

Genome Stability Is in the Eye of the Beholder: CR1 Retrotransposon Activity Varies Significantly across Avian Diversity

James D. Galbraith¹, Robert Daniel Kortschak², Alexander Suh^{3,4,*}, and David L. Adelson ^{1,*}

¹School of Biological Sciences, The University of Adelaide, South Australia, Australia

²Adelaide, South Australia, Australia

³School of Biological Sciences, University of East Anglia, Norwich, United Kingdom

⁴Department of Organismal Biology, Evolutionary Biology Centre (EBC), Science for Life Laboratory, Uppsala University, Sweden

*Corresponding authors: E-mails: alexander.suh@ebc.uu.se; david.adelson@adelaide.edu.au.

Accepted: 12 November 2021

Abstract

Since the sequencing of the zebra finch genome it has become clear that avian genomes, while largely stable in terms of chromosome number and gene synteny, are more dynamic at an intrachromosomal level. A multitude of intrachromosomal rearrangements and significant variation in transposable element (TE) content have been noted across the avian tree. TEs are a source of genome plasticity, because their high similarity enables chromosomal rearrangements through nonallelic homologous recombination, and they have potential for exaptation as regulatory and coding sequences. Previous studies have investigated the activity of the dominant TE in birds, chicken repeat 1 (CR1) retrotransposons, either focusing on their expansion within single orders, or comparing passerines with nonpasserines. Here, we comprehensively investigate and compare the activity of CR1 expansion across orders of birds, finding levels of CR1 activity vary significantly both between and within orders. We describe high levels of TE expansion in genera which have speciated in the last 10 Myr including kiwis, geese, and Amazon parrots; low levels of TE expansion in songbirds across their diversification, and near inactivity of TEs in the cassowary and emu for millions of years. CR1s have remained active over long periods of time across most orders of neognaths, with activity at any one time dominated by one or two families of CR1s. Our findings of higher TE activity in species-rich clades and dominant families of TEs within lineages mirror past findings in mammals and indicate that genome evolution in amniotes relies on universal TE-driven processes.

Key words: transposable element, genome evolution, birds.

Significance

Transposable elements (TEs) are mobile, self replicating DNA sequences within a species' genome, and are ubiquitous sources of mutation. The dominant group of TEs within birds is chicken repeat 1 (CR1) retrotransposons, making up 7–10% of the typical avian genome. Because past research has identified recent inactivity of CR1s within model birds such as the chicken and the zebra finch, this has fostered an erroneous view that all birds have low or no TE activity on recent timescales. Our analysis of numerous high-quality avian genomes across multiple orders identified both similarities and significant differences in how CR1s have expanded. Our results challenge the established view that CR1s in birds are largely inactive and instead suggest that their variable expansions and turnover have contributed to lineage-specific changes in genome structure. Many of the patterns we identify in birds have previously been seen in mammals, further highlighting parallels between the evolution of birds and mammals.

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction

Following rapid radiation during the Cretaceous-Paleogene transition, birds have diversified to be the most species-rich lineage of extant amniotes (Ericson et al. 2006; Jarvis et al. 2014; Wiens 2015). Birds are of particular interest in comparative evolutionary biology because of the convergent evolution of traits seen in mammalian lineages, such as vocal learning in songbirds and parrots (Petkov and Jarvis 2012; Pfenning et al. 2014; Bradbury and Balsby 2016), and potential consciousness in corvids (Nieder et al. 2020). However, in comparison to both mammals and non-avian reptiles, birds have much more compact genomes (Gregory et al. 2007). Within birds, smaller genome sizes correlate with higher metabolic rate and the size of flight muscles (Hughes and Hughes 1995; Wright et al. 2014). However, the decrease in avian genome size occurred in an ancestral dinosaur lineage over 200 Ma, well before the evolution of flight (Organ et al. 2007). A large factor in the smaller genome size of birds in comparison to other amniotes is a big reduction in repetitive content (Zhang et al. 2014).

The majority of transposable elements (TEs) in the chicken (*Gallus gallus*) genome are degraded copies of one superfamily of retrotransposons, chicken repeat 1 (CR1) (International Chicken Genome Sequencing Consortium 2004). The chicken has long been used as the model avian species, and typical avian genomes were believed to have been evolutionarily stable due to little variation in chromosome number and chromosomal painting showing little chromosomal rearrangement (Burt et al. 1999; Shetty et al. 1999). These initial, low-resolution comparisons of genome features, combined with the degraded nature of CR1s in the chicken genome, led to the assumption of a stable avian genome both in terms of karyotype and synteny but also in terms of little recent repeat expansion (International Chicken Genome Sequencing Consortium 2004; Wicker et al. 2005). The subsequent sequencing of the zebra finch (*Taeniopygia guttata*) genome supported the concept of a stable avian genome with little CR1 expansion, but revealed many intrachromosomal rearrangements and a significant expansion of endogenous retroviruses (ERVs), a group of long terminal repeat retrotransposons, since divergence from the chicken (Ellegren 2010; Warren et al. 2010). The subsequent sequencing of 48 bird genomes by the Avian Phylogenomics Project confirmed CR1s as the dominant TE in all non-passerine birds, with an expansion of ERVs in oscine passerines following their divergence from suboscine passerines (Zhang et al. 2014). The TE content of most avian genomes has remained between 7% and 10% not because of a lack of expansion, but due to the loss and decay of repeats and intervening noncoding sequence through nonallelic homologous recombination, canceling out genome size expansion that would have otherwise increased with TE expansion (Kapusta et al. 2017). Since then, hundreds of bird species have been sequenced, revealing

variation in karyotypes, and both intrachromosomal and interchromosomal rearrangements (Hooper and Price 2017; Damas et al. 2018; Feng et al. 2020; Kretschmer, Furo, et al. 2020; Kretschmer, Gunski, et al. 2020). This massive increase in genome sequencing has similarly revealed TEs to be highly active in various lineages of birds. Within the last 10 Myr ERVs have expanded in multiple lineages of songbirds, with the newly inserted retrotransposons acting as a source of structural variation (Suh et al. 2018; Boman et al. 2019; Weissensteiner et al. 2020). Recent CR1 expansion events have been noted in woodpeckers and hornbills, leading to strikingly more repetitive genomes than the “typical” 7–10%. Between 23% and 30% of woodpecker, hornbill, and hoopoe genomes are CR1s, however, their genome assembly size remains similar to that of other birds (Zhang et al. 2014; Manthey et al. 2018; Feng et al. 2020).

Although aforementioned research focusing on the chicken suggested CR1s have not recently been active in birds, research focusing on individual avian lineages has used both recent and ancient expansions of CR1 elements to resolve deep nodes in a wide range of orders including early bird phylogeny (Suh et al. 2011, 2015; Matzke et al. 2012), flamingos and grebes (Suh et al. 2012), landfowl (Kaiser et al. 2007; Kriegs et al. 2007), waterfowl (St John et al. 2005), penguins (Watanabe et al. 2006), ratites (Haddrath and Baker 2012; Baker et al. 2014; Cloutier et al. 2019), and perching birds (Treplin and Tiedemann 2007; Suh et al. 2017). These studies largely exclude terminal branches and, with the exception of a handful of CR1s in grebes (Suh et al. 2012) and geese (St John et al. 2005), the timing of very recent insertions across multiple species remains unaddressed.

An understanding of TE expansion and evolution is important as they generate genetic novelty by promoting recombination that leads to gene duplication and deletion, reshuffling of genes and major structural changes such as inversions and chromosomal translocations (Lim and Simmons 1994; Bailey et al. 2003; Zhou and Mishra 2005; Lee et al. 2008; Chuong et al. 2017; Underwood and Choi 2019). TEs also have the potential for exaptation as regulatory elements and both coding and noncoding sequences (Warren et al. 2015; Wang et al. 2017; Barth et al. 2020; Cosby et al. 2021). Ab initio annotation of repeats is necessary to gain a true understanding of genomic repetitive content, especially in nonmodel species (Platt et al. 2016). Unfortunately, many papers describing avian genomes (Cornetti et al. 2015; Laine et al. 2016; Jaiswal et al. 2018) only carry out homology-based repeat annotation using the Repbase (Bao et al. 2015) library compiled from often distantly related model avian genomes (mainly chicken and zebra finch). This lack of ab initio annotation can lead to the erroneous conclusion that TEs are inactive in newly sequenced species (Platt et al. 2016). Expectations of low repeat expansion in birds inferred from two model species, along with a lack of comparative TE analysis between lineages is the large knowledge gap we addressed here. As CR1s are

the dominant TE lineage in birds and present in all birds (Feng et al. 2020) unlike, for example, CR1-mobilized SINEs which exist in only some birds (Suh et al. 2017; Ottenburghs et al. 2021), we carried out comparative genomic analyses to investigate their diversity and temporal patterns of activity.

Results

Identifying Potential CR1 Expansion across Birds

From all publicly available avian genomes, we selected 117 representative assemblies not under embargo and with a scaffold N50 above 20,000 bp (available at July 2019) for analysis (supplementary table 1, Supplementary Material online). To find all CR1s that may have recently expanded in the 117 genomes, we first used the CARP ab initio TE annotation tool. From the output of CARP, we manually identified and curated CR1s with the potential for recent expansion based on the presence of protein domains necessary for retrotransposition, homology to previously described CR1s, and the presence of a distinctive 3' structure. To retrotranspose and hence expand, CR1s require endonuclease (EN) and reverse transcriptase (RT) domains within a single ORF, and a 3' structure containing a hairpin and microsatellite which potentially acts as a recognition site for the RT (Suh et al. 2014; Suh 2015). If a CR1 identified from homology contained both protein domains and the distinctive 3' structure, we classified it as a "full-length" CR1. We next classified a full-length CR1 as "intact" CR1 if the EN and RT were within a single intact ORF. Using the full-length CR1s and previously described avian and crocodylian CR1s in Repbase as queries (International Chicken Genome Sequencing Consortium 2004; Warren et al. 2010; Green et al. 2014), we performed iterative searches of the 117 genomes to identify divergent, low copy number CR1s which may not have been identified by ab initio annotation. We ensured the protein domains and 3' structures were present throughout the iterative searches. Assemblies with lower scaffold N50s generally contained fewer full-length CR1s and none in the lowest quartile contained intact CR1s (fig. 1). Outside of the lowest quartile, assembly quality appeared to have little impact on the proportion of intact, full-length repeats. The correlation of the low assembly quality with little to no full-length CR1s was seen both across all species and within orders.

Our iterative search identified high numbers of intact CR1s in kiwis, parrots, owls, shorebirds, and waterfowl (figs. 1 and 2). Only two of the 22 perching bird (Passeriformes) genomes contained intact CR1s, and all contained ten or fewer full-length CR1s. Similarly, of the seven landfowl (Galliformes) genomes, only the chicken contained intact CR1s and contained fewer than 20 full-length CR1s. High numbers of full-length and intact repeats were also identified in two woodpeckers, Anna's hummingbird, the chimney swift and the hoatzin, however, due to a lack of other genome sequences

from their respective orders, we were unable to perform further comparative within order analyses of these species to look for recent TE expansion, that is, within the last 10 Myr. Of all the lineages we examined, only four have high-quality assemblies of genera which have diverged within the last 10 Myr and, based on the number of full-length CR1s identified, the potential for very recent CR1 expansion: ducks (*Anas*), geese (*Anser*), Amazon parrots (*Amazona*), and kiwis (*Apteryx*) (Mitchell et al. 2014; Silva et al. 2017; Sun et al. 2017). A large number of full-length repeats were also identified in owls, however, we were unable to examine recent expansion in Strigiformes in detail due to the lack of a dated phylogeny. In addition to our genus scale analyses, we also examined CR1 expansion in parrots (Psittaciformes) overall, perching birds (Passeriformes) and shorebirds (Charadriiformes) since the divergence of each group, and compared the expansion in kiwis and their closest living relatives (Casuariiformes).

Order-Specific CR1 Annotations and a Phylogeny of Avian CR1s Reveal Diversity of Candidate Active CR1s in Neognaths

In order to perform comparative analyses of activity within orders, we created order-specific CR1 libraries. Instead of consensus sequences, all full-length CR1s identified within an order were clustered and the centroids of the clusters were used as cluster representatives for that avian order. To classify the order-specific centroids, we constructed a CR1 phylogeny from the centroids and full-length avian and crocodylian CR1s from Repbase (fig. 3 and supplementary fig. 1 and data 2, Supplementary Material online). From this tree, we partitioned CR1s into families to determine if groups of elements have been active in species concurrently. We partitioned the tree by eye based on the phylogenetic position of previously described CR1 families (Vandergon and Reitman 1994; Wicker et al. 2005; Warren et al. 2010; Bao et al. 2015) and long branch lengths rather than a cutoff for divergence, attempting to find the largest monophyletic groups containing as few previously defined CR1 families as possible. We took this "lumping" approach to our classification to avoid paraphyly and excessive splitting, resulting in some previously defined families being grouped together in one family (supplementary table 2, Supplementary Material online). For example, all full-length CR1s identified in songbirds were highly similar to the previously described CR1-K and CR1-L families and were nested deeply within the larger CR1-J family. As a result, CR1-K, CR1-L, and all full-length songbird CR1s were reclassified as subfamilies of the larger CR1-J family. Based on the position of high confidence nodes with long branch lengths and previously described CR1s in the phylogeny, we defined seven families of avian CR1s, with a new family, CR1-W, which was restricted to shorebirds. Interestingly, the 3' microsatellite of the CR1-W family is a 10-mer rather than

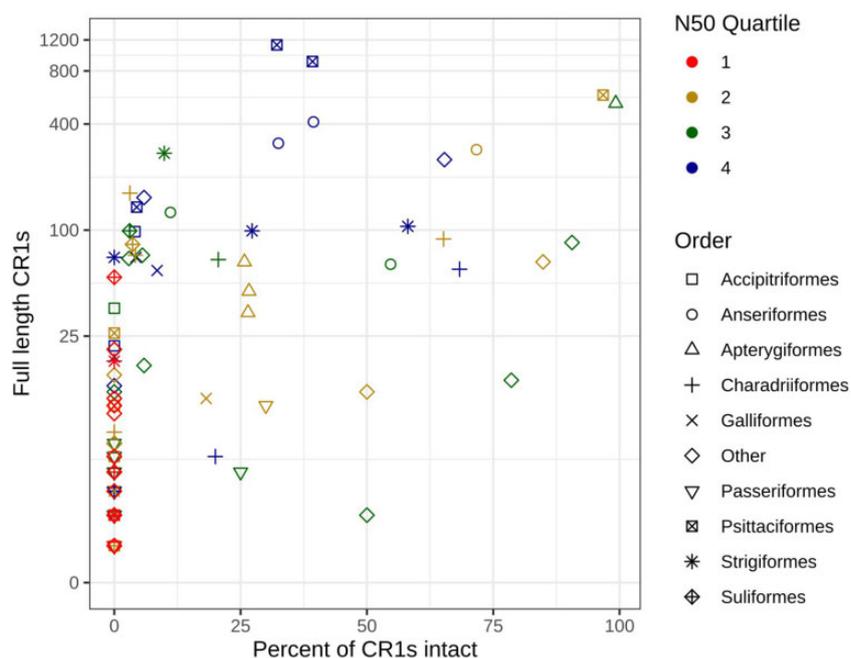


Fig. 1.—The impact of genome assembly quality on the identification of full-length and intact CR1s. CR1s containing both an endonuclease and reverse transcriptase domains were considered full length, and those containing both domains within a single ORF considered intact. Both across all orders and within individual orders, genomes with higher scaffold N50 values (quartiles 2 through 4) had higher numbers of full-length CR1s.

the octamer found in nearly all amniote CR1s (Suh 2015). With the exception of Palaeognathae (ratites and tinamous), all avian orders that contained large numbers of full-length CR1s also contained full-length CR1s from multiple CR1 families (fig. 3).

Variable Timing of Expansion Events across Avian Orders

We used the aforementioned order-specific centroid CR1s and avian and crocodilian Repbase sequences to create order-specific libraries. We used these in reciprocal searches to identify and classify 3' anchored CR1s (3' ends with homology to both the hairpin sequences and microsatellites) present within all orders in which we had identified full-length repeats. We used all 3' anchored CR1s identified above (both full length and truncated) and constructed divergence plots to gain a basic understanding of CR1 expansions within each genome (supplementary data 3 and 4, Supplementary Material online). At high Jukes–Cantor distances, divergence profiles in each order show little difference between species. However, at lower Jukes–Cantor distance, profiles differ significantly between species in some orders. For example, in songbirds at Jukes–Cantor distances higher than 0.1 the overall shape of the divergence plot curves and the proportions of the various CR1 families are nearly identical, whereas at distances lower than 0.1 higher numbers of the CR1-J family are present in *Sporophila hypoxantha* and *T. guttata* than in the three other species (supplementary fig. 2a, Supplementary Material online). CR1s most similar to all defined families

were present in all orders of Galloanserae and Neoaves examined, with the exception of CR1-X which was restricted to Charadriiformes. Almost all CR1s identified in Palaeognathae genomes were most similar to CR1-Y with a small number of truncated and divergent repeats most similar to crocodilian CR1s (supplementary data 3, Supplementary Material online).

Divergence plots may not accurately indicate the timing of repeat insertions as they assume uniform substitution rates across the noncoding portion of the genome. High divergence could be a consequence of either full-length CR1s being absent in a genome or the centroid identified by the clustering algorithm being distant from the CR1s present in a genome. To better determine when CR1 families expanded in avian genomes, we first identified regions orthologous to CR1 insertions sized 100–600 bp in related species (see Materials and Methods). We compared these orthologous regions and approximated the timing of insertion based on the presence or absence of the CR1 insertion in the other species. In most orders only long term trends could be estimated due to long branch lengths (cf., fig. 2) and high variability of the quality of genome assemblies (cf., fig. 1). Therefore, we focused our presence/absence analyses to reconstruct the timing of CR1 insertions in parrots, waterfowl, perching birds, and kiwis (fig. 4). We also applied the method to owls (supplementary figure 3, Supplementary Material online) and shorebirds (fig. 5), however, due to the lack of order-specific fossil-calibrated phylogenies of owls and long branch lengths of shorebirds, we could not determine how recent the CR1 expansions were.

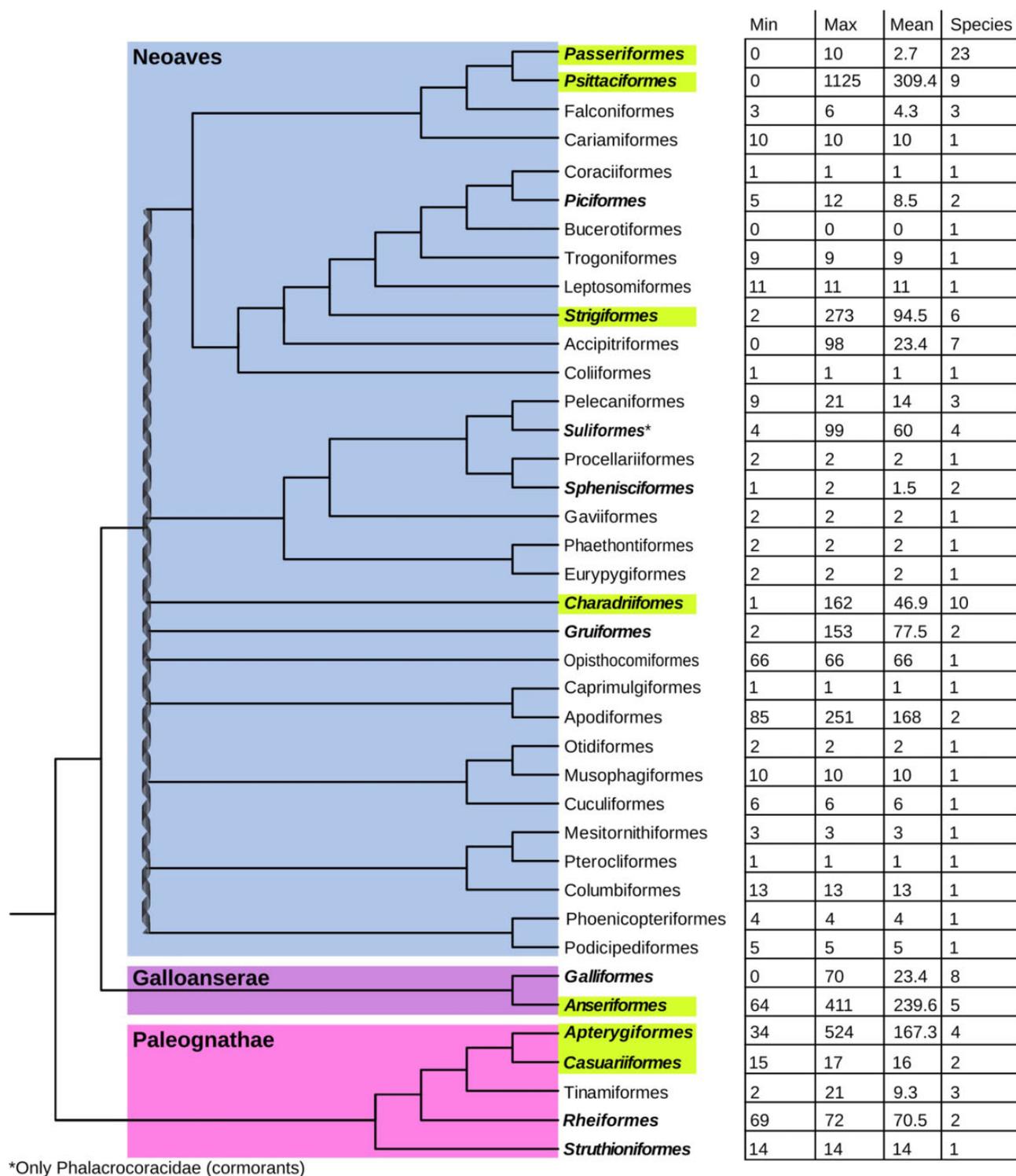


FIG. 2.—The number of full-length CR1s varies significantly across the diversity of birds sampled. Minimum, maximum, and mean number of full-length CR1 copies identified in each order of birds, and the number of species surveyed in each order. Largest differences are noticeable between sister clades such as parrots (Psittaciformes) and perching birds (Passeriformes), and landfowl (Galliformes) and waterfowl (Anseriformes). The double helix represents a putative hard polytomy at the root of Neoaves (Suh 2016). Orders bolded contain at least one intact and potentially active CR1 copy and those highlighted are the orders examined in detail. For coordinates of full-length CR1s within genomes, see [supplementary data 1, Supplementary Material](#) online. Tree adapted from Mitchell et al. (2014) and Suh (2016).

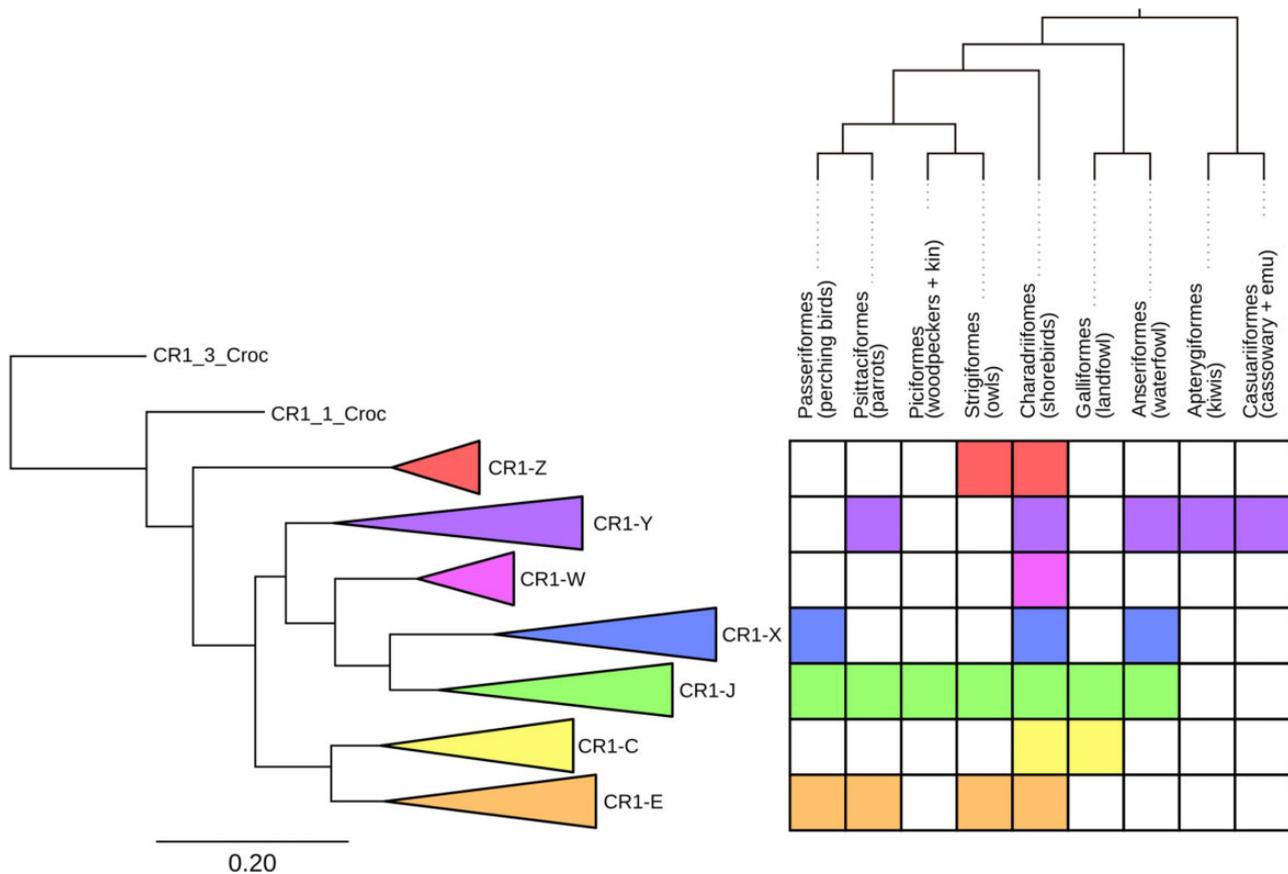


FIG. 3.—Collapsed tree of full-length CR1s and presence of full-length copies of CR1 families in selected avian orders. The name of each family is taken from a previously described CR1 present within the family (supplementary table 3, Supplementary Material online). The coloring of squares indicates the presence of full-length CR1s within the order. All orders shown were chosen due to the presence of high numbers of intact CR1 elements, except for Casuariiformes which are shown due to their recent divergence from Apterygiformes as well as Passeriformes due to their species richness and frequent use as model species (especially zebra finch). The full CR1 tree was constructed using FastTree from a MAFFT alignment of the nucleotide sequences. For the full tree and nucleotide alignment of 1,278 CR1s, see supplementary figure 1 and data 2, Supplementary Material online.

In analyzing the repeat expansion in the kiwi genomes, we used the closest living relatives, the cassowary and emu (Casuariiformes), as outgroups. Following the divergence of kiwis from Casuariiformes, CR1-Y elements expanded, both before and during the recent speciation of kiwis over the last few Myr. In contrast, there was little CR1 expansion in Casuariiformes, both following their divergence from kiwis, and more recently since their divergence approximately 28 Ma, with only one insertion found in the emu and three in the cassowary since they diverged (supplementary table 3, Supplementary Material online).

In the waterfowl species examined, both CR1-J and CR1-X families expanded greatly in both ducks and geese during the last 2 Myr. Expansion occurred in both examined genera, with greater expansions in the ducks (*Anas*) than the geese (*Anser*). Other CR1 families appear to have been active following the two groups' divergence approximately 30 Ma, but have not been active since each genus speciated.

Due to the high number of genomes available for passerines, we chose best quality representative genomes from

major groups sensu (Oliveros et al. 2019); New Zealand wrens (*Acanthisitta chloris*), Suboscines (*Manacus vitellinus*), Corvides (*Corvus brachyrhynchos*), and Muscicapida (*Sturnus vulgaris*), Sylvida (*Phylloscopus trochilus* and *Zosterops lateralis*), and Passerida (*T. guttata*, *S. hypoxantha*, and *Zonotrichia albicollis*). Between the divergence of Oscines (songbirds) and Suboscines from New Zealand wrens and the divergence of Oscines, there was a large spike in expansion of multiple families of CR1s, predominantly CR1-X. Since their divergence 30 Ma, only CR1-J remained active in oscines, though the degree of expansion varied between groups.

Of all avian orders examined, we found the highest levels of CR1 expansion in parrots. Because most branch lengths on the species tree were long, the timing of recent expansions could only be reconstructed in genus *Amazona*. The species from *Amazona* diverged 5 Ma and seem to vary significantly in their level of CR1 expansion. However, genome assembly quality might be a confounder as the number of insertions into a species of *Amazona* was highest in the best quality

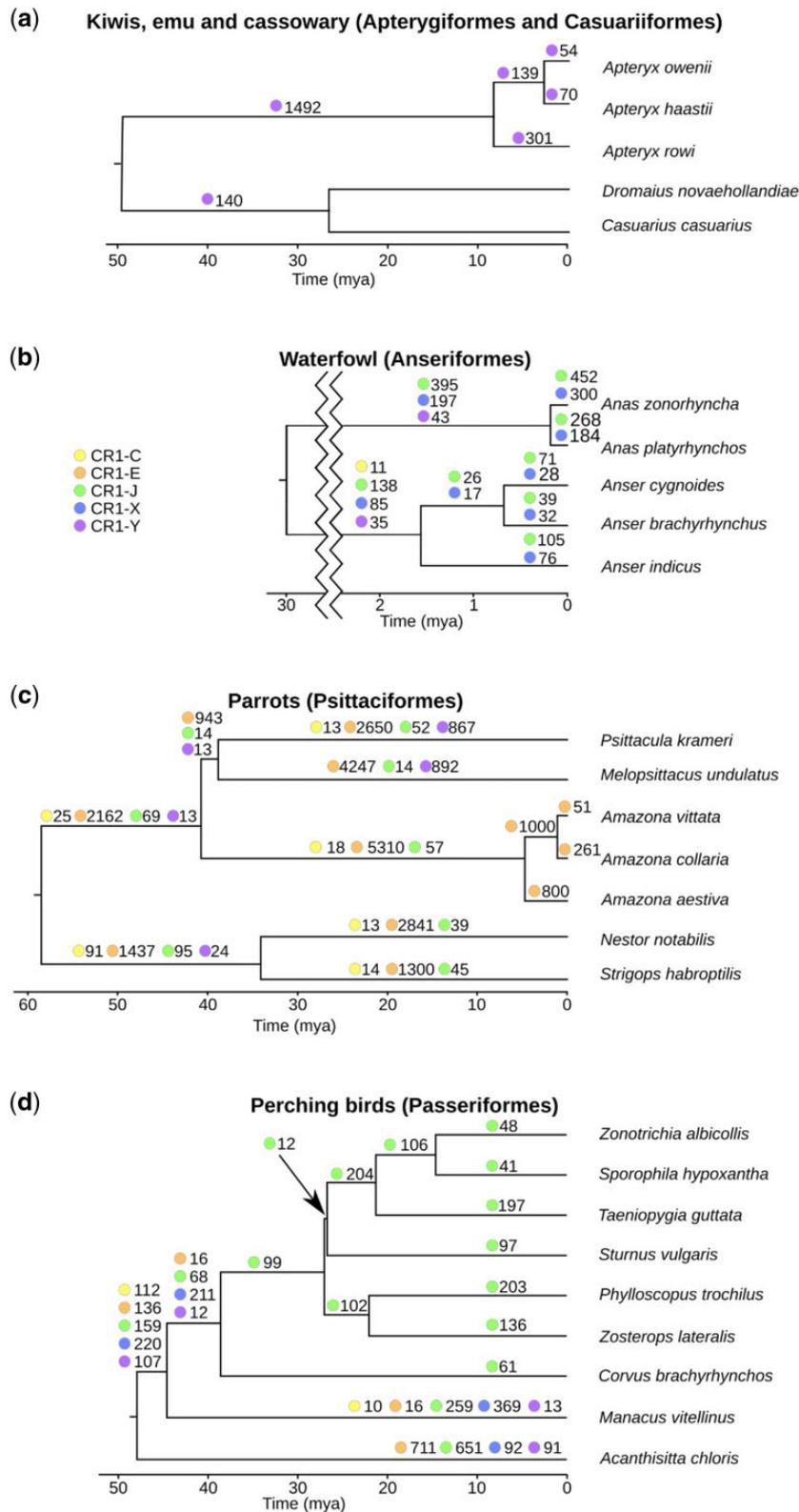


Fig. 4.—Presence/absence patterns reconstruct the timing of expansions of dominant CR1 families within five selected avian orders. The number next to the colored circle is the number of CR1 insertions found. Only CR1 families with more than ten CR1 presence/absence patterns (only CR1 insertions ranging between 100 and 600 bp were analyzed) are shown, for the complete number of insertions, see [supplementary table 3, Supplementary Material](#) online. Phylogenies adapted from Mitchell et al. (2014), Oliveros et al. (2019), Silva et al. (2017), and Sun et al. (2017).

Downloaded from <https://academic.oup.com/gbe/article/13/12/evab259/6433158> by University of East Anglia user on 15 December 2021

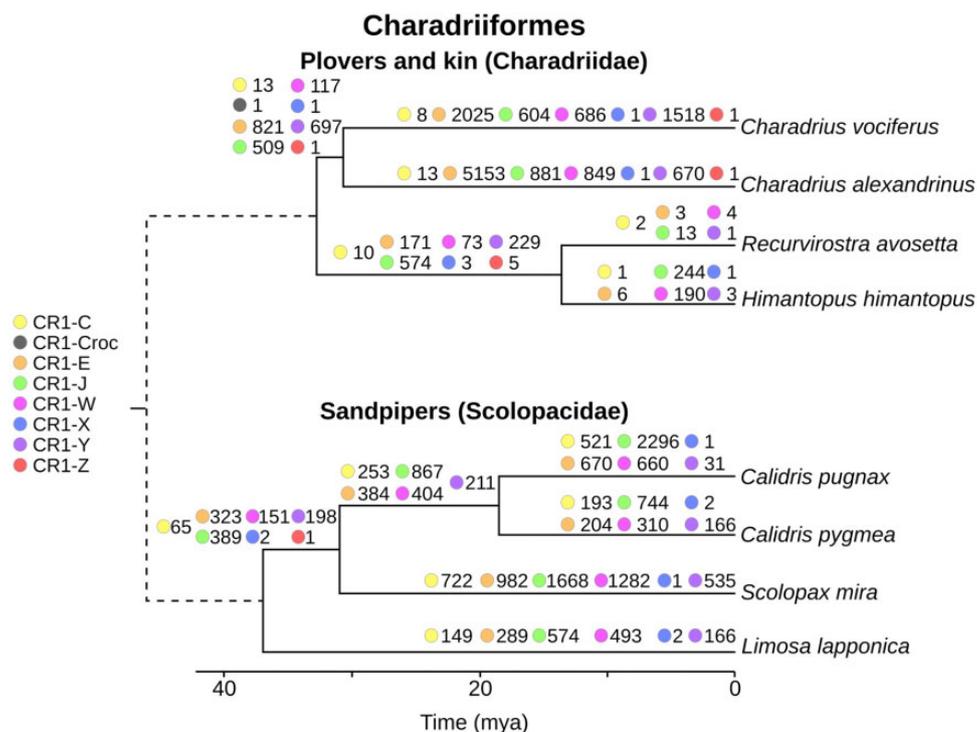


Fig. 5.—Presence/absence patterns reconstruct the timing of expansions of CR1 families in two lineages of shorebirds (Charadriiformes): plovers and sandpipers. The number next to the colored circle is the number of CR1 insertions identified and only CR1 insertions between 100 and 600 bp long were analyzed. Divergence dates between plovers and sandpiper clades may differ due to the source phylogenies (Paton et al. 2003; Baker et al. 2007; Barth et al. 2013) being constructed using different approaches.

genome (*Amazona collaria*), and lowest in the worst quality genome (*Amazona vittata*). In all parrots, CR1-E was the predominant expanding CR1 family, however, CR1-Y expanded in the *Melopsittacus–Psittacula* lineage, while remaining largely inactive in the other parrot lineages.

Multiple expansions of multiple families of CR1s have occurred in the two shorebird lineages examined; plovers (Charadriidae) and sandpipers (Scolopacidae) (fig. 5). The diversity of CR1 families that remained active through time was higher than in the other orders investigated, particularly in sandpipers, with four CR1 families showing significant expansion in *Calidris pugnax* and five in *Calidris pygmaea*, since their divergence. In all other orders examined in detail, CR1 expansions over similar time periods have been dominated by only one or two families, with insertions of fewer than ten CR1s from nondominant families (supplementary table 2, Supplementary Material online). Unfortunately, due to long branch lengths more precise timing of these expansions is not possible.

Finally, CR1s continuously expanded in true owls since divergence from barn owls, with almost all resolved insertions being CR1-E-like (supplementary fig. 3, Supplementary Material online). However, due to the lack of a genus-level timed phylogeny, the precise timing of these expansions cannot be determined.

Combined, our CR1 presence/absence analyses demonstrate that the various CR1 families have expanded at different rates both within and across avian orders. These differences are considerable, ranging from an apparent absence of CR1 expansion in the emu and cassowary to slow, continued expansion of a single CR1 family in songbirds, to recent rapid expansions of one or two CR1 families in kiwis, Amazon parrots and waterfowl, as well as a wide variety of CR1 families expanding concurrently in sandpipers.

To further examine the relative timing of the expansion of the various CR1 families in relation to each other, we performed transposition in transposition (TinT) analysis in species we have analyzed in detail above (supplementary data 5, Supplementary Material online). The TinT analysis largely confirmed the relative ages of insertions and activity profiles from the divergence and presence/absence analyses.

Discussion

Genome Assembly Quality Impacts Repeat Identification

The quality of a genome assembly has a large impact on the number of CR1s identified within it, both full-length and 5'-truncated. This is made clear when comparing the number of insertions identified within species in recently diverged genera. The three *Amazona* parrot species diverged

approximately 2 Ma (Silva et al. 2017) and the scaffold N50s of *A. vittata*, *A. aestiva*, and *A. collaria* are 0.18, 1.3, and 13 Mb, respectively. No full-length CR1s were identified in *A. vittata*, and only ten in *A. aestiva*, whereas 1,125 were identified in *A. collaria*. Similarly, in *Amazona* the total number of truncated insertions identified increased significantly with higher scaffold N50s. In contrast, the three species of kiwi compared, diverged approximately 7 Ma, and have similar N50s (between 1.3 and 1.7 Mb). This pattern of higher quality genome assemblies leading to higher numbers of both full-length and intact CR1s being identified is consistent across most orders examined, and is particularly true of the lowest N50 quartile (fig. 1). The lower number of repeats identified in lower quality assemblies is likely due to the sequencing technology used. Repeats are notoriously hard to assemble and are often collapsed, particularly when using short read Illumina sequencing, leading to fragmented assemblies (Alkan et al. 2011; Treangen and Salzberg 2011). The majority of the genomes we have used are of this data type. The recent sequencing of avian genomes using multiplatform approaches have resolved gaps present in short read assemblies, finding these gaps to be rich in interspersed, simple, and tandem repeats (Li et al. 2021; Peona et al. 2021). Of particular note, Li et al. (2021) resolved gaps in the assembly of *Anas platyrhynchos* which we analyzed here using long-read sequencing, and found the gaps to be dominated by the two CR1 families that have recently expanded in waterfowl (Anseriformes): CR1-J and CR1-X. Species with low-quality assemblies may have full-length repeats present in their genome, yet the sequencing technology used prevents the assembly of the repeats and hence detection. Thus, CR1 activity may be even more widespread in birds than we estimate here.

The Origin and Evolution of Avian CR1s

Avian CR1s are monophyletic in regards to other major CR1 lineages found in amniotes (Suh et al. 2014). For comparison, crocodylians contain some CR1 families more similar to those found in testudines and squamates than others in crocodylians. By searching for truncated copies of previously described CR1s in addition to our order-specific CR1s, we were able to uncover how CR1s have evolved in avian genomes as birds have diverged. CR1-Y is the only family with full-length CR1s present in Palaeognathae, Galloanserae, and Neoaves. The omnipresence of CR1-Y indicates it was present in the ancestor of all birds. A small number of highly divergent truncated copies of CR1s most similar to CR1-Z are found in ratites and CR1-J in tinamous (supplementary fig. 2b, Supplementary Material online). This is potentially indicative of an ancestral presence of CR1-J and CR1-Z in the common ancestor of all birds, or misclassification owing to the high divergence of these CR1 fragments. As mentioned above, we took a lumping approach to classification to CR1 classification to avoid paraphyly, thereby collapsing highly similar families elsewhere

considered as separate families. As CR1-C, CR1-E, and CR1-X are present in both Galloanserae and Neoaves but absent from Palaeognathae, we conclude these three families likely originated following the divergence of neognaths from paleognaths, but prior to the divergence of Neoaves and Galloanserae. In addition to having a 10-bp microsatellite instead of the typical 8-bp microsatellite, CR1-W is peculiar as it is unique to Charadriiformes but sister to CR1-J and CR1-X (fig. 3). This implies an origin in the neognath ancestor, followed by retention and activity in measurable numbers only in Charadriiformes.

A wide variety of CR1 families has expanded in all orders of neognaths, with many potential expansion events within the past 10 Myr present in many lineages. As mentioned in the results, it is not possible to conclude that insertions are ancient based on divergence plots alone. Some species with low-quality genome assemblies, such as *A. vittata*, contained very few full-length repeats compared with relatives (supplementary fig. 4, Supplementary Material online). As a result of full-length repeats not being assembled, the divergence of most or all truncated insertions identified in *A. vittata* would likely be calculated using CR1 centroids identified in *A. collaria*, leading to higher divergence values than those identified in *A. collaria*, and in turn an incorrect assumption of less recent expansion in *A. vittata* than *A. collaria*. In addition to fewer full-length repeats being assembled, fewer truncated repeats also appear to have been assembled in poorer quality genomes.

CR1 Family Expansions within Orders

Across all sampled neognaths, recent expansions appear to be largely restricted to one or two families of CR1. Our presence/absence analyses found this to be the case in waterfowl, parrots, songbirds, and owls, with shorebirds and the early passerine divergences the only exceptions. Similarly, based on the phylogeny of full-length elements, most orders only retain full-length CR1s from two or three families, whereas shorebirds retain full-length CR1s from across all seven families. Our presence/absence analysis revealed likely concurrent expansions of at least four CR1 families in two families of shorebirds: sandpipers of genus *Calidris* and plovers of genus *Charadrius*. In both genera four families of CR1s have significantly expanded since their divergence including the order-specific CR1-W (fig. 5). Although in both genera one family accounts for 40–50% of insertions, the other three families have hundreds of insertions each. This is highly different to the pattern seen in songbirds and waterfowl which, over a similar time period, have single digit insertions of nondominant CR1 families (supplementary table 3, Supplementary Material online).

This increase of CR1 diversity in shorebirds could be due to some CR1 families in shorebirds having 3' inverted repeat and microsatellite motifs which differ from the typical structure (Suh 2015) (supplementary fig., Supplementary Material

online). For example, the CR1-W family has an extended 10-bp microsatellite (5'-AAATTCYGTG-3') rather than the 8-bp microsatellite (5'-ATTCTRTG-3') seen in nearly all other avian CR1s. When transcribed the 3' structure upstream of the microsatellite is hypothesized to form a stable hairpin which acts as a recognition site for the cis-encoded RT (Luan et al. 1993; Suh 2015; Suh et al. 2017). The recently active CR1s we identified in other avian orders have 3' microsatellites and hairpins which closely resemble those previously described. Although the changes seen in shorebirds are minor, we speculate they could impact CR1 mobilization, allowing for more families to remain active than the typical one or two.

Rates of CR1 Expansion Can Vary Significantly within Orders

Based on the presence/absence of CR1 insertions and divergence plots and TinT analysis, rates of CR1 expansion within lineages appear to vary even across rather short evolutionary timescales. The expansion of CR1-Y in kiwis appears to be a recent large burst of expansion and accumulation, whereas since Passeriformes diverged CR1-J appear to have continued to expand slowly in all families, however, the number of new insertions seen in the American crow is much lower than that seen in the other oscine songbird species surveyed. The expansion of CR1-Y seen in the *Psittacula-Melopsittacus* lineage of parrots, following their divergence from the lineage leading to *Amazona*, appears to result from an increase in expansion, with little expansion in the period prior to divergence and none observed in other lineages of parrots. CR1s appear to have been highly active in all parrots examined since their divergence, however, due to the less dense sampling it is not clear if this has been continuous expansion as in songbirds or a burst of activity like that in kiwis. Finally, in sandpipers CR1s have continued to expand in both species of *Calidris* since divergence, however, the much lower number of new insertions in *C. pygmaea* suggests the rate of expansion differs significantly between the two species.

All full-length CR1s identified in ratites were CR1-Y, and almost all truncated copies found in ratites were most similar to either CR1-Y, or crocodylian CR1s typically not found in birds (Suh et al. 2014). This retention of ancient CR1s and the presence of full-length CR1s in species such as the southern cassowary (*Casuarus casuarus*) and emu (*Dromaius novaehollandiae*), yet without recent expansion, reflects the much lower substitution and deletion rates in ratites compared with Neoaves (Zhang et al. 2014; Kapusta et al. 2017). These crocodylian-like CR1s in ratites may be truncated copies of CR1s that were active in the common ancestor of crocodylians and birds (Suh et al. 2014), whereas we hypothesize that these have long since disappeared in Neoaves due to their higher deletion and substitution rates (Zhang et al. 2014; Kapusta et al. 2017).

Co-Occurrence of CR1 Expansion with Speciation

The four genera containing recent CR1 expansions we have examined co-occur with rapid speciation events. Of particular note, kiwis rapidly speciated into five distinct species composed of at least 16 distinct lineages arising due to significant population bottlenecks caused by Pleistocene glacial expansions (Weir et al. 2016). We speculate that the smaller population sizes might have allowed for CR1s to expand as a result of increased genetic drift (Szitenberg et al. 2016). This reflects previous findings of rapid fixation of TEs following population bottlenecks in birds (Matzke et al. 2012). Although we do not see CR1 expansion occurring alongside speciation in passerines, ERVs, which are rare in other birds, have expanded throughout their diversification (Warren et al. 2010; Boman et al. 2019). Investigating the potentially ongoing expansion of CR1s and its relationship to speciation in ducks, geese, and Amazon parrots will require a larger number of genomes from within the same and sister genera to be sequenced, especially in waterfowl due to the high rates of hybridization even between long diverged species (Ottenburghs et al. 2015).

Comparison to Mammals

As mentioned in the introduction, many parallels have been drawn between LINEs in birds and mammals, most notably the expansion of LINEs in both clades being balanced by a loss through purifying selection (Kapusta et al. 2017). Here, we have found additional trends in birds previously noted in mammals. The TE expansion during periods of speciation seen in *Amazona*, *Apteryx*, and *Anas* has previously been observed across mammals (Ricci et al. 2018). Similarly, the dominance of one or two CR1 families seen in most orders of birds resembles the activity of L1s in mammals (Ivancevic et al. 2016), however, the general persistence of activity of individual CR1 families seems to be more diverse (Kriegs et al. 2007; Suh et al. 2011).

Conclusion: The Avian Genome Is More Dynamic Than Meets the Eye

Although early comparisons of avian genomes were restricted to the chicken and zebra finch, where high level comparisons of synteny and karyotype led to the conclusion that bird genomes were largely stable compared with mammals (Ellegren 2010), the discovery of many intrachromosomal rearrangements across birds (Skinner and Griffin 2012; Zhang et al. 2014; Farré et al. 2016; Hooper and Price 2017) and interchromosomal recombination in falcons, parrots, and sandpipers (O'Connor et al. 2018; Coelho et al. 2019; Pinheiro et al. 2021) has shown that at a finer resolution for comparison, the avian genome is rather dynamic. The highly variable rate of TE expansion we have observed across birds extends knowledge from avian orders with "unusual"

repeat landscapes, that is, Piciformes (Manthey et al. 2018) and Passeriformes (Warren et al. 2010), and provides further evidence that the genome evolution of bird orders and species within orders differs significantly, even though synteny is often conserved. In our comprehensive characterization of CR1 diversity across 117 bird genome assemblies, we have identified significant variation in CR1 expansion rates, both within genera such as *Calidris* and between closely related orders such as kiwis and the cassowary and emu. As the diversity and quality of avian genomes sequenced continues to grow and whole-genome alignment methods improve (Feng et al. 2020; Rhie et al. 2020), further analysis of genome stability based on repeat expansions at the family and genus level will become possible. Although the chicken and zebra finch are useful model species, models do not necessarily represent diversity of evolutionary trajectories in nature. Our results indicate that recurrent, similar patterns of TE family expansion are seen across amniotes and suggest mechanisms of TE-driven genome evolution can be generalized across tetrapods.

Materials and Methods

Identification and Curation of Potentially Divergent CR1s

To identify potentially divergent CR1s, we processed 117 bird genomes downloaded from GenBank (Benson et al. 2015) with CARP (Zeng et al. 2018); see supplementary table, [Supplementary Material](#) online, for species names and assembly versions. We used RPSTBlastN (Altschul et al. 1997) with the CDD library (Marchler-Bauer et al. 2017) to identify protein domains present in the consensus sequences from CARP. Consensuses which contained both an EN and a RT domain were classified as potential CR1s. Using CENSOR (Kohany et al. 2006), we confirmed these sequences to be CR1s, removing others, more similar to different families of LINES, such as AviRTes, as necessary.

Confirmed CR1 CARP consensus sequences were manually curated through a “search, extend, align, trim” method as described in (Galbraith et al. 2020) to ensure that the 3′ hairpin and microsatellite were intact. Briefly, this curation method involves searching for sequences highly similar to the consensus with BlastN 2.7.1+ (Zhang et al. 2000), extending the coordinates of the sequences found by flanks of 600 bp, aligning these sequences using MAFFT v7.453 (Kato and Standley 2013) and trimming the discordant regions manually in Geneious Prime v2020.1. The final consensus sequences were generated in Geneious Prime from the trimmed multiple sequence alignments by majority rule.

Identification of More Divergent and Low Copy CR1s

To identify more divergent or low copy number CR1s which CARP may have failed to identify, we performed an iterative search of all 117 genomes. Beginning with a library of all avian

CR1s in Repbase (Bao et al. 2015) (see [supplementary table 2, Supplementary Material](#) online, for CR1 names and species names) and manually curated CARP sequences, we searched the genomes using BlastN (-task dc-megablast -max_target_seqs <number of scaffolds in respective genome>), selecting those over 2,700 bp and retaining 3′ hairpin and microsatellite sequences. Using RPSTBlastN, we then identified the full-length CR1s (those containing both EN and RT domains) and combined them with the previously generated consensus sequences. We clustered these combined sequences using VSEARCH 2.7.1 (Rognes et al. 2016) (-cluster_fast -id 0.9) and combined the cluster centroids with the Repbase CR1s to use as queries for the subsequent search iteration. This process was repeated until the number of CR1s identified did not increase compared with the previous round. From the output of the final round, order-specific clusters of CR1s were constructed and cluster centroids identified.

Tree Construction

To construct a tree of CR1s, the centroids of all order-specific CR1s were combined with all full-length avian and two crocodilian CR1s from Repbase and globally aligned using MAFFT (-thread 12 -localpair). We used FastTree 2.1.11 with default nucleotide parameters (Price et al. 2010) to infer a maximum likelihood phylogenetic tree from this alignment, and rooted the tree using the crocodilian CR1s. The crocodilian CR1s were used as an outgroup as all avian CR1s are nested within crocodilian CR1s (Suh et al. 2015). This tree was split into different families of CR1 by eye, based on the presence of long branches from high confidence nodes and the position of the previously described CR1 families from Repbase. To avoid excessive splitting and paraphyly of previously described families a lumping approach was taken resulting in some previously distinct families of CR1 from Repbase being treated as members of families they were nested within ([supplementary table 3, Supplementary Material](#) online).

Identification and Classification of CR1s within Species

To identify, classify, and quantify divergence of all 3′ anchored CR1s present within species, order-specific libraries were constructed from the order-specific clusters and the full-length avian and crocodilian Repbase CR1s. 3′-anchored sequences CR1s were defined as CR1s retaining the 3′ hairpin and microsatellite sequences. Using these libraries as queries, we identified 3′ anchored sequences CR1s present in assemblies using BlastN. The identified CR1s were then classified using a reciprocal BlastN search against the original query library.

Determination of Presence/Absence in Related Species

To reconstruct the timing of CR1 expansions, we selected the identified 3′ anchored CR1 copies of 100 and 600 bp length in a species of interest and at least 600 bp from the end of a

contig, extending the coordinates of the sequences by 600 bp to include the flanking region and extracting the corresponding sequences. If the flanking regions contained more than 25% unresolved nucleotides (“N” nucleotides) they were discarded.

Using BlastN, we identified homologous regions in species belonging to the same order as the species being analyzed, and through the following process of elimination identified the regions orthologous to CR1 insertions and their flanks in the related species. At each step of this process of elimination, if an initial query could not be satisfactorily resolved, we classified it as unscorable (unresolved) to reduce the chance of falsely classifying deletions or segmental duplications as new insertion events. First, we classified all hits containing the entire repeat and at least 150 bp of each flank as shared orthologous insertions. Following this, we discarded all hits with outer coordinates less than a set distance (150 bp) from the boundary of the flanks and CR1s to remove hits to paralogous CR1s insertions. This distance was chosen by testing the effect of a range of distances from 300 bp through to 50 bp in increments of 50 bp on a random selection of CR1s first identified in *Anser cygnoides* and *Corvus brachyrhynchos* and searched for in other species within the same order. Requiring outer coordinates to be higher values resulted in higher numbers of orthologous regions not being resolved, likely due to insertions or deletions within flanks since divergence. Allowing for boundaries of 50 or 100 bp resulted in many CR1s having multiple potential orthologous regions at 3′ flanks, many of which were false hits, only showing homology to the target site duplication and additional copies of the 3′ microsatellite sequence. Thus, 150 bp was chosen, as it was the shortest possible distance at which a portion of the flanking sequence was always present.

Based on the start and stop coordinates of the remaining hits, we determined the orientation the hit was in and discarded any queries without two hits in the same orientation. In addition, any queries with more than one hit to either strand were discarded. From the remaining data, we determined the distance between the two flanks. If the two flanks were within 16 bp of each other in the sister species and the distance between the flanks was near the same length of the query CR1, the insertion was classified as having occurred since divergence. If the distance between the ends of the flanks in both the original species and sister species were similar, the insertion was classified as shared. For a pictorial description of this process including the parameters used, see [supplementary figure 5, Supplementary Material](#) online. This process was conducted for other species in the same order as the original species. Finally, we determined the timing of each CR1 insertion event by reconciling the presence/absence of each CR1 insertion across sampled species with the most parsimonious placement on the species tree ([supplementary fig. 6, Supplementary Material](#) online).

Further Estimating Recent Activity by Identifying Transpositions in Transpositions

To further qualify timing the recent expansions of CR1 subfamilies in waterfowl, shorebirds, parrots, kiwis, cassowary, and emu, we performed “transposition in transposition” (TinT) analyses. We masked the relevant genomes using RepeatMasker (Smit 2004) and a library used consisting of the centroids of final output of the reciprocal search described above, combined with all avian and two crocodilian CR1s from Repbase. Using the TinT application (Churakov et al. 2010), we estimated the timing of CR1 subfamilies’ expansion relative to other subfamilies in each genome ([supplementary data 5, Supplementary Material](#) online).

Supplementary Material

[Supplementary data](#) are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank the anonymous reviewers for their constructive comments. Additionally, we thank Valentina Peona, Jesper Boman, Julie Blommaert, and Alastair Ludington for comments on an earlier version of this manuscript.

Data Availability

All scripts used are available at https://github.com/jamesdgalbraith/Avian_CR1_Activity. All genome details, genome coordinates, and sequence alignments are present in the [supplementary data, Supplementary Material](#) online. Raw sequence data used in alignments are available at <https://doi.org/10.5281/zenodo.5644944>.

Literature Cited

- Alkan C, Sajjadian S, Eichler EE. 2011. Limitations of next-generation genome sequence assembly. *Nat Methods*. 8(1):61–65.
- Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25(17):3389–3402.
- Bailey JA, Liu G, Eichler EE. 2003. An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet*. 73(4):823–834.
- Baker AJ, Haddrath O, McPherson JD, Cloutier A. 2014. Genomic support for a moa-tinamou clade and adaptive morphological convergence in flightless ratites. *Mol Biol Evol*. 31(7):1686–1696.
- Baker AJ, Pereira SL, Paton TA. 2007. Phylogenetic relationships and divergence times of Charadriiformes genera: multigene evidence for the Cretaceous origin of at least 14 clades of shorebirds. *Biol Lett*. 3(2):205–209.
- Bao W, Kojima KK, Kohany O. 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*. 6:11.
- Barth JMI, Matschiner M, Robertson BC. 2013. Phylogenetic position and subspecies divergence of the endangered New Zealand Dotterel (*Charadrius obscurus*). *PLoS One* 8(10):e78068.

- Barth NKH, Li L, Taher L. 2020. Independent transposon exaptation is a widespread mechanism of redundant enhancer evolution in the mammalian genome. *Genome Biol Evol.* 12(3):1–17.
- Benson DA, et al. 2015. GenBank. *Nucleic Acids Res.* 43(Database issue):D30–D35.
- Boman J, et al. 2019. The genome of blue-capped Cordon-Bleu uncovers hidden diversity of LTR retrotransposons in zebra finch. *Genes* 10(4):301.
- Bradbury JW, Balsby TJS. 2016. The functions of vocal learning in parrots. *Behav Ecol Sociobiol.* 70(3):293–312.
- Burt DW, et al. 1999. The dynamics of chromosome evolution in birds and mammals. *Nature* 402(6760):411–413.
- Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet.* 18(2):71–86.
- Churakov G, et al. 2010. A novel web-based TinT application and the chronology of the primate Alu retroposon activity. *BMC Evol Biol.* 10:376.
- Cloutier A, et al. 2019. Whole-genome analyses resolve the phylogeny of flightless birds (Palaeognathae) in the presence of an empirical anomaly zone. *Syst Biol.* 68(6):937–955.
- Coelho LA, Musher LJ, Cracraft J. 2019. A multireference-based whole genome assembly for the obligate ant-following antbird, *Rhegmatorhina melanosticta* (Thamnophilidae). *Diversity* 11(9):144.
- Cornetti L, et al. 2015. The genome of the ‘great speciator’ provides insights into bird diversification. *Genome Biol Evol.* 7(9):2680–2691.
- Cosby RL, et al. 2021. Recurrent evolution of vertebrate transcription factors by transposase capture. *Science* 371(6531):eabc6405. doi: 10.1126/science.abc6405.
- Damas J, Kim J, Farré M, Griffin DK, Larkin DM. 2018. Reconstruction of avian ancestral karyotypes reveals differences in the evolutionary history of macro- and microchromosomes. *Genome Biol.* 19(1):155.
- Ellegren H. 2010. Evolutionary stasis: the stable chromosomes of birds. *Trends Ecol Evol.* 25(5):283–291.
- Ericson PGP, et al. 2006. Diversification of Neoaves: integration of molecular sequence data and fossils. *Biol Lett.* 2(4):543–547.
- Farré M, et al. 2016. Novel insights into chromosome evolution in birds, archosaurs, and reptiles. *Genome Biol Evol.* 8(8):2442–2451.
- Feng S, et al. 2020. Dense sampling of bird diversity increases power of comparative genomics. *Nature* 587(7833):252–257.
- Galbraith JD, Ludington AJ, Suh A, Sanders KL, Adelson DL. 2020. New environment, new invaders—repeated horizontal transfer of LINEs to sea snakes. *Genome Biol Evol.* 12(12):2370–2383.
- Green RE, et al. 2014. Three crocodylian genomes reveal ancestral patterns of evolution among archosaurs. *Science* 346(6215):1254449.
- Gregory TR, et al. 2007. Eukaryotic genome size databases. *Nucleic Acids Res.* 35(Database issue):D332–D338.
- Haddrath O, Baker AJ. 2012. Multiple nuclear genes and retroposons support vicariance and dispersal of the palaeognaths, and an early Cretaceous origin of modern birds. *Proc Biol Sci.* 279(1747):4617–4625.
- Hooper DM, Price TD. 2017. Chromosomal inversion differences correlate with range overlap in passerine birds. *Nat Ecol Evol.* 1(10):1526–1534.
- Hughes AL, Hughes MK. 1995. Small genomes for better flyers. *Nature* 377(6548):391.
- International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432:695–716.
- Ivancevic AM, Kortschak RD, Bertozzi T, Adelson DL. 2016. LINEs between species: evolutionary dynamics of LINE-1 retrotransposons across the eukaryotic tree of life. *Genome Biol Evol.* 8(11):3301–3322.
- Jaiswal SK, et al. 2018. Genome sequence of peacock reveals the peculiar case of a glittering bird. *Front Genet.* 9:392.
- Jarvis ED, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346(6215):1320–1331.
- Kaiser VB, van Tuinen M, Ellegren H. 2007. Insertion events of CR1 retrotransposable elements elucidate the phylogenetic branching order in galliform birds. *Mol Biol Evol.* 24(1):338–347.
- Kapusta A, Suh A, Feschotte C. 2017. Dynamics of genome size evolution in birds and mammals. *Proc Natl Acad Sci U S A.* 114(8):E1460–E1469.
- Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30(4):772–780.
- Kohany O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* 7:474.
- Kretschmer R, Furo IdO, et al. 2020. A comprehensive cytogenetic analysis of several members of the family Columbidae (Aves, Columbiformes). *Genes* 11(6):632.
- Kretschmer R, Gunski RJ, et al. 2020. Chromosomal analysis in *Crotophaga ani* (Aves, Cuculiformes) reveals extensive genomic reorganization and an unusual Z-autosome Robertsonian translocation. *Cells* 10(1):4.
- Kriegs JO, et al. 2007. Waves of genomic hitchhikers shed light on the evolution of gamebirds (Aves: Galliformes). *BMC Evol Biol.* 7:190.
- Laine VN, et al. 2016. Evolutionary signals of selection on cognition from the great tit genome and methylome. *Nat Commun.* 7:10474.
- Lee J, Han K, Meyer TJ, Kim H-S, Batzer MA. 2008. Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons. *PLoS One* 3(12):e4047.
- Li J, et al. 2021. A new duck genome reveals conserved and convergently evolved chromosome architectures of birds and mammals. *Gigascience* 10(1):giaa142. doi: 10.1093/gigascience/giaa142.
- Lim JK, Simmons MJ. 1994. Gross chromosome rearrangements mediated by transposable elements in *Drosophila melanogaster*. *Bioessays* 16(4):269–275.
- Luan DD, Korman MH, Jakubczak JL, Eickbush TH. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* 72(4):595–605.
- Manthey JD, Moyle RG, Boissinot S. 2018. Multiple and independent phases of transposable element amplification in the genomes of Piciformes (Woodpeckers and Allies). *Genome Biol Evol.* 10(6):1445–1456.
- Marchler-Bauer A, et al. 2017. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* 45(D1):D200–D203.
- Matzke A, et al. 2012. Retroposon insertion patterns of neoavian birds: strong evidence for an extensive incomplete lineage sorting era. *Mol Biol Evol.* 29(6):1497–1501.
- Mitchell KJ, et al. 2014. Ancient DNA reveals elephant birds and kiwi are sister taxa and clarifies ratite bird evolution. *Science* 344(6186):898–900.
- Nieder A, Wagener L, Rinnert P. 2020. A neural correlate of sensory consciousness in a corvid bird. *Science* 369(6511):1626–1629.
- O'Connor RE, et al. 2018. Chromosome-level assembly reveals extensive rearrangement in saker falcon and budgerigar, but not ostrich, genomes. *Genome Biol.* 19(1):171.
- Oliveros CH, et al. 2019. Earth history and the passerine superradiation. *Proc Natl Acad Sci U S A.* 116(16):7916–7925.
- Organ CL, Shedlock AM, Meade A, Pagel M, Edwards SV. 2007. Origin of avian genome size and structure in non-avian dinosaurs. *Nature* 446(7132):180–184.
- Ottenburghs J, Ydenberg RC, Van Hooft P, Van Wieren SE, Prins HHT. 2015. The Avian Hybrids Project: gathering the scientific literature on avian hybridization. *Ibis* 157(4):892–894.
- Ottenburghs J, Geng K, Suh A, Kutter C. 2021. Genome size reduction and transposon activity impact tRNA gene diversity while ensuring translational stability in birds. *Genome Biol Evol.* 13:evab016.

- Paton TA, Baker AJ, Groth JG, Barrowclough GF. 2003. RAG-1 sequences resolve phylogenetic relationships within Charadriiform birds. *Mol Phylogenet Evol.* 29(2):268–278.
- Peona V, et al. 2021. Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Mol Ecol Resour.* 21(1):263–286.
- Petkov CI, Jarvis ED. 2012. Birds, primates, and spoken language origins: behavioral phenotypes and neurobiological substrates. *Front Evol Neurosci.* 4:12.
- Pfenning AR, et al. 2014. Convergent transcriptional specializations in the brains of humans and song-learning birds. *Science* 346(6215):1256846.
- Pinheiro MLS, et al. 2021. Chromosomal painting of the sandpiper (*Actitis macularius*) detects several fissions for the Scolopacidae family (Charadriiformes). *BMC Ecol Evol.* 21(1):8.
- Platt RN, 2nd Blanco-Berdugo L, Ray DA. 2016. Accurate transposable element annotation is vital when analyzing new genome assemblies. *Genome Biol Evol.* 8(2):289–295.
- Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* 5(3):e9490.
- Rhie A, et al. 2020. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 592:737–746.
- Ricci M, Peona V, Guichard E, Taccioli C, Boattini A. 2018. Transposable elements activity is positively related to rate of speciation in mammals. *J Mol Evol.* 86(5):303–310.
- Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4:e2584.
- Shetty S, Griffin DK, Graves JA. 1999. Comparative painting reveals strong chromosome homology over 80 million years of bird evolution. *Chromosome Res.* 7(4):289–295.
- Silva T, Guzmán A, Urantówka AD, Mackiewicz P. 2017. A new parrot taxon from the Yucatán Peninsula, Mexico-its position within genus Amazona based on morphology and molecular phylogeny. *PeerJ* 5:e3475.
- Skinner BM, Griffin DK. 2012. Intrachromosomal rearrangements in avian genome evolution: evidence for regions prone to breakpoints. *Heredity (Edinb).* 108(1):37–41.
- Smit A. 2004. Repeat-Masker Open-3.0. Available from: <https://ci.nii.ac.jp/naid/10029514778/> ><http://www.repeatmasker.org>. <https://ci.nii.ac.jp/naid/10029514778/> Date accessed October 18, 2021.
- St John J, Cotter J-P, Quinn TW. 2005. A recent chicken repeat 1 retrotransposition confirms the Coscoroba-Cape Barren goose clade. *Mol Phylogenet Evol.* 37(1):83–90.
- Suh A. 2015. The specific requirements for CR1 retrotransposition explain the scarcity of retrogenes in birds. *J Mol Evol.* 81(1–2):18–20.
- Suh A. 2016. The phylogenomic forest of bird trees contains a hard polytomy at the root of Neoaves. *Zool Scr.* 45(S1):50–62.
- Suh A, et al. 2011. Mesozoic retrotransposons reveal parrots as the closest living relatives of passerine birds. *Nat Commun.* 2:443.
- Suh A, et al. 2014. Multiple lineages of ancient CR1 retrotransposons shaped the early genome evolution of amniotes. *Genome Biol Evol.* 7(1):205–217.
- Suh A, et al. 2017. De-novo emergence of SINE retrotransposons during the early evolution of passerine birds. *Mob DNA.* 8:21.
- Suh A, Kriegs JO, Donnellan S, Brosius J, Schmitz J. 2012. A universal method for the study of CR1 retrotransposons in nonmodel bird genomes. *Mol Biol Evol.* 29(10):2899–2903.
- Suh A, Smeds L, Ellegren H. 2015. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. *PLoS Biol.* 13(8):e1002224.
- Suh A, Smeds L, Ellegren H. 2018. Abundant recent activity of retrovirus-like retrotransposons within and among flycatcher species implies a rich source of structural variation in songbird genomes. *Mol Ecol.* 27(1):99–111.
- Sun Z, et al. 2017. Rapid and recent diversification patterns in Anseriformes birds: inferred from molecular phylogeny and diversification analyses. *PLoS One* 12(9):e0184529.
- Szitenberg A, et al. 2016. Genetic drift, not life history or RNAi, determine long-term evolution of transposable elements. *Genome Biol Evol.* 8(9):2964–2978.
- Treangen TJ, Salzberg SL. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet.* 13(1):36–46.
- Treplin S, Tiedemann R. 2007. Specific chicken repeat 1 (CR1) retrotransposon insertion suggests phylogenetic affinity of rockfowls (genus *Picathartes*) to crows and ravens (Corvidae). *Mol Phylogenet Evol.* 43(1):328–337.
- Underwood CJ, Choi K. 2019. Heterogeneous transposable elements as silencers, enhancers and targets of meiotic recombination. *Chromosoma* 128(3):279–296.
- Vandergon TL, Reitman M. 1994. Evolution of chicken repeat 1 (CR1) elements: evidence for ancient subfamilies and multiple progenitors. *Mol Biol Evol.* 11(6):886–898.
- Wang D, et al. 2017. Transposable elements (TEs) contribute to stress-related long intergenic noncoding RNAs in plants. *Plant J.* 90(1):133–146.
- Warren IA, et al. 2015. Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates. *Chromosome Res.* 23(3):505–531.
- Warren WC, et al. 2010. The genome of a songbird. *Nature* 464(7289):757–762.
- Watanabe M, et al. 2006. The rise and fall of the CR1 subfamily in the lineage leading to penguins. *Gene* 365:57–66.
- Weir JT, Haddrath O, Robertson HA, Colbourne RM, Baker AJ. 2016. Explosive ice age diversification of kiwi. *Proc Natl Acad Sci U S A.* 113(38):E5580–E55807.
- Weissensteiner MH, et al. 2020. Discovery and population genomics of structural variation in a songbird genus. *Nat Commun.* 11(1):3403.
- Wicker T, et al. 2005. The repetitive landscape of the chicken genome. *Genome Res.* 15(1):126–136.
- Wiens JJ. 2015. Explaining large-scale patterns of vertebrate diversity. *Biol Lett.* 11(7):20150506.
- Wright NA, Gregory TR, Witt CC. 2014. Metabolic ‘engines’ of flight drive genome size reduction in birds. *Proc R Soc Proc Biol Sci.* 281(1779):20132780.
- Zeng L, Kortschak RD, Raison JM, Bertozzi T, Adelson DL. 2018. Superior ab initio identification, annotation and characterisation of TEs and segmental duplications from genome assemblies. *PLoS One* 13(3):e0193588.
- Zhang G, et al. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* 346(6215):1311–1320.
- Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. *J Comput Biol.* 7(1–2):203–214.
- Zhou Y, Mishra B. 2005. Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling. *Proc Natl Acad Sci U S A.* 102(11):4051–4056.

Associate editor: Adam Eyre-Walker