Molecular Biology and Evolution, 2025, **42**, 1–13
https://doi.org/10.1093/molbev/msaf171
Advance access publication 10 July 2025
Discoveries

MBE OXFORD

Phylogenomics Unveils the Complex Evolution of Retroviruses in Birds

Qinliu He [1], **Qiyan Liu** [1], **Zhen Gong** [1], **Guan-Zhu Han** [1,*]

[1]College of Life Sciences, Nanjing Normal University, Nanjing, Jiangsu 210023, China

***Corresponding author:** E-mail: guanzhu@njnu.edu.cn.
Associate editor: Adi Stern

Abstract

The rise of birds represents one of the major evolutionary transitions in the history of life. Yet, much remains obscure about the origins and diversification of viruses in birds. Endogenous retroviruses (ERVs), relics of past retroviral infections, provide molecular fossils for interrogating the evolution and ecology of retroviruses. Here, we perform phylogenomic mining of ERVs within the genomes of 758 bird species and identify more than 470,000 ERVs, revealing a highly diverse and complex retrovirus repertoire in birds. These ERVs greatly expand the diversity of retroviruses in birds, indicating that exogenous retroviruses characterized in birds to date are highly underestimated. The evolution of retroviruses in birds is shaped by both coevolution and cross-species transmission. Tens of retrovirus lineages originated during the early evolution of birds, four of which contribute to more than 90% of complete ERVs in birds. We also observe recent ERV activity across the bird phylogeny (particularly in Passeriformes). Moreover, we find that ERVs can mediate genome rearrangements, potentially facilitating the genome evolution of birds. Many bird retroviruses recruited genes of cellular provenience, which might drive the evolution of the genome complexity of retroviruses. Together, these results unveil a diverse and complex retrovirosphere in birds and provide insights into the intricate evolution of retrovirus–bird interaction.

Keywords: birds, endogenous retroviruses, virus evolution, phylogenetics

Introduction

Birds (the class Aves), with more than 10,000 extant species, evolved from theropod dinosaurs probably during the Middle–Late Jurassic period (Jarvis et al. 2014; Xu et al. 2014; Brusatte et al. 2015). Modern birds are classified into two major groups, namely Palaeognathae (the volant tinamous and flightless ratites) and Neognathae (Jarvis et al. 2014; Brusatte et al. 2015; Prum et al. 2015; Feng et al. 2020). Neognathae can be further divided into two subgroups, namely Galloanserae (the fowls, including ducks, chicken, and pheasants) and Neoaves (a diverse clade encompassing all the other living birds) (Jarvis et al. 2014; Brusatte et al. 2015; Prum et al. 2015; Feng et al. 2020). The rise of birds represents one of the major evolutionary transitions in the history of life on Earth. Over hundreds of millions of years of evolution, birds have achieved exceptional diversity in morphology, behaviors, ecological niches, and microbe associations (Xu et al. 2014; Brusatte et al. 2015). Whereas diverse viruses have been reported and isolated in birds, it remains largely obscure how these viruses originated and diversified during the rise and evolution of birds.

Retroviruses have long been known to infect birds, as exemplified by the landmark discovery of Rous sarcoma virus (RSV) more than 100 years ago (Rous 1911; Weiss and Vogt 2011). Yet, only tens of exogenous retroviruses (XRVs) have been characterized in birds, manly in the fowls. Among them, about 12 retroviruses have been formally classified by the International Committee on Taxonomy of Viruses (ICTV) (Coffin et al. 2021). Therefore, our current knowledge on the diversity of retroviruses remains highly fragmental and biased in birds.

The replication of retroviruses requires reverse transcription and integration into host genomes. When retroviruses infect germ line cells, the integrated retroviruses can be transmitted vertically, passing on to offspring and becoming endogenous retroviruses (ERVs) (Stoye 2012; Johnson 2019; Zheng et al. 2022). ERVs thus partially record retroviral infections that occurred in past and provide molecular fossils for studying the diversity and evolution of retroviruses as well as the evolution of interaction between retroviruses and their hosts (Stoye 2012; Johnson 2019; Zheng et al. 2022). ERVs have been found to be ubiquitously distributed in the genomes of vertebrates (Hayward et al. 2013; Hayward et al. 2015; Xu et al. 2018; Zheng et al. 2021; Chu et al. 2023; Wang and Han 2023), but ERVs were analyzed only in the genomes of a very limited number of birds or certain groups (such as Darwin's finches) of birds (Bolisetty et al. 2012; Cui et al. 2014; Hayward et al. 2015; Hill et al. 2022; Chen et al. 2024). The recent dense sequencing of bird genomes (Jarvis et al. 2014; Feng et al. 2020; Stiller et al. 2024) provides opportunities to interrogate the origin, diversification, and evolution of retroviruses during the rise and evolution of birds.

In this study, we leveraged phylogenomic approaches to mine ERVs in the genomes of 758 bird species, covering 97.5% of the orders of modern birds, and identified more than 470,000 ERVs. These ERVs greatly expand the diversity of retroviruses infecting or having infected birds. Our analyses

Received: March 1, 2025. **Revised:** May 21, 2025. **Accepted:** July 7, 2025

reveal that the evolution of retroviruses in birds is shaped by both coevolution and host switching. We reconstructed the possible evolutionary scenarios for many retroviral lineages. Moreover, we found that ERVs can mediate host genome rearrangements, potentially shaping the genome evolution in birds. ERVs also captured diverse cellular genes, which might drive the evolution of the genome complexity of retroviruses. Together, these results provide a snapshot of the highly diverse and complex retrovirosphere in birds and have implications in understanding the evolution of bird–retrovirus interaction.

Results

A Highly Diverse and Complex Retrovirosphere in Birds

We used a similarity search and phylogenetic analysis combined approach (Xu et al. 2018; Zheng et al. 2021) to mine ERVs in the genomes of 758 birds, including Palaeognathae (6 orders, 20 species), Galloanserae (2 orders, 51 species), and Neoaves (31 orders, 687 species) (supplementary material dataset S1, Supplementary Material online). These birds cover ~97.50% of bird orders and ~7.12% of bird species. We identified a total of 478,274 ERVs, among which 38,720 are complete or nearly complete ERVs with canonical retrovirus genes (*gag*, *pol*, and *env*) and long terminal repeats (LTRs) flanking at their 5′- and 3′-termini. Due to difference in the quality of genome sequencing, these numbers should be taken with caution. Nevertheless, our phylogenomic mining shows that ERVs are present in the genomes of all the birds analyzed in this study, which is consistent with the ubiquitous distribution of ERVs in vertebrates previously demonstrated through phylogenomic analyses (Bolisetty et al. 2012; Hayward et al. 2013; Cui et al. 2014; Hayward et al. 2015; Xu et al. 2018; Zheng et al. 2021; Hill et al. 2022; Chu et al. 2023; Wang and Han 2023; Chen et al. 2024).

To classify bird ERVs, we further identified ERVs that are phylogenetically closely related to bird ERVs or nest within the diversity of bird ERVs across the genomes of 1,371 vertebrates outside birds, including 446 mammals, 118 reptiles, 41 amphibians, 2 lobe-finned fishes, 742 ray-finned fishes, and 13 cartilaginous fishes (supplementary material dataset S2, Supplementary Material online). Phylogenetic analyses show that bird ERVs cluster into numerous groups of various size dispersed with ERVs from nonavian vertebrates (Fig. 1 and supplementary figs. S1 to S3, Supplementary Material online; see supplementary fig. S2, Supplementary Material online for several examples). Based on their phylogenetic relationship to retroviral genera and ERV classes, bird ERVs were broadly classified into seven major groups in this study, designated ERV-apha/beta-related-Aves (A/B for short), ERV-gamma-related-Aves (G), ERV-epsilon-related-Aves (E), ERV-ERV-L-related.1-Aves (L1), ERV-ERV-L-related.2-Aves (L2), ERV-ERV-L-related.3-Aves (L3), and ERV-spuma-related-Aves (S) (Fig. 1 and supplementary fig. S1, Supplementary Material online). The beta-like 1, beta-like 2, gamma, and SnRV-like clades of ERVs identified in Darwin's finches were classified into A/B, A/B, G, and L1, respectively (Hill et al. 2022). Most of the bird ERVs lack known XRV counterparts, indicating that the XRVs characterized in birds to date are highly underestimated or many XRVs have gone extinct already. Bird ERVs can be classified into 145, 1,390, and 31,646 clusters based on the reverse transcriptase (RT) amino acid identity (AAI) thresholds of 0.3, 0.5, and 0.7, respectively (Nayfach et al. 2021) (Fig. 2b). Rarefaction curves show that the number of ERV clusters based on the AAI thresholds of 0.5 and 0.7 increases monotonically with the number of bird genomes, whereas the number of ERV clusters based on the AAI threshold of 0.3 appears to be nearly saturated (Fig. 2b). The number of viral clusters varies greatly among different bird orders (supplementary fig. S4, Supplementary Material online). Together, our results reveal an unexpectedly diverse and complex retrovirosphere in birds.

We mapped the host distribution of complete ERVs across the bird phylogeny (Fig. 2a). While the S, L2, and L3 major group ERVs appear to be only sporadically present in the bird genomes, the A/B, G, E, and L1 major group ERVs were identified in a wide variety of birds (Fig. 2a). The number of complete ERVs varies among different bird orders (Fig. 2a), but the copy number should be taken with caution due to different genome assembly qualities. Interestingly, the number of A/B, L1, and G major group ERVs seems to exhibit similar trends across the bird phylogeny. To statistically verify this pattern, we selected 91 avian species with a high genome assembly quality (scaffold N50 > 10 Mb and all at the chromosome level) (supplementary table S1, Supplementary Material online) and tested the correlation in copy number among ERVs of the A/B, L1, G, and E major groups. We found significant positive correlation between the A/B and the G major groups (Pearson's $r = 0.480$, P-value $= 1.73 \times 10^{-6}$), between the L1 and the G major groups (Pearson's $r = 0.455$, P-value $= 6.60 \times 10^{-6}$), and between the L1 and the A/B major groups (Pearson's $r = 0.344$, P-value $= 8.87 \times 10^{-4}$) (Fig. 2c). However, no significant correlation was observed between the number of complete ERVs and the N50 values, indicating genome assembly quality has limited impact in the number of complete ERVs detected. Moreover, no significant correlation was found between major groups E and A/B (Pearson's $r = 0.159$, P-value $= 0.13$), L1 (Pearson's $r = 0.171$, P-value $= 0.11$), or G (Pearson's $r = 0.021$, P-value $= 0.84$), further confirming that the observed correction might be not attributed to genome assembly quality. Together, these results indicate that the A/B, G, and L1 major groups might have undergone expansion and contraction in concert during the evolutionary course of birds.

Evolutionary Dynamics of ERVs in Birds

Upon the integration of an ERV into the host genome, its 5′- and 3′-LTRs are identical. Then, the two LTRs accumulate mutations independently. Therefore, the genetic distance between 5′- and 3′-LTRs of an ERV can be used to estimate its integration time (Johnson 2019; Zheng et al. 2022). To explore the evolutionary dynamics of ERVs in birds, we estimated the distance between 5′- and 3′-LTRs flanking a complete ERV. Overall, the activity of ERVs varies greatly among different bird species, even among closely related species within the same genus (Fig. 3b). For instance, ERV activity exhibits contrasting patterns in seven species pairs with high genome quality (scaffold N50 > 10 Mb), each of which belong to the same genus (*Ammodramus caudacutus* vs. *A. savannarum pratensis*, *Melospiza georgiana* and *M. melodia*, *Amazona collaria* vs. *A. aestiva*, *Spheniscus humboldti* vs. *S. demersus*, *Falco naumanni* vs. *F. tinnunculus*, *Accipiter gentilis* vs. *A. cirrocephalus*, and *Anser cygnoides* vs. *A. indicus*). In these species pairs, the LTR–LTR distance distribution indicates that ERVs are actively proliferating in one species, but not in the other (Fig. 3b). The LTR–LTR distance

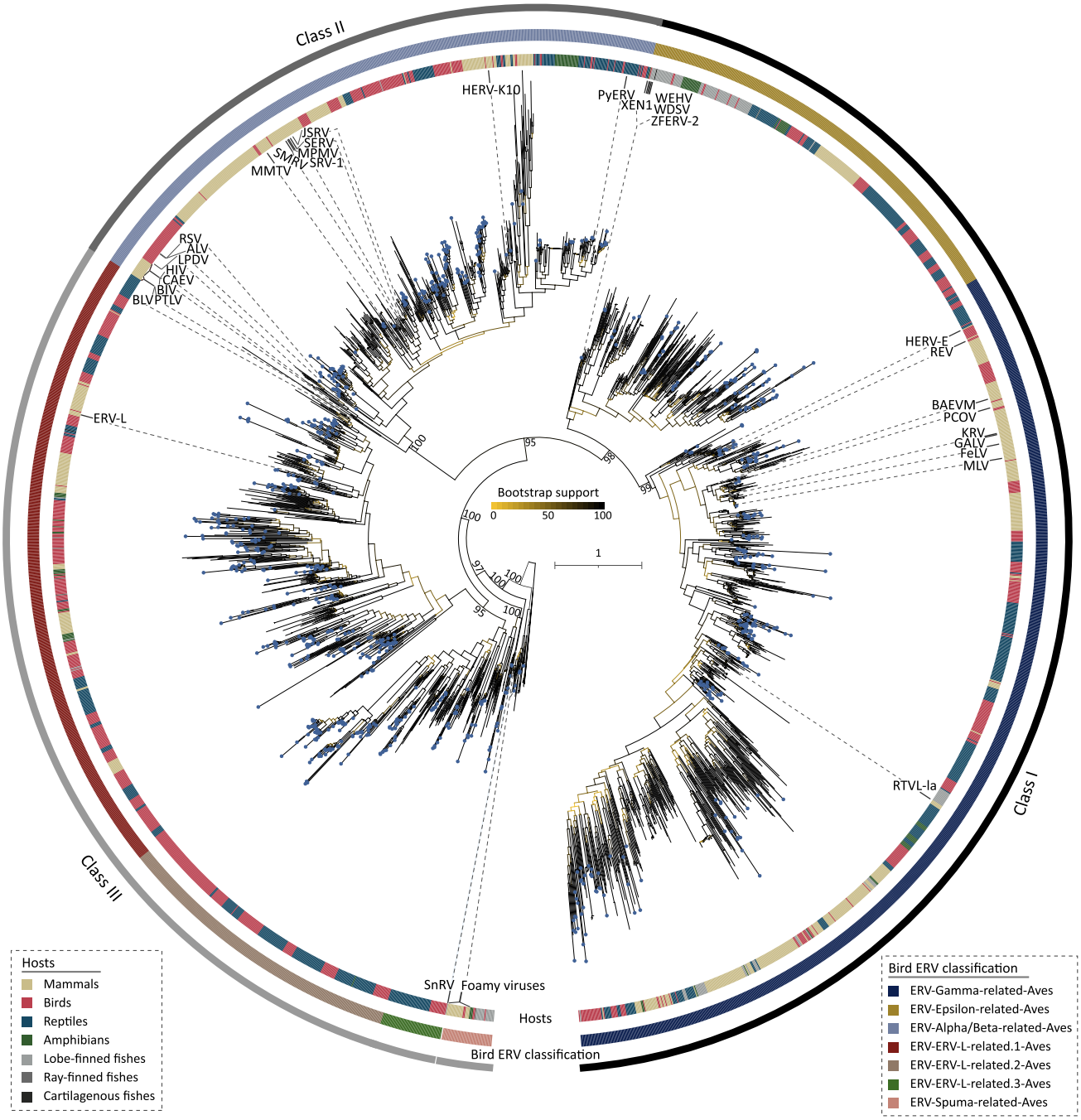

Fig. 1. The diversity of ERVs in birds. The phylogenetic relationship of representative ERVs from birds and nonavian vertebrates was reconstructed based on RT protein sequences and rooted with Odin retrotransposons and lokiretroviruses as outgroups. The inner circle indicates the host distribution of ERVs, the middle circle represents the classification of bird ERVs defined in this study, and the outer circle represents the traditional ERV classification. Representative retroviruses were specified, and their abbreviations are provided in supplementary table S4, Supplementary Material online. Bird ERVs are labeled using filled circles. Ultrafast bootstrap values are shown as a gradient and shown near the selected nodes. The full, detailed tree is provided in supplementary fig. S1, Supplementary Material online.

distribution largely follows power law, indicating that new ERVs are frequently generated, and meanwhile, older ERVs are removed from the host genome through recombination between LTRs or degradation due to mutations (Fig. 3; see supplementary fig. S5, Supplementary Material online for distribution at the species level). We also observed recent activity of ERVs across the bird phylogeny (Fig. 3a and supplementary fig. S5, Supplementary Material online). The LTR–LTR distance for ~6.58% of complete ERVs is equal to 0, and the peak of LTR–LTR distance distribution for the majority of avian orders also lies at 0 (examples are shown in Fig. 3c), suggesting that many retroviruses might be circulating or have been circulating very recently in birds. However, the estimation of integration time based on LTR–LTR distance comes with caveats: (i) Recombination between ERVs influences the calculation of LTR–LTR distance (Zheng et al. 2022). Therefore, all recombinational ERVs were excluded for LTR–LTR distance analyses in this study. (ii) The actual mutation rate of ERVs may be different from the mutation rate of hosts (Zheng et al. 2022), and mutation rates also vary

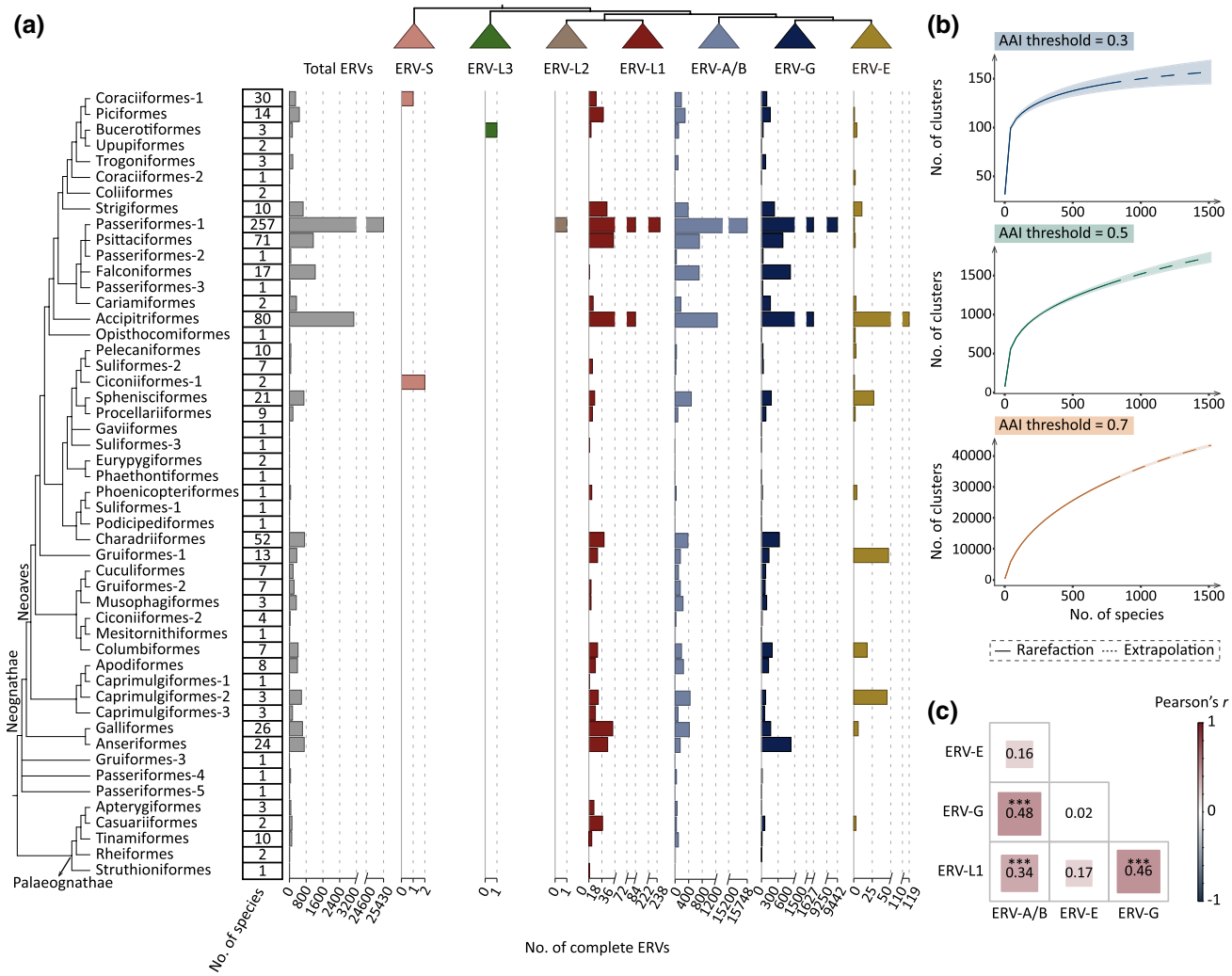

Fig. 2. The distribution of complete ERVs in birds. a) The distribution of complete ERVs in birds at the order level. The bird phylogeny is based on the studies (Avise and Walker 1998; Amaral et al. 2006; Brumfield et al. 2008; Jiang 2019; Pan et al. 2019; Feng et al. 2020; Cuevas-Caballé et al. 2022; Wilcox et al. 2022; Zhao et al. 2023). The number of species and the number of complete ERVs from different major groups are shown. b) The rarefaction curve of ERVs in birds based on RT AAI thresholds of 0.3, 0.5, and 0.7. 95% CIs are shown in shadow. c) The correlation in copy number among four ERV major groups, namely ERV-A/B, E, G, and L1. The number in the box represents the correlation coefficient (Pearson's *r*). The asterisks indicate significance levels: *** represents *P*-value < 0.001; ** represents *P*-value < 0.01; and * represents *P*-value < 0.05.

across host species (Zhang et al. 2014). Therefore, we directly compared LTR–LTR distance rather than converting it to time. Together, our results reveal a highly lineage- or species-specific activity of ERVs during the evolution of birds, which can explain the variation in ERV copy number across birds (Fig. 2a).

Bursts of ERV Activity in Passeriformes

Our initial analyses indicate that birds of the Passeriformes order seem to host an overall higher number of ERVs than other birds (Fig. 2a), which is consistent with the recent research (Cui et al. 2014). Further analyses confirm that Passeriformes birds harbor a significantly higher number of complete ERVs than non-Passeriformes birds ($P < 2.2e^{-16}$, Mann–Whitney *U* test) (supplementary fig. S6a, Supplementary Material online) and birds of other orders ($P = 3.11\times e^{-11}$ for comparing with Accipitriformes; $P = 3.27\times e^{-14}$ for Psittaciformes; $P < 2.2\times e^{-16}$ for Charadriiformes; Mann–Whitney *U* test) (Fig. 4a). Within Passeriformes, the number of complete ERVs also varies greatly among different families ($P = 9\times 10^{-3}$, Kruskal–Wallis test) (supplementary fig. S6b and c, Supplementary Material online). The families with higher ERV number are dispersed across the Passeriformes phylogeny, revealing that bursts of ERV activity are lineage-specific. We then analyzed the evolutionary dynamics of ERVs within one species from each of the top four families with the highest number of complete ERVs (Fig. 4c). Within each species, we observed several independent amplifications of ERVs that pertain to different viral lineages (Fig. 4d-g). Therefore, ERVs in these four Passeriformes species are not derived from a single ancient retrovirus invasion, but from independent integration and amplification of multiple retroviruses. The invasion of ERVs into the genomes of these four species appears to be ongoing (Fig. 4c to g).

Origins and Evolution of Retroviruses in Birds

Two evolutionary scenarios for the origins of retroviruses in birds could be conceived: (i) retroviruses were circulating in the progenitor of modern birds and become long associated with their bird hosts and (ii) retroviruses infect birds through subsequent cross-species transmission after the rise of modern birds. Our phylogenetic analyses show that bird ERVs and

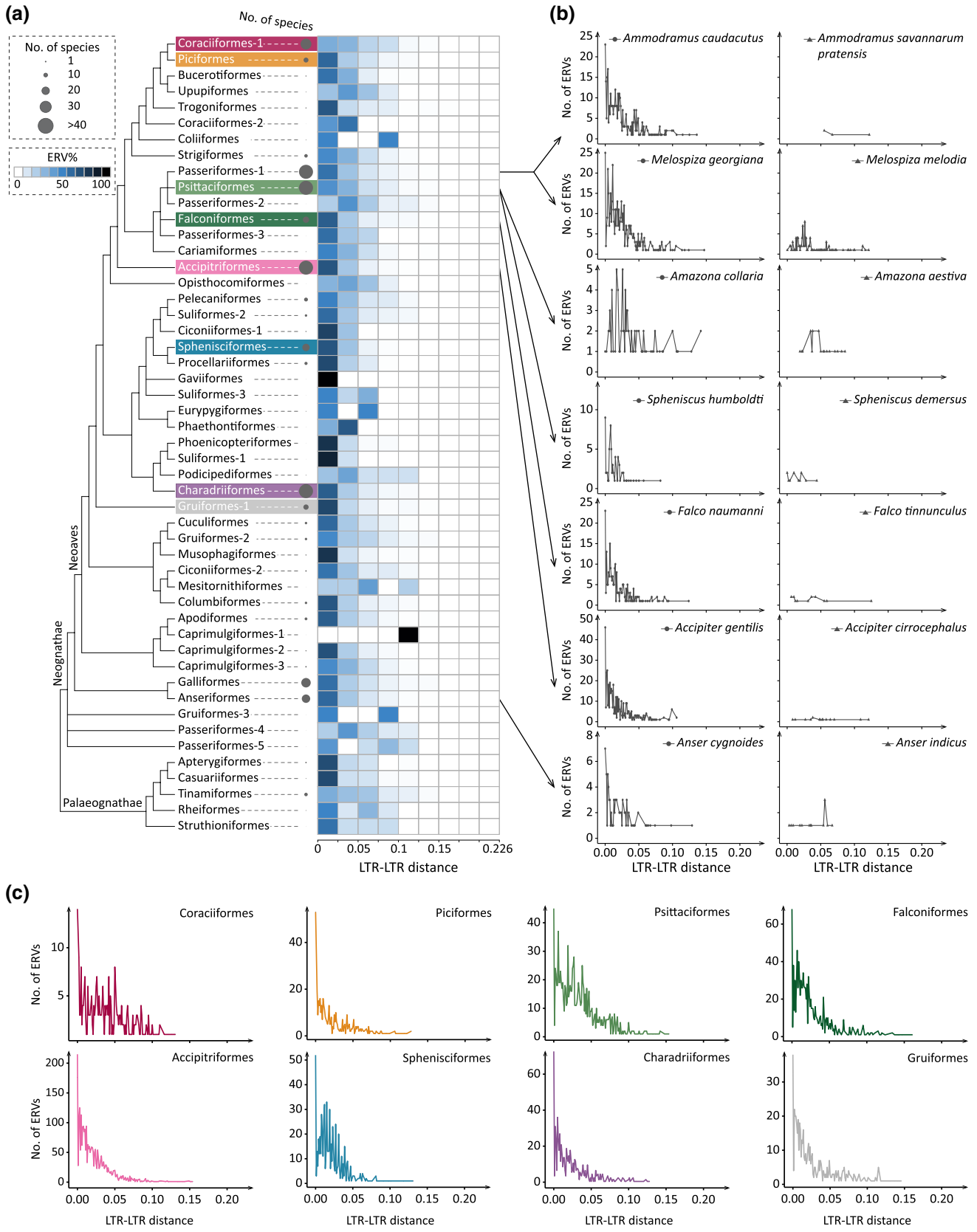

Fig. 3. The evolutionary dynamics of ERVs in birds. a) The distribution of LTR–LTR distance of complete ERVs in birds at the level of orders. The bird phylogeny is based on the studies (Avise and Walker 1998; Amaral et al. 2006; Brumfield et al. 2008; Jiang 2019; Pan et al. 2019; Feng et al. 2020; Cuevas-Caballé et al. 2022; Wilcox et al. 2022; Zhao et al. 2023). The gradient represents the proportion of ERVs (0% to 100%) in the corresponding LTR–LTR distance interval for a given bird order. b) The distribution of LTR–LTR distance of ERVs for species pairs from seven bird genera. c) The distribution of LTR–LTR distance of ERVs for eight representative bird orders. The orders are highlighted in a) in colors.

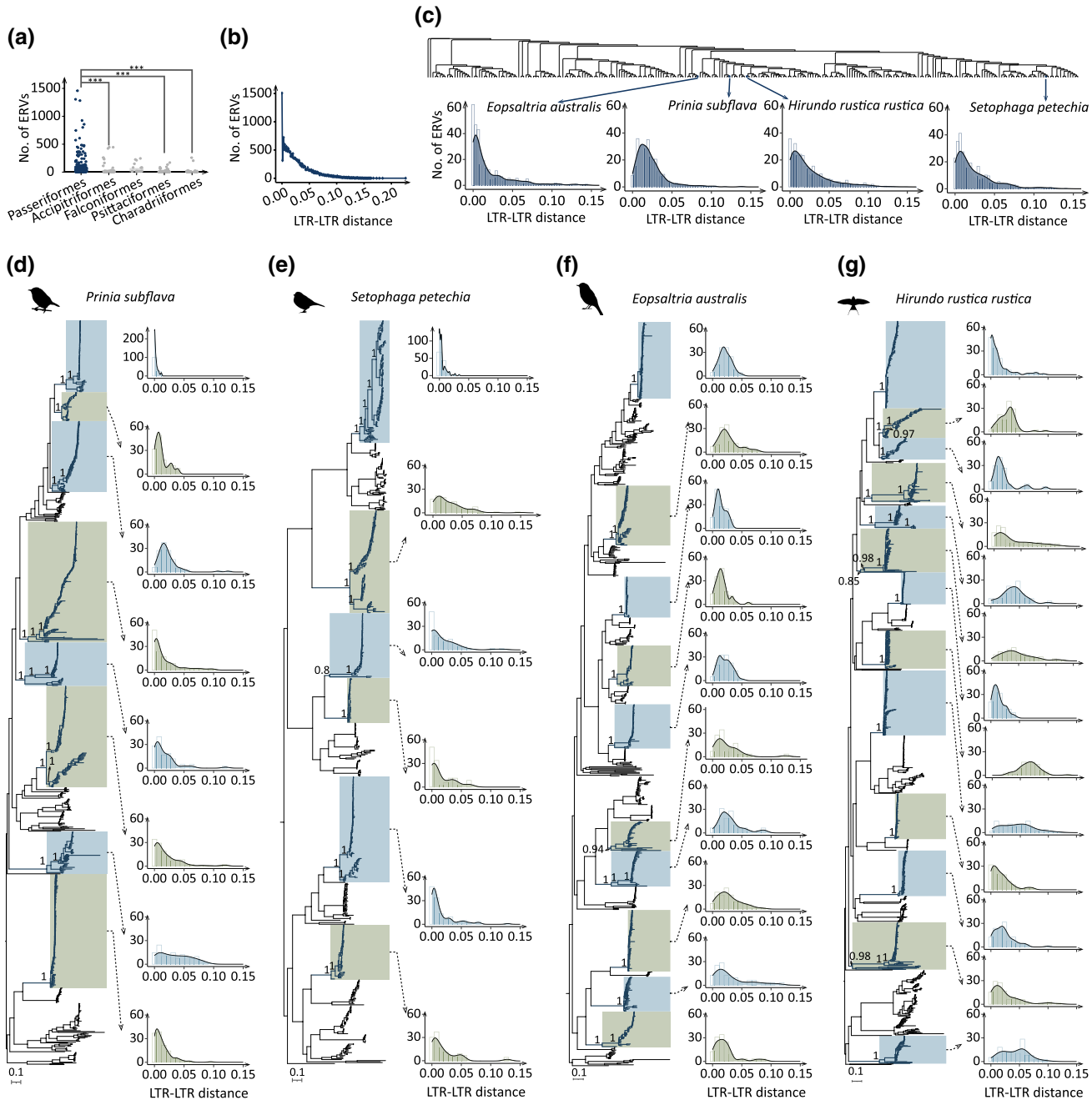

Fig. 4. The evolutionary dynamics of ERVs in Passeriformes. a) The comparison of the number of complete ERVs from Passeriformes, Accipitriformes, Falconiformes, Psittaciformes, and Charadriiformes. Dots indicate the numbers of ERVs in a species. The asterisks *** represents *P*-value < 0.001. b) The distribution of LTR–LTR distance of ERVs in Passeriformes. c) The distribution of LTR–LTR distance of ERVs from four species, which is indicated in the phylogeny of Passeriformes. The curves fitting of Gaussian mixture models are also shown. The phylogenetic trees of ERVs and the distribution of LTR–LTR distance of specific ERV clades are shown for *Prinia subflava* (d), *Setophaga petechia* (e), *Eopsaltria australis* (f), and *Hirundo rustica rustica* (g). The phylogenetic tree was reconstructed based on the full-length sequences of ERVs. The ultrafast bootstrap values are shown near the selected nodes. The LTR–LTR distance distribution of specific ERV clades along with the curve fitting of Gaussian mixture models are shown.

retroviruses from nonavian vertebrates exhibit complex mixing (Figs. 1 and 5a), suggesting that retroviruses infecting birds might have numerous independent origins. Based on phylogenetic analysis, we further classified the majority of avian ERVs (~99.08% of ERVs and ~95% of complete ERVs) into 61 groups (Fig. 5a). We identified orthologous ERV insertions across the bird phylogeny, which reveals in the last common ancestor of which species ERVs integrated, providing the minimum estimate for the age of a certain ERV group (supplementary table S3, Supplementary Material online). For seven groups (groups 5, 7, 14, 25, 28, 41, and 48), orthologous ERV insertions were identified in species from both Neognathae and Palaeognathae. For three groups (groups 55, 56, and 57), statistically significant congruence was observed between ERV and host phylogenies across the whole bird phylogeny at the level of order (supplementary table S2 and supplementary fig. S7, Supplementary Material online). With the exception of groups 5, 55, and 57, all the other groups were identified in over 80% of avian orders (Fig. 5b). These results indicate that at least 10 ERV groups might

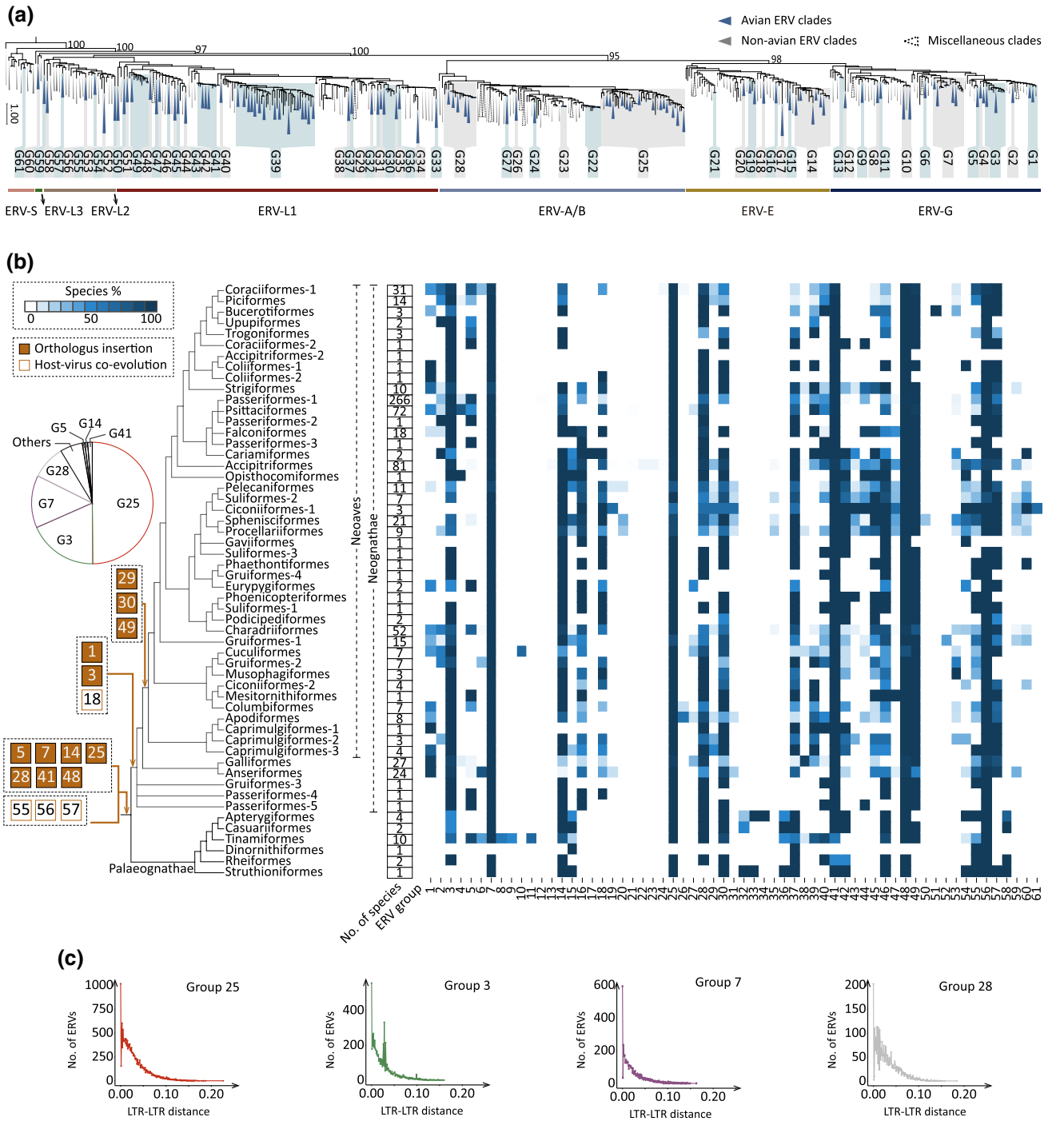

Fig. 5. Origins of retroviruses in birds. a) The phylogenetic tree of ERVs from birds and nonavian vertebrates, which is the same as Fig. 1. Groups and classification defined in this study are indicated. G is the abbreviation for Group. The full, detailed tree is provided in supplementary fig. S1, Supplementary Material online. b) Host distribution of different ERVs groups. The minimum time (indicated as the time to the last common ancestors of corresponding bird groups) was shown for specific ERV groups along the bird phylogeny. The minimum time for ERV groups in filled and empty squares was inferred based on orthologous ERV insertions and on virus–host coevolution, respectively. Phylogenetic congruence tests and orthologous ERV insertions are provided in supplementary tables S2 and S3, Supplementary Material online. The pie chart indicates the proportion of complete ERVs from a specific group in all the complete ERVs from birds. The heatmap on the right shows the host range of different ERV groups at the level of bird orders. The gradient indicates the proportion of species with a specific ERV group in a specific bird order. c) The distribution of LTR–LTR distance in certain ERV groups. The colors are the same as those in the pie chart of b).

have been circulating at least before the last common ancestor of modern birds and have long coevolved with birds. In addition, we identified orthologous insertions in species across Neognathae for groups 1 and 3 and across Neoaves for groups 29, 30, and 49, indicating that these groups originated deep in the bird evolution, namely before the last common ancestor of Neognathae and Neoaves, respectively (Fig. 5b and supplementary table S3, Supplementary Material online). ERVs of all these ancient groups account for ~94.22% complete ERVs in birds, and notably ~91.36% of complete ERVs belong to groups 3, 7, 25, and 28. Interestingly, we found that many of these ancient ERV groups (for example,

groups 3, 7, 25, and 28) are still or have been recently amplifying within the bird genomes (Fig. 5c). Together, our results suggest that the majority of ERVs in birds are derived from retroviruses that infected the progenitor of modern birds.

Retroviruses can infect birds through cross-species transmission after the rise of modern birds. The host distribution shows that 27 groups are only sporadically present in certain bird lineages, covering <15% of the bird orders (Fig. 5b), and some nest within the diversity of retroviruses from nonavian vertebrates (mostly reptiles or mammals). For instance, group 6 exhibits a closer phylogenetic relationship to reptilian ERVs, while group 27 is more closely related to rodent ERVs, and group 8 shows a closer affinity to ERVs from various mammalian species (supplementary fig. S1, Supplementary Material online). These findings suggest that the groups likely originated through cross-species transmission of retroviruses from other nonavian vertebrates. However, it is challenging to reconstruct the exact evolutionary history for many bird retrovirus lineages due to incomplete sampling of species outside birds. Nevertheless, our analyses indicate that both coevolution and cross-species transmission shape the diversity and evolution of retroviruses in birds.

Birds Act as a Source for Retrovirus Transmission

Interestingly, we also found some ERVs from nonavian vertebrates fall within the diversity of bird ERVs. We deciphered the evolutionary history for three ERV groups (groups 10, 14, and 25) (supplementary fig. S8, Supplementary Material online). For all the three groups, ERVs from nonavian source (carnivores, turtles, and monotremes) fall within the diversity of bird ERVs, indicating that these nonavian retroviruses might have arisen through cross-species transmission from birds. A single cross-species transmission event from birds to carnivores can be inferred for group 10. For group 14, at least two cross-species transmission events from birds to turtles might have taken place. The evolutionary history for group 25 is complex, involving multiple cross-species transmission from birds to turtles, carnivores, and monotremes. However, the actual scenarios may be more complex, because the genomes of relevant host species might not be sequenced, or relevant retroviruses might not form endogenous counterparts. Nevertheless, our study suggests that birds may act as source for retrovirus transmission to nonavian vertebrates.

ERVs Mediated Genome Rearrangements in Birds

ERVs have the potential to mediate genome rearrangements through recombination between ERVs at distant loci (Hughes and Coffin 2001). Because it is technically challenging to identify recombination through phylogenetic analysis of LTRs from such a large number of ERVs, we first clustered LTRs based on their sequence identity. We identified at least 3,044 ERVs whose LTRs belong to different sequence clusters, which might be generated through recombination between ERVs at different loci. It should be noted that this method is conservative because recombination between ERVs with high sequence identity could be overlooked. We then validated recombination events through RDP4 (Martin et al. 2020) and discordant target site duplication (TSD). Together, we identified recombination signals in 998 ERVs (supplementary table S5, Supplementary Material online). These ERVs were distributed in 118 bird species that belong to 18 out of 39 orders. Interestingly, we identified recombination events that occurred during the evolutionary course of *Accipiter* (before the last common ancestor of *A. gentilis* and *A. melanochlamys*) and *Buteo* (before the last common ancestor of *B. lagopus*, *B. japonicus*, *B. lineatus*, and *B. regalis*). The resulted genome rearrangements have been maintained for millions of years (supplementary fig. S9, Supplementary Material online) (Kumar et al. 2022). All these results suggest that recombination occurred between ERVs, which can mediate genome rearrangements in birds, potentially facilitating the evolution of bird genome complexity.

ERVs Captured Diverse Cellular Genes

Retroviruses typically encode three proteins, namely Gag, Pol, and Env, and some retroviruses also possess dUTPase (Hizi and Herzig 2015). Interestingly, we identified 107 host-related domains in 254 ERVs from 121 bird species, including cold-shock domain (CSD) (accession no. pfam00313), major facilitator superfamily (accession no. cl28910), and chromosome segregation ATPase (accession no. cl34174). The host-related domains were integrated at different locations within the retrovirus genomes (Fig. 6 and supplementary table S6, Supplementary Material online). Based on phylogenetic analysis of RT proteins, these ERVs belong to ERV-A/B, ERV-G, and ERV-L major groups. Interestingly, some ERVs that carry a common host-related domain belong to different major groups. For example, 46 ERVs encode CSD, but these ERVs belong to ERV-A/B or ERV-G major groups, and their positions also vary among different ERVs (Fig. 6a and c and supplementary table S6, Supplementary Material online). Phylogenetic analysis of CSD sequences shows that CSDs were captured by retroviruses multiple times (supplementary fig. S10, Supplementary Material online). Similarly, seven transmembrane G protein-coupled receptors superfamily (7mt_GPCRs) (accession no. cl28897) encoded by ERVs from diverse birds clustered into multiple lineages, suggesting that it was repeatedly acquired during the evolution of retroviruses in birds (supplementary fig. S10, Supplementary Material online). Together, our results suggest that bird retroviruses occasionally capture cellular genes.

Discussion

Meta-transcriptomic studies have facilitated virus discovery, reframing our understanding of the diversity and evolution of viruses (Shi et al. 2016, 2018; Harvey and Holmes 2022). However, retroviruses have been routinely excluded in meta-transcriptomic studies, mainly because it is technically challenging to distinguish XRVs and expressed ERVs, and the significance of ERVs has been unappreciated. On the contrary, ERVs document past retroviral infections, providing molecular fossils for interrogating the diversity and evolution of retroviruses and their relationship to their hosts (Stoye 2012; Johnson 2019; Zheng et al. 2022). In this study, we performed phylogenomic mining of ERVs in the genomes of 758 birds that cover 97.50% of bird orders and 7.12% of bird species. Birds were previously thought to harbor limited diversity of retroviruses, with only 12 retroviruses in birds officially recognized by the ICTV that belong to *Alpharetrovirus* or *Gammaretrovirus* (Coffin et al. 2021). All the nine viral species within *Alpharetrovirus* exclusively infect birds, including avian leucosis virus and RSV, which can cause immunosuppression, growth retardation, and tumorigenesis, leading to significant economic losses (Fandiño et al. 2023). Recent studies with a limited number of birds or

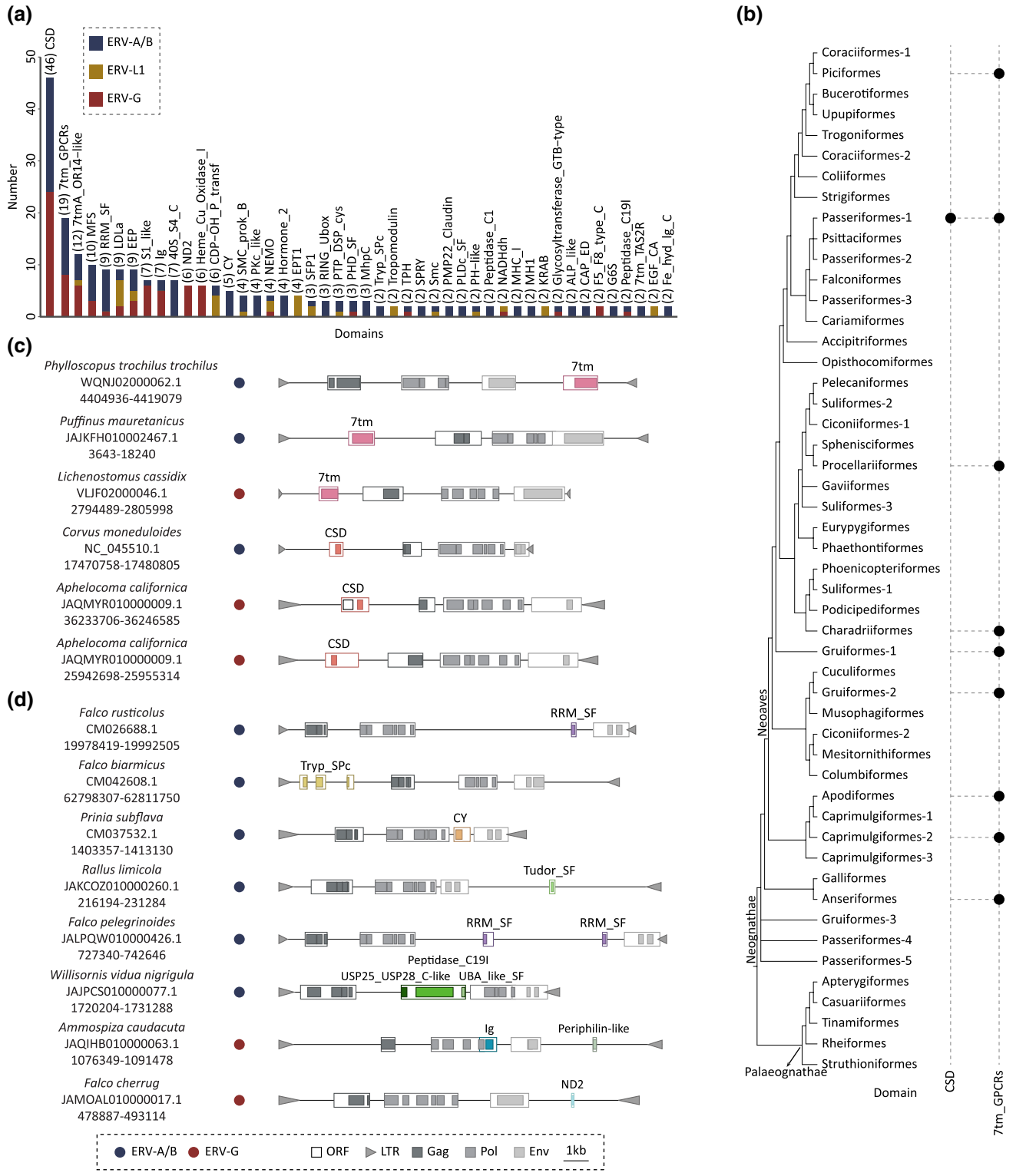

Fig. 6. Host-related domains in bird ERVs. a) The type and number of host-related domains identified in avian ERVs, with only those having a count >1 displayed. b) The host distribution of CSD (including ribosomal protein S1-like RNA-binding domain (S1_like) [accession no. cl09927]) and 7mt_GPCRs (including 7tmA_OR14-like [accession no. cd15227]). c) Genomic structure and open reading frames (ORFs) of representative avian ERVs with CSD or 7mt_GPCRs domains. d) Genomic structure and ORFs of representative avian ERVs with host-related domains.

certain groups of birds (such as Darwin's finches) suggest that birds might host diverse ERVs related to *Alpharetrovirus*, *Betaretrovirus*, *Gammaretrovirus*, *Epsilonretrovirus*, ERV-L, and *Spumaretrovirinae* (Bolisetty et al. 2012; Cui et al. 2014; Hill et al. 2022; Zheng et al. 2022). Our analyses further reveal that bird ERVs cluster into numerous distinct groups, greatly expanding the diversity of retroviruses associated with birds. Many ERVs lack XRV counterparts, suggesting that the diversity of XRVs might be highly underestimated in birds, which is consistent with the finding that at least 76 retroviruses are invading the bird genomes (Wang and Han 2023). The possibility that some XRVs have gone extinct already cannot be excluded.

One caveat is that some, if any, ERVs identified in this study might be XRVs. However, the overall likelihood that an XRV integrated into the genome of an individual and the individual was sequenced is very low. Nevertheless, our pan-Aves phylogenomic mining reveals a diverse and complex retrovirosphere in birds.

The activity of ERVs varies greatly among bird species, leading to extensive variation in ERV copy number among bird species. We also observed recent bursts of ERV activity in Passeriformes, which were resulted from independent integration and amplification of diverse retroviruses in Passeriformes. The invasion of ERVs appears to be ongoing in many species within the Passeriformes order and perhaps across birds (Figs. 2 and 3) (Wang and Han 2023). The amplification of multiple ERV lineages within small time intervals in the Passeriformes order and the amplification of ERVs of different clades in concert across birds might be attributed to host biology rather than retroviruses themselves. Nevertheless, our results reveal the highly dynamic and lineage-specific nature of ERV activity in birds, as observed in mammals (Tarlinton et al. 2006; Anai et al. 2012; Elleder et al. 2012; Tsangaras et al. 2015; Blyton et al. 2022).

Bird ERVs and retroviruses from nonavian vertebrates exhibit extremely complex phylogenetic mixing. We conclude that both coevolution and cross-species transmission shaped the origins and evolution of retroviruses in birds. Our analyses indicate at least 10 ERV groups originated before the last common ancestor of birds. It should be noted that establishing orthologous ERV insertions might be conservative, because ERVs evolved rapidly, and thus, the number of ancient ERV groups was underestimated. Nevertheless, the activity of these ancient ERV groups contributed to the majority of ERVs in the bird genomes. Moreover, we found a small portion of ERVs can undergo cross-species transmission between birds and nonavian vertebrates, which might be relevant to zoonoses in the wild. However, ERVs identified in this study represent only a subset of XRV that circulated or are circulating in by birds, and actual transmission is likely more complex.

Mutations frequently occur in ERVs, leading to their degradation. However, recent studies report numerous cases of cooption of ERV genes or regulatory sequences implicated in diverse cellular processes (Stoye 2012; Chuong et al. 2016; Wang and Han 2020; Frank et al. 2022; Zheng et al. 2022). In this study, we found evidence for recombination between ERVs at different loci, which can mediate genome rearrangements (Hughes and Coffin 2001). In some cases, the genome arrangement mediated by ERV recombination has been maintained for a long time, indicating putative beneficial effects. It should be noted that the method we used to identify recombination between ERVs is rather conservative. Thus, genome rearrangements medicated by recombination between ERVs might be much more prevalent during the evolution of birds.

It is long been appreciated that many retroviruses can acquire oncogenes that stimulate cell growth or division to induce cancers from their hosts, for example, *src* gene in RSV (Stehelin et al. 1976). On the one hand, the host-related domains or proteins identified in this study provide a list of putative oncogenes. Several domains are associated with known protooncogenes, such as ETS_PEA3_*N*, and some are relevant to cell cycle regulation, such as smc and RING-H2_TRAIP. On the other hand, several domains were acquired convergently, such as CSDs that function in posttranscriptional gene regulation (Mihailovich et al. 2010). Convergent evolution has long been taken as evidence of adaptation (Losos 2011). It follows that convergent acquisition of these host-related domains might contribute to the proliferation and infection of retroviruses, regardless of the host fitness. Together, the capture of cellular genes likely further complicates the interaction of retroviruses and their bird hosts.

Materials and Methods

ERV Identification

The bird genomes used in this study were retrieved from NCBI genome resources (https://www.ncbi.nlm.nih.gov/genome/). A total of 758 bird species were analyzed, including Palaeognathae (6 orders, 20 species), Galloanserae (2 order, 51 species), and Neoaves (31 orders, 687 species) (supplementary material dataset S1, Supplementary Material online). We used a similarity search and phylogenetic analysis combined approach (Xu et al. 2018; Zheng et al. 2021) to mine ERVs. Briefly, we first used the tBLASTn algorithm to search RT homologs in the bird genomes with RT sequences of representative retroviruses as queries and an *e*-value threshold of 10^{-5}. Significant hits with a length of >150 amino acids were retrieved. Because RT sequences of retroviruses and retrotransposons share detectable sequence similarity, we performed phylogenetic analyses of significant hits and RT sequences of representative retroviruses and retrotransposons (Xu et al. 2018; Zheng et al. 2021; Wang and Han 2023). Significant hits that cluster within the diversity of retroviruses were retroviruses and retrieved. RT sequences were aligned using MAFFT 7.505 (Katoh and Standley 2013) and refined using trimAl v1.4 (Capella-Gutiérrez et al. 2009). Large-scale phylogenetic analyses were performed using the approximate maximum likelihood method implemented in FastTree 2.1.10 (Price et al. 2010). Each identified ERV was defined as a separate ERV.

To mine complete or nearly complete ERVs, LTRs were identified using LTRharvest (Ellinghaus et al. 2008). Bird genomes with a scaffold N50 value of <10 kb and a genome size of <0.9 Gb were excluded due to their low sequencing quality (Feng et al. 2020). A total of 732 birds from 38 orders were used to mine complete ERVs. ERVs were identified through phylogenetic analysis of RT proteins as aforementioned. ERV genomes were annotated using CD-search (Wang et al. 2023) and the BLASTp algorithm. We defined complete ERVs as ERVs with canonical retrovirus genes (*gag*, *pol*, and *env*) and LTRs flanking at their 5'- and 3'-termini.

ERV Classification

To classify bird ERVs and explore the origins of bird ERVs, we also mined ERVs in the genome of 1,371 vertebrates outside birds, including 446 mammals, 118 reptiles, 41 amphibians, 2 lobe-finned fishes, 742 ray-finned fishes, and 13 cartilaginous fishes (supplementary material dataset S2, Supplementary Material online). We defined nonavian vertebrate ERVs that cluster with avian ERVs as the "closely related" ERVs, and these ERVs were retrieved for further analyses. Then, we performed phylogenetic analysis of representative RT sequences of bird ERV, ERVs from nonavian vertebrates, and XRVs. RT sequences were aligned using MAFFT (Katoh and Standley 2013) with the L-INS-i strategy. Alignments were refined using trimAl (Capella-Gutiérrez et al. 2009). Phylogenetic analyses were performed using the maximum likelihood

method implemented in IQ-tree 2 (Minh et al. 2020). ModelFinder was used to select the best amino acid model (Kalyaanamoorthy et al. 2017). The ultrafast bootstrap approximation (UFBoot) with 1,000 replicates was used to assess the support for nodes (Minh et al. 2013; Hoang et al. 2018). We identified ERV groups with UFBoot values of >70% and the number of host species of >2.

Correlation Analysis of ERV Groups

To explore the correlation among ERVs from different groups, we analyzed the number of complete ERVs for 91 bird species with genome assembly at the chromosome level and a scaffold N50 value of >10 Mb (supplementary table S1, Supplementary Material online). The correlation among the copy number of ERVs of different major groups was analyzed using the phylogenetically independent contrast method implemented in the ape package (Paradis and Schliep 2019).

ERV Clustering

MMseqs2 (Steinegger and Söding 2017) was used to cluster the RT sequences of ERVs into clusters based on their AAI threshold of 0.3, 0.5 and 0.7, with the *e*-value threshold of 10^{-5} and the coverage model of 5, which ensures that the covered region of the shorter sequence is entirely covered by the longer sequence.

Evolutionary Dynamics of ERVs

We estimated the genetic distance between 5'- and 3'-LTRs flanking a complete ERV, which reflects its integration time (Johnson 2019; Zheng et al. 2022). Phylogenetic analyses of LTRs from complete ERVs were performed, and only 5'- and 3'-LTRs that cluster together were used for analyses. LTRs were aligned using MAFFT (Katoh and Standley 2013). Genetic distance of 5'- and 3'-LTRs flanking a complete ERV was calculated using Kimura two-parameter substitution model (Kimura 1980).

Congruence Test Between ERV and Bird Phylogenies

Jane 4 (Conow et al. 2010) was used to compare the congruence between avian ERVs and phylogeny at order level. The cost values for cospeciation, duplication, duplication with host switch, loss, and failure to diverge were set at 0, 1, 2, 1, and 1, respectively (Zheng et al. 2021). *P*-values were estimated using random parasite tree and random tip mapping methods with a sample size of 50.

Identification of Orthologous ERV Insertions

To determine the minimum insertion time of bird ERVs, we retrieved the 1,000-bp sequences flanking complete ERVs within each group and performed similarity search against the whole genomes of one species from different orders using BLASTn with the flanking sequences and LTRs as queries, an *e*-value threshold of 10^{-5}, and a coverage threshold of 50%. Within 10,000-bp regions spanning both upstream and downstream of each significant hit of flanking sequences, we used tBLASTn to identify RT proteins. We then performed phylogenetic analysis of the identified RT proteins and representative bird ERVs. If RT proteins that share homologous flanking sequences cluster together or LTRs and their flanking regions are homologous within the same group, they are orthologous ERV insertions. We used MAFFT 7.505 (Katoh and Standley 2013) to align the sequences and FastTree 2.1.10 (Price et al. 2010) to perform phylogenetic analyses.

Recombination Analysis of ERVs

LTRs of complete ERVs were first subject to all-to-all BLASTn with an *e*-value threshold of 10^{-5}. Clustering analysis was then performed using the Markov clustering algorithm (Dongen 2008) based on the *e*-values of BLASTn with inflation of 2. Clusters with only one sequence were excluded for further analyses. Putative recombination was detected when two LTRs of an ERV do not cluster together. Recombinant ERVs were further confirmed using RDP4 (Martin et al. 2020) and discordant TSD. To reconstruct phylogenetic relationship of LTRs, target LTRs were searched against all the bird genomes using BLASTn. LTRs were aligned using MAFFT with the L-INS-i strategy (Katoh and Standley 2013). Phylogenetic analysis was performed using IQ-tree 2 as aforementioned (Minh et al. 2020).

Identification of Host-Related Domains in ERVs

All the complete bird ERVs were annotated using CD-search (Wang et al. 2023) and HMMER 3.3 (Potter et al. 2018). We also investigated the source of certain cellular domains. The domain sequences were retrieved from the Pfam (Paysan-Lafosse et al. 2023) and nonredundant protein sequence (nr) databases. Domain sequences were aligned using MAFFT (Katoh and Standley 2013). Phylogenetic analyses were performed using the approximate maximum likelihood method implemented in FastTree (Price et al. 2010).

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (32270684 and 32470652 to G.-Z.H. and 32300511 to Z.G.).

Data Availability

All the genomes used in this study were retrieved from the NCBI, and detailed information is given in Datasets S1 and S2. The sequences and alignments of bird ERVs generated in this study were deposited to Figshare, available at https://figshare.com/s/58373b8e0fcdc71eb1ec.

References

Amaral FS, Miller MJ, Silveira LF, Bermingham E, Wajntal A. Polyphyly of the hawk genera *Leucopternis* and *Buteogallus* (Aves, Accipitridae): multiple habitat shifts during the Neotropical buteonine diversification. *BMC Evol Biol*. 2006:6(1):10. https://doi.org/10.1186/1471-2148-6-10

Anai Y, Ochi H, Watanabe S, Nakagawa S, Kawamura M, Gojobori T, Nishigaki K. Infectious endogenous retroviruses in cats and emergence of recombinant viruses. *J Virol*. 2012:86(16):8634–8644. https://doi.org/10.1128/JVI.00280-12

Avise JC, Walker D. Pleistocene phylogeographic effects on avian populations and the speciation process. *Proc Biol Sci*. 1998:265(1395):457–463. https://doi.org/10.1098/rspb.1998.0317

Blyton MDJ, Young PR, Moore BD, Chappell KJ. Geographic patterns of koala retrovirus genetic diversity, endogenization, and subtype distributions. *Proc Natl Acad Sci U S A*. 2022:119(33): e2122680119. https://doi.org/10.1073/pnas.2122680119

Boliseti M, Blomberg J, Benachenhou F, Sperber G, Beemon K. Unexpected diversity and expression of avian endogenous retroviruses. *mBio*. 2012:3(5):e00344-12. https://doi.org/10.1128/mBio.00344-12

Brumfield RT, Liu L, Lum DE, Edwards SV. Comparison of species tree methods for reconstructing the phylogeny of bearded manakins (Aves: Pipridae, *Manacus*) from multilocus sequence data. *Syst Biol*. 2008:57(5):719–731. https://doi.org/10.1080/10635150802422290

Brusatte SL, O'Connor JK, Jarvis ED. The origin and diversification of birds. *Curr Biol*. 2015:25(19):R888–R898. https://doi.org/10.1016/j.cub.2015.08.003

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. Trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009:25(15):1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Chen G, Yu D, Yang Y, Li X, Wang X, Sun D, Lu Y, Ke R, Zhang G, Cui J, *et al*. Adaptive expansion of ERVK solo-LTRs is associated with Passeriformes speciation events. *Nat Commun*. 2024:15(1):3151. https://doi.org/10.1038/s41467-024-47501-3

Chu L, Su F, Han GZ, Wang J. Jawless vertebrates do not escape retrovirus infection. *Virology*. 2023:583:52–55. https://doi.org/10.1016/j.virol.2023.04.010

Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science*. 2016:351(6277):1083–1087. https://doi.org/10.1126/science.aad5497

Coffin J, Blomberg J, Fan H, Gifford R, Hatziioannou T, Lindemann D, Mayer J, Stoye J, Tristem M, Johnson W, *et al*. ICTV virus taxonomy profile: retroviridae 2021. *J Gen Virol*. 2021:102(12):001712. https://doi.org/10.1099/jgv.0.001712

Conow C, Fielder D, Ovadia Y, Libeskind-Hadas R. Jane: a new tool for the cophylogeny reconstruction problem. *Algorithms Mol Biol*. 2010:5(1):16. https://doi.org/10.1186/1748-7188-5-16

Cuevas-Caballé C, Ferrer Obiol J, Vizueta J, Genovart M, Gonzalez-Solís J, Riutort M, Rozas J. The first genome of the balearic shearwater (*Puffinus mauretanicus*) provides a valuable resource for conservation genomics and sheds light on adaptation to a pelagic lifestyle. *Genome Biol Evol*. 2022:14(5):evac067. https://doi.org/10.1093/gbe/evac067

Cui J, Zhao W, Huang Z, Jarvis ED, Gilbert MT, Walker PJ, Holmes EC, Zhang G. Low frequency of paleoviral infiltration across the avian phylogeny. *Genome Biol*. 2014:15(12):539. https://doi.org/10.1186/s13059-014-0539-3

Dongen SV. Graph clustering via a discrete uncoupling process. *SIAM J. Matrix Anal*. 2008:30(1):121–141. https://doi.org/10.1137/040608635

Elleder D, Kim O, Padhi A, Bankert JG, Simeonov I, Schuster SC, Wittekindt NE, Motameny S, Poss M. Polymorphic integrations of an endogenous gammaretrovirus in the mule deer genome. *J Virol*. 2012:86(5):2787–2796. https://doi.org/10.1128/JVI.06859-11

Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*. 2008:9(1):18. https://doi.org/10.1186/1471-2105-9-18

Fandiño S, Gomez-Lucia E, Benítez L, Doménech A. Comparison of endogenous alpharetroviruses (ALV-like) across galliform Species: new distant proviruses. *Microorganisms*. 2023:12(1):86. https://doi.org/10.3390/microorganisms12010086

Feng S, Stiller J, Deng Y, Armstrong J, Fang Q, Reeve AH, Xie D, Chen G, Guo C, Faircloth BC, *et al*. Dense sampling of bird diversity increases power of comparative genomics. *Nature*. 2020:587(7833): 252–257. https://doi.org/10.1038/s41586-020-2873-9

Frank JA, Singh M, Cullen HB, Kirou RA, Benkaddour-Boumzaouad M, Cortes JL, Garcia Pérez J, Coyne CB, Feschotte C. Evolution and antiviral activity of a human protein of retroviral origin. *Science*. 2022:3789(6618):422–428. https://doi.org/10.1126/science.abq7871

Harvey E, Holmes EC. Diversity and evolution of the animal virome. *Nat Rev Microbiol*. 2022:20(6):321–334. https://doi.org/10.1038/s41579-021-00665-x

Hayward A, Cornwallis CK, Jern P. Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. *Proc Natl Acad Sci U S A*. 2015:112(2):464–469. https://doi.org/10.1073/pnas.1414980112

Hayward A, Grabherr M, Jern P. Broad-scale phylogenomics provides insights into retrovirus-host evolution. *Proc Natl Acad Sci U S A*. 2013:110(50):20146–20151. https://doi.org/10.1073/pnas.1315419110

Hill J, Lillie M, Pettersson ME, Rubin CJ, Grant BR, Grant PR, Andersson L, Jern P. Spatiotemporal variations in retrovirus-host interactions among Darwin's finches. *Nat Commun*. 2022:13(1): 6033. https://doi.org/10.1038/s41467-022-33723-w

Hizi A, Herzig E. dUTPase: the frequently overlooked enzyme encoded by many retroviruses. *Retrovirology*. 2015:12(1):70. https://doi.org/10.1186/s12977-015-0198-9

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol*. 2018:35(2):518–522. https://doi.org/10.1093/molbev/msx281

Hughes JF, Coffin JM. Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. *Nat Genet*. 2001:29(4):487–489. https://doi.org/10.1038/ng775

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B, Howard JT, *et al*. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*. 2014:346(6215): 1320–1331. https://doi.org/10.1126/science.1253451

Jiang JQ. Characterization of the complete mitochondrial genome of *Corvus corone orientalis*. *Mitochondrial DNA B Resour*. 2019:4(2): 2102–2103. https://doi.org/10.1080/23802359.2019.1622467

Johnson WE. Origins and evolutionary consequences of ancient endogenous retroviruses. *Nat Rev Microbiol*. 2019:17(6):355–370. https://doi.org/10.1038/s41579-019-0189-2

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 2017:14(6):587–589. https://doi.org/10.1038/nmeth.4285

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013:30(4):772–780. https://doi.org/10.1093/molbev/mst010

Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*. 1980:16(2):111–120. https://doi.org/10.1007/BF01731581

Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, Stecher G, Hedges SB. TimeTree 5: an expanded resource for Species divergence times. *Mol Biol Evol*. 2022:39(8):msac174. https://doi.org/10.1093/molbev/msac174

Losos JB. Convergence, adaptation, and constraint. *Evolution*. 2011:65(7):1827–1840. https://doi.org/10.1111/j.1558-5646.2011.01289.x

Martin DP, Varsani A, Roumagnac P, Botha G, Maslamoney S, Schwab T, Kelz Z, Kumar V, Murrell B. RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evol*. 2020:7(1): veaa087. https://doi.org/10.1093/ve/veaa087

Mihailovich M, Militti C, Gabaldón T, Gebauer F. Eukaryotic cold shock domain proteins: highly versatile regulators of gene expression. *Bioessays*. 2010:32(2):109–118. https://doi.org/10.1002/bies.200900122

Minh BQ, Nguyen MA, von Haeseler A. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol*. 2013:30(5):1188–1195. https://doi.org/10.1093/molbev/mst024

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):1530–1534. https://doi.org/10.1093/molbev/msaa015

Nayfach S, Páez-Espino D, Call L, Low SJ, Sberro H, Ivanova NN, Proal AD, Fischbach MA, Bhatt AS, Hugenholtz P, *et al*. Metagenomic

compendium of 189,680 DNA viruses from the human gut microbiome. *Nat Microbiol*. 2021:6(7):960–970. https://doi.org/10.1038/s41564-021-00928-6

Pan H, Cole TL, Bi X, Fang M, Zhou C, Yang Z, Ksepka DT, Hart T, Bouzat JL, Argilla LS, *et al*. High-coverage genomes to elucidate the evolution of penguins. *Gigascience*. 2019:8(9):giz117. https://doi.org/10.1093/gigascience/giz117

Paradis E, Schliep K. Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*. 2019:35(3):526–528. https://doi.org/10.1093/bioinformatics/bty633

Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, Bileschi ML, Bork P, Bridge A, Colwell L, *et al*. InterPro in 2022. *Nucleic Acids Res*. 2023:51(D1):D418–D427. https://doi.org/10.1093/nar/gkac993

Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. *Nucleic Acids Res*. 2018:46(W1):W200–W204. https://doi.org/10.1093/nar/gky448

Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*. 2010:5(3):e9490. https://doi.org/10.1371/journal.pone.0009490

Prum RO, Berv JS, Dornburg A, Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, Lemmon AR. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature*. 2015:526(7574):569–573. https://doi.org/10.1038/nature15697

Rous P. A sarcoma of the fowl transmissible by an agent separable from the tumor cells. *J Exp Med*. 1911:13(4):397–411. https://doi.org/10.1084/jem.13.4.397

Shi M, Lin XD, Chen X, Tian JH, Chen LJ, Li K, Wang W, Eden JS, Shen JJ, Liu L, *et al*. The evolutionary history of vertebrate RNA viruses. *Nature*. 2018:556(7700):197–202. https://doi.org/10.1038/s41586-018-0012-7

Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, Qin XC, Li J, Cao JP, Eden JS, *et al*. Redefining the invertebrate RNA virosphere. *Nature*. 2016:540(7634):539–543. https://doi.org/10.1038/nature20167

Stehelin D, Varmus HE, Bishop JM, Vogt PK. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature*. 1976:260(5547):170–173. https://doi.org/10.1038/260170a0

Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*. 2017:35(11):1026–1028. https://doi.org/10.1038/nbt.3988

Stiller J, Feng S, Chowdhury AA, Rivas-González I, Duchêne DA, Fang Q, Deng Y, Kozlov A, Stamatakis A, Claramunt S, *et al*. Complexity of avian evolution revealed by family-level genomes. *Nature*. 2024:629(8013):851–860. https://doi.org/10.1038/s41586-024-07323-1

Stoye JP. Studies of endogenous retroviruses reveal a continuing evolutionary saga. *Nat Rev Microbiol*. 2012:10(6):395–406. https://doi.org/10.1038/nrmicro2783

Tarlinton RE, Meers J, Young PR. Retroviral invasion of the koala genome. *Nature*. 2006:442(7098):79–81. https://doi.org/10.1038/nature04841

Tsangaras K, Mayer J, Alquezar-Planas DE, Greenwood AD. An evolutionarily young polar bear (*Ursus maritimus*) endogenous retrovirus identified from next generation sequence data. *Viruses*. 2015:7(11):6089–6107. https://doi.org/10.3390/v7112927

Wang J, Chitsaz F, Derbyshire MK, Gonzales NR, Gwadz M, Lu S, Marchler GH, Song JS, Thanki N, Yamashita RA, *et al*. The conserved domain database in 2023. *Nucleic Acids Res*. 2023:51(D1):D384–D388. https://doi.org/10.1093/nar/gkac1096

Wang J, Han GZ. Frequent retroviral gene co-option during the evolution of vertebrates. *Mol Biol Evol*. 2020:37(11):3232–3242. https://doi.org/10.1093/molbev/msaa180

Wang J, Han GZ. Genome mining shows that retroviruses are pervasively invading vertebrate genomes. *Nat Commun*. 2023:14(1):4968. https://doi.org/10.1038/s41467-023-40732-w

Weiss RA, Vogt PK. 100 years of rous sarcoma virus. *J Exp Med*. 2011:208(12):2351–2355. https://doi.org/10.1084/jem.20112160

Wilcox JJS, Arca-Ruibal B, Samour J, Mateuta V, Idaghdour Y, Boissinot S. Linked-read sequencing of eight falcons reveals a unique genomic architecture in flux. *Genome Biol Evol*. 2022:14(6):evac090. https://doi.org/10.1093/gbe/evac090

Xu X, Zhao H, Gong Z, Han GZ. Endogenous retroviruses of non-avian/mammalian vertebrates illuminate diversity and deep history of retroviruses. *PLoS Pathog*. 2018:14(6):e1007072. https://doi.org/10.1371/journal.ppat.1007072

Xu X, Zhou Z, Dudley R, Mackem S, Chuong C-M, Erickson GM, Varricchio DJ. An integrative approach to understanding bird origins. *Science*. 2014:346(6215):1253293. https://doi.org/10.1126/science.1253293

Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW, *et al*. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*. 2014:346(6215):1311–1320. https://doi.org/10.1126/science.1251385

Zhao M, Kurtis SM, White ND, Moncrieff AE, Leite RN, Brumfield RT, Braun EL, Kimball RT. Exploring conflicts in whole genome phylogenetics: a case study within manakins (aves: pipridae). *Syst Biol*. 2023:72(1):161–178. https://doi.org/10.1093/sysbio/syac062

Zheng J, Wang J, Gong Z, Han GZ. Molecular fossils illuminate the evolution of retroviruses following a macroevolutionary transition from land to water. *PLoS Pathog*. 2021:17(7):e1009730. https://doi.org/10.1371/journal.ppat.1009730

Zheng J, Wei Y, Han GZ. The diversity and evolution of retroviruses: perspectives from viral "fossils". *Virol Sin*. 2022:37(1):11–18. https://doi.org/10.1016/j.virs.2022.01.019