1

1 **Genome stability is in the eye of the beholder: recent** 2 **retrotransposon activity varies significantly across avian** 3 **diversity**

4

5

6 James D. Galbraith¹, R. Daniel Kortschak¹, Alexander Suh^{2,3,*}, David L. Adelson^{1,*}.

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

8 2)School of Biological Sciences, University of East Anglia, Norwich, UK

9 3)Department of Organismal Biology, Evolutionary Biology Centre (EBC), Science for Life

10 Laboratory, Uppsala University, Uppsala, Sweden

11 *)Corresponding author

12

13 Short title: Variable retrotransposon activity in birds

3

14 **Abstract:**

15 Since the sequencing of the zebra finch genome it has become clear the avian genome, while
 16 largely stable in terms of chromosome number and gene synteny, is more dynamic at an
 17 intrachromosomal level. A multitude of intrachromosomal rearrangements and significant variation
 18 in transposable element content have been noted across the avian tree. Transposable elements
 19 (TEs) are a source of genome plasticity, because their high similarity enables chromosomal
 20 rearrangements through non-allelic homologous recombination, and they have potential for
 21 exaptation as regulatory and coding sequences. Previous studies have investigated the activity of
 22 the dominant TE in birds, CR1 retrotransposons, either focusing on their expansion within single
 23 orders, or comparing passerines to non-passerines. Here we comprehensively investigate and
 24 compare the activity of CR1 expansion across orders of birds, finding levels of CR1 activity vary
 25 significantly both between and with orders. We describe high levels of TE expansion in genera
 26 which have speciated in the last 10 million years including kiwis, geese and Amazon parrots; low
 27 levels of TE expansion in songbirds across their diversification, and near inactivity of TEs in the
 28 cassowary and emu for millions of years. CR1s have remained active over long periods of time
 29 across most orders of neognaths, with activity at any one time dominated by one or two families of
 30 CR1s. Our findings of higher TE activity in species-rich clades and dominant families of TEs within
 31 lineages mirror past findings in mammals.

32

33 **Author Summary:**

34 Transposable elements (TEs) are mobile, self replicating DNA sequences within a species'
 35 genome, and are ubiquitous sources of mutation. The dominant group of TEs within birds are
 36 chicken repeat 1 (CR1) retrotransposons, making up 7-10% of the typical avian genome. Because
 37 past research has examined the recent inactivity of CR1s within model birds such as the chicken
 38 and the zebra finch, this has fostered an erroneous view that all birds have low or no TE activity on
 39 recent timescales. Our analysis of numerous high quality avian genomes across multiple orders
 40 identified both similarities and significant differences in how CR1s expanded. Our results challenge
 41 the established view that TEs in birds are largely inactive and instead suggest that their variation in
 42 recent activity may contribute to lineage-specific changes in genome structure. Many of the

4

2

5

43 patterns we identify in birds have previously been seen in mammals, highlighting parallels between
44 the evolution of birds and mammals.

46

47 **Introduction:**

48 Following rapid radiation during the Cretaceous-Paleogene transition, birds have diversified to be
49 the most species-rich lineage of extant amniotes (Jarvis et al. 2014; Ericson et al. 2006; Wiens
50 2015). Birds are of particular interest in comparative evolutionary biology because of the
51 convergent evolution of traits seen in mammalian lineages, such as vocal learning in songbirds and
52 parrots (Bradbury and Balsby 2016; Petkov and Jarvis 2012; Pfenning et al. 2014), and potential
53 consciousness in corvids (Nieder et al. 2020). However in comparison to both mammals and non-
54 avian reptiles, birds have much more compact genomes (Gregory et al. 2007). Within birds,
55 smaller genome sizes correlate with higher metabolic rate and the size of flight muscles (Hughes
56 and Hughes 1995; Wright et al. 2014). However, the decrease in avian genome size occurred in an
57 ancestral dinosaur lineage over 200 Mya, well before the evolution of flight (Organ et al. 2007). A
58 large factor in the smaller genome size of birds in comparison to other amniotes is a big reduction
59 in repetitive content (Zhang et al. 2014).

60

61 The majority of transposable elements (TEs) in the chicken (*Gallus gallus*) genome are degraded
62 copies of one superfamily of retrotransposons, chicken repeat 1 (CR1) (International Chicken
63 Genome Sequencing Consortium 2004). The chicken has long been used as the model avian
64 species, and typical avian genomes were believed to have been evolutionarily stable due to little
65 variation in chromosome number and chromosomal painting showing little chromosomal
66 rearrangement (Burt et al. 1999; Shetty et al. 1999). These initial, low resolution comparisons of
67 genome features, combined with the degraded nature of CR1s in the chicken genome, led to the
68 assumption of a stable avian genome both in terms of karyotype and synteny but also in terms of
69 little recent repeat expansion (International Chicken Genome Sequencing Consortium 2004;
70 Wicker et al. 2005). The subsequent sequencing of the zebra finch (*Taeniopygia guttata*) genome
71 supported the concept of a stable avian genome with little repeat expansion, but revealed many

6

7
72 intrachromosomal rearrangements and a significant expansion of endogenous retroviruses (ERVs),
73 a group of long terminal repeat (LTR) retrotransposons, since divergence from the chicken (Warren
74 et al. 2010; Ellegren 2010). The subsequent sequencing of 48 bird genomes by the Avian
75 Phylogenomics Project confirmed CR1s as the dominant TE in all non-passerine birds, with an
76 expansion of ERVs in oscine passerines following their divergence from suboscine passerines
77 (Zhang et al. 2014). The TE content of most avian genomes has remained between 7-10% not
78 because of a lack of expansion, but due to the loss and decay of repeats and intervening non-
79 coding sequence through non-allelic homologous recombination, cancelling out genome size
80 expansion that would have otherwise increased with TE expansion (Kapusta et al. 2017). Since
81 then, hundreds of bird species have been sequenced, revealing variation in karyotypes, and both
82 intrachromosomal and interchromosomal rearrangements (Hooper and Price 2017; Damas et al.
83 2018; Feng et al. 2020; Kretschmer et al. 2020a, 2020b). This massive increase in genome
84 sequencing has similarly revealed TEs to be highly active in various lineages of birds. Within the
85 last 10 million years ERVs have expanded in multiple lineages of songbirds, with the newly
86 inserted retrotransposons acting as a source of structural variation (Suh et al. 2018; Boman et al.
87 2019; Weissensteiner et al. 2020). Recent CR1 expansion events have been noted in
88 woodpeckers and hornbills, leading to strikingly more repetitive genomes than the “typical” 7-10%.
89 Between 23% to 30% of woodpecker and hoopoe genomes are CR1s, however their genome size
90 remains similar to that of other birds (Feng et al. 2020; Manthey et al. 2018; Zhang et al. 2014).
91 While aforementioned research focusing on the chicken suggested CR1s have not recently been
92 active in birds, research focusing on individual avian lineages has used both recent and ancient
93 expansions of CR1 elements to resolve deep nodes in a wide range of orders including early bird
94 phylogeny (Suh et al. 2011; Matzke et al. 2012; Suh et al. 2015), flamingos and grebes (Suh et al.
95 2012), landfowl (Kriegs et al. 2007; Kaiser et al. 2007), waterfowl (St John et al. 2005), penguins
96 (Watanabe et al. 2006), ratites (Haddrath and Baker 2012; Baker et al. 2014; Cloutier et al. 2019)
97 and perching birds (Treplin and Tiedemann 2007; Suh et al. 2017). These studies largely exclude
98 terminal branches and, with the exception of a handful of CR1s in grebes (Suh et al. 2012) and
99 geese (St John et al. 2005), the timing of very recent insertions across multiple species remains
100 unaddressed.

9

101

102 An understanding of TE expansion and evolution is important as they generate genetic novelty by
103 promoting recombination that leads to gene duplication and deletion, reshuffling of genes and
104 major structural changes such as inversions and chromosomal translocations (Zhou and Mishra
105 2005; Bailey et al. 2003; Lim and Simmons 1994; Underwood and Choi 2019; Lee et al. 2008;
106 Chuong et al. 2017). TEs also have the potential for exaptation as regulatory elements and both
107 coding and noncoding sequences (Warren et al. 2015; Wang et al. 2017; Barth et al. 2020). *Ab*
108 *initio* annotation of repeats is necessary to gain a true understanding of genomic repetitive content,
109 especially in non-model species (Platt et al. 2016). Unfortunately, many papers describing avian
110 genomes (Cornetti et al. 2015; Jaiswal et al. 2018; Laine et al. 2016) only carry out homology-
111 based repeat annotation using the Repbase (Bao et al. 2015) library compiled from often distantly
112 related model avian genomes (mainly chicken and zebra finch. This lack of *ab initio* annotation can
113 lead to the erroneous conclusion that TEs are inactive in newly sequenced species (Platt et al.
114 2016). Expectations of low repeat expansion in birds inferred from two model species, along with a
115 lack of comparative TE analysis between lineages is the large knowledge gap we addressed here.
116 As CR1s are the dominant TE lineage in birds, we carried out comparative genomic analyses to
117 investigate their diversity and temporal patterns of activity.

118

119 **Results**

120 *Identifying potential CR1 expansion across birds*

121 From all publicly available avian genomes, we selected 117 representative assemblies not under
122 embargo and with a scaffold N50 above 20,000 bp (available at July 2019) for analysis (SI Table
123 1). To find all CR1s that may have recently expanded in the 117 genomes, we first used the CARP
124 *ab initio* TE annotation tool. From the output of CARP, we manually identified and curated CR1s
125 with the potential for recent expansion based on the presence of protein domains necessary for
126 retrotransposition, homology to previously described CR1s, and the presence of a distinctive 3'
127 structure. To retrotranspose and hence expand, CR1s require endonuclease and reverse
128 transcriptase domains within a single ORF, and a 3' structure containing a hairpin and
129 microsatellite which potentially acts as a recognition site for the reverse transcriptase (Suh et al.

10

5

11

130 2014; Suh 2015). If a CR1 identified from homology contained both protein domains and the

131 distinctive 3' structure, we classified it as a "full length" CR1. We next classified a full length CR1

132 as "intact" CR1 if the endonuclease and reverse transcriptase were within a single intact ORF.

133 Using the full length CR1s and previously described avian and crocodilian CR1s in Repbase as

134 queries (Green et al. 2014; International Chicken Genome Sequencing Consortium 2004; Warren

135 et al. 2010), we performed iterative searches of the 117 genomes to identify divergent, low copy

136 number CR1s which may not have been identified by *ab initio* annotation. We ensured the protein

137 domains and 3' structures were present throughout the iterative searches. Assemblies with lower

138 scaffold N50s generally contained fewer full length CR1s and none in the lowest quartile contained

139 intact CR1s (Figure 1). Outside of the lowest quartile, assembly quality appeared to have little

140 impact on the proportion of intact, full length repeats. The correlation of the low assembly quality

141 with little to no full length CR1s was seen both across all species and within orders.

142

143 Our iterative search identified high numbers of intact CR1s in kiwis, parrots, owls, shorebirds and

144 waterfowl (Figures 1 and 2). Only 2 of the 22 perching bird (Passeriformes) genomes contained

145 intact CR1s, and all contained 10 or fewer full length CR1s. Similarly, of the 7 landfowl

146 (Galliformes) genomes, only the chicken contained intact CR1s and contained fewer than 20 full

147 length CR1s. High numbers of full length and intact repeats were also identified in two

148 woodpeckers, Anna's hummingbird, the chimney swift and the hoatzin, however, due to a lack of

149 other genome sequences from their respective orders, we were unable to perform further

150 comparative within order analyses of these species to look for recent TE expansion, i.e., within the

151 last 10 million years. Of all the lineages we examined, only four have high quality assemblies of

152 genera which have diverged within the last 10 million years and, based on the number of full length

153 CR1s identified, the potential for very recent CR1 expansion: ducks (*Anas*), geese (*Anser*),

154 Amazon parrots (*Amazona*) and kiwis (*Apteryx*) (Silva et al. 2017; Mitchell et al. 2014; Sun et al.

155 2017). While the large number of full length repeats identified in owls is also high, we were unable

156 to examine recent expansion in Strigiformes in detail due to the lack of a dated phylogeny. In

157 addition to our genus scale analyses, we also examined CR1 expansion in parrots (Psittaciformes)

158 overall, perching birds (Passeriformes) and shorebirds (Charadriiformes) since the divergence of

Figure 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.

15

169

170

171

172

173

174

175

Figure 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

16

8

17

176 CR1 copy and those highlighted in yellow are the orders examined in detail. For coordinates of full

177 length CR1s within genomes, see SI Data 1. Tree adapted from (Mitchell et al. 2014; Suh 2016).

178

179 *Order-specific CR1 annotations and a phylogeny of avian CR1s reveal diversity of candidate active*

180 *CR1s in neognaths*

181 In order to perform comparative analyses of activity within orders, we created order-specific CR1

182 libraries. Instead of consensus sequences, all full length CR1s identified within an order were

183 clustered and the centroids of the clusters were used as cluster representatives for that avian

184 order. To classify the order-specific centroids, we constructed a CR1 phylogeny from the centroids

185 and full length avian and crocodilian CR1s from Repbase (Figure 3, SI Figure 1, SI Data 2). From

186 this tree, we partitioned CR1s into families to determine if groups of elements have been active in

187 species concurrently. We partitioned the tree by eye based on the phylogenetic position of

188 previously described CR1 families (Vandergon and Reitman 1994; Wicker et al. 2005; Warren et

189 al. 2010; Bao et al. 2015) and long branch lengths rather than a cutoff for divergence, attempting to

190 find the largest monophyletic groups containing as few previously defined CR1 families as

191 possible. We took this “lumping” approach to our classification to avoid paraphyly and excessive

192 splitting, resulting in some previously defined families being grouped together in one family (SI

193 Table 2). For example, all full length CR1s identified in songbirds were highly similar to the

194 previously described CR1-K and CR1-L families and were nested deeply within the larger CR1-J

195 family. As a result, CR1-K, CR1-L and all full length songbird CR1s were reclassified as

196 subfamilies of the larger CR1-J family. Based on the position of well resolved, deep nodes and

197 previously described CR1s in the phylogeny, we defined 7 families of avian CR1s, with a new

198 family, CR1-W, which was restricted to shorebirds. Interestingly, the 3' microsatellite of the CR1-W

199 family is a 10-mer rather than the octamer found in nearly all amniote CR1s (Suh 2015). With the

200 exception of Palaeognathae (ratites and tinamous), all avian orders that contained large numbers

201 of full length CR1s also contained full length CR1s from multiple CR1 families (Figure 3).

202

19

203

204

205

206 Figure 3: Collapsed tree of full length CR1s and presence of full length copies of CR1 families in
 207 selected avian orders. The name of each family is taken from a previously described CR1 present
 208 within the family (SI Table 3). The colouring of squares indicates the presence of full length CR1s
 209 within the order. All orders shown were chosen due to the presence of high numbers of intact CR1
 210 elements, except for Casuariiformes which are shown due to their recent divergence from
 211 Apterygiformes as well as Passeriformes due to their species richness and frequent use as model
 212 species (especially zebra finch). The full CR1 tree was constructed using FastTree from a MAFFT
 213 alignment of the nucleotide sequences. For the full tree and nucleotide alignment of 1278 CR1s
 214 see SI Figure 1 and SI Data 2.

215

216 *Variable timing of expansion events across avian orders*

217 We used the aforementioned order-specific centroid CR1s and avian and crocodilian Repbase
 218 sequences to create order-specific libraries. Throughout the following analysis we ensured CR1
 219 copies identified were 3' anchored, i.e. retain 3' ends with homology to both the hairpin sequences

20

10

21
220 and microsatellites. We used the order-specific libraries in reciprocal searches to identify and
221 classify 3' anchored CR1s present within all orders in which we had identified full length repeats.
222 Using the classified CR1s we searched for all 3' anchored CR1s (both full length and truncated)
223 and constructed divergence plots to gain a basic understanding of CR1 expansions within each
224 genome (SI Data 3). At high Jukes-Cantor distances, divergence profiles in each order show little
225 difference between species. However, at lower Jukes-Cantor distances divergence, profiles differ
226 significantly between species in some orders. For example, in songbirds at Jukes-Cantor distances
227 higher than 0.1 the overall shape of the divergence plot curves and the proportions of the various
228 CR1 families are nearly identical, while at distances lower than 0.1 higher numbers of the CR1-J
229 family are present in some passerines than others (SI Figure 2a). CR1s most similar to all defined
230 families were present in all orders of Galloanserae and Neoaves examined, with the exception of
231 CR1-X which was restricted to Charadriiformes. Almost all CR1s identified in Palaeognathae
232 genomes were most similar to CR1-Y with a small number of truncated and divergent repeats most
233 similar to crocodilian CR1s (SI Data 3).
234
235 Divergence plots may not accurately indicate the timing of repeat insertions as they assume
236 uniform substitution rates across the non-coding portion of the genome. High divergence could be
237 a consequence of either full length CR1s being absent in a genome or the centroid identified by the
238 clustering algorithm being distant from the CR1s present in a genome. To better determine when
239 CR1 families expanded in avian genomes, we first identified regions orthologous to CR1 insertions
240 sized 100-600 bp in related species (see Methods). We compared these orthologous regions and
241 approximated the timing of insertion based on the presence or absence of the CR1 insertion in the
242 other species. In most orders only long term trends could be estimated due to long branch lengths
243 (cf. Figure 2) and high variability of the quality of genome assemblies (cf. Figure 1). Therefore, we
244 focused our presence/absence analyses to reconstruct the timing of CR1 insertions in parrots,
245 waterfowl, perching birds, and kiwis (Figure 4). We also applied the method to owls (SI Figure 3)
246 and shorebirds (Figure 5), however due to the lack of order-specific fossil calibrated phylogenies of
247 owls and long branch lengths of shorebirds, we could not determine how recent the CR1
248 expansions were.

23

249

250 Figure 4: Presence/absence patterns reconstruct the timing of expansions of dominant CR1

251 families within five selected avian orders. The number next to the coloured circle is the number of

252 CR1 insertions found. Only CR1 families with more than 10 CR1 presence/absence patterns (only

24

12

25

253 CR1 insertions ranging between 100 and 600 bp were analyzed) are shown, for the complete

254 number of insertions see SI Table 3. Phylogenies adapted from (Mitchell et al., 2014; Oliveros et

255 al., 2019; Silva et al., 2017; Sun et al., 2017).

256

257 In analysing the repeat expansion in the kiwi genomes, we used the closest living relatives, the

258 cassowary and emu (Casuariiformes), as outgroups. Following the divergence of kiwis from

259 Casuariiformes, CR1-Y elements expanded, both before and during the recent speciation of kiwis

260 over the last few My. In contrast, there was little CR1 expansion in Casuariiformes, both following

261 their divergence from kiwis, and more recently since their divergence ~28 Mya, with only 1

262 insertion found in the emu and 3 in the cassowary since they diverged (SI Table 3).

263

264 In the waterfowl species examined, both CR1-J and CR1-X families expanded greatly in both

265 ducks and geese during the last 2 million years. Expansion occurred in both examined genera, with

266 greater expansions in the ducks (*Anas*) than the geese (*Anser*). Other CR1 families appear to have

267 been active following the two groups' divergence ~30 Mya, but have not been active since each

268 genus speciated.

269

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

271 genomes from major groups *sensu* (Oliveros et al. 2019); New Zealand wrens (*Acanthisitta*

272 *chloris*), Suboscines (*Manacus vitellinus*), Corvides (*Corvus brachyrhynchos*), and Muscicapida

273 (*Sturnus vulgaris*), Sylvida (*Phylloscopus trochilus* and *Zosterops lateralis*) and Passerida

274 (*Taeniopygia guttata*, *Sporophila hypoxantha* and *Zonotrichia albicollis*). Between the divergence

275 of Oscines (songbirds) and Suboscines from New Zealand wrens and the divergence of Oscines,

276 there was a large spike in expansion of multiple families of CR1s, predominantly CR1-X. Since

277 their divergence 30 Mya, only CR1-J remained active in oscines, though the degree of expansion

278 varied between groups.

279

280 Of all avian orders examined, we found the highest levels of CR1 expansion in parrots. Because

281 most branch lengths on the species tree were long, the timing of recent expansions could only be

27

282 reconstructed in genus *Amazona*. The species from *Amazona* diverged 5 Mya ago and seem to
 283 vary significantly in their level of CR1 expansion. However, genome assembly quality might be a
 284 confounder as the number of insertions into a species of *Amazona* was highest in the best quality
 285 genome (*Amazona collaria*), and lowest in the worst quality genome (*Amazona vittata*). In all
 286 parrots, CR1-E was the predominant expanding CR1 family, however CR1-Y expanded in the
 287 *Melopsittacus-Psittacula* lineage, while remaining largely inactive in the other parrot lineages.

288

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

295

29

296 Multiple expansions of multiple families of CR1s have occurred in the two shorebird lineages

297 examined; plovers (Charadriidae) and sandpipers (Scolopacidae) (Figure 5). The diversity of CR1

298 families that remained active through time was higher than in the other orders investigated,

299 particularly in sandpipers, with four CR1 families showing significant expansion in *Calidris pugnax*

300 and five in *Calidris pygmaea*, since their divergence. In all other orders examined in detail, CR1

301 expansions over similar time periods have been dominated by only one or two families, with

302 insertions of fewer than 10 CR1s from non-dominant families (SI Table 2). Unfortunately, due to

303 long branch lengths more precise timing of these expansions is not possible.

304

305 Finally, CR1s continuously expanded in true owls since divergence from barn owls, with almost all

306 resolved insertions being CR1-E-like (SI Figure 3). However, due to the lack of a genus-level timed

307 phylogeny, the precise timing of these expansions cannot be determined.

308 Combined, our CR1 presence/absence analyses demonstrate that the various CR1 families have

309 expanded at different rates both within and across avian orders. These differences are

310 considerable, ranging from an apparent absence of CR1 expansion in the emu and cassowary to

311 slow, continued expansion of a single CR1 family in songbirds, to recent rapid expansions of one

312 or two CR1 families in kiwis, Amazon parrots and waterfowl, as well as a wide variety of CR1

313 families expanding concurrently in sandpipers.

31

314

315 Discussion

316 *Genome assembly quality impacts repeat identification*

317 The quality of a genome assembly has a large impact on the number of CR1s identified within it,
 318 both full length and 5'-truncated. This is made clear when comparing the number of insertions
 319 identified within species in recently diverged genera. The three *Amazona* parrot species diverged
 320 approximately ~2 Mya (Silva et al. 2017) and the scaffold N50s of *A. vittata*, *A. aestiva* and *A.*
 321 *collaria* are 0.18, 1.3 and 13 Mbp respectively. No full length CR1s were identified in *A. vittata*, and
 322 only 10 in *A. aestiva*, while 1125 were identified in *A. collaria*. Similarly, in *Amazona* the total
 323 number of truncated insertions identified increased significantly with higher scaffold N50s. In
 324 contrast the three species of kiwi compared, diverged ~7 Mya and have similar N50s (between 1.3
 325 and 1.7 Mbp). This pattern of higher quality genome assemblies leading to higher numbers of both
 326 full length and intact CR1s being identified is consistent across most orders examined, and is
 327 particularly true of the lowest N50 quartile (Figure 1). The lower number of repeats identified in
 328 lower quality assemblies is likely due to the sequencing technology used. Repeats are notoriously
 329 hard to assemble and are often collapsed, particularly when using short read Illumina sequencing,
 330 leading to fragmented assemblies (Alkan et al. 2011; Treangen and Salzberg 2011). The majority
 331 of the genomes we have used are of this data type. The recent sequencing of avian genomes
 332 using multiplatform approaches have resolved gaps present in short read assemblies, finding these
 333 gaps to be rich in interspersed, simple and tandem repeats (Peona et al. 2021; Li et al. 2021). Of
 334 particular note (Li et al. 2021) resolved gaps in the assembly of *Anas platyrhynchos* which we
 335 analyzed here using long read sequencing, and found the gaps to be dominated by the two CR1
 336 families that have recently expanded in waterfowl (Anseriformes): CR1-J and CR1-X. Species with
 337 low quality assemblies may have full length repeats present in their genome, yet the sequencing
 338 technology used prevents the assembly of the repeats and hence detection. Thus TE activity may
 339 be even more widespread in birds than we estimate here.

340

341 *The origin and evolution of avian CR1s*

33

342 Avian CR1s are monophyletic in regards to other major CR1 lineages found in amniotes (Suh et al.

343 2014). For comparison, crocodilians contain some CR1 families more similar to those found in

344 testudines and squamates than others in crocodilians. By searching for truncated copies of

345 previously described CR1s in addition to our order-specific CR1s, we were able to uncover how

346 CR1s have evolved in avian genomes as birds have diverged. CR1-Y is the only family with full

347 length CR1s present in Paleognathae, Galloanserae and Neoaves. The omnipresence of CR1-Y

348 indicates it was present in the ancestor of all birds. A small number of highly divergent truncated

349 copies of CR1s most similar to CR1-Z are found in ratites and CR1-J in tinamous (SI Figure 2b).

350 This is potentially indicative of an ancestral presence of CR1-J and CR1-Z in the common ancestor

351 of all birds, or misclassification owing to the high divergence of these CR1 fragments. As

352 mentioned above, we took a lumping approach to classification to CR1 classification to avoid

353 paraphyly, thereby collapsing highly similar families elsewhere considered as separate families. As

354 CR1-C, CR1-E, and CR1-X are present in both Galloanserae and Neoaves but absent from

355 Palaeognathae, we conclude these 4 families likely originated following the divergence of

356 neognaths from paleognaths, but prior to the divergence of Neoaves and Galloanserae. In addition

357 to having a 10 bp microsatellite instead of the typical 8 bp microsatellite, CR1-W is peculiar as it is

358 unique to Charadriiformes but sister to CR1-J and CR1-X (Figure 3). This implies an origin in the

359 neognath ancestor, followed by retention and activity in measurable numbers only in

360 Charadriiformes.

361

362 A wide variety of CR1 families has expanded in all orders of neognaths, with many potential

363 expansion events within the past 10 My present in many lineages. As mentioned in the results, it is

364 not possible to conclude that insertions are ancient based on divergence plots alone. Some

365 species with low quality genome assemblies, such as *A. vittata*, contained very few full length

366 repeats compared to relatives (SI Figure 4). As a result of full length repeats not being assembled,

367 the divergence of most or all truncated insertions identified in *A. vittata* would likely be calculated

368 using CR1 centroids identified in *A. collaria*, leading to higher divergence values than those

369 identified in *A. collaria*, and in turn an incorrect assumption of less recent expansion in *A. vittata*

35

370 than *A. collaria*. In addition to fewer full length repeats being assembled, fewer truncated repeats
371 also appear to have been assembled in poorer quality genomes.

372

373 *CR1 family expansions within orders*

374 Across all sampled neognaths, recent expansions appear to be largely restricted to one or two
375 families of CR1. Our presence/absence analyses found this to be the case in waterfowl, parrots,
376 songbirds and owls, with shorebirds and the early passerine divergences the only exceptions.
377 Similarly, based on the phylogeny of full length elements, most orders only retain full length CR1s
378 from two or three families, while shorebirds retain full length CR1s from across all seven families.
379 Our presence/absence analysis revealed likely concurrent expansions of at least four CR1 families
380 in two families of shorebirds: sandpipers of genus *Calidris* and plovers of genus *Charadrius*. In
381 both genera four families of CR1s have significantly expanded since their divergence including the
382 order-specific CR1-W (Figure 5). While in both genera one family accounts for 40 to 50% of
383 insertions, the other three families have hundreds of insertions each. This is highly different to the
384 pattern seen in songbirds and waterfowl which, over a similar time period, have single digit
385 insertions of non-dominant CR1 families (SI Table 3).

386

387 This increase of CR1 diversity in shorebirds could be due to some CR1 families in shorebirds
388 having 3' inverted repeat and microsatellite motifs which differ from the typical structure (Suh 2015)
389 (SI Fig). For example, the CR1-W family has an extended 10 bp microsatellite (5'-AAATTCYGTG-
390 3') rather than the 8 bp microsatellite (5'-ATTCTRTG-3') seen in nearly all other avian CR1s. When
391 transcribed the 3' structure upstream of the microsatellite is hypothesized to form a stable hairpin
392 which acts as a recognition site for the cis-encoded reverse transcriptase (Suh 2015; Suh et al.
393 2017; Luan et al. 1993). The recently active CR1s we identified in other avian orders have 3'
394 microsatellites and hairpins which closely resemble those previously described. While the changes
395 seen in shorebirds are minor we speculate they could impact CR1 mobilisation, allowing for more
396 families to remain active than the typical one or two.

397

398 *Rates of CR1 expansion can vary significantly within orders*

36

37

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

413

414 All full length CR1s identified in ratites were CR1-Y, and almost all truncated copies found in ratites
415 were most similar to either CR1-Y, or crocodilian CR1s typically not found in birds (Suh et al.
416 2014). This retention of ancient CR1s and the presence of full length CR1s in species such as the
417 southern cassowary (*Casuarus casuarus*) and emu (*Dromaius novaehollandiae*), yet without
418 recent expansion, reflects the much lower substitution and deletion rates in ratites compared to
419 Neoaves (Zhang et al. 2014; Kapusta et al. 2017). These crocodilian-like CR1s in ratites may be
420 truncated copies of CR1s that were active in the common ancestor of crocodilians and birds (Suh
421 et al. 2014) while we hypothesise that these have long since disappeared in Neoaves due to their
422 higher deletion and substitution rates (Kapusta et al. 2017; Zhang et al. 2014).

423

424 *Co-occurrence of CR1 expansion with speciation*

425 The four genera containing recent CR1 expansions we have examined co-occur with rapid
426 speciation events. Of particular note, kiwis rapidly speciated into 5 distinct species composed of at
427 least 16 distinct lineages arising due to significant population bottlenecks caused by Pleistocene

38

39
428 glacial expansions (Weir et al. 2016). We speculate that the smaller population sizes might have
429 allowed for CR1s to expand as a result of increased genetic drift (Szitenberg et al. 2016). While we
430 do not see CR1 expansion occurring alongside speciation in passerines, ERVs, which are rare in
431 other birds, have expanded throughout their diversification (Boman et al. 2019; Warren et al.
432 2010). Investigating the potentially ongoing expansion of CR1s and its relationship to speciation in
433 ducks, geese, and Amazon parrots will require a larger number of genomes from within the same
434 and sister genera to be sequenced, especially in waterfowl due to the high rates of hybridisation
435 even between long diverged species (Ottenburghs et al. 2015).

436

437 *Comparison to mammals*

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

446

447 *Conclusion: the avian genome is more dynamic than meets the eye*

448 While early comparisons of avian genomes were restricted to the chicken and zebra finch, where
449 high level comparisons of synteny and karyotype led to the conclusion that bird genomes were
450 largely stable compared to mammals (Ellegren 2010), the discovery of many intrachromosomal
451 rearrangements across birds (Hooper and Price 2017; Skinner and Griffin 2012; Zhang et al. 2014;
452 Farre et al. 2016) and interchromosomal recombination in falcons, parrots and sandpipers
453 (O'Connor et al. 2018; Coelho et al. 2019; Pinheiro et al. 2021) has shown that at a finer resolution
454 for comparison, the avian genome is rather dynamic. The highly variable rate of TE expansion we
455 have observed across birds extends knowledge from avian orders with “unusual” repeat
456 landscapes, i.e., Piciformes (Manthey et al. 2018) and Passeriformes (Warren et al. 2010), and

41

457 provides further evidence that the genome evolution of bird orders and species within orders differs
458 significantly, even though synteny is often conserved. In our comprehensive characterization of
459 CR1 diversity across 117 bird genome assemblies, we have identified significant variation in CR1
460 expansion rates, both within genera such as *Calidris* and between closely related orders such as
461 kiwis and the cassowary and emu. As the diversity and quality of avian genomes sequenced
462 continues to grow and whole genome alignment methods improve (Feng et al. 2020; Rhie et al.
463 2020), further analysis of genome stability based on repeat expansions at the family and genus
464 level will become possible. While the chicken and zebra finch are useful model species, models do
465 not necessarily represent diversity of evolutionary trajectories in nature.

466

467 **Methods and Materials**

468 *Identification and curation of potentially divergent CR1s*

469 To identify potentially divergent CR1s we processed 117 bird genomes downloaded from Genbank
470 (Benson et al. 2015) with CARP (Zeng et al. 2018); see SI Table for species names and assembly
471 versions. We used RPSTBLASTN (Altschul et al. 1997) with the CDD library (Marchler-Bauer et al.
472 2017) to identify protein domains present in the consensus sequences from CARP. Consensuses
473 which contained both an endonuclease and a reverse transcriptase domain were classified as
474 potential CR1s. Using CENSOR (Kohany et al. 2006) we confirmed these sequences to be CR1s,
475 removing others, more similar to different families of LINEs, such as AviRTes, as necessary.

476

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

485

42

43

486 *Identification of more divergent and low copy CR1s*

487 To identify more divergent or low copy number CR1s which CARP may have failed to identify, we
 488 performed an iterative search of all 117 genomes. Beginning with a library of all avian CR1s in
 489 Repbase (Bao et al. 2015) (see SI Table 2 for CR1 names and species names) and manually
 490 curated CARP sequences we searched the genomes using BLASTN (-task dc-megablast -
 491 max_target_seqs <number of scaffolds in respective genome>), selecting those over 2700 bp and
 492 retaining 3' hairpin and microsatellite sequences. Using RPSTBLASTN we then identified the full
 493 length CR1s (those containing both endonuclease (EN) and reverse transcriptase (RT) domains)
 494 and combined them with the previously generated consensus sequences. We clustered these
 495 combined sequences using VSEARCH 2.7.1 (Rognes et al. 2016) (--cluster_fast --id 0.9) and
 496 combined the cluster centroids with the Repbase CR1s to use as queries for the subsequent
 497 search iteration. This process was repeated until the number of CR1s identified did not increase
 498 compared to the previous round. From the output of the final round, order-specific clusters of
 499 CR1s were constructed and cluster centroids identified.

500

501 *Tree construction*

502 To construct a tree of CR1s, the centroids of all order-specific CR1s were combined with all full
 503 length avian and two crocodilian CR1s from Repbase and globally aligned using MAFFT (--thread
 504 12 --localpair). We used FastTree 2.1.11 with default nucleotide parameters (Price et al. 2010) to
 505 infer a maximum likelihood phylogenetic tree from this alignment, and rooted the tree using the
 506 crocodilian CR1s. The crocodilian CR1s were used as an outgroup as all avian CR1s are nested
 507 within crocodilian CR1s (Suh et al. 2015). This tree was split into different families of CR1 by eye
 508 based on the presence of long branches from high confidence nodes and the position of the
 509 previously described CR1 families from Repbase. To avoid excessive splitting and paraphyly of
 510 previously described families a lumping approach was taken resulting in some previously distinct
 511 families of CR1 from Repbase being treated as members of families they were nested within (SI
 512 Table 3).

513

514 *Identification and classification of CR1s within species*

44

45

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

521

522 *Determination of presence/absence in related species*

523 To reconstruct the timing of CR1 expansions we selected the identified 3' anchored CR1 copies of
524 100 and 600 bp length in a species of interest and at least 600 bp from the end of a contig,
525 extending the coordinates of the sequences by 600 bp to include the flanking region and extracting
526 the corresponding sequences. If the flanking regions contained more than 25% unresolved
527 nucleotides ('N' nucleotides) they were discarded.

528

529 Using BLASTN we identified homologous regions in species belonging to the same order as the
530 species being analysed, and through the following process of elimination identified the regions
531 orthologous to CR1 insertions and their flanks in the related species. At each step of this process
532 of elimination, if an initial query could not be satisfactorily resolved, we classified it as unscorable
533 (unresolved) to reduce the chance of falsely classifying deletions or segmental duplications as new
534 insertion events. First, we classified all hits containing the entire repeat and at least 150 bp of each
535 flank as shared orthologous insertions. Following this, we discarded all hits with outer coordinates
536 less than a set distance (150 bp) from the boundary of the flanks and CR1s to remove hits to
537 paralogous CR1s insertions. This distance was chosen by testing the effect of a range of distances
538 from 300 bp through to 50 bp in increments of 50 bp on a random selection of CR1s first identified
539 in *Anser cygnoides* and *Corvus brachyrhynchos* and searched for in other species within the same
540 order. Requiring outer coordinates to higher values resulted in higher numbers orthologous regions
541 not being resolved, likely due to insertions or deletions within flanks since divergence. Allowing for
542 boundaries of 50 or 100 bp resulted in many CR1s having multiple potential orthologous regions at
543 3' flanks, many of were false hits, only showed homology to the target site duplication and

46

47

544 additional copies of the 3' microsatellite sequence. Thus 150 bp was chosen, as it was the shortest
545 possible distance at which a portion of the flanking sequence was always present.

546

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

559

560 **Acknowledgments**

561 We thank Valentina Peona, Jesper Boman, Julie Blommaert and Alastair Ludington for comments
562 on an earlier version of this manuscript.

563

564 **References**

- 565 Alkan C, Sajjadian S, Eichler EE. 2011. Limitations of next-generation genome sequence
566 assembly. *Nat Methods* **8**: 61–65.
- 567 Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped
568 BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic
569 Acids Res* **25**: 3389–3402.
- 570 Bailey JA, Liu G, Eichler EE. 2003. An Alu transposition model for the origin and expansion of
571 human segmental duplications. *Am J Hum Genet* **73**: 823–834.
- 572 Baker AJ, Haddrath O, McPherson JD, Cloutier A. 2014. Genomic support for a moa-tinamou
573 clade and adaptive morphological convergence in flightless ratites. *Mol Biol Evol* **31**: 1686–
574 1696.
- 575 Baker AJ, Pereira SL, Paton TA. 2007. Phylogenetic relationships and divergence times of

48

49

- 576 Charadriiformes genera: multigene evidence for the Cretaceous origin of at least 14 clades of
577 shorebirds. *Biol Lett* **3**: 205–209.
- 578 Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in
579 eukaryotic genomes. *Mob DNA* **6**: 11.
- 580 Barth JMI, Matschiner M, Robertson BC. 2013. Phylogenetic position and subspecies divergence
581 of the endangered New Zealand Dotterel (*Charadrius obscurus*). *PLoS One* **8**: e78068.
- 582 Barth NKH, Li L, Taher L. 2020. Independent Transposon Exaptation Is a Widespread Mechanism
583 of Redundant Enhancer Evolution in the Mammalian Genome. *Genome Biol Evol* **12**: 1–17.
- 584 Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2015. GenBank. *Nucleic
585 Acids Res* **43**: D30–5.
- 586 Boman J, Frankl-Vilches C, da Silva Dos Santos M, de Oliveira EHC, Gahr M, Suh A. 2019. The
587 Genome of Blue-Capped Cordon-Bleu Uncovers Hidden Diversity of LTR Retrotransposons in
588 Zebra Finch. *Genes* **10**. <http://dx.doi.org/10.3390/genes10040301>.
- 589 Bradbury JW, Balsby TJS. 2016. The functions of vocal learning in parrots. *Behav Ecol Sociobiol*
590 **70**: 293–312.
- 591 Burt DW, Bruley C, Dunn IC, Jones CT, Ramage A, Law AS, Morrice DR, Paton IR, Smith J,
592 Windsor D, et al. 1999. The dynamics of chromosome evolution in birds and mammals. *Nature*
593 **402**: 411–413.
- 594 Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from
595 conflicts to benefits. *Nat Rev Genet* **18**: 71–86.
- 596 Cloutier A, Sackton TB, Grayson P, Clamp M, Baker AJ, Edwards SV. 2019. Whole-Genome
597 Analyses Resolve the Phylogeny of Flightless Birds (Palaeognathae) in the Presence of an
598 Empirical Anomaly Zone. *Syst Biol* **68**: 937–955.
- 599 Coelho LA, Musher LJ, Cracraft J. 2019. A Multireference-Based Whole Genome Assembly for the
600 Obligate Ant-Following Antbird, *Rhegmatorhina melanosticta* (Thamnophilidae). *Diversity* **11**:
601 144.
- 602 Cornetti L, Valente LM, Dunning LT. 2015. The genome of the “great speciator” provides insights
603 into bird diversification. *Genome Biol*.
604 <https://academic.oup.com/gbe/article-abstract/7/9/2680/592400>.
- 605 Damas J, Kim J, Farré M, Griffin DK, Larkin DM. 2018. Reconstruction of avian ancestral
606 karyotypes reveals differences in the evolutionary history of macro- and microchromosomes.
607 *Genome Biol* **19**: 155.
- 608 Ellegren H. 2010. Evolutionary stasis: the stable chromosomes of birds. *Trends Ecol Evol* **25**: 283–
609 291.
- 610 Ericson PGP, Anderson CL, Britton T, Elzanowski A, Johansson US, Källersjö M, Ohlson JI,
611 Parsons TJ, Zuccon D, Mayr G. 2006. Diversification of Neoaves: integration of molecular
612 sequence data and fossils. *Biol Lett* **2**: 543–547.
- 613 Farre M, Narayan J, Slavov GT, Damas J. 2016. Novel insights into chromosome evolution in
614 birds, archosaurs, and reptiles. *Genome Biol*.
615 <https://academic.oup.com/gbe/article-abstract/8/8/2442/2198198>.
- 616 Feng S, Stiller J, Deng Y, Armstrong J, Fang Q, Reeve AH, Xie D, Chen G, Guo C, Faircloth BC, et
617 al. 2020. Dense sampling of bird diversity increases power of comparative genomics. *Nature*
618 **587**: 252–257.

50

25

51

- 619 Galbraith JD, Ludington AJ, Suh A. 2020. New Environment, New Invaders—Repeated Horizontal
620 Transfer of LINEs to Sea Snakes. *Genome Biol.* [https://academic.oup.com/gbe/article-](https://academic.oup.com/gbe/article-abstract/12/12/2370/5918459)
621 [abstract/12/12/2370/5918459](https://academic.oup.com/gbe/article-abstract/12/12/2370/5918459).
- 622 Green RE, Braun EL, Armstrong J, Earl D, Nguyen N, Hickey G, Vandeweghe MW, St John JA,
623 Capella-Gutiérrez S, Castoe TA, et al. 2014. Three crocodilian genomes reveal ancestral
624 patterns of evolution among archosaurs. *Science* **346**: 1254449.
- 625 Gregory TR, Nicol JA, Tamm H, Kullman B, Kullman K, Leitch IJ, Murray BG, Kapraun DF,
626 Greilhuber J, Bennett MD. 2007. Eukaryotic genome size databases. *Nucleic Acids Res* **35**:
627 D332–8.
- 628 Haddrath O, Baker AJ. 2012. Multiple nuclear genes and retroposons support vicariance and
629 dispersal of the palaeognaths, and an Early Cretaceous origin of modern birds. *Proc Biol Sci*
630 **279**: 4617–4625.
- 631 Hooper DM, Price TD. 2017. Chromosomal inversion differences correlate with range overlap in
632 passerine birds. *Nat Ecol Evol* **1**: 1526–1534.
- 633 Hughes AL, Hughes MK. 1995. Small genomes for better flyers. *Nature* **377**: 391.
- 634 International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative
635 analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*
636 **432**: 695–716.
- 637 Ivancevic AM, Kortschak RD, Bertozzi T, Adelson DL. 2016. LINEs between Species: Evolutionary
638 Dynamics of LINE-1 Retrotransposons across the Eukaryotic Tree of Life. *Genome Biol Evol*
639 **8**: 3301–3322.
- 640 Jaiswal SK, Gupta A, Saxena R, Prasoodanan VPK, Sharma AK, Mittal P, Roy A, Shafer ABA,
641 Vijay N, Sharma VK. 2018. Genome Sequence of Peacock Reveals the Peculiar Case of a
642 Glittering Bird. *Front Genet* **9**: 392.
- 643 Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard
644 JT, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern
645 birds. *Science* **346**: 1320–1331.
- 646 Kaiser VB, van Tuinen M, Ellegren H. 2007. Insertion events of CR1 retrotransposable elements
647 elucidate the phylogenetic branching order in galliform birds. *Mol Biol Evol* **24**: 338–347.
- 648 Kapusta A, Suh A, Feschotte C. 2017. Dynamics of genome size evolution in birds and mammals.
649 *Proc Natl Acad Sci U S A* **114**: E1460–E1469.
- 650 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
651 improvements in performance and usability. *Mol Biol Evol* **30**: 772–780.
- 652 Kohany O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of
653 repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* **7**: 474.
- 654 Kretschmer R, Furo I de O, Gomes AJB, Kiazim LG, Gunski RJ, Del Valle Garnero A, Pereira JC,
655 Ferguson-Smith MA, Corrêa de Oliveira EH, Griffin DK, et al. 2020a. A Comprehensive
656 Cytogenetic Analysis of Several Members of the Family Columbidae (Aves, Columbiformes).
657 *Genes* **11**. <http://dx.doi.org/10.3390/genes11060632>.
- 658 Kretschmer R, Gunski RJ, Garnero ADV, de Freitas TRO, Toma GA, Cioffi M de B, Oliveira EHC
659 de, O'Connor RE, Griffin DK. 2020b. Chromosomal Analysis in *Crotophaga ani* (Aves,
660 Cuculiformes) Reveals Extensive Genomic Reorganization and an Unusual Z-Autosome
661 Robertsonian Translocation. *Cells* **10**. <http://dx.doi.org/10.3390/cells10010004>.

52

26

53

- 662 Kriegs JO, Matzke A, Churakov G, Kuritzin A, Mayr G, Brosius J, Schmitz J. 2007. Waves of
663 genomic hitchhikers shed light on the evolution of gamebirds (Aves: Galliformes). *BMC Evol*
664 *Biol* **7**: 190.
- 665 Laine VN, Gossmann TI, Schachtschneider KM, Garroway CJ, Madsen O, Verhoeven KJF, de
666 Jager V, Megens H-J, Warren WC, Minx P, et al. 2016. Evolutionary signals of selection on
667 cognition from the great tit genome and methylome. *Nat Commun* **7**: 10474.
- 668 Lee J, Han K, Meyer TJ, Kim H-S, Batzer MA. 2008. Chromosomal inversions between human and
669 chimpanzee lineages caused by retrotransposons. *PLoS One* **3**: e4047.
- 670 Li J, Zhang J, Liu J, Zhou Y, Cai C, Xu L, Dai X, Feng S, Guo C, Rao J, et al. 2021. A new duck
671 genome reveals conserved and convergently evolved chromosome architectures of birds and
672 mammals. *Gigascience* **10**. <http://dx.doi.org/10.1093/gigascience/giaa142>.
- 673 Lim JK, Simmons MJ. 1994. Gross chromosome rearrangements mediated by transposable
674 elements in *Drosophila melanogaster*. *Bioessays* **16**: 269–275.
- 675 Luan DD, Korman MH, Jakubczak JL, Eickbush TH. 1993. Reverse transcription of R2Bm RNA is
676 primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition.
677 *Cell* **72**: 595–605.
- 678 Manthey JD, Moyle RG, Boissinot S. 2018. Multiple and Independent Phases of Transposable
679 Element Amplification in the Genomes of Piciformes (Woodpeckers and Allies). *Genome Biol*
680 *Evol* **10**: 1445–1456.
- 681 Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, Chitsaz F, Derbyshire MK, Geer RC,
682 Gonzales NR, et al. 2017. CDD/SPARCLE: functional classification of proteins via subfamily
683 domain architectures. *Nucleic Acids Res* **45**: D200–D203.
- 684 Matzke A, Churakov G, Berkes P, Arms EM, Kelsey D, Brosius J, Kriegs JO, Schmitz J. 2012.
685 Retroposon insertion patterns of neoavian birds: strong evidence for an extensive incomplete
686 lineage sorting era. *Mol Biol Evol* **29**: 1497–1501.
- 687 Mitchell KJ, Llamas B, Soubrier J, Rawlence NJ, Worthy TH, Wood J, Lee MSY, Cooper A. 2014.
688 Ancient DNA reveals elephant birds and kiwi are sister taxa and clarifies ratite bird evolution.
689 *Science* **344**: 898–900.
- 690 Nieder A, Wagener L, Rinnert P. 2020. A neural correlate of sensory consciousness in a corvid
691 bird. *Science* **369**: 1626–1629.
- 692 O'Connor RE, Farré M, Joseph S, Damas J, Kiazim L, Jennings R, Bennett S, Slack EA, Allanson
693 E, Larkin DM, et al. 2018. Chromosome-level assembly reveals extensive rearrangement in
694 saker falcon and budgerigar, but not ostrich, genomes. *Genome Biol* **19**: 171.
- 695 Oliveros CH, Field DJ, Ksepka DT, Barker FK, Aleixo A, Andersen MJ, Alström P, Benz BW, Braun
696 EL, Braun MJ, et al. 2019. Earth history and the passerine superradiation. *Proc Natl Acad Sci*
697 *U S A* **116**: 7916–7925.
- 698 Organ CL, Shedlock AM, Meade A, Pagel M, Edwards SV. 2007. Origin of avian genome size and
699 structure in non-avian dinosaurs. *Nature* **446**: 180–184.
- 700 Ottenburghs J, Ydenberg RC, Van Hooft P, Van Wieren SE, Prins HHT. 2015. The Avian Hybrids
701 Project: gathering the scientific literature on avian hybridization. *Ibis* **157**: 892–894.
- 702 Paton TA, Baker AJ, Groth JG, Barrowclough GF. 2003. RAG-1 sequences resolve phylogenetic
703 relationships within Charadriiform birds. *Mol Phylogenet Evol* **29**: 268–278.
- 704 Peona V, Blom MPK, Xu L, Burri R, Sullivan S, Bunikis I, Liachko I, Haryoko T, Jønsson KA, Zhou

54

27

55

- 705 Q, et al. 2021. Identifying the causes and consequences of assembly gaps using a
706 multiplatform genome assembly of a bird-of-paradise. *Mol Ecol Resour* **21**: 263–286.
- 707 Petkov CI, Jarvis ED. 2012. Birds, primates, and spoken language origins: behavioral phenotypes
708 and neurobiological substrates. *Front Evol Neurosci* **4**: 12.
- 709 Pfenning AR, Hara E, Whitney O, Rivas MV, Wang R, Roulhac PL, Howard JT, Wirthlin M, Lovell
710 PV, Ganapathy G, et al. 2014. Convergent transcriptional specializations in the brains of
711 humans and song-learning birds. *Science* **346**: 1256846.
- 712 Pinheiro MLS, Nagamachi CY, Ribas TFA, Diniz CG, O’Brien PCM, Ferguson-Smith MA, Yang F,
713 Pieczarka JC. 2021. Chromosomal painting of the sandpiper (*Actitis macularius*) detects
714 several fissions for the Scolopacidae family (Charadriiformes). *BMC Ecology and Evolution* **21**:
715 8.
- 716 Platt RN 2nd, Blanco-Berdugo L, Ray DA. 2016. Accurate Transposable Element Annotation Is
717 Vital When Analyzing New Genome Assemblies. *Genome Biol Evol* **8**: 403–410.
- 718 Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for
719 large alignments. *PLoS One* **5**: e9490.
- 720 Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W,
721 Fungtammasan A, Gedman GL, et al. 2020. Towards complete and error-free genome
722 assemblies of all vertebrate species. *bioRxiv* 2020.05.22.110833.
723 <https://www.biorxiv.org/content/10.1101/2020.05.22.110833v1.full-text> (Accessed March 31,
724 2021).
- 725 Ricci M, Peona V, Guichard E, Taccioli C, Boattini A. 2018. Transposable Elements Activity is
726 Positively Related to Rate of Speciation in Mammals. *J Mol Evol* **86**: 303–310.
- 727 Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool
728 for metagenomics. *PeerJ* **4**: e2584.
- 729 Salter JF, Oliveros CH, Hosner PA, Manthey JD. 2020. Extensive paraphyly in the typical owl
730 family (Strigidae). *Auk*. <https://academic.oup.com/auk/article-abstract/137/1/ukz070/5673551>.
- 731 Shetty S, Griffin DK, Graves JA. 1999. Comparative painting reveals strong chromosome
732 homology over 80 million years of bird evolution. *Chromosome Res* **7**: 289–295.
- 733 Silva T, Guzmán A, Urantówka AD, Mackiewicz P. 2017. A new parrot taxon from the Yucatán
734 Peninsula, Mexico-its position within genus *Amazona* based on morphology and molecular
735 phylogeny. *PeerJ* **5**: e3475.
- 736 Skinner BM, Griffin DK. 2012. Intrachromosomal rearrangements in avian genome evolution:
737 evidence for regions prone to breakpoints. *Heredity* **108**: 37–41.
- 738 St John J, Cotter J-P, Quinn TW. 2005. A recent chicken repeat 1 retrotransposition confirms the
739 Coscoroba-Cape Barren goose clade. *Mol Phylogenet Evol* **37**: 83–90.
- 740 Suh A. 2016. The phylogenomic forest of bird trees contains a hard polytomy at the root of
741 Neoaves. *Zool Scr* **45**: 50–62.
- 742 Suh A. 2015. The Specific Requirements for CR1 Retrotransposition Explain the Scarcity of
743 Retrogenes in Birds. *J Mol Evol* **81**: 18–20.
- 744 Suh A, Bachg S, Donnellan S, Joseph L, Brosius J, Kriegs JO, Schmitz J. 2017. De-novo
745 emergence of SINE retrotransposons during the early evolution of passerine birds. *Mob DNA* **8**: 21.
- 746 Suh A, Churakov G, Ramakodi MP, Platt RN 2nd, Jurka J, Kojima KK, Caballero J, Smit AF, Vliet

56

28

57

- 747 KA, Hoffmann FG, et al. 2014. Multiple lineages of ancient CR1 retroposons shaped the early
748 genome evolution of amniotes. *Genome Biol Evol* **7**: 205–217.
- 749 Suh A, Kriegs JO, Donnellan S, Brosius J, Schmitz J. 2012. A universal method for the study of
750 CR1 retroposons in nonmodel bird genomes. *Mol Biol Evol* **29**: 2899–2903.
- 751 Suh A, Paus M, Kieffmann M, Churakov G, Franke FA, Brosius J, Kriegs JO, Schmitz J. 2011.
752 Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds. *Nat*
753 *Commun* **2**: 443.
- 754 Suh A, Smeds L, Ellegren H. 2018. Abundant recent activity of retrovirus-like retrotransposons
755 within and among flycatcher species implies a rich source of structural variation in songbird
756 genomes. *Mol Ecol* **27**: 99–111.
- 757 Suh A, Smeds L, Ellegren H. 2015. The Dynamics of Incomplete Lineage Sorting across the
758 Ancient Adaptive Radiation of Neoavian Birds. *PLoS Biol* **13**: e1002224.
- 759 Sun Z, Pan T, Hu C, Sun L, Ding H, Wang H, Zhang C, Jin H, Chang Q, Kan X, et al. 2017. Rapid
760 and recent diversification patterns in Anseriformes birds: Inferred from molecular phylogeny
761 and diversification analyses. *PLoS One* **12**: e0184529.
- 762 Szitenberg A, Cha S, Opperman CH, Bird DM, Blaxter ML, Lunt DH. 2016. Genetic Drift, Not Life
763 History or RNAi, Determine Long-Term Evolution of Transposable Elements. *Genome Biol*
764 *Evol* **8**: 2964–2978.
- 765 Treangen TJ, Salzberg SL. 2011. Repetitive DNA and next-generation sequencing: computational
766 challenges and solutions. *Nat Rev Genet* **13**: 36–46.
- 767 Treplin S, Tiedemann R. 2007. Specific chicken repeat 1 (CR1) retrotransposon insertion suggests
768 phylogenetic affinity of rockfowls (genus *Picathartes*) to crows and ravens (*Corvidae*). *Mol*
769 *Phylogenet Evol* **43**: 328–337.
- 770 Underwood CJ, Choi K. 2019. Heterogeneous transposable elements as silencers, enhancers and
771 targets of meiotic recombination. *Chromosoma* **128**: 279–296.
- 772 Vandergon TL, Reitman M. 1994. Evolution of chicken repeat 1 (CR1) elements: evidence for
773 ancient subfamilies and multiple progenitors. *Mol Biol Evol* **11**: 886–898.
- 774 Wang D, Qu Z, Yang L, Zhang Q, Liu Z-H, Do T, Adelson DL, Wang Z-Y, Searle I, Zhu J-K. 2017.
775 Transposable elements (TEs) contribute to stress-related long intergenic noncoding RNAs in
776 plants. *Plant J* **90**: 133–146.
- 777 Warren IA, Naville M, Chalopin D, Levin P, Berger CS, Galiana D, Volff J-N. 2015. Evolutionary
778 impact of transposable elements on genomic diversity and lineage-specific innovation in
779 vertebrates. *Chromosome Res* **23**: 505–531.
- 780 Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Künstner A, Searle S, White S, Vilella
781 AJ, Fairley S, et al. 2010. The genome of a songbird. *Nature* **464**: 757–762.
- 782 Watanabe M, Nikaido M, Tsuda TT, Inoko H, Mindell DP, Murata K, Okada N. 2006. The rise and
783 fall of the CR1 subfamily in the lineage leading to penguins. *Gene* **365**: 57–66.
- 784 Weir JT, Haddrath O, Robertson HA, Colbourne RM, Baker AJ. 2016. Explosive ice age
785 diversification of kiwi. *Proc Natl Acad Sci U S A* **113**: E5580–7.
- 786 Weissensteiner MH, Bunikis I, Catalán A, Francoijs K-J, Knief U, Heim W, Peona V, Pophaly SD,
787 Sedlazeck FJ, Suh A, et al. 2020. Discovery and population genomics of structural variation in
788 a songbird genus. *Nat Commun* **11**: 3403.

58

29

59

- 789 Wicker T, Robertson JS, Schulze SR, Feltus FA, Magrini V, Morrison JA, Mardis ER, Wilson RK,
790 Peterson DG, Paterson AH, et al. 2005. The repetitive landscape of the chicken genome.
791 *Genome Res* **15**: 126–136.
- 792 Wiens JJ. 2015. Explaining large-scale patterns of vertebrate diversity. *Biol Lett* **11**.
793 <http://dx.doi.org/10.1098/rsbl.2015.0506>.
- 794 Wright NA, Ryan Gregory T, Witt CC. 2014. Metabolic “engines” of flight drive genome size
795 reduction in birds. *Proceedings of the Royal Society B: Biological Sciences* **281**: 20132780.
796 <http://dx.doi.org/10.1098/rspb.2013.2780>.
- 797 Zeng L, Kortschak RD, Raison JM, Bertozzi T, Adelson DL. 2018. Superior ab initio identification,
798 annotation and characterisation of TEs and segmental duplications from genome assemblies.
799 *PLoS One* **13**: e0193588.
- 800 Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW, et
801 al. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation.
802 *Science* **346**: 1311–1320.
- 803 Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences.
804 *J Comput Biol* **7**: 203–214.
- 805 Zhou Y, Mishra B. 2005. Quantifying the mechanisms for segmental duplications in mammalian
806 genomes by statistical analysis and modeling. *Proc Natl Acad Sci U S A* **102**: 4051–4056.

807

808

809

61

810 **SI Information**

811 **Figures**

812 SI Figure 1. Phylogenetic tree of newly identified full length CR1s and full length avian CR1s from
813 Repbase. The full length CR1s used are the centroids of order specific clusters constructed using
814 VSEARCH at 90% identity. Phylogeny constructed using FastTree from a MAFFT alignment of the
815 nucleotide sequences.

816

817 SI Figure 2. Scaled divergence of 3' anchored CR1s identified in a) selected passerines and b)
818 selected paleognaths. CR1s were initially identified using a reciprocal BLAST search based on
819 libraries consisting of RepBase avian and crocodilian repeats and the centroids of full length
820 sequences identified within the order clustered in VSEARCH.

821

822 SI Figure 3. Number of high confidence insertions of dominant CR1 families in owls approximated
823 by presence/absence patterns of orthologous CR1 insertions between 100 and 600 bp in length.
824 CR1 subfamilies are labeled by colour (see legend). Phylogeny adapted from (Salter et al. 2020).

825

826 SI Figure 4. Scaled divergence of 3' anchored CR1s identified in species of Amazon parrot
827 (*Amazona*). CR1s were initially identified using a reciprocal BLAST search based on a consisting
828 of RepBase avian and crocodilian repeats and the centroids of full length sequences identified
829 within parrots clustered in VSEARCH.

830

831 SI Figure 5. Presence/absence workflow. 3' anchored CR1 insertions in a genome between 100
832 and 600 bp (1) were identified with BLASTN and had coordinates extended to include 600 bp of
833 flanking sequence at both the 5' and 3' ends (2). The resulting 1300-1800 bp long sequences were
834 searched for in a related genome using BLASTN. Hits containing the entire insertion and at least
835 150 bp of each flank were treated as ancestral insertions (3). Hits to insertion not containing any
836 flanking region, with hits to the flanking sequence on differing strands or multiple hits to a single
837 flanking sequence far from each other were treated as unresolvable and discarded. Insertions
838 having at least 150 bp of each flank in close proximity and one flank containing at least 90 bp of

62

63

839 the insertion were treated as ancestral insertions of which part was deleted in the species being
840 searched (4). Sequences remaining were either flanks in close proximity or flanks plus a portion of
841 the CR1 insertion. The distance between the flanks potentially containing part of the insertion was
842 calculated in both species, qdist in the query species and sdist in the related species (5). If qdist
843 was greater or equal to the length of the original CR1 insertion (olen) minus the length of 3x the 3'
844 microsatellite monomer and sdist was within the length of 2x the 3' microsatellite monomer the
845 insertion was treated as since divergence (6). If qdist was within the length of 2x the 3'
846 microsatellite monomer and the sdist was greater than 90 bp the insertion was treated as ancestral
847 (7). Any insertions not fitting these criteria were treated as unresolvable and discarded. This strict
848 process was calibrated through adjusting variables and viewing resulting pairwise alignments
849 between regions identified as orthologous, using the presence of target site duplications in the
850 query species and if part of the CR1 insertion was present in the related species to determine if
851 insertions had truly occurred in an orthologous region, erring on the side of discarding new
852 insertions over misclassifying partially deleted ancestral insertions as new insertions.

853

854 SI Figure 6. Presence/absence resolution - Example of the method we used to resolve the
855 presence/absence, and hence insertion timing, of each CR1 in a species (species a), two related
856 species (species b and c) and an outgroup (species d). The CR1 insertion in question is
857 represented in green, the flanking regions in black and the branches labelled 1-3. The branch in
858 bold italics is the branch on which the insertion occurred. If an CR1 was present in species a
859 through c we considered the repeat to have been inserted at branch 1 (i), if in a and b at branch 2
860 (ii) and if in species a alone to be since the divergence from the immediate sister species and on
861 branch 3 (iii). If present in all three species and the outgroup species examined we consider the
862 repeat to be ancestral (iv). If a CR1 was absent from an immediate sister species but present in the
863 more distant related species we considered this to be a result of deletion in the immediate sister
864 species (v). Finally, if the orthologous region was present in a species or group of species but
865 could not be resolved in the immediate sister species we considered the timing of insertion to be
866 unresolvable (vi).

867

64

65

868 **Tables**

869 SI Table 1. Genome assemblies used throughout this analysis. All genomes were downloaded
870 from GenBank.

871

872 SI Table 2 - Reclassification of previously described full length avian CR1s based on their position
873 within our CR1 phylogeny (SI Figure 1; same color coding).

874

875 SI Table 3. Resolution of presence or absence of orthologous CR1 insertions between 100 and
876 600 bp in related species in waterfowl, shorebirds, perching birds, parrots, owls, and kiwis +
877 cassowary + emu genomes. Cells highlighted in yellow are the values used to construct Figures 4
878 and 5 and SI Figure 3.

879

880

881

882 **Data**

883 SI Data 1 - Coordinates of full length CR1s identified in each genome in BED format. For the
884 appropriate genome version see SI Table 1.

885

886 SI Data 2 - Multiple sequence alignment used to create the CR1 phylogeny (SI Figure 1) and
887 Newick tree of said phylogeny.

888

889 SI Data 3. Divergence plots of 3' anchored CR1s identified in each species of bird belonging to
890 orders in which we detected full length CR1s. CR1s were identified using a reciprocal BLAST
891 search based on libraries consisting of Repbase avian and crocodilian repeats and the centroids of
892 full length sequences identified within the order clustered in VSEARCH. Jukes-Cantor distance
893 was calculated from the reciprocal BLAST search output.

*Only Phalacrocoracidae (cormorants)

Charadriiformes

Plovers and kin (Charadriidae)

Sandpipers (Scolopacidae)