Received: 25 July 2017 | Revised: 1 November 2017 | Accepted: 6 November 2017

DOI: 10.1111/mec.14439

ORIGINAL ARTICLE

WILEY MOLECULAR ECOLOGY

Abundant recent activity of retrovirus-like retrotransposons within and among flycatcher species implies a rich source of structural variation in songbird genomes

Alexander Suh | Linnéa Smeds | Hans Ellegren

Department of Evolutionary Biology, Evolutionary Biology Centre (EBC), Uppsala University, Uppsala, Sweden

Correspondence
Alexander Suh, Department of Evolutionary Biology, Evolutionary Biology Centre (EBC), Uppsala University, Uppsala, Sweden.
Email: alexander.suh@ebc.uu.se

Funding information
Vetenskapsrådet; Knut och Alice Wallenbergs Stiftelse

Abstract

Transposable elements (TEs) are genomic parasites capable of inserting virtually anywhere in the host genome, with manifold consequences for gene expression, DNA methylation and genomic stability. Notably, they can contribute to phenotypic variation and hence be associated with, for example, local adaptation and speciation. However, some organisms such as birds have been widely noted for the low densities of TEs in their genomes and this has been attributed to a potential dearth in transposition during their evolution. Here, we show that avian evolution witnessed diverse and abundant transposition on very recent timescales. First, we made an in-depth repeat annotation of the collared flycatcher genome, including identification of 23 new, retrovirus-like LTR retrotransposon families. Then, using whole-genome resequencing data from 200 *Ficedula* flycatchers, we detected 11,888 polymorphic TE insertions (TE presence/absence variations, TEVs) that segregated within and among species. The density of TEVs was one every 1.5–2.5 Mb per individual, with heterozygosities of 0.12–0.16. The majority of TEVs belonged to some 10 different LTR families, most of which are specific to the flycatcher lineage. TEVs were validated by tracing the segregation of hundreds of TEVs across a three-generation pedigree of collared flycatchers and also by their utility as markers recapitulating the phylogenetic relationships among flycatcher species. Our results suggest frequent germline invasions of songbird genomes by novel retroviruses as a rich source of structural variation, which may have had underappreciated phenotypic consequences for the diversification of this species-rich group of birds.

KEYWORDS

bird, insertion polymorphism, pedigree, retrotransposon, retrovirus, structural variation, transposon

1 | INTRODUCTION

Transposable elements (TEs) are genomic parasites, which propagate either via a copy-and-paste (class I elements or retrotransposons) or via a cut-and-paste (most class II elements or DNA transposons) mechanism (Kazazian, 2004; Kidwell, 2005). These selfish genetic elements are present in the genomes of virtually all eukaryotic organisms and may, in addition to having direct effects on genome size by their propagation (Elliott & Gregory, 2015; Kapusta, Suh, & Feschotte, 2017), influence the evolution of transcriptional regulation (Lowe, Bejerano, & Haussler, 2007; Slotkin & Martienssen, 2007; Steige, Laenen, Reimegård, Scofield, & Slotte, 2017), differential DNA methylation (Grandi et al., 2015), 3D genome folding (Schmidt et al., 2012; Wang et al., 2015) and chromosome stability (Farré et al., 2016; Konkel & Batzer, 2010). A growing body of evidence suggests that individual TE insertions may have large phenotypic effects, including both deleterious effects such as disease (reviewed by Hancks & Kazazian, 2012) and beneficial traits involved with adaptation (Guio, Barrón, & González, 2014; van't Hof et al., 2016; Mateo, Ullastres, & González, 2014; Rey, Danchin, Mirouze, Loot, & Blanchet, 2016; Stapley, Santure, & Dennis, 2015). An excellent example is provided by the classroom case of adaptation in form of the industrial melanism mutation identified in the peppered moth (*Biston betularia*), which is a TE insertion (van't Hof et al., 2016). Understanding past and ongoing TE activity is thus an important aspect of elucidating the molecular basis of phenotypic changes.

Birds have the smallest genomes among amniotes (Gregory, 2017), ranging from an estimated 0.9 Gb in the black-chinned hummingbird to 2.1 Gb in the ostrich (Gregory, Andrews, McGuire, & Witt, 2009; Wright, Gregory, & Witt, 2014) among some 758 species for which genome size has been estimated (Gregory, 2017). Their average genome size is 1.33 Gb and thus approximately a third of that of mammals (Hillier et al., 2004; Kapusta & Suh, 2017). Importantly, the diminutiveness of avian genomes coincides with significantly lower densities of transposable elements (TEs) than in other land vertebrates (~10% in birds vs. ~30% nonavian reptiles and ~50% in mammals; Chalopin, Naville, Plard, Galiana, & Volff, 2015; Kapusta & Suh, 2017; Kordiš, 2009; Sotero-Caio, Platt, Suh, & Ray, 2017). Most bird species exhibit relatively high degree of chromosomal synteny (Ellegren, 2010) and stability of chromosome numbers with many species showing $2n \approx 80$ (Ruiz-Herrera, Farré, & Robinson, 2012). A common explanation for this stability is the scarcity of TEs in birds under the assumption that chromosomal rearrangements are related to or promoted by TE activity (Janes, Organ, Fujita, Shedlock, & Edwards, 2010; Shedlock, 2006; Shedlock et al., 2007). Moreover, it has often been suggested that the small size of avian genomes (and thereby indirectly low TE densities) is the result of constraints on cell size (and thus genome size) associated with the metabolic requirements of powered flight (Gregory et al., 2009; Hughes & Hughes, 1995; Vinogradov, 1997; Wright et al., 2014; Zhang & Edwards, 2012).

While both phylogenetic analyses of TE presence/absence in different species (Baker, Haddrath, McPherson, & Cloutier, 2014; Kaiser, van Tuinen, & Ellegren, 2007; Suh, Smeds, & Ellegren, 2015; Suh et al., 2011, 2016) and analyses of TE divergence landscapes (Kapusta & Suh, 2017; Zhang et al., 2014) demonstrate that there was TE activity across the breadth of the avian phylogeny, the low densities of TEs in avian genomes imply that overall TE activity was likely lower than in other amniotes (Chalopin et al., 2015; Kapusta & Suh, 2017; Kapusta et al., 2017; Kordiš, 2009; Sotero-Caio et al., 2017). However, it is not known whether this is also a characteristic of contemporary bird species such that TE insertions would play a minor role in ongoing trait evolution and adaptation of birds. To address this question, there is a need for large-scale analyses of TE presence/absence variation (TEV) in avian genomes to quantify recent TE activity. Previously, only a single TEV has been reported in grebes and chickens, respectively (Lee et al., 2017; Suh, Kriegs, Donnellan, Brosius, & Schmitz, 2012).

The recent availability of massive whole-genome resequencing data paves the way for estimating the amount and character of TEVs segregating within populations. Studies of model organisms including human (Rishishwar, Tellez Villa, & Jordan, 2015; Sudmant et al., 2015), mouse (Nellåker et al., 2012), *Drosophila melanogaster* (Barrón, Fiston-Lavier, Petrov, & González, 2014; Kim et al., 2014), yeast (Jeffares et al., 2015; Liti et al., 2009) and *Arabidopsis thaliana* (Quadrana et al., 2016; Stuart et al., 2016) have shown that rates and diversity of transpositions vary significantly between and within populations. This is likely of significance to evolutionary processes in population and species differentiation.

Here, we investigated the quantity and diversity of recent TE activity in *Ficedula* flycatchers, a well-established system for speciation genomic research (Burri et al., 2015; Ellegren et al., 2012; Nadachowska-Brzyska, Burri, Smeds, & Ellegren, 2016; Nadachowska-Brzyska et al., 2013). This avian system is highly suitable for the inference of TEVs because of an existing chromosome-level genome assembly of the collared flycatcher *Ficedeula albicollis* (Ellegren et al., 2012; Kawakami et al., 2014) and extensive prior genomic information on the phylogenetic relationships among species (Nater, Burri, Kawakami, Smeds, & Ellegren, 2015) and genome-wide levels of nucleotide diversity (Dutoit, Burri, Nater, Mugal, & Ellegren, 2017). The latter is particularly important as it serves as baseline information on the amount of genetic variation present within these species. We in-depth annotated the TE content of the flycatcher genome and scanned whole-genome resequencing data of 200 genomes from 10 populations of four *Ficedula* flycatcher species for TEVs. In addition, we screened a three-generation pedigree of collared flycatchers for TEVs and followed the inheritance of segregating variants. This provides, to our knowledge, one of the first large-scale assessments of TEVs segregating in natural populations of nonmodel species.

2 | MATERIALS AND METHODS

2.1 | Identification of flycatcher-specific repeats

We manually curated a de-novo prediction of repeats in the collared flycatcher genome generated by RepeatModeler (version 1.0.5, Smit & Hubley, 2010) using standard procedures for repeat consensus curation (Lavoie, Platt, Novick, Counterman, & Ray, 2013; Suh, Churakov et al., 2015). This included MAFFT (version 6; Katoh & Toh, 2008) alignment of the 20 best BLASTN (Altschul, Gish, Miller, Myers, & Lipman, 1990) hits (with 2-kb flanks) per repeat candidate and manual curation of majority-rule consensus sequences. Most repeat candidates had been automatically classified as “unknown” by RepeatModeler. We classified nearly all of these as LTR retrotransposons because their curated consensus sequences had canonical 5′-TG…CA-3′ termini (Wicker et al., 2007) and were flanked by target site duplications of 4 bp (endogenous retrovirus 1; ERV1), 6 bp (endogenous retrovirus K or 2; ERVK or ERV2) or 5 bp (endogenous retrovirus L or 3; ERVL or

ERV3; Kapitonov & Jurka, 2008). Finally, we compared the curated consensus sequences (Data S1) to avian repeats from Repbase (Bao, Kojima, & Kohany, 2015; mostly from chicken *Gallus gallus* and zebra finch *Taeniopygia guttata*) in CENSOR (http://www.girinst.org/censor/index.php) and named them following a nomenclature similar to the one used for the zebra finch (Warren et al., 2010). Consensus sequences with sequence similarity to known repeats across their entire length in CENSOR were given the name of the known repeat + suffix "_fAlb," while consensus sequences with partial sequence similarity to known repeats were named with the suffix "-L_fAlb" ("L" denoting "like"). Consensus sequences with no significant sequence similarity to known repeats were considered as belonging to novel repeat families, which were named with the prefix "fAlb," followed by the respective superfamily name (e.g., new families fAlbLTR1, fAlbLTRK1 and fAlbLTRL1; Table S1; Data S2 and S3). The new CR1 subfamilies were classified (Figure S1) by alignment to Repbase CR1 subfamilies from chicken and zebra finch using MAFFT (Data S4), followed by maximum-likelihood phylogenetic analysis using RAxML (version 8.1.11, GTRCAT model, 1,000 bootstrap replicates; Stamatakis, Hoover, & Rougemont, 2008) on the CIPRES Science Gateway (Miller, Pfeiffer, & Schwartz, 2010).

2.2 | Generation of collared flycatcher TE landscapes

We merged the consensus sequences of the finalized repeat curation and classification (Data S1; also deposited in Dfam_consensus, http://dfam-consensus.org/) with all avian repeat consensus sequences present in Repbase (mostly from chicken and zebra finch). This custom library was used for annotation of the male collared flycatcher reference genome fAlb15 (in-house version of FicAlb1.5; Kawakami et al., 2014) and female collared flycatcher W chromosome sequence (Smeds et al., 2015) using RepeatMasker (version 4.0.1; Smit, Hubley, & Green, 1996–2010; parameters -a -xsmall -gccalc). We then generated TE landscapes from the resultant .align files as described elsewhere (Suh, Churakov et al., 2015). The male reference individual was from Sweden (Gotland, Baltic Sea; B), and the female used for W chromosome assembly was from Hungary (H).

2.3 | TEV prediction in 200-flycatcher and pedigree resequencing data

Using .bam files from whole-genome resequencing data (Burri et al., 2015), we predicted TE insertions absent from the reference genome assembly (nonreference TEVs) from discordant read mapping (Figure 1) against the fAlb15 reference using RetroSeq (version 1.41; Keane, Wong, & Adams, 2013). We chose RetroSeq over other available TEV detection programs (reviewed in Rishishwar, Mariño-Ramírez, & Jordan, 2017) because it is, to our knowledge, the most widely used program and was among the best performing in human TEV benchmarking (Rishishwar et al., 2017) after MELT (Gardner et al., 2017) and Mobster (Thung et al., 2014). TEV mapping was carried out for 200 resequenced *Ficedula* flycatchers including 79 collared flycatchers from four populations [Italy (I), Hungary (H), Czech Republic (CZ), and Sweden, Baltic Sea (B)], 79 pied flycatchers *Ficedula hypoleuca* also from four populations [Spain (E), Sweden mainland (S), Czech Republic (CZ), and Sweden, Baltic Sea (B)], 20 Atlas flycatchers *Ficedula speculigera* from Morocco, 20 semicollared flycatchers *Ficedula semitorquata* from Bulgaria, one red-breasted flycatcher *Ficedula parva* from Sweden and one snowy-browed flycatcher *Ficedula hyperythra* from Indonesia. It was also carried out for an additional 11 resequenced collared flycatchers from a three-generation pedigree (Smeds, Mugal, Qvarnström, & Ellegren, 2016). Discordant and split reads were identified by RetroSeq and mapped against our library of collared flycatcher repeat consensus sequences (Data S1). We ran RetroSeq separately for each individual and merged the resulting VCF (Danecek et al., 2011) files of TEV calls into one for the 200-flycatcher data and one for the pedigree data. Following recent observations of the basepair accuracy of insertion site prediction by RetroSeq (Hénaff, Zapata, Casacuberta, & Ossowski, 2015), we considered TE insertions of copies from the same TE family detected in different individuals as orthologous if their breakpoint predictions were within 100 bp of each other.

We employed a series of very strict filters to make sure that only the most confident TEV loci remained for downstream analyses. (1) We excluded TEV loci from resequenced reads that

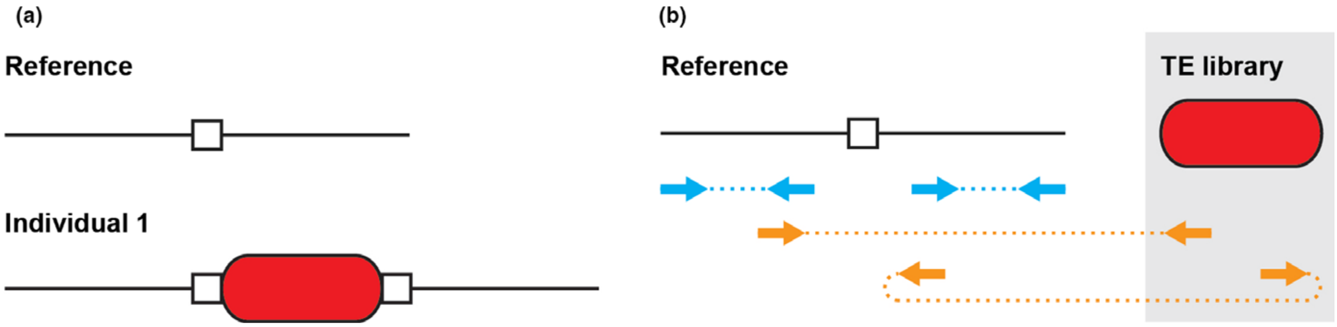

FIGURE 1 Schematic overview of TEV discovery from resequencing data. (a) A situation where the reference genome contains an empty target site (white square; i.e., TEV absence) at a unique genomic locus, and a resequenced individual contains a TE insertion (red) flanked by a duplicated target site (i.e., TEV presence), (b) mapping of resequencing reads against the reference will yield properly mapped read pairs (blue) and discordant read pairs (orange) where one of the reads instead maps to a library of transposable elements consensus sequences [Colour figure can be viewed at wileyonlinelibrary.com]

overlapped with other fAlb15 reference repeats including 200-bp flanks. The 200-bp flanks correspond to two times the read length, which we consider necessary to minimize the number of false TEV calls due to erroneous paired-end read mapping in a short-read assembly. This filter is even more conservative than other studies using RetroSeq (e.g., 100-bp flanks; Lammers, Gallus, Janke, & Nilsson, 2017). For this step, we conducted fAlb15 repeat-masking with the same input library as for the TE landscape analyses (i.e., flycatcher repeats + avian repeats from Repbase) but added a partially curated library of hooded crow repeats (Vijay et al., 2016) to mask as many potential repeat regions as possible; (2) We excluded TEV loci that overlapped with fAlb15 reference tandem repeats. We ran Tandem Repeats Finder (Benson et al., 2013) on a hard-masked version of fAlb15 (RepeatMasker annotation masked as "N" nucleotides) to maximize the masking of tandem repeats; (3) We excluded TEV loci that overlapped with "N" assembly gaps including 200-bp flanks (Lammers et al., 2017) and scaffold ends including 200-bp flanks, as well as scaffolds smaller than 1 kb; (4) We excluded TEV loci that overlapped with regions that remained unmapped in SAMtools mpileup (Li et al., 2009) when mapping resequencing data from the reference individual against the fAlb15 reference assembly; (5) Finally, we excluded TEV loci where none of the TEV calls passed all internal RetroSeq filters (i.e., call status FL = 8; FL < 8 indicates calls which failed a particular filter: "1—depth too high in region, 2—not enough reads in cluster, 3—not enough total flanking reads, 4—not enough inconsistently mapped reads, 5—neither side passes ratio test, 6—one side passes ratio test, 7—distance too large at breakpoint"; Keane et al., 2013). The resultant filtered VCF files contained 11,888 TEV loci in the 200-flycatcher data (Data S5) and 713 TEV loci in the pedigree data (Data S6). The genomic locations of these TEVs were annotated using the collared flycatcher ENSEMBL annotation as intergenic regions, introns, coding sequences and untranslated regions.

2.4 | Comparison of TEV vs. SNP heterozygosities and derived allele frequencies

Given that coverage in the resequencing data ranges from 5 to 27× among the 200 flycatcher genomes (Burri et al., 2015), we considered the coverage too low to allow for reliable homozygous/heterozygous TEV genotype calls in RetroSeq. TEVs were therefore treated as dominant markers, that is, markers where homozygotes and heterozygotes cannot be distinguished, as classically has been the case for data with presence/absence alleles such as amplified fragment length polymorphisms (AFLPs; Excoffier & Heckel, 2006). As there is limited evidence for inbreeding in collared flycatchers (Dutoit et al., 2017), we imputed heterozygosity levels assuming Hardy–Weinberg equilibrium by letting the frequency of absence calls be q^2 in $p^2 + 2pq + q^2 = 1$. Additionally, we imputed heterozygosities using the software AFLP-SURV (Vekemans, Beauwens, Lemaire, & Roldán-Ruiz, 2002), which is based on a Bayesian method with a nonuniform prior distribution (Zhivotovsky, 1999), under Hardy–Weinberg equilibrium or observed levels of inbreeding ($F_{IS} = 0.004$; Dutoit et al., 2017).

3 | RESULTS

3.1 | In-depth TE annotation of the collared flycatcher genome

We conducted an in-depth annotation of repetitive elements in the collared flycatcher reference genome assembly fAlb15 by combining automated de-novo prediction of repetitive elements in RepeatModeler, and manual curation and classification of repeat consensus sequences. The manual curation was particularly important for identifying LTR retrotransposons (or their solo-LTRs) because RepeatModeler automatically classified most of these as "unknown" TEs. Following the general classification systems for TEs (Kapitonov & Jurka, 2008; Wicker et al., 2007), we defined our manually curated TE consensus sequences as TE subfamilies (Table S1). If these subfamilies exhibited sequence similarity to known TEs in CENSOR or to other subfamilies from our annotation, we considered these to belong to the same TE family (Table S1). TE subfamilies or groups of similar TE subfamilies were then defined as novel TE families (prefix "fAlb"; Table S1) if they had no significant sequence similarity to known TE families in CENSOR. We identified a total of six satellite repeat subfamilies (some of these LTR-derived) and 85 TE subfamilies, the latter grouping into five CR1 long interspersed element (LINE) families (see Figure S1 for their phylogenetic relationships), nine endogenous retrovirus 1 (ERV1) long-terminal repeat (LTR) families, 19 ERVK LTR families and 11 ERVL families (Table S1). Among these, five ERV1 LTR families, 14 ERVK LTR families and four ERVL LTR families (i.e., 23 in total) represented previously unknown families (Figure S2).

In total, we identified 193,611 TEs or TE fragments in the collared flycatcher genome assembly using RepeatMasker (Table 1), representing 5.5% of the genome (Table 1). Also including all forms of tandem repeats, the total repeat content of the assembly was 11.7%.

3.2 | Timescale of TE activity in the collared flycatcher genome

To estimate the relative timescales of activity of different repeat families/subfamilies, we analysed the distribution of pairwise Kimura 2-parameter distance between TEs and their respective consensus sequence, also called TE landscapes (Figure 2). For autosomal TEs, we dated their approximate age of insertion using the estimated mutation rate of collared flycatcher of 2.3×10^{-9} mutations per site per year (Smeds, Qvarnstrom, & Ellegren, 2016). The dated TE landscape suggested that the vast majority of detectable TE sequences are ancient, with most TE accumulation (especially CR1 retrotransposons) occurring between 33 and 54 million years ago (MYA). Note that there is an upper limit to how far back in time TE activity can be reliably detected, as at some point mutations will have accumulated such that sequences are beyond recognition as TEs. However, the distribution of pairwise distances also suggested that there has been TE activity all the way until in the very recent past, indicating that TEs may segregate in contemporary populations. In fact, the

TABLE 1 RepeatMasker annotation of the male fAlb15 assembly and the female W chromosome assembly from collared flycatcher using a manually curated collared flycatcher repeat library merged with avian repeats from Repbase

	Autosomes (fAlb15)			Z chromosome (fAlb15)			W chromosome		
Repeat type	Copies	Total bp	Total %	Copies	Total bp	Total %	Copies	Total bp	Total %
SINE	6,722	790,514	0.07	392	45,132	0.07	20	2,037	0.03
LINE	113,197	30,879,733	2.94	8,515	3,782,019	5.50	1,105	567,451	8.09
LTR	47,428	19,986,049	1.90	5,136	2,681,766	3.90	5,621	2,746,777	39.15
DNA	3,121	471,889	0.04	143	23,400	0.03	3	592	0.01
Unclassified	2,112	359,824	0.03	95	15,555	0.02	1	78	0.00
Total interspersed repeats	172,580	52,488,009	4.98	14,281	6,547,872	9.52	6,750	3,316,935	47.28
Small RNA	494	48,005	0.00	29	2,136	0.00	9	956	0.01
Satellites	2,602	1,197,524	0.11	100	13,635	0.02	85	41,985	0.60
Simple repeats	306,888	61,255,371	5.84	20,348	3,589,682	5.22	815	43,628	0.62
Low complexity	47,533	2,728,128	0.26	3,505	197,242	0.29	159	16,382	0.23
Total tandem repeats	357,517	65,229,028	6.21	23,982	3,802,695	5.53	1,068	102,951	1.46
Total repeats	530,097	117,717,037	11.19	38,263	10,350,567	15.05	7,818	3,419,886	48.74
Assembly		1,047,509,136			68,824,682			7,015,564	
Gap ("N") bp		13,281,900			718,894			86,899	

Repeat copy numbers and total basepairs were taken from the RepeatMasker .tbl file where copy numbers are estimated after merging repeat fragments derived from the same insertion event. Note that the collared flycatcher assembly is based on short reads, and repeat copy numbers may be overestimated for young TEs interrupted by assembly gaps.

extent of very recent TE activity is likely to be underestimated from the occurrence of young TEs in genome assemblies due to issues leading to collapsed or unassembled repeats.

Notably, over 90% of autosomal TE sequences presumably younger than 11 MY (Figure 2; that is, <5% distance to consensus) belonged to the group of LTR retrotransposons (mostly from the 23 novel LTR families; Figure S2). CR1 retrotransposons, the "typical" TEs of bird genomes (Hillier et al., 2004; Kapusta & Suh, 2017; Warren et al., 2010; Zhang et al., 2014), only made up <10% of the bp and underwent a decrease in accumulation towards present times; very few CR1 retrotransposons accumulated within the last 5 MY. The TE landscape of the Z sex chromosome was overall similar in shape but with slightly higher TE densities. The female-specific W chromosome had very high densities of retrovirus-like LTR retrotransposons (Figure 2; cf. Smeds et al., 2015). The trend of relatively recent replacement of CR1 activity by LTR activity was also visible on the Z and W sex chromosomes (Figure 2, cf. Smeds et al., 2015).

We then compared the diversity of LTR retrotransposon families among three bird lineages, collared flycatcher, zebra finch (both belonging to Passeriformes, with an estimated divergence time of 21.2 MYA; Moyle et al., 2016) and chicken, by BLASTn searches of the respective species' LTR subfamilies against the other genomes (Data S2 and S3). This showed that among the 23 LTR families newly described from collared flycatcher, 14 families (comprising 33 subfamilies) were specific to the flycatcher lineage, while nine were present in low copy numbers also in the zebra finch (and absent in chicken) but not identified in previous repeat annotations of that species. This corroborates the observation of recent LTR activity in the flycatcher lineage. Reciprocally, 17 families (comprising 175 subfamilies) were specific to zebra finch (Figure 3).

3.3 | Abundance and diversity of TE presence/absence variation (TEV)

Next, we used the in-depth repeat annotation of the fAlb15 assembly as a reference for mapping of whole-genome resequencing data from 200 *Ficedula* flycatcher genomes using RetroSeq (see Materials and Methods). These were from 10 populations of four closely related black-and-white flycatcher species (collared flycatcher, pied flycatcher, semicollared flycatcher and Atlas flycatcher; divergence time <1–2 MYA; Nater et al., 2015), plus single individuals from two outgroup species (red-breasted flycatcher and snowy-browed flycatcher). We focused on identifying TEs present in any of the 200 genomes that were absent in the collared flycatcher genome assembly ("nonreference TEVs"; see Figure 1), potentially representing TE presence/absence variation (TEV). Given the challenge in confidently distinguishing between homozygote and heterozygote TEV calls (see Materials and Methods), it should be noted that "presence" means "presence of one or two allelic copies." We refrained from analysing TEVs representing TE insertions present in the reference genome assembly ("reference TEVs") but absent in individuals from the resequencing data. This was mainly because short-read assemblies (such as collared flycatcher) typically contain many assembly gaps in TE sequences, especially for copies belonging to young TE families, which we assumed problematic for presence/absence mapping. Also, and more generally, demonstrating the presence of a sequence in resequencing data is more straightforward than proving its absence (cf. below, concerning a dependence of coverage on the ability to detect TE insertions).

After very stringent filtering of TEV calls predicted by RetroSeq (Keane et al., 2013), we found a total of 11,888 TEV loci where the

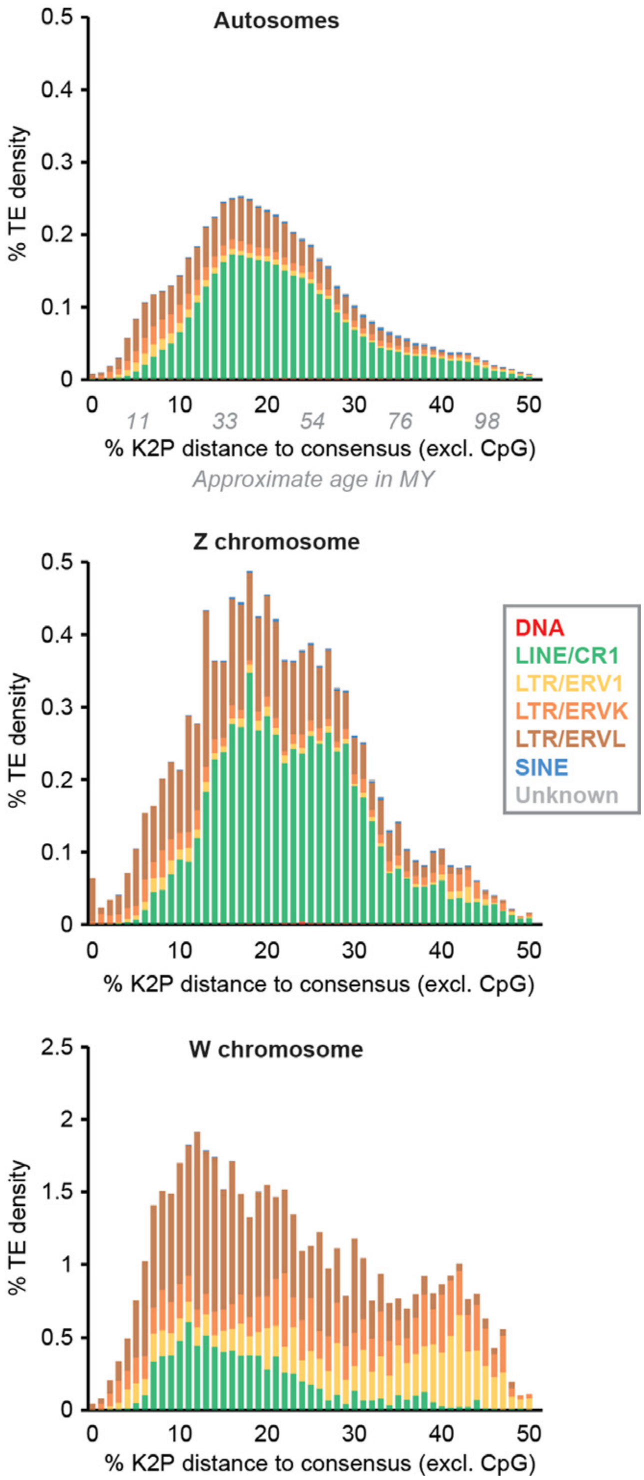

FIGURE 2 Transposable element (TE) landscapes of collared flycatcher autosomes, and the Z and W sex chromosomes. Approximate ages of TE insertions were calculated by dividing the Kimura 2-parameter (K2P) distance by two times the mutation rate of collared flycatcher, namely 2.3 × 10^{-9} mutations per site per year (Smeds, Qvarnstrom et al., 2016). The age distributions suggest dominance of LTR retrotransposon activity within the last 22 MY (i.e., 0-10% K2P distance) [Colour figure can be viewed at wileyonlinelibrary.com]

TE insertion was absent from the collared flycatcher assembly. Of these TEVs, 10,211 segregated within or among the four black-and-white species (see further below; Figure S3, Data S5). Most TEVs belonged to eight different LTR families, notably from all three major groups of endogenous retroviruses ERV1, ERVK and ERVL (Table 2). The finding of >10,000 polymorphic TEs among four closely related flycatcher species provides evidence for significant TE activity in the recent past in this avian lineage and means that TE insertions can potentially have played a role in recent trait evolution. We analysed the distribution TEs in relation to other genomic features and, assuming a random distribution in the genome, found a slight overrepresentation in intergenic regions (ratio of observed number to expected number = 1.12; 95% confidence interval: 1.11–1.13) and a slight underrepresentation in introns (0.82; 0.80–0.83), potentially indicating that TE insertions on average have a somewhat more deleterious effect in introns than in intergenic regions. Moreover, an overall deleterious effect on coding sequence was quite evident from a strong underrepresentation in this sequence category (0.22; 0.18–0.28), and a similar effect was also seen for untranslated regions (0.63; 0.54–0.71; Table S2).

Phylogenetic network analysis (Huson, 1998) of population-pooled TEVs recapitulated the phylogenetic relationships among the four black-and-white *Ficedula* flycatcher species given by analyses of genome-wide sequence data (Nater et al., 2015; Figure S4). These relationships included detection of the collared/pied/Atlas flycatcher clade, the pied/Atlas flycatcher clade and the divergence of Spanish pied flycatchers from the remaining pied flycatcher populations (Nater et al., 2015). Furthermore, the distribution and extent of reticulations within the phylogenetic network were in agreement with the previously noted high prevalence of phylogenetic discordance due to incomplete lineage sorting at the base of this rapid four-species radiation (Nater et al., 2015). This demonstrates that TEVs, at least in these species, behave in a similar way to nucleotide sequence polymorphisms in terms of how they segregate within and between species. However, we note that as our analysis is based on nonreference TEVs, the amount of phylogenetically concordant TEVs may be underestimated for collared flycatchers.

For the species with population samples, we identified between 13 and 630 TEVs per individual. Importantly, both within each of these species and for all species taken together, there was a strong positive relationship between genomic sequence coverage and number of TEVs detected per individual (Figure 4). A near-linear increase in the number of TEVs was apparent across the whole range of sequence coverage, from 2× to 27× (Table S3).

Not surprisingly given the observed relationship between coverage and TEV detection, we found the largest number of TEVs in the population with the highest mean sequencing coverage (Atlas flycatcher, n = 20 individuals, 2,446 TEVs, mean coverage = 20.9×; Table 3) and the lowest numbers in the two populations with the lowest coverage (Spanish collared flycatcher, n = 20 individuals, 1,489 TEVs, mean coverage = 13.0×; Baltic Sea collared flycatcher, n = 19 individuals, 1,561 TEVs, mean coverage = 12.9×). However, as the total number of sampled individuals was much higher for

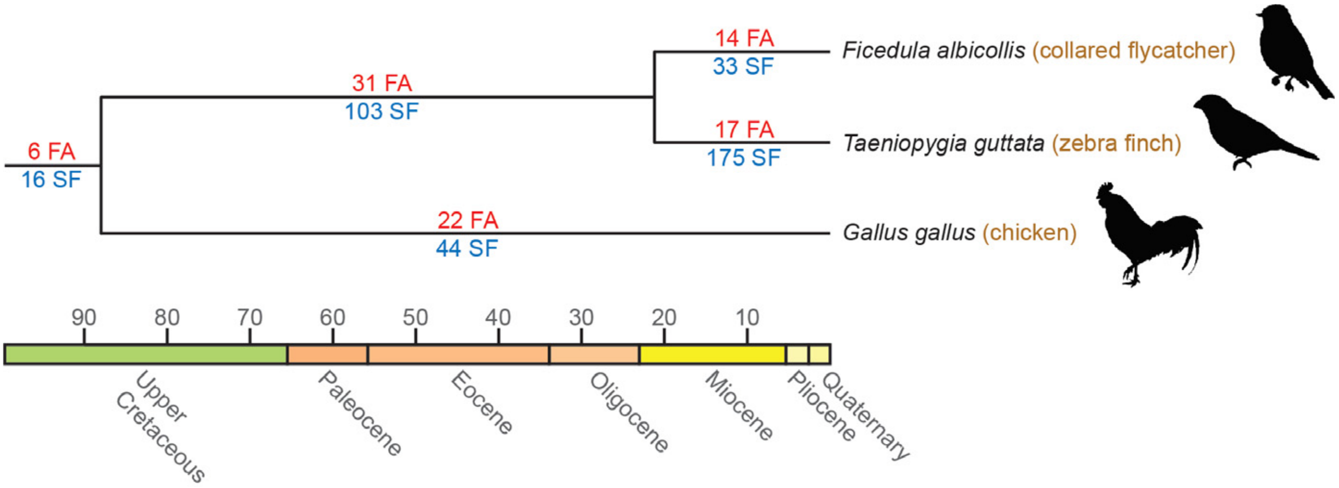

FIGURE 3 Emergence of novel LTR retrotransposon families (FA) and subfamilies (SF) in the three avian genomes with manually curated TE annotations. The divergence estimates are based on timetrees of major avian taxa (Jarvis et al., 2014) and passerines (Moyle et al., 2016). Each LTR family is defined as a group of similar LTR subfamilies with no sequence similarity to LTR subfamilies from a different LTR family. For details on the distribution of LTR families and subfamilies across the three genomes, see Data S2 and S3 [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Number of polymorphic TE insertions (TEVs) from the 10 most abundant TE families in 200 flycatcher genomes and the number of segregating insertions traced across a three-generation pedigree of collared flycatchers

TE family	TE classification	Number of population samples	Number traced across pedigree
TguERVL2a2_LTR	LTR/ERV3	2,993	120
fAlbLTRK12	LTR/ERV2	2,809	86
fAlbLTRK10	LTR/ERV2	2,169	60
TguLTRL1b	LTR/ERV3	1,666	74
TguLTR13c	LTR/ERV1	794	14
fAlbLTRL3	LTR/ERV3	440	12
TguLTRK3d	LTR/ERV2	310	7
fAlbLTRL1	LTR/ERV3	102	2
CR1-E	LINE/CR1	82	18
TguLTRL2a8	LTR/ERV3	63	1

collared flycatcher and pied flycatcher ($n = 79$ for both species) than for Atlas flycatcher and semicollared flycatcher ($n = 20$ for both), the total number of TEVs per species was highest in collared flycatchers (4,161 TEVs, 5.6 TEVs/Mb) and in pied flycatchers (3,806 TEVs, 5.1 TEVs/Mb). There was a high degree of TEVs shared among the sampled ingroup species (Figure 5a), confirming that lineage sorting of ancestral genetic variation is still incomplete in this recent species radiation.

The observed number of TEVs per individual corresponded to an average density of one TEV every 1.5–2.5 Mb in flycatcher genomes (Table 3). Per population sample of 19–20 individuals, the density was one every 0.3–0.5 Mb. To quantify levels of polymorphism per locus, we treated TEVs as dominant markers (i.e., where the presence or absence of an allele is scored similar to, for example, AFLPs; Excoffier & Heckel, 2006) and imputed heterozygosity and derived allele frequency under Hardy–Weinberg equilibrium from the frequency of absence calls for each of the sampled flycatcher populations (Figure S5, Table 4). Mean heterozygosity per population was between 0.12 and 0.16, and the mean derived allele frequency between 0.09 and 0.15 (Table 4). We obtained similar heterozygosity estimates when using a Bayesian method with nonuniform prior distribution (Zhivotovsky, 1999) (Table S4), and this was the case irrespective of assuming Hardy–Weinberg equilibrium or taking inbreeding into account. Indeed, inbreeding levels in flycatcher populations are low ($F_{IS} = 0.004$; Dutoit et al., 2017; Table S4). When we limited diversity estimation to only include individuals with >15× genomic coverage (Table 4), estimates of heterozygosity as well as derived allele frequency were on average slightly higher than when using the full data set. This is not unexpected given the observed relationship between coverage and ability to detect TEVs. The allele frequency spectrum indicated that about half of all TEVs were private, that is, present in only a single individual (Figure S5b). On top of the overall challenges in population genetic analysis of dominant markers, it is difficult to compare these diversity estimates with data from SNPs due to the very different procedures and protocols for genotyping. However, the results were largely in agreement with previous SNP data from the same flycatcher populations, which show similar polymorphism levels and allele frequency spectra (Burri et al., 2015; Dutoit et al., 2017).

We attempted to infer the timescales of retrotransposition events leading to TEVs within and among *Ficedula* flycatchers by parsimony-based mapping of TEVs on the phylogenetic tree of the sampled populations and species. This revealed frequent retrotransposition across initial (representing TEVs shared among two or more species), shallow (TEVs shared among two or more populations within species), as well as terminal (species- or population-specific TEVs) branches of the phylogeny (Figure 5b). Given the rapid divergence of these species, and limited differentiation among populations within species, sequencing

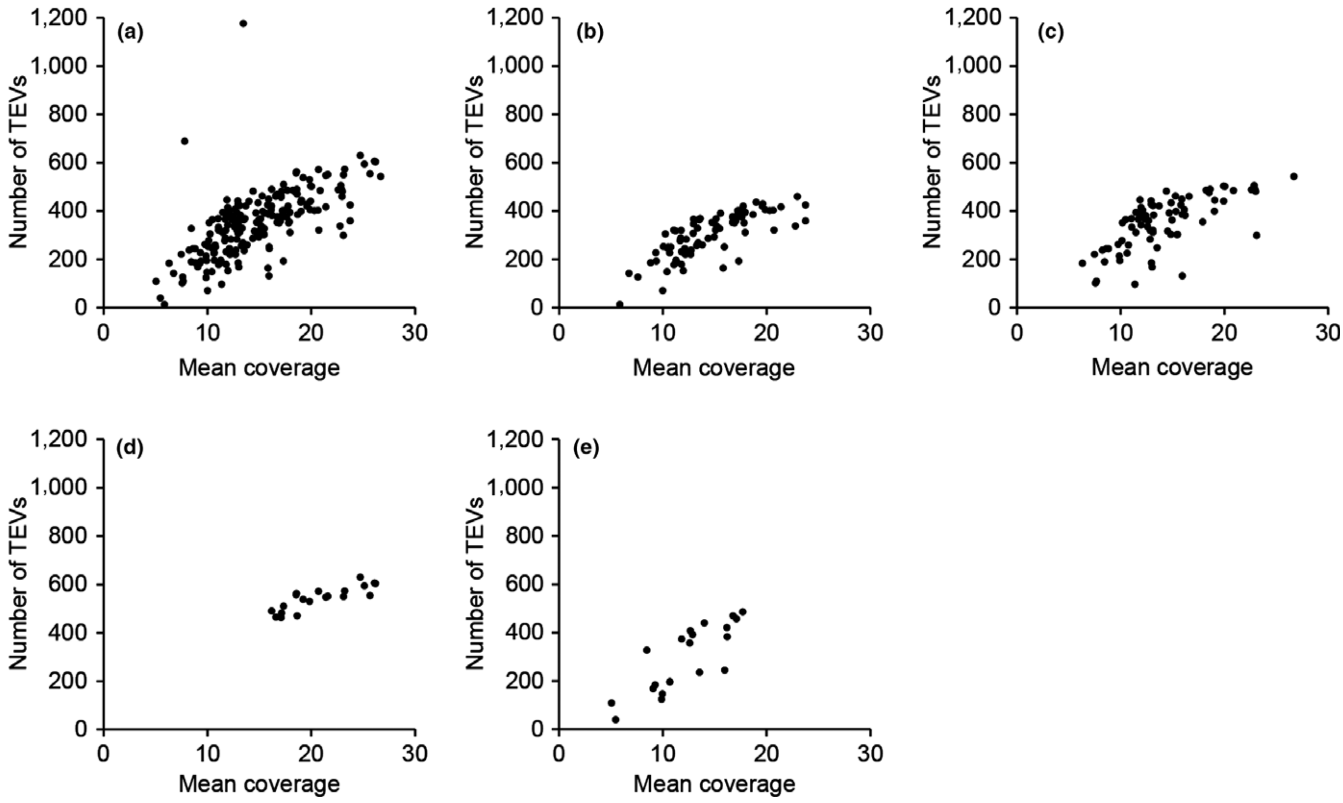

FIGURE 4 Number of TEVs plotted against mean genomic coverage of the 200 resequenced flycatchers. Shown are (a) all 200 sampled individuals, (b) 79 collared flycatchers, (c) 79 pied flycatchers, (d) 20 Atlas flycatchers and (e) 20 semicollared flycatchers. Note the two outliers in panel A, which are the outgroups red-breasted flycatcher (1,176 TEVs, 13.5× mean coverage) and snowy-browed flycatcher (689 TEVs, 7.8× mean coverage)

TABLE 3 Distribution of TEVs across the sampled species and populations

	Collared				Pied								
	H	CZ	I	B	CZ	E	B	S	Atlas	Semicollared	Red-breasted	Snowy-browed	All
Number of individuals	20	20	20	19	20	20	19	20	20	20	1	1	200
Mean coverage	13.4	15.0	13.9	15.8	14.4	13.0	12.9	16.0	20.9	12.3	13.5	7.8	16.1
SD of coverage	3.7	3.5	3.8	4.8	3.4	4.7	3.5	4.6	3.4	3.8	N/A	N/A	4.5
Mean number of TEVs/individual	281	297	299	314	366	304	318	422	542	298	1,176	689	350
SD of number of TEVs/individual	88	113	80	92	88	89	132	74	49	139	N/A	N/A	138
Total number of TEVs/population	1,821	1,885	1,775	1,610	1,610	1,489	1,561	1,881	2,446	1,452	N/A	N/A	11,888
Number of TEVs/Mb/individual	0.4	0.4	0.4	0.4	0.5	0.4	0.4	0.6	0.7	0.4	1.6	0.9	0.5
Number of TEVs/Mb/population	2.4	2.5	2.4	2.2	2.2	2.0	2.1	2.5	3.3	1.9	N/A	N/A	15.9

SD, standard deviation.

The TEV data come from 747 Mb of the genome remaining after filtering (see Material and Methods). Collared flycatcher populations are Italy (I), Hungary (H), Czech Republic (CZ) and Sweden Baltic Sea (B); pied flycatcher populations are Spain (E), Sweden mainland (S), Czech Republic (CZ) and Sweden Baltic Sea (B).

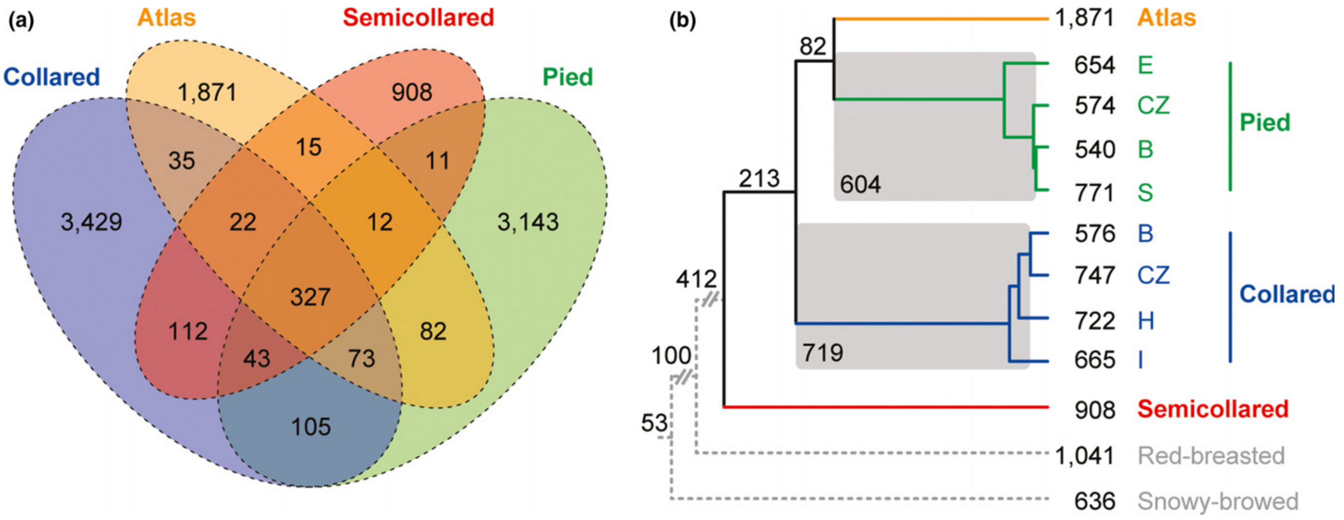

FIGURE 5 Distribution of 11,888 TEVs across *Ficedula* flycatcher species and populations. (a) Venn diagram of species-level distributions of TEVs present only in the ingroup. (b) Number of retrotransposition events per lineage inferred from TEVs parsimoniously mapped on the population-level phylogeny of Burri et al. (2015). Grey dashed lines indicate the phylogenetic position of the two outgroups (cf. Moyle, Hosner, Jones, & Outlaw, 2015). Collared flycatcher populations are Italy (I), Hungary (H), Czech Republic (CZ) and Sweden Baltic Sea (B); pied flycatcher populations are Spain (E), Sweden mainland (S), Czech Republic (CZ) and Sweden Baltic Sea (B) [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 4 Mean heterozygosity (*h*) and derived allele frequency (*daf*) estimates of TEVs from ten populations of *Ficedula* flycatchers using all ingroup individuals (n = 198) or only those with >15× genomic coverage (n = 87; Table S3)

	Collared				Pied					
Population	H	CZ	I	B	CZ	E	B	S	Atlas	Semicollared
h (all)	0.12	0.13	0.13	0.15	0.16	0.15	0.16	0.15	0.14	0.16
	0.06				0.07					
h (>15x)	0.10				0.13				0.14	0.27
daf (all)	0.09	0.09	0.10	0.12	0.14	0.13	0.12	0.15	0.15	0.12
	0.04				0.06					
daf (>15x)	0.08				0.12				0.15	0.27

of additional samples would likely move some insertion events further back on the phylogeny. The highest number of insertion events (1,871) mapped to the Atlas flycatcher lineage, consistent with the highest number of TEVs seen in this species. This was mostly due to higher numbers of TEVs from two LTR families, namely TguERV-L2a2_LTR and TguLTRL1b (Data S5).

3.4 | Verification of TEVs via a three-generation pedigree

We finally sought to validate TEVs and confirm stable Mendelian inheritance by tracing the segregation of TEVs across a three-generation pedigree of collared flycatchers (sequenced to 36–45× coverage; Smeds, Mugal et al., 2016) consisting of paternal and maternal grandparents (i.e., four individuals in the P generation), mother and father (two F_1 individuals) and five full-siblings (F_2 individuals; Figure S6a). We identified 713 nonreference TEVs in the pedigree (Data S6) and most of these belonged to the eight most common TEV families identified in the 200-genomes data set (Table 2). For most of the 96% (686) pedigree-concordant TEVs, the observed pattern was either consistent with stable inheritance across all three generations of the pedigree and in concordance with it (417 TEVs, Table S5), or constituted insertions present in the P generation that were not transmitted to the F_1 generation (248 TEVs; this is not unexpected if they represent P heterozygotes). Moreover, to directly be able to follow the inheritance of heterozygous loci, we focused on TEVs present in one P and one F_1 individual, as these markers should be very strong candidates for heterozygosity in the F_1 generation. We then followed their inheritance to the F_2 generation and estimated the binomial probabilities of inheriting TEV presence in zero to five F_2 individuals. The observed and expected distributions were very similar (Table S6). Overall, this suggests a low rate of false positives in the pipeline used for TEV detection and filtering in this study.

Furthermore, the pedigree information permitted us to estimate a minimum false-negative rate from those TEVs present in one or more individuals in a fashion that disagreed with the pedigree. We found 27 TEVs where a missing TEV presence call would explain the observed TEV call distribution across the pedigree, which would translate into a minimum rate of 3.8% false negatives (Table S5, Figure S6). Another 14 TEVs were either false negatives (in which case the overall false negative would be 5.8%) or constituted de-novo retrotransposition events (mutations) in the pedigree (Table S5, Figure S6).

4 | DISCUSSION

Our in-depth annotation of repeats in the collared flycatcher reference genome assembly permitted a detailed investigation of both ancient and recent TE activities. Transposition in the last 33 MY of the flycatcher lineage was characterized by LINE and LTR retrotransposition, but with virtually no activity of SINEs or DNA transposons, and an increased dominance of LTR activity over CR1 LINE activity within the last 11 MY (Figure 2; i.e., 0–5% distance to consensus). The highest TE activity occurred 33–54 MYA and then declined. The recent LTR dominance in the flycatcher lineage is very similar to the situation seen in the zebra finch lineage (Kapusta & Suh, 2017) but must have evolved largely independently as the flycatcher and zebra finch lineages are estimated to have diverged 21 MYA (Moyle et al., 2016). In line with this, both the flycatcher and the zebra finch lineages exhibit a high diversity of novel, retrovirus-like LTR retrotransposons, which are specific to each of the two lineages (Figure 3). With in-depth repeat annotations from only two songbird lineages, it is of course difficult to generalize from these observations, but we hypothesize that songbirds in general are frequently exposed to germline infiltration by retroviruses. If so, the vast majority of songbird retrovirus diversity might yet await discovery.

This genome-scale study provides novel insights into how recent transposition activity in an avian lineage sets the stage for a rich source of structural variation within species as well as between closely related species. We consider the identification of 11,888 nonreference TEVs from 200 flycatcher genomes to be a conservative estimate for the amount of TE-derived structural variation in this sample, mainly because we only analysed nonreference TEVs. Reference TEVs likely constitute an additional source of TEVs in the sample but were not analysed due to methodological challenges. Furthermore, we required the TEV loci to have >200-bp distance to assembly gaps and repeats in the collared flycatcher reference (thus excluding 369 Mb or 33% of the reference genome) and filtered TEVs for call quality more conservatively than originally proposed for the RetroSeq program (Keane et al., 2013). Finally, we note that the ability to detect nonreference TEVs in resequencing data was highly sensitive to coverage. Confidently scoring the full breadth of TEVs present in an individual is likely to require >30× of genomic coverage, at least with the present pipeline for scoring TEVs. This is consistent with recent observations from human TEV benchmarking (Rishishwar et al., 2017).

We independently verified TEV calls by tracing 713 TEVs across a three-generation pedigree of collared flycatchers. Only 3.8% of these TEVs were inconsistent with the pedigree (Figure S6), and 425 TEVs were traced across all three generations, suggesting that false negatives and false positives did not significantly contribute to the observed patterns of TEV abundance and diversity. This is further supported by inheritance patterns of putatively heterozygous TEVs from the F_1 generation to the five F_2 individuals (Table S6). Furthermore, population-pooled analysis of the 11,888 TEVs as phylogenetic markers (Figure S4) recapitulated the phylogenetic relationships among the flycatcher species and populations (Burri et al., 2015), including the previously noted high degree of phylogenetic discordance in the deepest branching event between the four *Ficedula* species (Nater et al., 2015).

Importantly, our data highlight the potential importance of TEs in adaptation and speciation in birds. First, the high LTR retrotransposon activity during the last 20–30 MY in the lineage leading to flycatcher (as well as in the independent lineage leading to zebra finch) suggests that LTR insertions might have influenced the evolution of many traits. Second, the frequent occurrence of polymorphic TEs among and within contemporary flycatcher species may very well imply that TEVs have contributed to recent speciation events and currently contribute to fitness variation among individuals. These observations therefore suggest that genome scans aimed at identification of loci or genomic regions involved in speciation and adaptation need to integrate screening for TEVs in candidate regions.

Few thoroughly repeat-annotated genomes are available for birds. The only in-depth genome annotations that so far are comparable to our manually curated collared flycatcher repeat annotation are for chicken (Hillier et al., 2004) and zebra finch (Warren et al., 2010). The high diversity of newly discovered retrovirus-like LTR families in collared flycatcher, most of which are lineage-specific, suggests a high and yet largely unexplored diversity of songbird-infecting retroviruses. We note that this unusual diversity of retroviruses coincides with the massive diversification of songbirds into thousands of species. Future research on population dynamics of retrovirus-like TEs will likely reveal the extent to which arms races with these genomic parasites impacted the population and species differentiation of songbirds. Additionally, our population-scale TEV data pave the way to elucidate how TEVs contributed to phenotypic variation through their manifold effects on transcriptional regulation, 3D genome folding and chromosomal stability.

ACKNOWLEDGEMENTS

We thank Homa Papoli Yazdi, Alexander Nater, Fritjof Lammers, Takeshi Kawakami, Toby Fountain, Ludovic Dutoit, Rory Craig and Reto Burri for helpful discussions. We thank two anonymous reviewers for their valuable comments on an earlier version of this manuscript. This work was supported by grants from the Knut and Alice Wallenberg Foundation, and the Swedish Research Council. Computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

DATA ACCESSIBILITY

Repeat consensus sequences and VCF files of TEV calls are available as data files in supporting information.

AUTHOR CONTRIBUTIONS

A.S. and H.E. conceived and designed the study. A.S. and L.S. performed the experiments. A.S. analysed the data. A.S. and H.E. wrote the manuscript.

ORCID

Alexander Suh http://orcid.org/0000-0002-8979-9992
Hans Ellegren http://orcid.org/0000-0002-5035-1736

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*, 403–410.

Baker, A. J., Haddrath, O., McPherson, J. D., & Cloutier, A. (2014). Genomic support for a moa-tinamou clade and adaptive morphological convergence in flightless ratites. *Molecular Biology and Evolution*, *31*, 1686–1696.

Bao, W., Kojima, K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, *6*, 11.

Barrón, M. G., Fiston-Lavier, A.-S., Petrov, D. A., & González, J. (2014). Population genomics of transposable elements in *Drosophila*. *Annual Review of Genetics*, *48*, 561–581.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, *41*, D36–D42.

Burri, R., Nater, A., Kawakami, T., Mugal, C. F., Olason, P. I., Smeds, L., . . . Hogner, S. (2015). Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum in *Ficedula* flycatchers. *Genome Research*, *25*, 1656–1665.

Chalopin, D., Naville, M., Plard, F., Galiana, D., & Volff, J.-N. (2015). Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biology and Evolution*, *7*, 567–580.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . McVean, G. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*, 2156–2158.

Dutoit, L., Burri, R., Nater, A., Mugal, C. F., & Ellegren, H. (2017). Genomic distribution and estimation of nucleotide diversity in natural populations: Perspectives from the collared flycatcher (*Ficedula albicollis*) genome. *Molecular Ecology Resources*, *17*, 586–597.

Ellegren, H. (2010). Evolutionary stasis: The stable chromosomes of birds. *Trends in Ecology & Evolution*, *25*, 283–291.

Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backström, N., Kawakami, T., . . . Uebbing, S. (2012). The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature*, *491*, 756–760.

Elliott, T. A., & Gregory, T. R. (2015). What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philosophical Transactions of the Royal Society B*, *370*, 20140331.

Excoffier, L., & Heckel, G. (2006). Computer programs for population genetics data analysis: A survival guide. *Nature Reviews Genetics*, *7*, 745.

Farré, M., Narayan, J., Slavov, G. T., Damas, J., Auvil, L., Li, C., . . . Larkin, D. M. (2016). Novel insights into chromosome evolution in birds, archosaurs, and reptiles. *Genome Biology and Evolution*, *8*, 2442–2451.

Gardner, E. J., Lam, V. K., Harris, D. N., Chuang, N. T., Scott, E. C., Pittard, W. S., . . . 1000 Genomes Project Consortium (2017). The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Research*, *27*(11), 1916–29. https://doi.org/10.1101/gr.218032.116

Grandi, F. C., Rosser, J. M., Newkirk, S. J., Yin, J., Jiang, X., Xing, Z., . . . Ye, P. (2015). Retrotransposition creates sloping shores: A graded influence of hypomethylated CpG islands on flanking CpG sites. *Genome Research*, *25*, 1135–1146.

Gregory, T. R. (2017). *Animal genome size database*. Retrieved from http://www.genomesize.com

Gregory, T. R., Andrews, C. B., McGuire, J. A., & Witt, C. C. (2009). The smallest avian genomes are found in hummingbirds. *Proceedings of the Royal Society of London B: Biological Sciences*, *276*, 3753–3757.

Guio, L., Barrón, M. G., & González, J. (2014). The transposable element Bari-Jheh mediates oxidative stress response in Drosophila. *Molecular Ecology*, *23*, 2020–2030.

Hancks, D. C., & Kazazian, H. H. Jr (2012). Active human retrotransposons: Variation and disease. *Current Opinion in Genetics & Development*, *22*, 191–203.

Hénaff, E., Zapata, L., Casacuberta, J., & Ossowski, S. (2015). Jitterbug: Somatic and germline transposon insertion detection at single-nucleotide resolution. *BMC Genomics*, *16*, 768.

Hillier, L. W., Miller, W., Birney, E., Warren, W., Hardison, R. C., Ponting, C. P., . . . Dodgson, J. B. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, *432*, 695–716.

Hughes, A. L., & Hughes, M. K. (1995). Small genomes for better flyers. *Nature*, *377*, 391.

Huson, D. H. (1998). SplitsTree: Analyzing and visualizing evolutionary data. *Bioinformatics*, *14*, 68–73.

Janes, D. E., Organ, C. L., Fujita, M. K., Shedlock, A. M., & Edwards, S. V. (2010). Genome evolution in Reptilia, the sister group of mammals. *Annual Review of Genomics and Human Genetics*, *11*, 239–264.

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., . . . Suh, A. (2014). Whole genome analyses resolve the early branches in the tree of life of modern birds. *Science*, *346*, 1320–1331.

Jeffares, D. C., Rallis, C., Rieux, A., Speed, D., Převorovský, M., Mourier, T., . . . Pracana, R. (2015). The genomic and phenotypic diversity of *Schizosaccharomyces pombe*. *Nature Genetics*, *47*, 235–241.

Kaiser, V. B., van Tuinen, M., & Ellegren, H. (2007). Insertion events of CR1 retrotransposable elements elucidate the phylogenetic branching order in galliform birds. *Molecular Biology and Evolution*, *24*, 338–347.

Kapitonov, V. V., & Jurka, J. (2008). A universal classification of eukaryotic transposable elements implemented in Repbase. *Nature Reviews Genetics*, *9*, 411–412.

Kapusta, A., & Suh, A. (2017). Evolution of bird genomes—a transposon's-eye view. *Annals of the New York Academy of Sciences*, *1389*, 164–185.

Kapusta, A., Suh, A., & Feschotte, C. (2017). Dynamics of genome size evolution in birds and mammals. *Proceedings of the National Academy of Sciences of the United States of America*, *114*, E1460–E1469.

Katoh, K., & Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*, *9*, 286–298.

Kawakami, T., Smeds, L., Backström, N., Husby, A., Qvarnström, A., Mugal, C. F., . . . Ellegren, H. (2014). A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. *Molecular Ecology*, *23*, 4035–4058.

Kazazian, H. H. Jr (2004). Mobile elements: Drivers of genome evolution. *Science*, *303*, 1626–1632.

Keane, T. M., Wong, K., & Adams, D. J. (2013). RetroSeq: Transposable element discovery from next-generation sequencing data. *Bioinformatics*, *29*, 389–390.

Kidwell, M. (2005). Transposable elements. In T. R. Gregory (Ed.), *The evolution of the genome* (pp. 165–221). Burlington, MA: Elsevier Academic Press.

Kim, Y. B., Oh, J. H., McIver, L. J., Rashkovetsky, E., Michalak, K., Garner, H. R., … Michalak, P. (2014). Divergence of *Drosophila melanogaster* repeatomes in response to a sharp microclimate contrast in Evolution Canyon, Israel. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 10630–10635.

Konkel, M. K., & Batzer, M. A. (2010). A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. *Seminars in Cancer Biology*, *20*, 211–221.

Kordiš, D. (2009). Transposable elements in reptilian and avian (Sauropsida) genomes. *Cytogenetic and Genome Research*, *127*, 94–111.

Lammers, F., Gallus, S., Janke, A., & Nilsson, M. A. (2017). Phylogenetic conflict in bears identified by automated discovery of transposable element insertions in low-coverage genomes. *Genome Biology and Evolution*, *9*, 2862–2878.

Lavoie, C., Platt, R., Novick, P., Counterman, B., & Ray, D. (2013). Transposable element evolution in *Heliconius* suggests genome diversity within Lepidoptera. *Mobile DNA*, *4*, 21.

Lee, J., Mun, S., Kim, D. H., Cho, C. S., Oh, D. Y., & Han, K. (2017). Chicken (*Gallus gallus*) endogenous retrovirus generates genomic variations in the chicken genome. *Mobile DNA*, *8*, 2.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*, 2078–2079.

Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., … Tsai, I. J. (2009). Population genomics of domestic and wild yeasts. *Nature*, *458*, 337–341.

Lowe, C. B., Bejerano, G., & Haussler, D. (2007). Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 8005–8010.

Mateo, L., Ullastres, A., & González, J. (2014). A transposable element insertion confers xenobiotic resistance in *Drosophila*. *PLoS Genetics*, *10*, e1004560.

Miller, M. A., Pfeiffer, W., & Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *Proceedings of the Gateway Computing Environments Workshop (GCE)*, 1–8.

Moyle, R. G., Hosner, P. A., Jones, A. W., & Outlaw, D. C. (2015). Phylogeny and biogeography of *Ficedula* flycatchers (Aves: Muscicapidae): Novel results from fresh source material. *Molecular Phylogenetics and Evolution*, *82*(Part A), 87–94.

Moyle, R. G., Oliveros, C. H., Andersen, M. J., Hosner, P. A., Benz, B. W., Manthey, J. D., … Faircloth, B. C. (2016). Tectonic collision and uplift of Wallacea triggered the global songbird radiation. *Nature Communications*, *7*, 12709.

Nadachowska-Brzyska, K., Burri, R., Olason, P. I., Kawakami, T., Smeds, L., & Ellegren, H. (2013). Demographic divergence history of pied flycatcher and collared flycatcher inferred from whole-genome re-sequencing data. *PLoS Genetics*, *9*, e1003942.

Nadachowska-Brzyska, K., Burri, R., Smeds, L., & Ellegren, H. (2016). PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white *Ficedula* flycatchers. *Molecular Ecology*, *25*, 1058–1072.

Nater, A., Burri, R., Kawakami, T., Smeds, L., & Ellegren, H. (2015). Resolving evolutionary relationships in closely related species with whole-genome sequencing data. *Systematic Biology*, *64*, 1000–1017.

Nellåker, C., Keane, T. M., Yalcin, B., Wong, K., Agam, A., Belgard, T. G., … Ponting, C. P. (2012). The genomic landscape shaped by selection on transposable elements across 18 mouse strains. *Genome Biology*, *13*, R45.

Quadrana, L., Silveira, A. B., Mayhew, G. F., LeBlanc, C., Martienssen, R. A., Jeddeloh, J. A., & Colot, V. (2016). The *Arabidopsis thaliana* mobilome and its impact at the species level. *Elife*, *5*, e15716.

Rey, O., Danchin, E., Mirouze, M., Loot, C., & Blanchet, S. (2016). Adaptation to global change: A transposable element–epigenetics perspective. *Trends in Ecology & Evolution*, *31*, 514–526.

Rishishwar, L., Mariño-Ramírez, L., & Jordan, I. K. (2017). Benchmarking computational tools for polymorphic transposable element detection. *Briefings in Bioinformatics*, *18*, 908–918.

Rishishwar, L., Tellez Villa, C., & Jordan, I. (2015). Transposable element polymorphisms recapitulate human evolution. *Mobile DNA*, *6*, 21.

Ruiz-Herrera, A., Farré, M., & Robinson, T. J. (2012). Molecular cytogenetic and genomic insights into chromosomal evolution. *Heredity*, *108*, 28–36.

Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Gonçalves, Â., Kutter, C., … Odom, D. T. (2012). Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, *148*, 335–348.

Shedlock, A. M. (2006). Phylogenomic investigation of CR1 LINE diversity in reptiles. *Systematic Biology*, *55*, 902–911.

Shedlock, A. M., Botka, C. W., Zhao, S., Shetty, J., Zhang, T., Liu, J. S., … Edwards, S. V. (2007). Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 2767–2772.

Slotkin, R. K., & Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics*, *8*, 272–285.

Smeds, L., Mugal, C. F., Qvarnström, A., & Ellegren, H. (2016). High-resolution mapping of crossover and non-crossover recombination events by whole-genome re-sequencing of an avian pedigree. *PLoS Genetics*, *12*, e1006044.

Smeds, L., Qvarnstrom, A., & Ellegren, H. (2016). Direct estimate of the rate of germline mutation in a bird. *Genome Research*, *26*, 1211–1218.

Smeds, L., Warmuth, V., Bolivar, P., Uebbing, S., Burri, R., Suh, A., … Moreno, J. (2015). Evolutionary analysis of the female-specific avian W chromosome. *Nature Communications*, *6*, 7330.

Smit, A., & Hubley, R. (2010). *RepeatModeler Open-1.0*. Retrieved from http://www.repeatmasker.org

Smit, A., Hubley, R., & Green, P. (1996–2010). *RepeatMasker Open-3.3.0*. Retrieved from http://www.repeatmasker.org

Sotero-Caio, C., Platt, R. N. II, Suh, A., & Ray, D. A. (2017). Evolution and diversity of transposable elements in vertebrate genomes. *Genome Biology and Evolution*, *9*, 161–177.

Stamatakis, A., Hoover, P., & Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML web servers. *Systematic Biology*, *75*, 758–771.

Stapley, J., Santure, A. W., & Dennis, S. R. (2015). Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. *Molecular Ecology*, *24*, 2241–2252.

Steige, K. A., Laenen, B., Reimegård, J., Scofield, D. G., & Slotte, T. (2017). Genomic analysis reveals major determinants of cis-regulatory variation in *Capsella grandiflora*. *Proceedings of the National Academy of Sciences of the United States of America*, *114*, 1087–1092.

Stuart, T., Eichten, S. R., Cahn, J., Karpievitch, Y., Borevitz, J. O., & Lister, R. (2016). Population scale mapping of novel transposable element diversity reveals links to gene regulation and epigenomic variation. *Elife*, *5*, e20777.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., … Konkel, M. K. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*, 75–81.

Suh, A., Bachg, S., Donnellan, S., Joseph, L., Brosius, J., Kriegs, J. O., & Schmitz, J. (2016). *De-novo* emergence and template switching of SINE retroposons during the early evolution of passerine birds. *bioRxiv*. https://doi.org/10.1101/081950

Suh, A., Churakov, G., Ramakodi, M. P., Platt, R. N., Jurka, J., Kojima, K. K., … Brosius, J. (2015). Multiple lineages of ancient CR1 retroposons shaped the early genome evolution of amniotes. *Genome Biology and Evolution*, *7*, 205–217.

Suh, A., Kriegs, J. O., Donnellan, S., Brosius, J., & Schmitz, J. (2012). A universal method for the study of CR1 retroposons in nonmodel bird genomes. *Molecular Biology and Evolution*, *29*, 2899–2903.

Suh, A., Paus, M., Kiefmann, M., Churakov, G., Franke, F. A., Brosius, J., . . . Schmitz, J. (2011). Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds. *Nature Communications*, *2*, 443.

Suh, A., Smeds, L., & Ellegren, H. (2015). The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. *PLoS Biology*, *13*, e1002224.

Thung, D. T., de Ligt, J., Vissers, L. E., Steehouwer, M., Kroon, M., de Vries, P., . . . Hehir-Kwa, J. Y. (2014). Mobster: Accurate detection of mobile element insertions in next generation sequencing data. *Genome Biology*, *15*, 488.

van't Hof, A. E., Campagne, P., Rigden, D. J., Yung, C. J., Lingley, J., Quail, M. A., . . . Saccheri, I. J. (2016). The industrial melanism mutation in British peppered moths is a transposable element. *Nature*, *534*, 102–105.

Vekemans, X., Beauwens, T., Lemaire, M., & Roldán-Ruiz, I. (2002). Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size. *Molecular Ecology*, *11*, 139–151.

Vijay, N., Bossu, C. M., Poelstra, J. W., Weissensteiner, M. H., Suh, A., Kryukov, A. P., & Wolf, J. B. (2016). Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. *Nature Communications*, *7*, 13195.

Vinogradov, A. E. (1997). Nucleotypic effect in homeotherms: Body-mass independent resting metabolic rate of passerine birds is related to genome size. *Evolution*, *51*, 220–225.

Wang, J., Vicente-García, C., Seruggia, D., Moltó, E., Fernandez-Miñán, A., Neto, A., . . . Jordan, I. K. (2015). MIR retrotransposon sequences provide insulators to the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, E4428–E4437.

Warren, W. C., Clayton, D. F., Ellegren, H., Arnold, A. P., Hillier, L. W., Künstner, A., . . . Heger, A. (2010). The genome of a songbird. *Nature*, *464*, 757–762.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., . . . Paux, E. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, *8*, 973–982.

Wright, N. A., Gregory, T. R., & Witt, C. C. (2014). Metabolic 'engines' of flight drive genome size reduction in birds. *Proceedings of the Royal Society B: Biological Sciences*, *281*, 20132780.

Zhang, Q., & Edwards, S. V. (2012). The evolution of intron size in amniotes: A role for powered flight? *Genome Biology and Evolution*, *4*, 1033–1043.

Zhang, G., Li, C., Li, Q., Li, B., Larkin, D. M., Lee, C., . . . Ödeen, A. (2014). Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, *346*, 1311–1320.

Zhivotovsky, L. A. (1999). Estimating population structure in diploids with multilocus dominant DNA markers. *Molecular Ecology*, *8*, 907–913.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Suh A, Smeds L, Ellegren H. Abundant recent activity of retrovirus-like retrotransposons within and among flycatcher species implies a rich source of structural variation in songbird genomes. *Mol Ecol.* 2018;27:99–111. https://doi.org/10.1111/mec.14439