retroviruses. Phylogenetic trees of the pol and env regions of amphibian retroviruses were constructed by aligning the translated ORFs using MAFFT, trimming the alignment with TRIMal and generating the tree using iqTree2 with 1,000 UF bootstraps. Support values over 70 are shown. Branches are coloured based on host as follows: yellow for piscine, blue for amphibian, pink for mammalian and green for avian. Orange lines denote recombination events between the pol and env regions. Viruses identified in this study are denoted with red text. Clade I and Clade II polymerase trees were generated using 420 AA and 850 AA respectively covering the reverse transcriptase domain. The clade I envelope tree was generated using 247 AA of the transmembrane domain of the envelope region

To determine if these viral transcripts came from ERVs or possible exogenous infecting retroviruses, we mapped them back to their host genomes where possible. Of the fifteen species with Clade I retroviral transcripts, six genomes were available and allowed us to confirm the endogenous origin of transcripts in *Dendrobates tinctorius* (ERV-Epsilon-Dti), *Pleurodeles waltl* (ERV-Epsilon-Pwa), *Lithobates pipiens* (ERV-Epsilon-Lpi), *Rhinella marina* (ERV-Epsilon-Rma), *Rhinatrema bivittatum* (ERV-Gamma-Rbi) and *Ichthyophis bannanicus* (ERV-Gamma-Iba) (Tables 1 and S2). Of the other retroviruses where no host genome was available, one had a truncated *env* ORF (350 nt) and therefore was also endogenous in origin (ERV-Epsilon-Npo), however the other seven transcripts contained intact ORFs and functional motifs, and may represent infecting XRVs (Table 1).

Endogenous clade I amphibian retroviruses

For Clade I retroviral transcripts that were endogenous in origin, we further analysed genomic representatives of each ERV to understand more about the age of these viral lineages. A representative ERV was chosen from each host genome for dating analysis (Table S1). The ERVs are all estimated to be between 0 and 9.3 million years old based on LTR divergence (Tables 1 and S3). The number of ERV insertions varied greatly from two copies of ERV-Gamma-Rbi in the two-lined caecilian to 9,219 copies of ERV-Gamma-Iba in the Banna caecilian (Tables 1 and S2). There was no obvious correlation between number of ERV copies and either viral genus (*epsilonretrovirus* vs. *gammaretrovirus*) or host (*Anura* vs. *Gymnophiona*).

The two-lined caecilian genome (aRhiBiv1.1) contained two full-length genomic insertions of the ERV-Gamma-Rbi retroviral transcript with 97.9 and 98.6% nt identity to the transcript respectively, both of which retained intact *gag*, *pol* and *env* ORFs (Table S2). The endogenous representative of ERV-Gamma-Rbi (NC_042616.1:304740150–304756914) had LTRs with 99% identity over their 508 bp, indicating this virus infected and integrated into the caecilian genome less than one million years ago (Table S3).

The Lithobates pipiens genome (aLitPip1_p1.0) contained 142 copies of ERV-Epsilon-Lpi sharing between 92.1 and 98.9% identity to the transcript (Table S2). The genomic ERV representative (JAZDVW010000070.1:4624284–4632130) shared 100% identity over the 438 bp LTRs (Table S3) indicating this viral lineage has integrated recently and may still be circulating as an XRV in leopard frogs.

The dyeing poison dart frog genome (ASM3965494v1) contained 147 copies of the ERV-Epsilon-Dti ranging from 99.1 to 99.9% identity to the retroviral transcript (Table S2). The representative ERV (JAYMGB0100000 35.1:13169521 – 13162600) had 495 bp LTRs with 98.0% identity, giving a time of approximately 2.0 million years since integration (Table S3).

Following the recent discovery of rhinella marina endogenous retrovirus (RMERV) in the cane toad [19], a second retrovirus ERV-Epsilon-Rma identified in this study had 20 genomic insertions, ranging from 96.2 to 99.3% nt pairwise identity to the transcript (Table S2). The 397 bp LTRs shared 97.0% identity, placing the estimated integration time at approximately 2.47 million years ago (Table S3).

The Iberian ribbed newt genome contained 953 ERVs with >90% identity to the retroviral transcript of ERV-Epsilon-Pwa, with the top hit only sharing 94.41% identity over the 8,012 transcript (Table S2). The low identity either means the transcript is from a related infecting virus, or the ERV is not yet fixed within the newt population (leading to heterogeneity in ERV integrations between the RNA-Seq individual and the genome individual). The closest genomic hit (NC_090438.1:284189964–284197968) shared 98.0% identity over the LTRs indicating it is approximately 9.3 million years old (Table S3).

The Banna caecilian ERV-Gamma-Iba contained 9,219 ERV copies with >90% identity to the retroviral transcript. The high copy number and varying levels of nucleotide identity to the ERV-Gamma-Iba transcript may stem from post-integration divergence or may indicate many similar retroviruses have integrated into the Banna caecilian recently. The ERV representative (JAW IIG010000005.1:718113171–718133617) has 98.0% identical 637 bp LTRs, indicating an integration time of approximately 4.4 million years ago (Table S3).

Phylogeny of clade I amphibian retroviruses

To investigate the evolution of Clade I amphibian retroviruses and identify signs of recombination, we compared the phylogeny of the *pol* and *env* regions of each retrovirus. Four of the fifteen Clade I retroviruses showed signs of recombination between the *pol* and *env* genomic regions, three of which were in *Gymnophiona* hosts (Fig. 3).

Of the Clade I retroviruses identified, twelve shared closest *pol* identity with previously characterised amphibian epsilonretroviruses [17, 18, 33]. These epsilonretroviruses fell in one of two phylogenetic clades, grouping either with RMERV and Xenopus laevis ERV Xen1, or grouping with Xenopus tropicalus endogenous retrovirus (XTERV) (Fig. 3). The XTERV branch, containing the two retroviruses Hong Kong paddy frog virus and ERV-Epsilon-Pwa, is situated between the Walleye dermal viruses in fish and basal to the other amphibian epsilonretrovirus RMERV clade (Fig. 3). The RMERV clade contained the other ten epsilonretroviruses from this study, forming an anuran-specific clade (Fig. 3). With the exception of

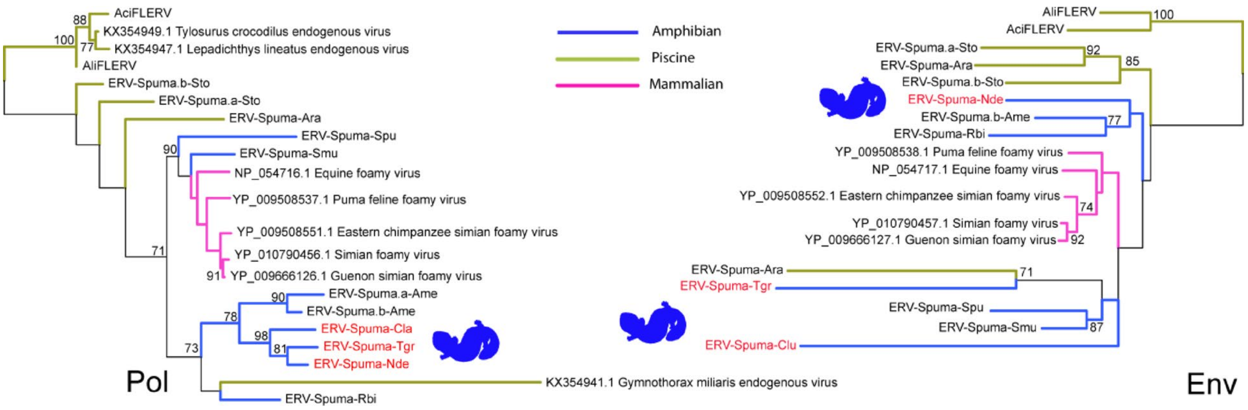

Fig. 4 Phylogeny of Clade III amphibian retroviruses. Phylogenetic trees of the pol and env regions of amphibian retroviruses were constructed by aligning the translated ORFs using MAFFT, trimming the alignment with TRIMal and generating the tree using iqTree2 with 1,000 UF bootstraps. Support values over 70 are shown. Branches are coloured based on host as follows: yellow for piscine, blue for amphibian and pink for mammalian. Viruses identified in this study are denoted with red text. The polymerase and envelope trees were constructed with 535 AA and 363 AA respectively. Blue silhouettes denote Urodela hosts

ERV-Epsilon-Lpi and ERV-Epsilon-Pwa, the epsilonretroviruses showed no signs of recombination, with their *env* phylogeny closely matching that of the *pol* (Fig. 3). The two recombinants ERV-Epsilon-Lpi from the Northern leopard frog and ERV-Epsilon-Pwa from the Iberian ribbed newt had *env* regions with homology to *Alpharetrovirus* envelopes (Fig. 3), indicating at least one recombination event occurred in their evolutionary history.

The three other Clade I retroviruses discovered shared highest *pol* identity with viruses from the *gammaretrovirus* genus and also fell into two phylogenetic clusters: ERV-Gamma-Iba clustered within endotherm gammaretroviruses, whilst the two-lined caecilian and typhlonectus compressucauda retroviruses formed their own lineage placed between characterised *Gammaretrovirus* and *Epsilonretrovirus* (Fig. 3). All three of these retroviruses had differing *pol* and *env* phylogenies, indicating they were all recombinants between gammaretrovirus *pol* and alpharetrovirus-like *env* (Fig. 3). The ERV-Gamma-Rbi *env* is closely related to the two amphibian betaretrovirus *envs*, which together form a novel sister clade to avian alpharetroviruses.

Interestingly, *Typhlonectes compressicauda* retrovirus shared 95.6% identity with ERV-Gamma-Rbi over the *gag*/*pol* regions, indicating they both originated from a similar retroviral progenitor (Fig. 3). However, the *env* regions of *Typhlonectus compressicauda* retrovirus and ERV-Gamma-Rbi only shared 38.4% nt identity, providing evidence of one or more recombination events to acquire diverse envelopes within caecilian Clade I retroviruses (Fig. 3).

Clade III - *Spumaretrovirinae*

Three retroviruses discovered in this study clustered within the *Spumaretrovirinae* subfamily with previously characterised endogenous spumaviruses from amphibians [13]. These viruses were present in geographically distributed hosts from Europe (Gold-striped salamander), Middle east (Kurdistan spotted newt) and the USA (Rough skinned newt). The two newt retrovirus transcripts had highest BLAST identity to ERVs from the eastern newt (*Notophthalmus viridescens*) genome [13], sharing between 54.6 and 66.7% nt identity over the 4,086 nt *pol* region.

All three spumaviruses had interrupted or truncated ORFs, indicating they were endogenous transcripts and had lost their coding capability. These three ERVs formed an amphibian-specific *Spumaretrovirinae* clade with ERV-Spuma.a-Ame and ERV-Spuma.b-Ame from the axolotl (*A. mexicanum*) genome [21] (Fig. 4). This clade is basal to and distinct from mammalian spumaviruses, suggesting a wider unsampled diversity of amphibian spumaviruses and a novel amphibian spumavirus lineage (Fig. 4). The phylogenies between *pol* and *env* did differ, however the lack of amphibian spumaviruses sampling leaves the recombinant nature of these ERVs ambiguous. Further sampling will aid in resolving the evolutionary history of these retroviruses.

ERV-Spuma-Clu contained a hypothetical accessory protein (502 aa) on the 3’ end of the genome similar to other *Urodela* spumavirus ERVs (Fig. 2) [13]. This ORF did not have BLAST homology to any proteins in the NCBI non-redundant database (BLASTp, e-value < 1) and therefore could not be assigned a putative function.

Discussion

Despite the diversity of amphibians globally, very little is known about retroviruses that infect them, with only eight well-characterised endogenous retroviruses known [15–19, 21]. Recent studies have begun to characterise and understand the retroviral diversity in amphibians [13, 15, 16], however knowledge of their retroviruses still significantly lags behind that of mammals and birds. This study surveyed amphibian RNA-sequencing data to capture both infecting XRV sequences and transcriptionally active ERVs to better understand the viral load of retroviruses in amphibians and the evolution of *Retroviridae* in amphibia. We identified 20 retroviruses, 6 of which were full-length and may represent infecting XRVs. The other 14 retroviruses were endogenised within amphibian genomes and integrated between 0 and 9.93 million years ago. Despite their age, all 14 ERVs were still transcriptionally active and detectable in a variety of different tissues. Whilst their transcription may be a by-product of the unsilenced ERV loci, it may also indicate that the ERVs are functional in some amphibian cells: actively transposing, causing pathologies or restricting viral infection. Overall, this study provides insight into the retroviruses that have co-evolved with amphibians over millions of years and highlights the unsampled diversity of *Retroviridae* in ectothermic hosts.

An important limitation of this study is the availability and distribution of datasets covering amphibian taxa. We selected RNA-sequencing spanning 102 species and the three amphibian orders to best represent sampling efforts to date, however many host lineages remain unsampled. Improved sampling of diverse taxa would allow further exploration of the trends observed in this study. In the case of endangered amphibian populations, better sampling will allow better understanding of the retroviral challenges these populations face.

Amphibians can harbour retroviruses from all three clades

Exogenous retroviruses are only a fraction of historical retrovirus diversity, and endogenous retroviruses often do not fit into modern XRV genera. To better incorporate both exogenous and endogenous retroviruses, a class classification system based on the *pol* region is used wherein Clade I comprises modern *Gammaretrovirus* and *Epsilonretroviruses*, Clade II comprises modern *Alpharetroviruses*, *Betaretroviruses*, *Deltaretroviruses* and *Lentiviruses*, and Clade III comprises the *Spumaretrovirinae* subfamily [6]. In this study we identified retroviruses from all three clades in amphibians, substantially expanding the known diversity of the amphibian retrovirome.

The majority of retroviruses identified in this study were epsilonretroviruses (12/20), the genus most commonly associated with amphibians (Fig. 1; Table 1). Indeed, the first two amphibian retroviruses discovered in *Xenopus* frogs were both epsilonretroviruses, followed by more recent discoveries in *Rhinella* toads [17–19]. This study identified 12 novel epsilonretroviruses from *Anuran* and newt hosts which phylogenetically separate into two distinct amphibian clades (Fig. 3). These clades are not distinguished by host type or geographic region, instead suggesting multiple lineages of epsilonretroviruses that have evolved and circulate within the same areas concurrently. Both amphibian *epsilonretrovirus* clades are distinct from the canonical epsilonretroviruses from walleye fish and likely represent new genera.

The discovery of a *betaretrovirus* ERV within caecilians was the first evidence of a Clade II retrovirus infecting anamniotes [15]. This study confirmed the presence and active transcription of that ERV not only in the cayenne caecilian, but also a similar retrovirus in another caecilian *Microcaecilia dermatophaga*. These two species are estimated to have diverged >50 million years ago [28], however the high identity between their retroviruses suggests the presence of an exogenous retrovirus that can infect both species and has since integrated into their genomes. The high identity of their LTRs also supports recent retroviral integration and suggests extant infecting amphibian betaretroviruses are still circulating. Interestingly, no Clade II retroviruses were identified outside *Gymnophiona* hosts (Fig. 1), suggesting that caecilians have recently and uniquely acquired Clade II retroviruses rather than coevolving with them. Further exploration of the host biology underpinning betaretroviral infection in caecilians will aid our understanding of the history and evolution of this retroviral genus.

We also identified Clade III spumaretroviruses exclusively within *Urodela* amphibians, all of which were endogenised within their respective genomes. Endogenous spumaretroviruses have been identified in amphibians previously [16, 21], however this study identifies the first full-length endogenous spumaretroviruses from *Urodela*. Spumaretroviruses have been well studied in mammalian hosts, but only recently were they discovered in non-mammalian hosts and their origins better estimated [13, 21, 34]. The three *Urodela* spumaretroviruses cluster with previously identified *Gymnophiona* spumaretroviruses, comprising an amphibian spumaretrovirus clade that sits basally to amniote spumaretroviruses (Fig. 4). This indicates a much wider diversity of amphibian spumaretroviruses which have not yet been sampled, and likely are representatives of larger anamniotic spumaretrovirus lineages. The limited degradation of spumaretrovirus retroviruses in both *Gymnophiona* and *Urodela* genomes suggest that they are recent integrations rather than remnants of ancient retroviruses [21], and the modern exogenous forms may still be circulating within amphibian populations.

Gamma-type envelopes are ubiquitous in amphibians

The retroviral envelope plays an important role in facilitating interactions between the virion and its host and harbours viral glycoproteins which are important for cell entry [35]. As such, understanding how different retroviral envelope phylogenies interact across varying vertebrae species is vital in developing our understanding of retroviral evolution, infection and transmission. Therefore, this project aimed to identify recombination events in amphibian retroviruses to elucidate information about retroviral genetic exchange across the *Amphibia* family.

Gamma-type envelopes (GTEs), comprising envelopes with homology to alpharetroviruses, gammaretroviruses and epsilonretroviruses, are ubiquitous within amphibian and fish genomes [36]. We also observed this envelope bias, with all amphibian retroviruses identified in this study except the *Spumaretrovirinae* (17/20) containing GTEs, despite their differing pol phylogenies (Fig. 3). This is most apparent in *Gymophiona* hosts which contained beta- and gammaretroviruses (based on pol phylogeny) but contained GTEs which formed a monophyletic sister clade to avian alpharetroviruses (Fig. 3). The observed envelope convergence suggests strong evolutionary selection on the envelope region within amphibians favouring GTEs. The ubiquity of GTEs in aquatic hosts may indicate they confer an advantage for aquatic transmission, possibly by enhancing virion stability in environments with varying osmotic pressure. Further studies exploring the biochemical properties of GTEs will help us understand this apparent retroviral bias in the aquavirome.

The similarity between the caecilian retrovirus envelopes and those of avian retroviruses suggests previous vertebrate inter-class recombination. Two evolutionary scenarios could give rise to the observed envelope similarities: one or more host jumps introducing the avian-like *env* into caecilians or caecilians evolving with an avian-like *env* lineage derived from a common ancestor. Due to the long divergence time between birds and caecilians and the lack of Clade II ERVs in any other amphibian order, scenario one is more likely. Avians and caecilians have many opportunities to interact, both directly through predator-prey relationships, and indirectly through avian faecal or fomite contamination of caecilian habitats. As more genomic and transcriptomic data becomes available, future studies focussing on the evolution and host range of these alpharetrovirus-like envelopes will allow us to better understand the movement of retroviruses between hosts in an ecosystem.

The retrovirome of amphibians differs greatly between orders

The class *Amphibia* is taxonomically divided into three orders: *Gymnophiona* (caecilians), *Anura* (frogs and toads) and *Urodela* (salamanders and newts). These orders diverged ~ 320 million years ago and have since diversified into over 8,000 current species [28]. This study screened 102 amphibian species from all three orders to identify novel retroviruses and better understand the amphibian virome. The retroviromes of each amphibian order was quite distinct: *Anura* (frogs and toads) retroviruses were exclusively e*psilonretrovirus*, *Urodela* (salamanders and newts) retroviruses were a mix of *Gammaretrovirus* and *Spumaretrovirinae*, whilst *Gymnophiona* (caecilians) were both *Gammaretrovirus* and *Betaretrovirus* (Fig. 1).

This difference suggests that, although amphibian orders often geographically overlap and share environments, the retroviruses that infect them are order-specific. This observed distinction may be due to a variety of factors, including differences between their cellular retroviral defences, adult habitat and lifestyle and diet. The activity of APOBEC proteins, known vertebrate anti-retroviral defences, differs between amphibian species and may influence amphibian susceptibility to certain retroviral infections [37]. Further studies focusing on amphibian immunobiology would help determine what drives their retrovirome distinctions and limits their host ranges.

ERVs remain transcriptionally active within amphibian genomes for millions of years

We identified transcriptionally active ERVs in a range of amphibian species, ranging from recent integrations to ERVs ~ 9.26 million years old. Compared to mammals, where ERVs are rapidly inactivated and degraded through methylation and anti-retroviral APOBEC activity [38, 39], amphibian ERVs remain transcribed and relatively intact within their respective genomes. Previously characterised amphibian retroviruses are between 1.2 and 14 million years old [16, 17, 33] and together with this study suggest an increase in retrovirus invasion and amplification within amphibian genomes in the last 15 million years. Further paleovirological studies of ERVs within amphibian genomes will elucidate infection, integration and persistence patterns of amphibian retroviruses, which in turn will better inform us of the pathogenic pressures amphibians have co-evolved with.

In addition, amphibian ERVs are often present as multiple full-length copies within their host genomes – ERVs-AB-Mun betaretrovirus within the cayenne caecilian has over 100 full-length integrated copies which retain protein coding capabilities [15] and RMERV epsilonretrovirus in the cane toad retains eight full-length copies [19]. Overall, this also suggests a lack of negative selection on integrated amphibian ERVs, suggesting they may remain transcriptionally active long after they have integrated and fixated within a population. As ERV transcription can be correlated with pathologies [40, 41], studies investigating the burden of ERV transcription in amphibians would be useful to better understand amphibian immunology.

We identified many ERVs with >100 genomic copies (Table 1), one of which had 9,217 copies, indicating extensive activity post-integration and limited transposon control/regulatory mechanisms. Uneven ERV distributions have been previously observed in mammalian species [8, 10, 42, 43], however whether these are host- or viral-driven processes remains to be determined. The ERV heterogeneity reported in this study could provide insights into historical retrovirus pandemics, species bottlenecks or simply reflect differences in species anti-retroviral genomic mechanisms. Further work exploring correlations between high-copy ERV lineages and species with high ERV burdens will aid our understanding of the processes involved in endogenization and ERV persistence in vertebrates.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12977-025-00669-y.

Supplementary material 1.

Acknowledgements

This research includes computations using the computational cluster Katana supported by Research Technology Services at UNSW Sydney. EFH, LKM and TS acknowledge funding through the Australian Government Research Training Program.

Author contributions

E.F.H: Conceptualisation, Admin, Writing - original draft, Curation, Investigation, Validation, Writing - review and editing, Formal analysis, Software, B.J.W: Writing - original draft, Curation, Validation, Formal analysis, Methodology, Software, Visualisation, T.S: Writing - review and editing, L.K.M: Curation, Writing - review and editing, Methodology, Software, VisualisationP.A.W: Conceptualisation, Funding, Admin, Supervision, Writing - review and editing.

Data availability

The retroviral genomes described in this study are deposited on NCBI Genbank under the accession numbers BK071386-BK071405. All datasets used are publicly available on the NCBI Sequence Read Archive (SRA) and their accessions are listed in Supplementary Table 4.

Declarations

Competing interests

The authors declare no competing interests.

Received: 29 August 2025 / Accepted: 17 November 2025

References

1. Arnaud F, Caporale M, Varela M, Biek R, Chessa B, Alberti A, Golder M, Mura M, Zhang Y-p, Yu L, Pereira F, DeMartini JC, Leymaster K, Spencer TE, Palmarini M. A paradigm for Virus–Host coevolution: sequential Counter-Adaptations between endogenous and exogenous retroviruses. PLoS Pathog. 2007;3(11):e170.
2. Kaján GL, Doszpoly A, Tarján ZL, Vidovszky MZ, Papp T. Virus–Host Coevolution with a focus on animal and human DNA viruses. J Mol Evol. 2020;88(1):41–56.
3. Weiss RA. The discovery of endogenous retroviruses. Retrovirology. 2006;3(1):67.
4. Coffin JM, Varmus HS. HE The Place of Retroviruses in Biology, in *Retroviruses* 1997, Cold Spring Harbour Laboratory Press Cold Spring Harbour.
5. Greenwood AD, O'Brien SP, Roca AL, Eiden MV. Transmission, evolution, and endogenization: lessons learned from recent retroviral invasions. Microbiol Mol Biol Rev, 2018; 82(1).
6. Gifford R, Tristem M. The Evolution, distribution and diversity of endogenous retroviruses. Virus Genes. 2003;26(3):291–315.
7. Henzy JE, Coffin JM. Betaretroviral envelope subunits are noncovalently associated and restricted to the mammalian class. J Virol. 2013;87(4):1937–46.
8. Hayward A, Cornwallis CK, Jern P. Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. Proc Natl Acad Sci. 2015;112(2):464–9.
9. Hayward JA, Tachedjian M, Cui J, Field H, Holmes EC, Wang L-F, Tachedjian G. Identification of diverse full-length endogenous betaretroviruses in megabats and microbats. Retrovirology. 2013;10(1):35.
10. Harding EF, Mercer LK, Yan GJH, Waters PD, White PA. Invasion and amplification of endogenous retroviruses in dasyuridae marsupial genomes. Volume 41. Molecular Biology and Evolution; 2024. 8.
11. Parry R, Wille M, Turnbull OMH, Geoghegan JL, Holmes EC. Divergent influenza-like Viruses Amphibians Fish Support Anc Evolutionary Association Viruses, 2020;12(9).
12. Harding EF, Russo AG, Yan GJH, Mercer LK, White PA. Revealing the uncharacterised diversity of amphibian and reptile viruses. ISME Commun. 2022;2(1):95.
13. Aiewsakun P, Katzourakis A. Marine origin of retroviruses in the early palaeozoic era. Nat Commun. 2017;8(1):13954.
14. Naville M, Volff JN. Endogenous retroviruses in fish genomes: from relics of past infections to evolutionary innovations? Front Microbiol. 2016;7:1197.
15. Chen M, Guo X, Zhang L. Unexpected discovery and expression of amphibian class II endogenous retroviruses. J Virol, 2021;95(3).
16. Yedavalli VRK, Patil A, Parrish J, Kozak CA. A novel class III endogenous retrovirus with a class I envelope gene in African frogs with an intact genome and developmentally regulated transcripts in xenopus tropicalis. Retrovirology, 2021;18(1).
17. Sinzelle L, Carradec Q, Paillard E, Bronchain OJ, Pollet N. Characterization of a xenopus tropicalis endogenous retrovirus with developmental and Stress-Dependent expression. 2011;85(5): pp. 2167–79.
18. Kambol R, Kabat P, Tristem M. Complete nucleotide sequence of an endogenous retrovirus from the amphibian, xenopus laevis. Virology. 2003;311(1):1–6.
19. Russo AG, Eden JS, Tuipulotu DE, Shi M, Selechnik D, Shine R, Rollins LA, Holmes EC, White PA. Viral discovery in the invasive Australian cane Toad (Rhinella marina) using metatranscriptomic and genomic approaches. J Virol, 2018;92(17).
20. Tristem M, Herniou E, Summers K, Cook J. Three retroviral sequences in amphibians are distinct from those in mammals and birds. J Virol. 1996;70(7):4864–70.
21. Chen Y, Zhang YY, Wei X, Cuia J. Multiple infiltration and cross-species transmission of foamy viruses across the paleozoic to the cenozoic era. J Virol, 2021;95(14).
22. Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics; 2010.
23. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59–60.
24. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
25. 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.
26. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 2018;35(6):1547–9.
27. Crawford AJ. Relative rates of nucleotide substitution in frogs. J Mol Evol. 2003;57(6):636–41.
28. 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.
29. AmphibiaWeb. University of California, Berkeley, CA, USA. 2024.
30. Huder JB, Böni J, Hatt JM, Soldati G, Lutz H, Schüpbach J. Identification and characterization of two closely related unclassifiable endogenous retroviruses in pythons (Python molurus and python curtus). J Virol. 2002;76(15):7607–15.
31. Wilkinson M, Sherratt E, Starace F, Gower DJ. A new species of Skin-Feeding caecilian and the first report of reproductive mode in microcaecilia (Amphibia: gymnophiona: Siphonopidae). PLoS ONE. 2013;8(3):e57756.
32. Sharma PL, Nurpeisov V, Schinazi RF. Retrovirus reverse transcriptases containing a modified YXDD motif. Antiviral Chem Chemother. 2005;16(3):169–82.
33. Russo AG, Harding EF, Yan GJH, Selechnik D, Ducatez S, DeVore JL, Zhou J, Sarma RR, Lee YP, Richardson MF, Shine R, Rollins LA, White PA. Discovery of

novel viruses associated with the invasive cane Toad (Rhinella marina) in its native and introduced ranges. Frontiers in Microbiology; 2021;p. 12.
34. Wei X, Chen Y, Duan G, Holmes EC, Cui J. A reptilian endogenous foamy virus sheds light on the early evolution of retroviruses. Virus Evol, 2019;5(1).
35. Henzy JE, Johnson WE. Pushing the endogenous envelope. Philos Trans R Soc Lond B Biol Sci. 2013;368(1626):20120506.
36. Chen Y, Wang X, Liao ME, Song Y, Zhang YY, Cui J. Evolution Genetic Divers Retroviral Envelope Anamniotes J Virol, 2022. 96(8).
37. Wang W, Han GZ. A Long-Running arms race between APOBEC1 genes and retroviruses in tetrapods. J Virol. 2023;97(1):e0179522.
38. Modzelewski AJ, Gan Chong J, Wang T, He L. Mammalian genome innovation through transposon domestication. Nat Cell Biol. 2022;24(9):1332–40.
39. Feschotte C, Gilbert C. Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet. 2012;13(4):283–96.
40. Kitsou K, Lagiou P, Magiorkinis G. Human endogenous retroviruses in cancer: oncogenesis mechanisms and clinical implications. J Med Virol. 2023;95(1):e28350.
41. Buttler CA, Chuong EB. Emerging roles for endogenous retroviruses in immune epigenetic regulation*. Immunol Rev. 2022;305(1):165–78.
42. Tsangaras K, Mayer J, Greenwood AD. Crossing wallace's line: an evolutionarily young gibbon ape leukemia virus like endogenous retrovirus identified from the Philippine flying Lemur (Cynocephalus volans). Sci Rep. 2025;15(1):9790.
43. Magiorkinis G, Gifford RJ, Katzourakis A, De Ranter J, Belshaw R. Env-less endogenous retroviruses are genomic superspreaders. Proc Natl Acad Sci U S A. 2012;109(19):7385–90.

Publisher's note