The diversity and evolution of retroviruses: Perspectives from viral “fossils”

Jialu Zheng, Yutong Wei, Guan-Zhu Han

Jiangsu Key Laboratory for Microbes and Functional Genomics, College of Life Sciences, Nanjing Normal University, Nanjing, 210023, China

ARTICLE INFO

Keywords:
- Retroviruses
- Endogenous retroviruses (ERVs)
- Origin and evolution
- Diversity

ABSTRACT

Retroviruses exclusively infect vertebrates, causing a variety of diseases. The replication of retroviruses requires reverse transcription and integration into host genomes. When infecting germline cells, retroviruses become inherited vertically, forming endogenous retroviruses (ERVs). ERVs document past viral infections, providing molecular fossils for studying the evolutionary history of retroviruses. In this review, we summarize the recent advances in understanding the diversity and evolution of retroviruses from the perspectives of viral fossils, and discuss the effects of ERVs on the evolution of host biology.

1. Introduction

Retroviruses (the Retroviridae family) exclusively infect vertebrates and cause a variety of diseases, including acquired immune deficiency syndrome (AIDS) and cancers (Goff, 2007; Coffin et al., 1997; Telesnitsky, 2010). Unique among RNA viruses, retrovirus replication requires reverse transcription and integration into host genomes. Retroviruses usually infect somatic cells. But when retroviruses infect germline cells, the integrated retroviruses are passed on vertically to the next generation, forming endogenous retroviruses (ERVs). ERVs were first discovered in 1960s (Weiss, 2013). In the pre-genomic era, many ERVs had been identified by Southern blot and/or PCR approaches (e.g. Repaske et al., 1985; Tristem et al., 1995; Tristem et al., 1996). Thanks to the recent development of next-generation sequencing techniques, the genomes of many vertebrates have been sequenced, providing important resources for studying the extent and distribution of ERVs with the vertebrate genomes. ERVs are widespread and highly abundant in vertebrate genomes; for example, ERVs account for ~8% of the human genome (Hayward et al., 2013, 2015; Xu et al., 2018).

It is notoriously hard to study the deep evolution of viruses over geographic time scale due to high mutation rates of viruses (Duffy et al., 2008; Gago et al., 2009) and no physical fossils preserved. Retroviruses are no exception. Reverse transcriptase lacks the ability of proofreading, resulting in high mutation rates in retroviruses. However, ERVs evolve with their hosts as genomic loci, and thus have much lower evolutionary rates (Duffy et al., 2008; Gago et al., 2009). ERVs provide “molecular fossils” for studying the deep history of retroviruses and the ancient evolution of host-retrovirus interaction. More generally, many viruses, both RNA and DNA viruses, can integrate into host genomes, forming endogenous viral elements (EVEs) (Holmes, 2011; Feschotte and Gilbert, 2012). The discovery of EVEs in a wide range of eukaryotes lays the foundation of an emerging field, namely Paleovirology. Paleovirology is the study of the ancient evolution of viruses through analyzing endogenous viral elements or evolutionary signatures of host-virus conflicts left in the host genomes (Emerman and Malik, 2010; Katzourakis, 2013). In this review, we introduce methods to analyze ERVs, summarize the recent advances in understanding the diversity and evolution of retroviruses from the perspectives of viral fossils, and discuss the effects of ERVs on the evolution of host biology.

2. Retroviruses and endogenous retroviruses

Retroviruses are enveloped viruses with RNA genomes ~7–10 kb in length. All the retroviruses encode three common genes, namely gag, pol, and env. The gag gene encodes matrix (MA), capsid (CA), and nucleocapsid (NC) proteins, the pol gene encodes protease (PR), reverse transcriptase (RT), RNase H (RH), and integrase (IN) proteins, and the env gene encodes transmembrane (TM) and surface (SU) glycoproteins. Some complex retroviruses can also encode several accessory genes (Goff, 2007; Coffin et al., 1997; Telesnitsky, 2010). Unlike other RNA viruses, the replication of retroviruses requires reverse transcription of viral RNA genomes into double stranded DNA, and integration of the newly synthesized DNA into host genomes. Retrovirus infection typically occurs in the somatic cells of hosts. On occasion, retroviruses infect host germline cells, and the integrated retroviruses flank by two long terminal repeats (LTRs) become vertically inherited to next generation as part of the
host chromosomes, forming endogenous retroviruses (ERVs) (Stoye, 2012; Johnson, 2015) (Fig. 1).

Like new mutations, the fate of novel retrovirus insertions within the host population is governed by population genetics processes. As new mutations, most of new ERV insertions are deleterious and removed from host population by purifying selection (Wang and Han, 2020). The integration of exogenous retroviruses occurs preferentially near or within host genes, but ERVs are found mainly outside host genes, suggesting that a majority of ERVs have been removed by negative selection (Brady et al., 2009; Emerman and Malik, 2010). The fates of the ERVs that have little effect on host fitness and are selectively neutral are governed by random genetic drift. The probability that a newly retroviral insertion is fixed in the host population is 1/2N, where N is the effective population size of hosts (Hartl and Clark, 2006). These neutral ERVs might accumulate disruptive mutations (premature stop codons, frameshift mutations, and insertions or deletions), which eventually lead to the degradation of these ERVs. In extreme but frequent cases, recombination occurs between 5'- and 3-LTRs, generating a solo-LTR. In the rare cases that an ERV insertion is beneficial, it will be fixed in the host population by positive selection (Fig. 2). In the meantime, ERVs retain the ability to proliferate within the host genomes by retrotransposition in cis (ERVs use their own proteins for amplification), complementation in trans (ERVs hitchhike proteins produced by other ERVs or transposable elements within the same cell), or recombination (ERVs produce virus particles to infect other germline cells), increasing their copy numbers within the host genomes (Belshaw et al., 2005; Bannert and Kurth, 2006; Magiorkinis et al., 2012) (Fig. 3). Together, these processes generate complex insertion polymorphism pattern of ERVs within the host populations.

ERV insertion polymorphisms have been found in the populations of many vertebrates (Zhang et al., 2008; Turner et al., 2001; Roca et al., 2004; Polani et al., 2010; Ngo et al., 2019). An ongoing fixation of endogenous koala retrovirus (KoRV) had been observed in the wild koala (Phascolarctos cinereus) population: all the individuals in the north of Australia had endogenous KoRV, while KoRV was not detected in some individuals in the south of Australia (Tarlinton et al., 2006). Consistently, southern koala individuals carried fewer KoRV loci than northern ones (Ishida et al., 2015). The KoRV endogenization started before the late 1800s, but no more than 22,200–49,990 years ago (Avila-Arcos et al., 2013; Ishida et al., 2015). Tens of HERV-K elements were found to be polymorphic in the human population (Wildschutte et al., 2016), and several HERV-K elements were absent from modern human genomes and present in archaic (Denisovan and Neanderthal) hominids (Lee et al., 2014), suggesting the recent activity of the HERV-K elements in the human population.

3. Mining ERVs

ERVs often accumulate a large number of mutations, making them difficult to identify. Several main strategies have been commonly used to mine ERVs within the vertebrate genomes. (i) ERVs can be identified using similarity search with representative retroviruses as queries. However, representative retrovirus insertions might not cover the major diversity of retroviruses. (ii) Because ERVs are flanked by two LTRs, ERVs can be identified through first detecting candidate LTRs and then detecting and reconstructing retroviral proteins, as exemplified by RetroTector (Sperber et al., 2007). However, ERVs frequently accumulate large-scale deletions and become fragmented, and thus not all the ERVs are flanked by two LTRs. (iii) ERVs can also be identified through first searching homologs of retrovirus hallmark RT proteins. Because retrovirus RT proteins share detectable similarity with retrotransposon RT proteins, phylogenetic analyses are then performed to identify ERVs that cluster with representative retroviruses (Xu et al., 2018). Once again, fragmented ERVs that lack RT proteins cannot be identified using this approach. In general, the approach based on identifying RT homologs is more sensitive than the one based on detecting LTRs (Xu et al., 2018). Nevertheless, old ERVs are often more difficult to detect than young ERVs.

![Fig. 1. Pattern of retrovirus transmission.](image-url)

Retroviruses integrate into the chromosomes of germline cells, forming endogenous retroviruses (ERVs). The new ERVs might be fixed in the host population either by positive selection or random genetic drift. Moreover, ERVs might increase their copy numbers during the evolutionary course. The colored circles represent germline cells, and green and orange lines represent host genomes and ERVs, respectively.
4. Classification and taxonomy of retroviruses

Based on the latest release of International Committee on the Taxonomy of Viruses (ICTV), the Retroviridae family is classified into two subfamilies, namely Orthoretrovirinae and Spumaretrovirinae. The Orthoretrovirinae subfamily includes six genera (Alpharetrovirus, Betaretrovirus, Gammaretrovirus, Deltaretrovirus, Epsilonretrovirus, and Lentivirus), and the Spumaretrovirinae subfamily (also known as foamy viruses) includes five genera (Bovispumavirus, Equispumavirus, Felispumavirus, Prosimiispumavirus, Simiispumavirus) (Hendrickson et al., 2018) (Fig. 4A). However, the ICTV classification only accounts for exogenous retroviruses. ERVs have been traditionally classified into three classes based on their similarity and phylogenetic relationship with exogenous retroviruses. Class I ERVs cluster with Gammaretrovirus and Epsilonretrovirus, Class II ERVs cluster with Alpharetrovirus, Betaretrovirus, Deltaretrovirus, and Lentivirus, and Class III ERVs cluster with foamy viruses and ERV-L (Gifford and Tristem, 2003; Gifford et al., 2018) (Fig. 4A). For historical reasons, human ERVs (HERVs) had been grouped into families primarily based on the specificity of tRNA primer binding sites (PBS) (Bannert and Kurth, 2006; Jern and Cofrin, 2008). However, this classification system had problems: ERVs that belong to the same phylogenetic group might have different types of PBS, while unrelated ERVs might share the same type of PBS, and mutations and deletions were present in PBS regions (Bannert and Kurth, 2006; Jern and Coffin, 2008). Therefore, the classification of exogenous and endogenous retroviruses has not been well incorporated. Most of ERVs cannot be assigned into certain genus. It is also challenging to
classify and name ERVs, when ERVs arise from recent endogenization events are nested with the diversity of exogenous retroviruses (Blomberg et al., 2009). It is logic to include both exogenous and endogenous retroviruses in a taxonomy of retroviruses (Blomberg et al., 2009; Gifford et al., 2018). Based on phylogenetic analyses of Pol proteins from exogenous and endogenous retroviruses sampled across vertebrates, retroviruses have been classified into five major clades, designated Jin, Mu, Shui, Huo, and Tu (Xu et al., 2018). Clades Jin and Mu include viruses related to gammaretroviruses and epsilonretroviruses, respectively, and they include class I ERVs. Clade Shui includes viruses related to alpha-, beta-, delta-retroviruses, lentiviruses, and class II ERVs. Clade Huo is related to snakehead retrovirus, ERV-L, and class III ERVs. Clade Tu is related to foamy viruses (Fig. 4A). However, RH and IN proteins were acquired multiple times in retroviruses, making phylogenetic analyses based on Pol proteins should be taken with cautions (Wang and Han, 2021a). On the other hand, based on phylogenetic analyses of the hallmark RT proteins, the positions of some retroviruses are unstable, and the support values for some groups are low. The recent discovery of lokiretroviruses makes the situation more complicated (Wang and Han, 2021a). Lokiretroviruses are sister to all the known retroviruses, and might represent a novel subfamily of retroviruses. We anticipate a classification system that fully integrates endogenous and exogenous retroviruses and follows phylogeny will be developed in the future.

5. Dating ERV invasion

The time for ERV invasion into host genomes can help establish time scales for retrovirus evolution. Several methods have been developed to estimate the integration time of ERVs. (i) The most commonly used approach is based on the divergence of 5'- and 3'-LTRs of an ERV. When integrated into host genomes, 5'- and 3'-LTRs of an ERV are identical at sequence level (Coffin et al., 1997). Then, they accumulate mutations independently as distinct genomic loci. Therefore, the sequence

Fig. 4. Retrovirus taxonomy and distribution. A Classification and taxonomy of retroviruses. On the top is the phylogenetic relationship of retroviruses. The gray and yellow rectangles highlight Orthoretrovirinae and Spumaretrovirinae subfamily, respectively. The dark blue dashed rectangle represents the ICTV classification system. B Distribution of retrovirus clades in vertebrates under different taxonomic systems. On the left is the phylogenetic relationship of vertebrate groups, whereas on the top-right is the phylogenetic relationships among the major groups of retroviruses under different taxonomic systems. α, β, γ, δ, ε, Lentivirus, and Spuma-represent Alpharetrovirus, Betaretrovirus, Gammaretrovirus, Deltaretrovirus, Epsilonretrovirus, Lentivirus, and Spumaretrovirinae, respectively. The dark blue and orange circles indicate the presence of corresponding exogenous retroviruses and ERVs based on literature (Xu et al., 2018; Zheng et al., 2021; Hayward et al., 2015; Han and Worobey, 2012a, 2012b, 2012c, 2014, 2015; Chen et al. 2019, 2021a, 2021b; Wang and Han, 2021a).
divergence between 5'- and 3'-LTRs are proportional to the integration time (Johnson and Coffin, 1999). The integration time of an ERV, \( T = D/2\mu \), where \( D \) indicates the genetic distance between 5'- and 3'-LTR, and \( \mu \) indicates host neutral evolutionary rate (Fig. 5A). However, the approach comes with at least two caveats: (i) 5'- and 3'-LTRs may evolve at a rate different from host neutral evolutionary rate (Martins and Villoosen, 2011); gene conversion frequently occurs in LTR regions, which decreases their genetic distance (Johnson and Coffin, 1999). Moreover, given ERVs frequently accumulate large-scale deletion and become fragmented, not all the ERVs are flanked by two LTRs and thus their ages cannot be estimated using the approach based on the divergence between 5'- and 3'-LTRs. (ii) The age of an ERV can also be estimated based on its orthologous insertions within closely related species. The possibility that two ERVs integrated independently at an orthologous position of closely related species is extremely low (Stoye, 2001). Therefore, ERV insertion should occur in the last common ancestor of species who share orthologous ERVs, and the divergence time of their last common ancestor provides a minimum time for the ERV Insertion. This approach might sometimes underestimate the invasion time, if an orthologous insertion is lost in certain outgroup species (Fig. 5B). (iii) ERV insertions might undergo segmental duplication. ERVs loci generated by segmental duplication can be identified by ERVs flanking with highly similar host sequences. Calculating the divergence time of segmental duplicates provides a minimum estimate of the invasion time (Katzourakis et al., 2007) (Fig. 5C). (iv) Rarely, the invasion time of an ERV can be estimated based on the occurrence of stop codons within it. Under neutral evolution, stop codons are generated at a rate of \( \sim 1/2310 \) per codon per million years in primate coding sequences (Belyi et al., 2010). However, it is unclear whether ERVs evolve neutrally after integration, and the reliability of this method should be taken with caution (Fig. 5D).

6. Origin, evolution, and diversity of retroviruses

Systemic analyses of ERVs in vertebrates provide important insights into the origin, evolution, and diversity of retroviruses. Large-scale phylogenetic analyses show that ERVs are present in all the genomes of jawed vertebrates studied to date (Hayward et al. 2013, 2015; Xu et al., 2018; Zheng et al., 2021). Within jawless vertebrates, ERVs (later classified to endogenous loki retroviruses) were also identified in sea lamprey (Petromyzon marinus) and Arctic lamprey (Lebetheron camtschaticum) (Hayward et al., 2015; Xu et al., 2018; Wang and Han, 2021a). ERVs have not been identified in any species outside vertebrates (Xu et al., 2018; Wang and Han, 2021a). The distribution of retroviruses suggests that retroviruses may have originated during the early evolution of vertebrates >450 million years ago (Xu et al., 2018). However, the distribution might also be explained by a later origin and spread across vertebrate through frequent host switching. Foamy viruses, a subfamily of retroviruses, have been found to mainly co-diverge with their jawed vertebrate hosts for >450 million years (Han and Worobey, 2012a; Aiewsakun and Katzourakis, 2017; Xu et al., 2018; Wei et al., 2019; Chen et al., 2021a). Moreover, retroviruses of tetrapods (amphibians, reptiles, birds, and mammals) nest within the diversity of fish retroviruses, and fish retrovirus occupy basal positions of all major retrovirus clades. These findings reveal an ancient aquatic origin of retroviruses >450 million years ago (Xu et al., 2018).

The isolation and characterization of retroviruses have been highly biased to human and vertebrates of agricultural or medical importance. ERVs expand our understanding of the host distribution of major retrovirus groups (Fig. 4B). Clade Jin retroviruses (including gamma-related retroviruses) infect mammals, reptiles, and birds (Xu et al., 2018; Zheng et al., 2021). Clade Mu retroviruses (including epsilon-related retroviruses) infect almost all jawed vertebrates (Xu et al., 2018). Clade Shui retroviruses (including alpha-, beta-, delta-related retroviruses and lentiviruses) infect amphibians, reptiles, birds, and mammals; among them, delta-retroviruses and lentiviruses are mammal specific (Katzourakis et al., 2007; Gifford et al., 2008; Kekecsova et al., 2009; Gilbert et al., 2009; Han and Worobey, 2012b, 2015; Cui and Holmes, 2012; Hron et al. 2016, 2018; Farkasová et al., 2017; Xu et al., 2018; Chen et al. 2021a, 2021b; Zheng et al., 2021). Clade Hiu retroviruses (ERV-L and snakehead retrovirus related) infect all the major lineages of vertebrates (Xu et al., 2018). Clade Tu retroviruses (or fomay viruses) infect all major lineages of jawed vertebrates (Han and Worobey, 2012a; Han and Worobey, 2012b; Han and Worobey, 2014; Katzourakis et al., 2014; Ruboyianes and Worobey, 2016; Aiewsakun and Katzourakis, 2017; Xu et al., 2018; Wei et al., 2019; Aiewsakun et al., 2019; Chen et al., 2019; Aiewsakun et al., 2020; Chen et al., 2021a). Recently, phylogenomics has identified a novel major retrovirus subfamily known as loki retrovirus (Wang and Han, 2021a). Lokiretroviruses whose Env proteins share detectable similarity with fusion proteins of some negative sense single-stranded RNA viruses form a sister group to all known retroviruses. Endogenous loki retroviruses are widely distributed in vertebrates, including lampreys, ray-finned fishes, lobe-finned fishes, amphibians, and reptiles (Fig. 4B). The number of ERVs from different retrovirus groups is dramatically different. Fossil records of

![Fig. 5. ERV integration time estimation.](image-url)
delta-retroviruses, lentiviruses, and foamy viruses are extremely rare, which may reflect their inefficiency in entering germline cells, or inherent barriers of replication in germline cells (Johnson, 2015). Most of ERVs lack exogenous retrovirus counterparts, and known exogenous retroviruses cluster into several “small” groups within the large ERV phylogeny (Hayward et al. 2013, 2015; Xu et al., 2018), suggesting that an enormous number of exogenous retroviruses remain to be discovered.

The discoveries of ERVs also help establish or refine the time scales of retrovirus evolution. Molecular clock analyses of exogenous lentiviruses and delta-retroviruses indicate their evolutionary time scales in the unit of hundred years, and both have been thought to be young viral groups (Van Dooren et al., 2001; Wertheim and Worobey, 2009). Endogenous lentiviruses have been identified in lagomorphs, primates, carnivores, and Dermoptera (Katzourakis et al., 2007; Keckesova et al., 2009; Gifford et al., 2008; Gilbert et al., 2009; Han and Worobey, 2012b, 2015; Cui and Holmes, 2012). These endogenous lentiviruses invaded host genomes from 1.9 to 3.8 million years ago (Mya) (in Malagasy lemurs) to 21–40 Mya (in Colugo) (Han and Worobey, 2012b, 2015; Cui and Holmes, 2012; Hron et al., 2018). Endogenous delta-retroviruses have been identified in bats, carnivores, cetaceans, and insectivores (Farková et al., 2017; Hron et al., 2018; Zheng et al., 2021). A delta-retrovirus was estimated to integrate into carnivore genomes >24.6 Mya (Hron et al., 2018). These findings strongly support that both lentiviruses and delta-retroviruses are not young viral groups, and they have been circulating in mammals for millions of years.

7. Biological function of endogenous retroviruses

Upon infecting host germ cells, ERVs are inherited in the Mendelian way. ERVs have various effects on their hosts: deleterious, neutral, and beneficial (Fig. 2). If ERVs are inserted into host genes, they are likely to interrupt the normal open reading frames and function of the host genes, which are generally detrimental. Indeed, ERVs are mainly found outside of the host genes (Brady et al., 2009; Emerman and Malik, 2010). ERVs have been proposed to be responsible for a wide range of diseases, such as cancer, autoimmune diseases, and neurogenerative diseases, but these hypotheses remain controversial and unproven (Voisset et al., 2008; Stoye, 2012; Bhardwaj and Coffin, 2014; Küry et al., 2018). On the other hand, the transcripts or proteins of HERV s might represent biomarker markers for the diagnosis of those diseases (Song et al., 2021). The deleterious ERV insertions will be removed from the host population by purifying selection. Most of the remaining ERVs are neutral and will become degraded due to the accumulation of disruptive mutations, and insertions or deletions. Beneficial ERVs will be recruited by hosts for host biological functions, a process known as co-option or exaptation (Wang and Han, 2020, 2021b) (Fig. 2).

BROADLY, the functional diversity of ERVs can be divided into three categories, namely physiological, immunological, and regulatory functions (Frank and Feschotte, 2017; Feschotte, 2008; Chuong et al., 2016). (i) Physiological roles. Some ERVEnv proteins, termed as syncytins, were repurposed to function in placenta formation and development (Mi et al., 2000; Blaise et al., 2003). Syncytins originated tens of times in mammals and a viviparous lizard, which might be pivotal in the emergence of mammals (Dupressoir et al., 2005; Lavialle et al., 2013; Cornels et al., 2017). Moreover, Syncytins might also contribute to myoblast fusion and muscle sexual dimorphism in mice (Redelsperger et al., 2016). (ii) Immunological roles. Some ERV-derived proteins can protect the hosts from exogenous retrovirus infection. Expression of the ERV en gene is known to inhibit exogenous retrovirus infection by saturating receptors recognized by exogenous retroviruses (Aswad and Katzourakis, 2012; Kozak et al., 1984; McDougall et al., 1994; Palmarini et al., 2004; Tikhonenko and Lomonoskaya, 1990). Hosts also employ the ERV-derived proteins to block the replication of exogenous retroviruses. Some endogenous jaagsiekte sheep retroviruses (JSRVs) express defective Gag proteins, which prevent virion release and interfere with exogenous JSRV (Arnaud et al., 2007; Palmarini et al., 2004; Mura et al., 2004). Fv1 gene was derived from the gag gene of a murine retrovirus that was inserted into the genomes of species in Muroidae ~45 million years ago, and serves as the restriction factor to protect the hosts against infections of various retroviruses (Boso et al., 2018; Young et al., 2018; Best et al., 1996; Yap et al., 2014; Yan et al., 2009). Moreover, the level of ERV expression controls host’s response to the microbiota in homeostatic or inflammatory manner (Lima-Junior et al., 2021). (iii) Regulatory roles. Given that ERVs contain specific regulatory sequences modulating their RNA synthesis and processing, they may provide new regulatory elements and thus influence the expression pattern of nearby host genes (Feschotte, 2008; Stoye, 2012). ERVs have been found to shape the evolution of transcriptional networks underlying immune response and placental function (Finnegan, 2012; Chuong et al., 2016, 2017; Pasquesi et al., 2020; Sun et al., 2021). Moreover, ERV-derived long noncoding RNA, designated Inc-EPAV, serves as a positive regulator of innate immune responses (Zhou et al., 2019). HERVH might act as an enhancer associated IncRNA that interacts with mediator kinase proteins and pluripotency-associated transcription factors to regulate the transcription of pluripotency-associated genes (Sexton et al., 2022). Besides, homologous recombination of ERVs at different chromosome regions can mediates chromosome rearrangement, which might play a major role in vertebrate genome evolution (Hughes, 2001).

Recently, a comprehensive analysis of vertebrate genomes identified 177 independent retroviral gene (gag and env) co-option events in vertebrates, suggesting ERV genes had been frequently co-opted during the evolution of vertebrates (Wang and Han, 2020). The evolution of retroviral gene co-option followed a birth and death model, in which retrovirus genes were frequently co-opted, and co-opted retroviruses were frequently lost (Wang and Han, 2020, 2021b). These co-opted retroviral genes were subject to different selection pressure, implying potentially diverse cellular functionalities, but the specific cellular functionality remained to be studied (Wang and Han, 2020).

8. Outstanding questions

The discoveries of ERVs have greatly refined our understanding of the origin, deep evolution, and diversity of retroviruses. ERVs also shape the evolution of the genome complexity of their hosts. However, many outstanding questions remain: (i) Retroviruses have been thought to originate during the early evolution of vertebrates >450 Mya. However, it remains unclear how retroviruses originated (Hayward, 2017). (ii) Multiple classification systems of retroviruses have been developed, but all these systems have their own shortcomings. A classification system that fully integrates endogenous and exogenous retroviruses and follows phylogeny is eagerly needed. (iii) Retrovirus genes or regulatory sequences have been frequently co-opted in vertebrates, but their functionalities need to be experimentally characterized. (iv) The relationship between ERVs and human diseases should be fully evaluated. (v) ERVs are ubiquitously distributed in the vertebrate genomes, but only tens of exogenous retroviruses have been isolated. More efforts should be paid to isolate exogenous retroviruses in wild animals and assess their potential pathogenicity, which are crucial for predicting and controlling zoonotic disease outbreaks in the future. (vi) ERVs are highly abundant in vertebrates, but some vertebrate genomes harbor more ERVs than others (Hayward et al. 2015, 2015; Xu et al., 2018; Zheng et al., 2021). It remains obscure about the factors that shape the landscape of ERVs within a certain genome.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgement

This work was supported by National Natural Science Foundation of China (31922001).
References

Aiewsakun, P., 2020. Avian and serpentine endogenous foamy viruses, and new insights into the macroevolutionary history of foamy viruses. Virus Evol. 6, evoz057.

Aiewsakun, P., Katzourakis, A., 2017. Marine origin of retroviruses in the early Paleozoic 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.

Arnaud, F., Caporello, M., Varelis, M., Bieck, R., Chessa, B., Alberti, G., Golder, M., Mura, M., Zhang, Y.P., Yu, L., Pereira, F., Demartin, J.C., Leymaster, K., Spencer, T.E., Palese, P., 2007. A paradoxical vertical host-coevolution: sequential counter-adaptations between endogenous and exogenous retroviruses. PLoS Pathog. 3, e170.

Aswad, A., Katzourakis, A., 2012. Paleoviropology and virally derived immunity. Trends Genet. 27, 627–641.

Ávila-Arcos, M.C., Ho, S.Y., Ishida, Y., Nikolaidis, N., Tsangaras, K., Hulva, P., Benda, P., Gifford, R.J., Elleder, D., 2017. Extremely high mutation rate of a primate endogenous retrovirus. Proc. Natl. Acad. Sci. U. S. A. 114, 10901–10906.

Belyi, V.A., Levine, A.J., Skalka, A.M., 2010. Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes. PLoS Pathog. 6, e1001030.

Benoit, S., Teissier, P., Towers, G., Stoye, J.P., 1996. Positional cloning of the mouse retrovirus restriction gene Fv1. Nat. Genet. 832, 826-829.

Blaise, S., Cofﬁn, J.M., 2017. Endogenous retroviruses and human cancer: is there anything to the rumors? Cell Host Microbe 23, 155–259.

Blaise, S., de Parseval, N., Biek, R., Belyi, V., Cofﬁn, J.M., 2003. Genomewide screening for fosogenic human endogenous retrovirus envelopes identiﬁes syncytnia 2, a gene conserved on primate evolution. Proc. Natl. Acad. Sci. U.S.A. 100, 13013–13018.

Blumberg, J., Benachouche-Fouquet, S., Bilskast, V., Speghter, G., Mayer, J., 2009. Classiﬁcation and nomenclature of endogenous retroviral sequences (ERVs): problems and recommendations. Gene 448, 115–123.

Bocca, G., Buckler-White, A., Buckler-White, A., Kozak, C.A., 2018. Ancient evolutionary origin and positive selection of the retroviral restriction factor Fv1 in muroid rodents. J. Virol. 92, 10991–10998.

Cheng, X., Zhang, Y., Wei, X., Zhang, G., Holmes, E.C., Cui, J., 2019. The evolution of endogenous viral elements. Cell Host Microbe 10, 1083–1090.

Cofﬁn, J.M., 2013. Paleovirology: inferring viral evolution from host genome sequence data. Philos. Trans. R. Soc. Lond. B Biol. Sci. 368, 20120493.

Cornwell, W.K., Pendrak, J., Huffman, J.M., 2008. Constructing primate phylogenies from ancient virus sequence data. Proc. Natl. Acad. Sci. U. S. A. 105, 20367–20372.

Covas, J., Elleder, D., 2016. Life history of the oldest lentivirus: characterization of ELVyg integrations in the dermopteran genome. Mol. Biol. Evol. 33, 2659–2669.

Cypser, G., Gifford, R.J., Barna, R., Wang, D., Fields, S., Elleder, D., 2007. Rearrangement of a lentivirus in two Malagasy lemurs. PLoS Genet. 3, e217.

D Diallo, A.I., Enamorado, M., Cataisson, C., Gil, L., 2018. Implications for lentivirus evolution. Proc. Natl. Acad. Sci. U.S.A. 105, 20362–20367.

Diallo, A.I., Elleder, D., 2016. Origin of the retroviral genomes. Nat. Commun. 8, 13954.

Diallo, A.I., Elleder, D., 2016. Origin of the retroviral genomes. Nat. Commun. 8, e13954.

Diallo, A.I., Elleder, D., 2016. Origin of the retroviral genomes. Nat. Commun. 8, e13954.

Diallo, A.I., Elleder, D., 2016. Origin of the retroviral genomes. Nat. Commun. 8, e13954.

Diallo, A.I., Elleder, D., 2016. Origin of the retroviral genomes. Nat. Commun. 8, e13954.

Diallo, A.I., Elleder, D., 2016. Origin of the retroviral genomes. Nat. Commun. 8, e13954.

Diallo, A.I., Elleder, D., 2016. Origin of the retroviral genomes. Nat. Commun. 8, e13954.
