Natural selection beyond genes: Identification and analyses of evolutionarily conserved elements in the genome of the collared flycatcher (*Ficedula albicollis*)

Rory J. Craig1,2 | Alexander Suh1 | Mi Wang1 | Hans Ellegren1

1Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden
2Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, UK

Correspondence
Rory J. Craig and Hans Ellegren, Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden.
Emails: rory.craig@ed.ac.uk; hans.ellegren@ebc.uu.se

Funding information
Vetenskapsrådet, Grant/Award Number: 2013-8271; Knut och Alice Wallenbergs Stiftelse

Abstract
It is becoming increasingly clear that a significant proportion of the functional sequence within eukaryotic genomes is noncoding. However, since the identification of conserved elements (CEs) has been restricted to a limited number of model organisms, the dynamics and evolutionary character of the genomic landscape of conserved, and hence likely functional, sequence is poorly understood in most species. Moreover, identification and analysis of the full suite of functional sequence are particularly important for the understanding of the genetic basis of trait loci identified in genome scans or quantitative trait locus mapping efforts. We report that ~6.6% of the collared flycatcher genome (74.0 Mb) is spanned by ~1.28 million CEs, a higher proportion of the genome but a lower total amount of conserved sequence than has been reported in mammals. We identified >200,000 CEs specific to either the archosaur, avian, neornithian or passeridan lineages, constituting candidates for lineage-specific adaptations. Importantly, no less than ~71% of CE sites were nonexonic (52.6 Mb), and conserved nonexonic sequence density was negatively correlated with functional exonic density at local genomic scales. Additionally, nucleotide diversity was strongly reduced at nonexonic conserved sites (0.00153) relative to intergenic nonconserved sites (0.00427). By integrating deep transcriptome sequencing and additional genome annotation, we identified novel protein-coding genes, long noncoding RNA genes and transposon-derived (exapted) CEs. The approach taken here based on the use of a PROGRESSIVE CACTUS whole-genome alignment to identify CEs should be readily applicable to nonmodel organisms in general and help to reveal the rich repertoire of putatively functional noncoding sequence as targets for selection.

KEYWORDS
collared flycatcher, comparative genomics, conserved elements, PROGRESSIVE CACTUS, targets for selection, transposon exaptation

1 INTRODUCTION

The identification of selectively constrained sites within a genome, and analysis of their character and distribution, is an issue of central importance for the ability to link genotypes with phenotypes. As the number of species with sequenced genomes has rapidly expanded, it has become possible to utilize large-scale comparative genomics approaches to identify conserved sequence in several taxa, including...
mammals (Lindblad-Toh et al., 2005; Miller et al., 2007; Siepel et al., 2005), fish (Braasch et al., 2016; Jones et al., 2012; Lowe et al., 2011) and crucifer plants (Haudry et al., 2013; Williamson et al., 2014). In the best-studied organisms such as human, the detection of conserved elements (CEs) has recently been performed at increasingly fine-scale resolutions (Lindblad-Toh et al., 2011; Rosenbloom et al., 2015). While protein-coding sequence can be determined using various approaches, comparative analyses are key to identifying putatively functional (yet still unknown) noncoding sequence evolving under selective constraint. As an indication of the potential importance of this latter category of sequences, in humans conserved non-coding elements (CNEs) constitute ~3.5% of the genome, relative to ~1.5% for protein-coding sequence (Lindblad-Toh et al., 2011; Miller et al., 2007; Siepel et al., 2005). Many CNEs function as cis-regulatory elements and have been shown to be especially associated with genes involved in transcriptional regulation and development (Bejerano et al., 2004; Lowe et al., 2011; Woorfe et al., 2005).

The identification of CNEs has proved of considerable importance in enhancing genome annotation (Lindblad-Toh et al., 2011), investigating signatures of adaptive evolution in noncoding sequence (Halligan et al., 2013; Hernandez et al., 2011; Williamson et al., 2014), and for identifying putative trait loci (Marcovitz, Jia, & Bejerano, 2016) and transposable element (TE) exaptation events (Lowe & Haussler, 2012; Mikkelsen et al., 2007; Xie, Kamal, & Lander, 2006), the latter potentially being an underrated source of phenotypic novelty. Incorporating putatively functional noncoding annotation is particularly important in the context of ecological or evolutionary studies of natural populations of nonmodel organisms. In the absence of identified CNEs for essentially all such organisms, nomination of candidate loci from signals in genome scans or QTL mapping efforts is necessarily limited to protein-coding genes, which certainly may be misleading.

Furthermore, determining the density and distribution of genomic targets for selection is required to better understand within-genome heterogeneity in genetic diversity (Ellegren & Galtier, 2016). Alongside variation in the recombination rate, the density of such sites is expected to be the main determinant of variation in genetic diversity via the diversity-reducing effects of selection on linked sites (either by background selection or selective sweeps; Cutter & Payseur, 2013). In lieu of functional noncoding sequence data for a species of interest, exonic or coding sequence density alone has often been used to approximate targets for selection (Ellegren & Galtier, 2016). However, for many species this is unlikely to represent a suitable proxy, as coding sequence will underestimate the actual density of sites under selection (cf. above). Moreover, coding sequence may potentially fail to accurately capture the true distribution of sites under selection, as at least in vertebrates many CNEs are located within large “gene deserts” containing no protein-coding sequence (Ovcharenko et al., 2005; Siepel et al., 2005), and a negative correlation between the densities of conserved noncoding sequences and exons has been reported in human (Dermitzakis, Reymond, & Antonarakis, 2005). Thus, it is important to incorporate putatively functional noncoding sequence alongside protein-coding sequence when approximating the density and distribution of genomic targets for selection. While this is currently applied for a limited number of well-studied taxa such as human (Enard, Messer, & Petrov, 2014; Hernandez et al., 2011; Lohmueller et al., 2011; McVicker, Gordon, Davis, & Green, 2009) and Drosophila melanogaster (Comeron, 2014; Elyashiv et al., 2016; Halligan & Keightley, 2006; Sella, Petrov, Przeworski, & Andolfatto, 2009), the necessary annotation of noncoding sequence is typically unavailable for studies of nonmodel organisms.

The majority of multispecies whole-genome alignments currently available are referenced to a defined species, including the frequently used MULTIZ alignments (Blanchette et al., 2004) available from the UCSC Genome Browser (Karolchik et al., 2003). The reference species is most often a model organism of the group in question, for example human in the case of mammalian alignments (Lindblad-Toh et al., 2011; Miller et al., 2007; Siepel et al., 2005) or D. melanogaster for drosophilids (Siepel et al., 2005; Stark et al., 2007). Although it is possible to utilize CEs predicted from an alignment referenced to a species other than the focal species using tools such as UCSC’s LIFTOVER, such an approach is likely to underestimate conservation for two related reasons. First, any lineage-specific functional elements that arose on the lineage to the focal species after the branching event between the reference species lineage and focal species lineage will be absent from the alignment and hence any CE prediction. Likewise, any evolutionarily ancient elements lost on the reference species lineage but remaining functional on the focal species lineage will also be absent. The degree of underestimation of conserved sequence is therefore expected to increase in severity as the evolutionary distance between the reference and focal species increases. Although there is often insufficient power to detect many lineage-specific CEs at recent evolutionary timescales (Ponting, 2008; Rands, Meader, Ponting, & Lunter, 2014), even if the number detected is small, such elements are likely to be of specific interest as putative candidates for lineage-specific function or loss of function (Chan et al., 2010; Lowe, Clarke, Baker, Haussler, & Edwards, 2015; Marcovitz et al., 2016; McLean et al., 2011; Schmidt et al., 2010).

Advances in alignment algorithms have led to the production of whole-genome alignments without reference species (e.g., Dubchak, Poljakov, Kislyuk, & Brudno, 2009; Paten, Herrero, Beal, Fitzgerald, & Birney, 2008). A recent example of this approach is PROGRESSIVE CACTUS (Paten et al., 2011), which produces alignments in the hierarchical alignment format (HAL; Hickey, Paten, Earl, Zerbino, & Haussler, 2013). HAL is indexed on all genomes contained within the alignment and can be easily converted to the frequently used multi-alignment format (MAF) with any user-specified reference species. Here, we utilize this flexibility to predict CEs explicitly for a nonmodel organism, by converting an existing 23 sauropsid (birds and reptiles) whole-genome alignment produced by Green et al. (2014) to an alignment referenced to the collared flycatcher (Ficedula albicollis) genome. We explore the density and distribution of CEs and utilize population resequencing data to investigate patterns of genetic diversity at different classes of conserved sites. Using these approaches, we find that the majority of conserved sites in the genome are nonexonic.
that these conserved nonexonic sites have substantially reduced genetic diversity, and that CNE density is negatively correlated with functional exonic density at local genomic scales. We further incorporate deep transcriptome sequencing and TE annotation to improve genome annotation and identify thousands of CEs putatively derived from TE exaptation events.

2 | MATERIAL AND METHODS

2.1 | Whole-genome alignment

We performed analyses on the 23 sauropod whole-genome alignment produced using PROGRESSIVE CACTUS (https://github.com/glennhickey/progressiveCactus) by Green et al. (2014). The species included in the alignment are listed in Table S1. The alignment was downloaded as the file crocPaper_reestimated.hal from crocgenomes.org (ftp://ftp.crocgenomes.org/pub/ICGWG/Supplementary_Materials/main_page/Ancestral_genome_reconstruction_files/), and format conversion from HAL (https://github.com/ComparativeGenomicSpanKits/hal) to MAF (http://genome.ucsc.edu/FAQ/FAQformat.html#format5) was performed using the HAL tools command hal2maf (Hickey et al., 2013) with the following parameters: -refGenome ficAlb2 (to extract alignments referenced to the collared flycatcher genome assembly FicAlb1.5 [GenBank Accession GCA_000247815.2]), -noAncestors (to exclude unrequired ancestral reconstructions), -onlyOrthologs (to include only sequences orthologous to the collared flycatcher).

Further formatting was performed using MAFTOOLS (Earl et al., 2014), with the alignment split into individual files for the 33 assembled collared flycatcher chromosomes (Kawakami et al., 2014), and an additional file containing unassembled scaffolds (with any scaffolds containing fewer than 2,000 alignable sites excluded, where successful alignment was defined at any individual site as alignment between a collared flycatcher base and bases from at least two other species). An in-house Perl script that incorporates the EMBOSSTool CONSAMBIG (Rice, Longden, & Bleasby, 2000) was used to create consensus sequences where there were two or more sequences from the same species present in a MAF block.

2.2 | Prediction of conserved elements

PHASTCONS (Siepel et al., 2005) was used to predict CEs from the whole-genome alignment. The phast tool PHYLOFIT (REV model) was used to generate a neutral evolutionary model from fourfold degenerate (4D) sites extracted from the alignment, downloaded as the file WholeGenome_4Dsites_filter2.phy from GigaScience (https://doi.org/10.5524/100125). These 114,709 4D sites were previously defined using nonduplicated coding transcripts from the American alligator (Alligator mississippiensis), where codons were only included if they were present in all species and encoded the same amino acid (Green et al., 2014). The topology of the phylogeny was identical to that derived by Green et al. (2014). Although some 4D sites may be subject to evolutionary constraint (Chamary, Parmley, & Hurst, 2006), including in avian species (Kunstner, Nabholz, & Ellegren, 2011), any subsequent reductions in branch length are expected to reduce power to detect CEs, resulting in only somewhat more conservative predictions.

PHASTCONS was run using two alternative parameter sets: the “standard” parameters (SP) that are used to produce the “most conserved” tracks in the UCSC genome browser (expected length = 45, target coverage = 0.3, rho = 0.31; Miller et al., 2007), and a “tuned” parameter (TP) set that was specifically tuned for use with the alignment (expected length = 12, target coverage = 0.0225, rho = 0.27). Tuning criteria are detailed in the Supporting information.

Additional CE prediction was performed using GERPP++ (Davydov et al., 2010) on the same alignments and evolutionary model, using default parameters. For both PHASTCONS and GERPP++ predicted CEs, any bases falling within assembly gaps were filtered, and any CEs that comprised more that 20% of such bases were excluded entirely.

2.3 | Annotation of conserved elements by site class

Sequence coordinates for CDS, 5’ and 3’ UTRs, and pseudogenes were taken from the ENSEMBL version 73 collared flycatcher genome annotation (see Uebbing et al. (2016) for detailed description), translated from the FicAlb1.4 to the FicAlb1.5 assembly. RNA gene coordinates for miRNAs, rRNAs, snRNAs, snoRNAs and other miscellaneous RNAs were also taken from the ENSEMBL annotation and combined with additional rRNA, scRNA, snRNA and tRNA genes annotated by an in-house pipeline (see Repeat annotation). For protein-coding genes with annotated 5’ UTRs of at least 15 bp, promoters were annotated as 2 kb upstream from the transcription start site. Putatively novel coding sequence, UTR sequence and long noncoding RNA (IncRNA) genes were annotated based on expression data (Supporting information). Following Lindblad-Toh et al. (2011), the intersect between CEs and the various annotated classes was defined in a hierarchical format, such that if a base within a CE overlapped two or more different classes, it was assigned to a single class based on the first appearance in the following order: CDS, 5’ UTR, 3’ UTR, promoter, RNA gene, novel CDS, novel UTR, IncRNA, intronic, intergenic. This approach affected only 1.1% of CE bases.

Conserved noncoding elements were defined as CEs without any overlap with exon-associated site classes (CDS, 5’ and 3’ UTRs, promoters, novel CDS/UTRs, RNA genes and pseudogenes). Similarly defined elements have previously been more accurately termed “conserved nonexonic elements” (e.g., by Lowe et al. 2011), but we here use the more widely used term “conserved noncoding elements.” CEs overlapping IncRNAs but no other exonic site classes were defined as CNEs, as despite these being putatively exonic, their function (if any) is most often unknown, and hence, it is unclear whether the observed conservation is due to their transcription or some other function.
2.4 Distribution and density of conserved noncoding elements

CNE density and the density of CDS and any other exon-associated CEs (hereafter referred to as CDS + exonic CEs) were calculated for nonoverlapping 200-kb windows along the genome. Windows were discarded if they included assembly gaps between scaffolds, and window size was therefore selected to maximize the number of complete windows residing on single scaffolds (a size of 200 kb resulted in 5,000 windows covering 89.4% of the genome). Total CE densities were calculated per chromosome, and the percentage of alignable sites per chromosome was noted (where alignment between collared flycatcher and at least two other species at an individual site was considered successful). CE densities were then corrected to reflect the number of alignable bases per chromosome rather than total chromosome length. All correlations were performed on 31 of the 33 assembled chromosomes, with the two smallest chromosomes (LG34 and LG35; both <500 kb) excluded as <2% of their bases were alignable.

Gene deserts were defined as intergenic regions >200 kb and were further categorized as evolutionarily stable or variable based on collared flycatcher-human alignability (stable deserts >5% sites aligned, variable <5% sites aligned, Supporting information).

2.5 Comparison of conservation and polymorphism

We used the protocol for estimating nucleotide diversity described in Dutoit, Burri, Nater, Mugal, and Ellegren (2016), based on whole-genome resequencing data from 20 (10 males, 10 females) collared flycatchers from a single Italian population from the Apennine Mountains (available on the European Nucleotide Archive (ENA), Accession no. PRJEB7359). Information regarding the sampling, DNA analyses and bioinformatics methods used to produce these data can be found in Burri et al. (2015), and information regarding additional variant filtering and estimation of nucleotide diversity is given in Dutoit et al. (2016). Briefly, nucleotide diversity (π; Nei & Li, 1979), the average number of pairwise differences per site among chromosomes in a population, was estimated collectively for the 20 individuals (40 chromosomes) using in-house Python scripts and the Python package pyvcf 0.4.0 (https://py.vcf.readthedocs.org). This was done for all sites within each of the annotated site-class categories described above, for the intersect between CEs and each site-class category, and for the remaining sites not overlapped by CEs. Only autosomal sequences were analysed.

2.6 Update of genome annotation using transcriptome sequencing

To investigate if any conserved bases overlapped previously unannotated exonic sequences (CDS and UTRs) and to investigate patterns of sequence conservation in IncRNA genes, we updated the collared flycatcher genome annotation using deep transcriptome data. Briefly, putatively novel protein-coding sequences were identified using BRAKER1 (Hoff, Lange, Lomsadze, Borodovsky, & Stanke, 2016) with STAR alignments (Dobin et al., 2013), and novel UTR sequences were identified using MAKER (Holt & Yandell, 2011) with a de novo transcriptome assembly produced using TRINITY (Grabherr et al., 2011). A reference-based transcriptome assembly produced with CUFFLINKS (Roberts, Trapnell, Donaghey, Rinn, & Pachter, 2011; Trapnell et al., 2010) from the STAR alignments was used in conjunction with a pipeline of filtering steps to annotate IncRNA genes. Detailed methods are available in the Supporting information.

2.7 Repeat annotation

To identify CEs putatively originating from the exaptation of transposable elements, we utilized a recent in-depth annotation of collared flycatcher repeats (Suh, Smeds, & Ellegren, 2017). Briefly, repetitive elements and additional RNA genes were identified in the FicAlb1.5 genome assembly using REPEATMASKER v4.0.1 (Smit, Hubley, & Green, 1996–2010) and a repeat library of collared flycatcher repeats. This library contained consensus sequences of avian repeats from Repbase (Bao, Kojima, & Kohany, 2015) and collared flycatcher repeats. Consensus sequences of the latter were manually curated through visual inspection of multiple sequence alignments and classified based on characteristic sequence features or target site duplication. Based on this annotation, TEs (short interspersed elements or SINEs, DNA transposons, long interspersed elements or LINEs, long terminal repeats or LTRs) were intersected with CEs. Following the methodology of Lowe and Haussler (2012), CEs were defined as TE-derived if the majority of their bases overlapped a single TE, or TEs from a single superfamily.

2.8 Identification of lineage-specific conserved elements

For each collared flycatcher CE, presence/absence of orthologous sequence in the whole-genome alignment was scored for each species to determine lineage-specific CEs. Following Lowe et al. (2015), absence of a CE was defined in a species if less than one-third of orthologous bases were aligned. CEs were then defined as specific to a lineage if they were present in a species from the deepest branch of that lineage, but absent in all more distantly related species. As this approach could potentially lead to false positives due to gaps in genome assemblies and/or alignment errors, CEs were only defined as lineage-specific if there were at least three outgroup species (all with an absence of orthologous CE sequence) available at a particular branch for comparison. This resulted in four scales of lineage-specificity: archosaur-specific CEs (absent from anole lizard Anolis carolinensis and all four turtles, but present in at least one crocodilian), avian-specific CEs (absent from all reptiles, including the three crocodilian species, but present in ostrich Struthio camelus (representing the deepest branching avian clade)), neoavian-specific CEs (absent from all reptiles, ostrich and the three fowl species, but present in rock pigeon Columba livia) and passeridan-specific CEs.
3.1 the optimal alignment for our purpose of identifying CEs in the et al. (2014) is currently the most species-rich reference-free align-
ment that includes high-quality genome assemblies for species that maximize alignability (Miller et al., 2007). Hence, an optimal alignment will be increased by aligning as many distantly related species as possible including 8.2% (91.3 Mb) or 6.6% (74.0 Mb) of the collared flycatcher coding sequence (CDS) was alignable, with the average number of aligned species increasing to 21.5, and average total branch length to 2.43 subs/site, at such sites.

It should be noted that if the whole-genome alignment does not contain a representative species from the deepest lineage for one of the clades of interest (potentially for Neoaves, Suh, 2016; and almost certainly for Passerida, Moyle et al., 2016), then CEs specific to these clades are only defined herein in relation to the species included in this study. Also, for most lineages specificity is only defined based on presence in a single species, which could underes-
timate CE lineage-specificity (again due to assembly gaps/alignment errors). Finally, a small proportion of CEs assigned to a particular evolutionary branch may have a more ancient origin than inferred due to CE loss events. For example, a CE that originated prior to the split between crocodilians and birds that was then subsequently lost in the crocodilian lineage (and is hence absent from all extant crocodilians) would correctly be defined as avian specific, but its origin would be incorrectly assigned to the evolutionary branch between crocodilians and birds.

To investigate potential phenotypic consequences of CE innova-
tion, we assigned genes with the nearest upstream or downstream transcription start site to each neovian- and passeridan-specific CNE, as enhancers can influence gene expression over long dis-
tances in either orientation (Visel, Rubin, & Pennacchio, 2009). CEs specific to these lineages that had direct overlap with genes were assigned to those genes. GO term analysis for the associated genes was then performed as described for genes flanking gene deserts (Supporting information).

3 RESULTS

3.1 23 Sauropsid whole-genome alignment

The effectiveness of a whole-genome alignment for identifying CEs is a function of the total branch length of the phylogeny represent-
ing the aligned species, as well as the alignability of the genomes included (Cooper et al., 2003). While the power to detect CEs can be increased by aligning as many distantly related species as possible (maximizing total branch length), such increases in phylogenetic scope result in a diminishing proportion of the reference genome that can be aligned to all other species (resulting in only the most ancient CEs being identified with the full power available; Cooper et al., 2003; Miller et al., 2007). The quality of the included genomes is also important, as lower assembly qualities will further reduce alignability (Miller et al., 2007). Hence, an optimal alignment will include high-quality genome assemblies for species that maximize branch length at a reasonable phylogenetic scope.

The 23 sauropsid whole-genome alignment produced by Green et al. (2014) is currently the most species-rich reference-free align-
ment that includes the collared flycatcher genome. As such, it was the optimal alignment for our purpose of identifying CEs in the genome of a nonmodel organism, without the necessity of producing a new alignment. The phylogeny connecting the 23 sauropsids (Table S1; 15 birds, three crocodilians, four turtles and anole lizard) provided a total branch length of 2.63 substitutions per 4D site (subs/site), and the subphylogeny connecting the 15 avian species 1.24 subs/site (Figure 1a). Notably, four of the five high-quality assemblies were avian species (collared flycatcher, zebra finch Taeniopygia guttata, chicken Gallus gallus and turkey Meleagris gallopavo), providing higher quality genomes at the more recent phylo-
genetic scope of the alignment. Considering alignability, 88.6% of sites were alignable between collared flycatcher (the focal species) and at least two other species (Figure 1b), and at such sites, the average number of aligned species was 15.5, with an average total branch length of 1.43 subs/site. Comparatively, 96.0% of collared flycatcher coding sequence (CDS) was alignable, with the average number of aligned species increasing to 21.5, and average total branch length to 2.43 subs/site, at such sites.

3.2 Conserved elements

The phastcons analyses identified ~1.41 and ~1.28 million CEs, spanning 8.2% (91.3 Mb) or 6.6% (74.0 Mb) of the collared flycatcher genome, for the SP ("standard") and TP ("tuned") methods, respec-
tively (Table S2). The genomic regions corresponding to SP-predicted CEs had ~7 times as many bases overlapping both assembly gaps and flycatcher lineage-specific TEs relative to the TP-predicted CEs, and some of the longest SP-predicted CEs harboured long assembly gaps (e.g., the longest contained two gaps of 590 and 312 bp), indicative of either "coverage" being too high or over-"smoothing" in the phastcons model (Siepel et al., 2005). The TP-predicted CEs were almost entirely overlapped by the SP data set (~0.1% of TP-predicted CEs had no overlap), suggesting that they could be treated as a higher confidence data set contained within the larger SP data set. The TP data set presumably had a lower rate of false positives, although conceivably a higher rate of false negatives; we thus note that the SP data set can be treated as an upper bound for phast-
cons-predicted conservation in the collared flycatcher. We therefore focused all further analyses on the TP data set of CEs.

Gerpp++ analysis identified 817.511 CEs spanning 12.3% (137.3 Mb) of the genome. As is the case in human, phastcons-predicted CEs generally fell within the longer gerpp++ CEs (Davydov et al., 2010), with 88.0% and 94.5% of phastcons CEs (SP and TP methods, respectively) overlapped by gerpp++ CEs. The higher congru-
ence between the gerpp++ CEs and phastcons TP-predicted CEs (rela-
tive to SP-predicted CEs) also supported the decision to focus analyses only on the TP data set.

3.3 Annotation of conserved elements by site class

Incorporating annotation from CDS, 5’ and 3’ UTRs, promoters, RNA genes and novel putative CDS and UTRs, 31.7% of CEs (23.5 Mb) were found to be associated with exonic sequence. The majority of
putatively functional sequence in the collared flycatcher genome is thus neither protein-coding nor exonic. At the nucleotide level, 23.8% of CE bases overlapped CDS, 0.5% 5' UTRs, 3.1% 3' UTRs, 0.6% promoters, 0.1% RNA genes and 0.9% previously unannotated CDS and UTR sequence (Figure 2). With the exception of promoters, all of these site-class categories overlapped more CE bases than expected given a random distribution of conserved bases and exhibited the highest proportions of bases overlapped by CEs (Table 1). Unsurprisingly, CDS showed the most constraint among site-class categories, with 70.4% of bases overlapped by CEs, a 9.2-fold increase relative to random expectation. This corresponded to 89.6% of CDS exons having at least a partial overlap with CEs.

Despite the extensive conservation detected in exonic sequence, most CE bases were found within intronic and intergenic sequence (23.7% and 44.4%, respectively; Figure 2). Putative lncRNA genes also overlapped a substantial proportion of CE bases (2.9%), but fewer than expected by chance, consistent with high levels of evolutionary turnover/lineage-specificity and/or limited selective constraint (Supporting information). The majority of the conserved bases overlapping lncRNA genes, introns and intergenic sequence corresponded to over 900,000 CNEs covering 50.5 Mb of the genome, with these elements on average 14.8 bp shorter than exon-associated CEs (mean length 53.8 bp, relative to 68.6 bp).

### 3.4 Distribution and density of conserved noncoding elements

Considering all CDS and other exonic CEs collectively as a proxy for functional exonic sites, total density of such sequences was 2.6% of the collared flycatcher genome (28.9 Mb), relative to 4.5% (50.5 Mb)
as the above negative correlations could be artefacts caused by windows with high exonic density having lower CNE density due only to having less nonexonic sequence, CNE densities were recalculated based on the proportion of nonexonic sequence in each window (rather than the full 200 kb). Both correlations remained negative after this correction, although less so (corrected CNE and CDS + exonic CEs, \( R = -0.222 \); corrected CNE and CDS, \( R = -0.213 \)).

Examining the largest intergenic regions of the genome, 747 intervals longer than 200 kb contained no protein-coding genes and were considered gene deserts. These regions comprised 51.2% (298.8 Mb) of all intergenic sequence and covered 26.7% of the genome, containing 325,577 CNEs (54.8% of intergenic CNEs). Total CNE density in gene deserts was 6.2%, relative to 4.8% for the remaining intergenic regions, further highlighting the negative association between CNE density and CDS + exonic CE density. Dividing gene deserts based on their alignability to human, 378 gene deserts were defined as evolutionarily stable and 369 as variable (not alignable). Three of the largest (>1 Mb) stable gene deserts, with a range of associated CNE densities, can be observed on a representative 30-Mb section of chromosome 1 (Figure 3b).

As in human, CNE density in variable gene deserts was considerably lower (3.6%) than that observed in stable gene deserts (8.6%). CNEs located within variable gene deserts were more often lineage-specific, with 14.5% identifiable as avian-specific and 38.9% identifiable as present in the anole lizard, relative to 9.1% and 47.4% for stable gene desert CNEs. The latter result suggests that although CNE turnover may be reduced in stable gene deserts, it is still considerably higher than that observed for exon-associated CEs (90.6% of which are present in the anole lizard). Alternatively, genuinely orthologous CNEs may exist in the anole lizard, but these may be difficult to align in such large intergenic regions. Genes immediately flanking stable gene deserts were enriched for GO terms relating to transcriptional regulation and DNA binding (Table S3) and included the DACH1 and SOX2 genes that also flank human stable gene deserts (Ovcharenko et al., 2006).

As the above negative correlations could be artefacts caused by windows with high exonic density having lower CNE density due only to having less nonexonic sequence, CNE densities were recalculated based on the proportion of nonexonic sequence in each window (rather than the full 200 kb). Both correlations remained negative after this correction, although less so (corrected CNE and CDS + exonic CEs, \( R = -0.222 \); corrected CNE and CDS, \( R = -0.213 \)).

Examining the largest intergenic regions of the genome, 747 intervals longer than 200 kb contained no protein-coding genes and were considered gene deserts. These regions comprised 51.2% (298.8 Mb) of all intergenic sequence and covered 26.7% of the genome, containing 325,577 CNEs (54.8% of intergenic CNEs). Total CNE density in gene deserts was 6.2%, relative to 4.8% for the remaining intergenic regions, further highlighting the negative association between CNE density and CDS + exonic CE density. Dividing gene deserts based on their alignability to human, 378 gene deserts were defined as evolutionarily stable and 369 as variable (not alignable). Three of the largest (>1 Mb) stable gene deserts, with a range of associated CNE densities, can be observed on a representative 30-Mb section of chromosome 1 (Figure 3b).

As in human, CNE density in variable gene deserts was considerably lower (3.6%) than that observed in stable gene deserts (8.6%). CNEs located within variable gene deserts were more often lineage-specific, with 14.5% identifiable as avian-specific and 38.9% identifiable as present in the anole lizard, relative to 9.1% and 47.4% for stable gene desert CNEs. The latter result suggests that although CNE turnover may be reduced in stable gene deserts, it is still considerably higher than that observed for exon-associated CEs (90.6% of which are present in the anole lizard). Alternatively, genuinely orthologous CNEs may exist in the anole lizard, but these may be difficult to align in such large intergenic regions. Genes immediately flanking stable gene deserts were enriched for GO terms relating to transcriptional regulation and DNA binding (Table S3) and included the DACH1 and SOX2 genes that also flank human stable gene deserts (Ovcharenko et al., 2006).

As the above negative correlations could be artefacts caused by windows with high exonic density having lower CNE density due only to having less nonexonic sequence, CNE densities were recalculated based on the proportion of nonexonic sequence in each window (rather than the full 200 kb). Both correlations remained negative after this correction, although less so (corrected CNE and CDS + exonic CEs, \( R = -0.222 \); corrected CNE and CDS, \( R = -0.213 \)).

Examining the largest intergenic regions of the genome, 747 intervals longer than 200 kb contained no protein-coding genes and were considered gene deserts. These regions comprised 51.2% (298.8 Mb) of all intergenic sequence and covered 26.7% of the genome, containing 325,577 CNEs (54.8% of intergenic CNEs). Total CNE density in gene deserts was 6.2%, relative to 4.8% for the remaining intergenic regions, further highlighting the negative association between CNE density and CDS + exonic CE density. Dividing gene deserts based on their alignability to human, 378 gene deserts were defined as evolutionarily stable and 369 as variable (not alignable). Three of the largest (>1 Mb) stable gene deserts, with a range of associated CNE densities, can be observed on a representative 30-Mb section of chromosome 1 (Figure 3b).

As in human, CNE density in variable gene deserts was considerably lower (3.6%) than that observed in stable gene deserts (8.6%). CNEs located within variable gene deserts were more often lineage-specific, with 14.5% identifiable as avian-specific and 38.9% identifiable as present in the anole lizard, relative to 9.1% and 47.4% for stable gene desert CNEs. The latter result suggests that although CNE turnover may be reduced in stable gene deserts, it is still considerably higher than that observed for exon-associated CEs (90.6% of which are present in the anole lizard). Alternatively, genuinely orthologous CNEs may exist in the anole lizard, but these may be difficult to align in such large intergenic regions. Genes immediately flanking stable gene deserts were enriched for GO terms relating to transcriptional regulation and DNA binding (Table S3) and included the DACH1 and SOX2 genes that also flank human stable gene deserts (Ovcharenko et al., 2006).
2005), while genes flanking variable gene deserts were instead enriched for a wider range of ontology terms (Table S3).

Genetic diversity has previously been shown to correlate positively with chromosome size in the collared flycatcher (Dutoit et al., 2016), contrary to expectations based on the higher recombination rates of the smaller microchromosomes (defined here as those <10 Mb) and the predicted negative relationship between recombination rate and the diversity-reducing effects of linked selection. A major factor explaining this finding is likely to be the negative relationship between the density of targets for selection and chromosome size, with increased selection on linked sites leading to reduced genetic diversity on the more gene-rich microchromosomes (chromosome size and CDS density are strongly negatively correlated \( R = -0.714 \)). Chromosome size and CE density were also negatively correlated, although less strongly than CDS density \( R = -0.326 \); Figure 4a). After correcting CE density for the proportion of alignable sites on each chromosome, the correlation was, however, of a similar strength to CDS density \( R = -0.752 \); Figure 4b), which was largely as a result of the reduced alignability of microchromosomes (with positive correlations between chromosome size and both the proportion of aligned sites \( R = 0.493 \) and the average branch length at aligned sites \( R = 0.258 \)). Such a correction is only valid if the reduced alignability observed for microchromosomes affects functional sequence, as if the reductions solely affected nonfunctional sequence, then the original densities would be correct regardless of the overall reductions in alignability. There was some evidence for such an effect, as the proportion of aligned sites and proportion of aligned CDS (taken as an example of known functional sites) were positively correlated \( R = 0.517 \); Figure S1a), due to a marked decrease in the proportion of aligned CDS on microchromosomes (Figure S1b).

CNE density was, however, weakly positively correlated with chromosome size \( R = 0.052 \), and even when corrected for the proportion of alignable sites on each chromosome the correlation

![Figure 3](image1.png)

**FIGURE 3** Relationship between CNE density and CDS + exonic CE density. (a) Correlation between CNE density and CDS + exonic CE density for 200 kb windows \( R = -0.284 \). (b) Representative example of variation in CNE and CE + exonic CE densities for 200-kb windows for a 30-Mb region of chromosome 1. Three of the largest gene deserts are located between approximately 30.3–31.3 Mb, 38.4–39.9 Mb and 55.1–56.1 Mb.

![Figure 4](image2.png)

**FIGURE 4** Relationship between chromosome size and CE density. (a) Correlation between CE density and chromosome size \( R = -0.326 \). (b) Correlation between CE density corrected by the proportion of alignable sites per chromosome and chromosome size \( R = -0.752 \).
between size and density was only weakly negative ($R = -0.144$). This suggests that the higher gene density on microchromosomes is largely driving the correlation between chromosome size and CE density, and potentially the observed reductions in genetic diversity relative to macrochromosomes. Despite the differences in exonic densities, there may not be substantial differences in CNE densities between macro- and microchromosomes, although better assemblies and alignments of microchromosomes may be required to investigate any potential differences in CNE densities further.

### 3.5 Patterns of genetic diversity at conserved sites

Nucleotide diversity was reduced for all CEs ($\pi = 0.00142$) by 66.7% relative to nonconserved intergenic sequence (used as a putative neutral reference, $\pi = 0.00427$). Across all site-class categories (with the exception of pseudogenes, presumably due to nonfunctional collared flycatcher CEs being identified via alignment to still functional orthologs in other species) nucleotide diversity was reduced for bases overlapped by CEs, relative to neutral sequence as well as to sites within the same site class not overlapping CEs (Figure 5). Comparing sites within the same site class (and ignoring pseudogenes), the reduction between CE and non-CE sites ranged from 55.2% for CDS to 72.7% for 3' UTRs. Conserved exonic sequence generally exhibited the lowest nucleotide diversity, with 3' UTRs the least variable site class ($\pi = 0.000904$), followed by novel UTRs (which include 3' and 5' UTRs, $\pi = 0.00100$), RNA genes ($\pi = 0.00103$), CDS ($\pi = 0.00113$, which is presumably higher than most other exonic categories due to CEs containing both degenerate and nondegenerate sites), 5' UTRs ($\pi = 0.00116$) and novel CDS ($\pi = 0.00164$). CE bases overlapping IncRNA genes had a more similar nucleotide diversity ($\pi = 0.00143$) to intronic ($\pi = 0.00145$) and intergenic ($\pi = 0.00158$) CE-overlapped sites. Overall, CE bases overlapping both exon-associated ($\pi = 0.00110$) and nonexonic ($\pi = 0.00153$) sequences exhibited substantial reductions in nucleotide diversity (74.3% and 64.1%, respectively) relative to non-CE intergenic sequence.

### 3.6 Exaptation of transposable elements

Based on at least 50% of bases being overlapped by a TE (or TEs from the same superfamily), 10,287 CEs (~626 kb) were identified as putatively exapted from 7,298 TEs (some TEs harbour more than one CE). Of these CEs, the vast majority (9,980, ~607 kb) were CNEs derived from 7,085 TEs. This corresponds to 0.8% of CEs, and 1.1% of CNEs, putatively originating from TE exaptation events. Lineage-specific CEs may more accurately capture the proportion of TE-derived CEs (Lowe & Haussler, 2012), and 1.4% of avian-specific CEs (1.5% of CNEs) were identified as putatively exapted (Supporting information). While the exaptation of TEs as cis-regulatory elements can be summarized by solely considering CNEs, TEs have also been shown to contribute functional sequence within exonic regions (e.g., in UTR sequence and RNA genes, Chuong, Elde, & Feschotte, 2017; Mikkelsen et al., 2007, as well as in protein-coding sequence via exonization, Schmitz & Brosius, 2011). We therefore considered the full set of CEs for our analyses, although the results are largely identical when considering all CEs (Table 2) or only CNEs (Table S4).

The most abundant superfamily of TEs in the collared flycatcher (and most birds) are CR1 LINEs (Kapusta & Suh, 2016), and unsurprisingly, these elements have been involved in more than twice as many exaptation events as other TE superfamilies (Table 2). With the exception of LTRs (with only 174 exapted TEs), a similar number

![FIGURE 5 Nucleotide diversity ($\pi$) for each site-class category for collared flycatcher autosomes. Conserved sites are those overlapping CEs, while nonconserved sites are those with no CE overlap](image-url)
of TEs have contributed to CEs across the remaining three TE subclasses found in vertebrate genomes (LINEs 2,407, SINEs 2,607 and DNA transposons 2,110; Table 2). This is despite very large differences in the proportions of exapted TEs between subclasses, with only 0.3% of LTRs and 1.8% of LINEs exapted at one extreme and 36.3% of SINEs and 64.2% of DNA transposons at the other. Much of this variation likely reflects differences in the evolutionary period in which particular TE superfamilies were active. For example, only 119 (0.2%) of the most abundant LTR superfamilies (ERV1, ERVK and ERVL) have been exapted. While some families within these LTR superfamilies were active during early Neognathae (Neoaves and Galloanserae) evolution (e.g., TguLTR5d and TguLTR5e; Suh, Smeds, & Ellegren, 2015; Suh et al., 2011), LTR families have mostly undergone diversification specifically in the oscine passerines (Suh et al., 2017; Vijay et al., 2016; Warren et al., 2010). Hence, even if TEs from

| Subclass | Superfamily | TEs | TE (bp) | Exapted TEs | Exapted TEs (bp) | Exapted TEs (%) | TE-derived CEs | TE-derived CEs (bp) | CEs per TE |
|----------|-------------|-----|---------|------------|-----------------|----------------|-----------------|------------------|------------|
| LINE     |             | 132,249 | 35,295,800 | 2,407 | 570,668 | 1.82 | 3,301 | 210,652 | 1.37 |
| LINE/CRI |             | 132,042 | 35,277,260 | 2,347 | 558,488 | 1.78 | 3,191 | 202,161 | 1.36 |
| LINE/L1  |             | 119    | 9,447   | 0       | 0               | 0.00 | 0    | 0     | 0.00 |
| LINE/LTR |             | 22     | 2,550   | 13      | 1,521 | 59.09 | 16   | 851  | 1.23 |
| LINE/Penelope |       | 66     | 13,715  | 47      | 10,659 | 71.21 | 94   | 7,640 | 2.00 |
| SINE     |             | 7,175  | 858,040 | 2,607 | 376,356 | 36.33 | 3,620 | 189,066 | 1.39 |
| SINE     |             | 1,052  | 158,110 | 612    | 105,465 | 58.17 | 938  | 49,953 | 1.53 |
| SINE?    |             | 461    | 49,639  | 311    | 35,616 | 67.46 | 395  | 24,436 | 1.27 |
| SINE/Deu |             | 1,244  | 184,049 | 683    | 113,396 | 54.90 | 1,019 | 48,384 | 1.49 |
| SINE/MIR |             | 2,216  | 236,658 | 1,000  | 121,744 | 45.13 | 1,267 | 66,127 | 1.27 |
| SINE/tRNA-CR1 |       | 2,202  | 229,632 | 1     | 135  | 0.05 | 1    | 166  | 1.00 |
| DNA      |             | 3,288  | 495,231 | 2,110  | 326,482 | 64.17 | 3,125 | 213,896 | 1.48 |
| DNA      |             | 1,019  | 131,887 | 701    | 98,381 | 68.79 | 1,007 | 69,522 | 1.44 |
| DNA?     |             | 1,243  | 167,027 | 861    | 126,578 | 69.27 | 1,236 | 96,309 | 1.44 |
| DNA/hAT-Charlie |   | 166    | 53,330  | 1     | 993  | 0.60 | 1    | 26   | 1.00 |
| DNA/PiggyBac (?) |   | 476    | 75,038  | 262    | 46,343 | 55.04 | 406  | 19,934 | 1.55 |
| DNA/TcMar (?) |       | 150    | 30,013  | 129    | 26,895 | 86.00 | 240  | 17,131 | 1.86 |
| DNA/TcMar-Pogo |       | 60     | 7,430   | 38     | 4,996 | 63.33 | 45   | 1,724 | 1.18 |
| DNA/TcMar-Tigger |   | 174    | 30,506  | 118    | 22,296 | 67.82 | 190  | 9,250 | 1.61 |
| LTR      |             | 67,018 | 23,876,265 | 174 | 62,795 | 0.26 | 241  | 12,368 | 1.39 |
| LTR (?)  |             | 2,497  | 833,946 | 55     | 12,639 | 2.20 | 88   | 4,424 | 1.60 |
| LTR/ERV1 |             | 9,387  | 3,351,866 | 16   | 4,810  | 0.17 | 32   | 1,245 | 2.00 |
| LTR/ERVK |             | 15,045 | 5,604,321 | 15  | 4,121  | 0.10 | 18   | 616  | 1.20 |
| LTR/ERVL |             | 39,078 | 14,074,306 | 88  | 41,225 | 0.23 | 103  | 6,083 | 1.17 |
| LTR/ERVL (?) |       | 1,011  | 291,816 | 0     | 0    | 0.00 | 0    | 0    | 0.00 |
| All      |             | 209,730 | 60,477,606 | 7,298 | 1,336,301 | 3.48 | 10,287 | 625,982 | 1.41 |

TE superfamilies with "?" notation are uncertain classifications from REPEATMASKER.

| Lineage   | CEs | CEs (bp) | CNEs | CNEs (bp) | Proportion CNE (%) | Proportion TE-derived CEs (%) | Proportion TE-derived CNEs (%) |
|-----------|-----|----------|------|-----------|-------------------|-------------------------------|------------------------------|
| Sauropsida | 679,184 | 46,506,358 | 386,481 | 25,504,445 | 56.90 | 0.41 | 0.69 |
| Archosauria | 86,083 | 3,741,927 | 81,508 | 3,534,555 | 94.69 | 1.74 | 1.81 |
| Aves       | 119,923 | 5,849,300 | 115,194 | 5,613,845 | 96.06 | 1.41 | 1.45 |
| Neoaves    | 2,252 | 178,957 | 1,395 | 92,526 | 61.94 | 2.09 | 3.23 |
| Passerida  | 247 | 33,790 | 123 | 16,380 | 49.80 | 5.26 | 10.57 |
these families have contributed to the origin of functional sequence, such exaptation events are likely to be absent from the CE data due to a lack of power to detect conservation at this recent evolutionary timescale. In contrast, the superfamilies with the highest relative numbers of exaptations were ancient (Supporting information); for example, most SINE superfamilies exhibited high proportions of exapted TEs (45.1% – 67.5%, Table 2). These superfamilies contain the CORE-SINE families, copies of which are present in all amniote genomes and show no evidence of activity since the last common ancestor of birds (Kapusta & Suh, 2016).

3.7 | Lineage-specific conserved elements

Based on presence/absence of aligned bases, 679,184 CEs (53.0% of all CEs) were inferred as having an evolutionary origin before the emergence of the sauropod clade ~280 mya (Alfoldi et al., 2011), while 219,311 (17.1% of CEs) were defined as having an origin after the split between Testudines (turtles) and Archosauria (crocodilians and birds) ~257 mya (Wang et al., 2013; with the remaining CEs identifiable in at least one turtle genome, but not in the anole lizard genome). Of these lineage-specific CEs, 208,505 (95.1%) could be assigned to one of the four evolutionary branches considered (archosaur-, avian-, neoavian- and passeridan-specific) based on CE absence in at least three outgroup species (Table 3). Considering the CEs assigned to one of these four lineages, over 95% were CNEs, compared to 68.3% for all CEs and 56.9% for CEs present in the anole lizard genome. This result is broadly consistent with a higher rate of turnover of functional noncoding sequence relative to coding sequence (Mikkelsen et al., 2007; Rands et al., 2014).

We identified 41,820 CEs (~1.8 Mb) that were absent from the chicken genome. These presumably represent CE losses on the chicken lineage, or CE gains on the collared flycatcher lineage, after the split between Neoaves and Galloanserae. We consider these CEs as a conservative estimate of the amount of conserved sequence that would be missed by utilizing existing CE predictions from the chicken genome in a passerian species (e.g., using a LIPTOVER-like tool). Additionally, we further identified 2,252 neoavian-specific CEs and 247 passeridan-specific CEs. Genes associated with neoavian-specific CEs were enriched for GO terms related to regulatory functions, as well as terms involved specifically in the development of certain tissues/organs (e.g., the pituitary gland and pharynx; Table S6). None of the genes associated with passeridan-specific CEs were significantly enriched for GO terms; however, intriguingly several of the most significant terms prior to FDR correction were related to neural functions (Table S6). These regions may be interesting candidates for further research given the interest in songbird neuroscience (Warren et al., 2010).

4 | DISCUSSION

This study demonstrates how CEs can be identified and analysed in a nonmodel organism using a combination of comparative genomic, population genomic and transcriptomic approaches. We used progressive CACTUS alignments, which can efficiently extend the prediction of CEs to any species with an available genome sequence included in an alignment, in turn enabling the identification of CEs specific to previously unstudied lineages. In the following, we discuss specific aspects related to the identification and analysis of CEs in the collared flycatcher genome.

4.1 | Comparison with existing predictions of avian conserved elements

Two previous studies have used species-rich whole-genome alignments to predict CEs in the chicken genome, with Lowe et al. (2015) producing a 19-way vertebrate species alignment and predicting ~71 Mb of conserved sequence, and Zhang et al. (2014) producing a 48-way avian species alignment and predicting ~90 Mb of conserved sequence. While the total amount of conserved sequence is difficult to compare between studies due to differences in alignments and CE prediction methods, taken together with the ~74 Mb (and ~92 Mb using less conservative prediction parameters) reported herein, these studies suggest that birds have a lesser total amount of conserved sequence relative to mammals. For example, PHASTCONS predictions in human suggest there being ~112 Mb of conserved sequence (Lindblad-Toh et al., 2011; Miller et al., 2007). It should though be noted that birds exhibit a higher genomic coverage of CEs relative to mammals as a result of their smaller genome size (~6.5%–8.5% in birds vs. ~5% in mammals).

The difference in the estimated total amount of conserved sequence between these vertebrate lineages could potentially be explained by smaller avian genomes requiring less regulatory sequence involved with the organization of chromatic structure. Alternatively, there has been some evidence for substantial gene loss during avian evolution (Lovell et al., 2014; Meredith, Zhang, Gilbert, Jarvis, & Springer, 2014; Zhang et al., 2014), which would be expected to reduce the amount of conserved sequence as genic sequences comprise a large proportion of CEs (and intergenic CNEs specifically associated with a lost gene would in turn also be lost). However, the idea that birds would contain fewer genes than mammals has been questioned (Hron, Pajer, Paces, Bartunek, & Elleder, 2015) and may be an incorrect conclusion resulting from an under-representation of GC-rich genes in short-read avian assemblies (as there are inherent difficulties in sequencing and assembling GC-rich regions using Illumina sequencing, Bentley et al., 2008; Chaisson, Wilson, & Eichler, 2015). Moreover, while studies of mammalian CEs have been able to utilize the most complete vertebrate genome assemblies in human and mouse, as well as high-quality draft genomes for a number of other species (e.g., chimpanzee, rat, and dog), avian genome sequencing has a substantial challenge of assembling the GC-rich microchromosomes (numerous chromosomes smaller than 5–10 Mb). These are gene dense and so are expected to contribute disproportionately to the total density of functional sequence. All current chromosome-level avian assemblies lack several microchromosomes (including the most recent chicken genome
assembly, galGal5, Warren et al., 2017), and those microchromosomes that are assembled in the collared flycatcher exhibited reduced alignability (including in coding sequence) in this study. While we also scored unassembled scaffolds for CEs, these too showed considerably reduced alignability. Long-read assemblies have the potential to improve the quality of large and complex genomes (Gordon et al., 2016), and the further application of such approaches to avian species would thus be expected to lead to increases in the amount of assembled, and subsequently alignable, functional sequence. While such improvements are expected to be genomewide, the potential for increasing quality is perhaps greatest for microchromosomes.

It is also notable that promoters were the only exon-associated site class that contained fewer conserved bases than expected under a random distribution of CE coverage. This is at first glance surprising considering the high constraint observed in many human promoters regions (Lindblad-Toh et al., 2011), and existing evidence for conservation at transcription factor binding sites and short tandem repeats within avian promoters (Abe & Gemmell, 2016). A possible explanation for this result is that promotor sequences contain a high proportion of assembly gaps (21.8%) in the collared flycatcher assembly, potentially due to their high GC content. This again highlights the potential increases in resolution of CE prediction that could be achieved with improved genome assemblies. In the end, it cannot be excluded that the observation of birds having a lesser total amount of conserved sequence relative to mammals is mainly due to methodological issues.

4.2 Incorporating conserved noncoding elements when approximating genomic targets for selection

We found the proportion of noncoding elements among all CEs to be similar to that observed in mammals (~70%), where even in the human the majority of conserved noncoding bases remain functionally unannotated (Lindblad-Toh et al., 2011). Very little is known about the function of CNEs in avian genomes, with only one recent study utilizing chromatin immunoprecipitation sequencing (ChIP-seq) to investigate potential regulatory functions for avian-specific CNEs in chicken (Seki et al., 2017). Together with the observed reductions in nucleotide diversity of collared flycatcher CNEs, our data are consistent with CNEs evolving under purifying selection both over historical (conservation) and contemporary (reduced diversity) timescales. CNE density was negatively correlated with functional exonic density at local genomic scales, demonstrating that the latter alone is not an ideal proxy for the overall density of targets for selection. The collared flycatcher has been studied extensively in the field of speciation genomics (Burri et al., 2015; Ellegren et al., 2012; Nadachowska-Brzyska, Burri, Smeds, & Ellegren, 2016; Nadachowska-Brzyska et al., 2013; Nater, Burri, Kawakami, Smeds, & Ellegren, 2015), where it is becoming increasingly apparent that explicit modelling of the heterogeneity in genetic diversity is required in the interpretation of patterns of genetic differentiation between species, especially concerning the appearance of "divergence islands" (Cruickshank & Hahn, 2014; Wolf & Ellegren, 2017). Specifically, background selection in genomic regions with low rates of recombination and/or a high density of targets of selection will lead to reduced levels of genetic diversity and thereby increased levels of differentiation, mimicking signals from reproductive barrier loci and loci evolving under the influence of divergent selection. Incorporation of CNE density in modelling the expected effects of background selection on genetic diversity will therefore be important for future studies of both this and other species used in speciation genomics research.

4.3 Exaptation of transposable elements in avian evolution

Via exaptation events, TEs have been shown to be a major source of regulatory (and various other functional) elements in a wide array of species (Chuong et al., 2017). The availability of comprehensive repeat annotation for the collared flycatcher (Suh et al., 2017) allowed us to explore such putative exaptations for the first time in an avian species. In contrast to mammals, where TEs account for more than half of many species’ genomes (Sotero-Caio, Platt, Suh, & Ray, 2017), birds generally have far fewer TEs, with the number of copies ranging from 130,000 to 350,000 and genomic TE densities from 4.1% to 9.8% in most species so far investigated (Kaputa & Suh, 2016; Zhang et al., 2014). The current collared flycatcher genome assembly and annotation contains 209,730 TEs, spanning 5.4% of the genome. In comparison, a study of human TE exaptations by Lowe and Haussler (2012) included over 4 million TEs, spanning 44.9% of the genome. Additionally, birds have less diversity in TE superfamilies, with relatively few of the superfamilies prevalent in other amniotes present (Kaputa & Suh, 2016).

Lowe and Haussler (2012) identified 11% of human CNEs (~7 Mb) as putatively derived from TEs and inferred that as many as 20% may have been exapted by considering CNEs specific to the evolutionary branch between marsupials and eutherians. In contrast, using similar methods we found that <2% of collared flycatcher CNEs are TE-derived. Although it will be important to revisit such analyses once a sufficient number of long-read avian genome assemblies are available (as many TEs are expected to be missing from current assemblies, Gordon et al., 2016; Warren et al., 2017), our results for ancient TEs are not expected to be substantially biased as in most events only specific sections of individual TE copies are exapted (Lowe, Bejerano, & Haussler, 2007). With the remaining regions likely evolving neutrally, they would be expected to diverge relative to other copies of the parental TE, and thus be more amenable to assembly even when using short-read sequencing. Our results hence indicate that ancient TEs have not been a major evolutionary driver in the origin of novel functional sequence in birds to the same extent as in mammals. Exploring other vertebrate lineages for the presence of CEs derived from ancient TEs and sampling additional birds to increase power to detect CEs derived from more recently active TEs will be important directions for future research.
to investigate the generality of these contrasting results from birds and mammals.

Approximately a quarter of TE-derived CEs were present in the anole lizard genome and hence presumably predate the origin of the sauropsid clade ~280 mya (Alfoldi et al., 2011). This demonstrates that many TE-derived CEs have an ancient origin and have been conserved for a very long time, and is consistent with the finding that orthologous sequences from exapted TE s in the human genome can be identified in chicken (implying an origin before the common ancestor of all amniotes; Gentles et al., 2007) and that a small number have even been found to predate the origin of ray-finned fish (Lowe & Haussler, 2012). The majority of TE-derived CEs were, however, absent from the anole lizard genome, and over 3,000 were inferred to be specific to the archosaur and avian lineages. Taken together, these findings reinforce the view that the exaptation of TEs is an evolutionary ancient phenomenon, which has continued at least throughout the early evolutionary history of birds, albeit with reduced quantity to that observed in mammals.

### 4.4 Identification of lineage-specific conserved elements

We identified over 200,000 CEs specific to one of four evolutionary branches considered, namely archosaur-, avian-, neovian- and passeridan-specific CEs. Relative to CEs inferred to be present in the sauropsid common ancestor, a far higher proportion of these lineage-specific CEs were noncoding. Functional noncoding sequence is thought to have reduced pleiotropy relative to coding sequence (Carroll, 2008), and it follows that the loss of functional noncoding sequence should be less deleterious than that of coding sequence and is expected to be more common (Carroll, 2005). It has thus been hypothesized that the proportion of lineage-specific functional noncoding sequence is substantially higher than that of protein-coding sequence, with evidence for extensive turnover of functional noncoding sequence in mammalian evolution (Mikkelsen et al., 2007; Rands et al., 2014; Smith, Brandstrom, & Ellegren, 2004), and for loss events of hundreds of CNEs in specific mammalian lineages (Hiller, Schaar, & Bejerano, 2012). Presumably, many of the putative CE gains we have identified have not substantially increased the overall amount of conserved sequence due to loss events for other CEs; however, exploring the specific dynamics of such turnover would require a more thorough assessment of CE gains and losses than has been performed in this study.

We also identified ~2,500 CEs specific to the neovian and passeridan lineages. These CEs are expected to be only a subset of the total number of functional elements specific to these lineages, as only the longest and/or most conserved elements will have been detected based on the available power in these relatively short branches. For example, the proportions of lineage-specific CEs at these scales that are noncoding (~61.9% of neovian-specific elements are CNEs, 49.8% for passerid specific) is considerably lower than that observed for archosau and avian-specific CEs (94.7% and 96.1%, respectively). While this may reflect true biology, it seems more probable that this is a result of a bias towards identifying exonic CEs when the power to detect conservation is low. Generally, identifying functional sequence that has arisen at recent evolutionary timescales has been a major challenge for methods that rely on conservation between multiple species, due to a lack of power provided by the short branch lengths that connect the aligned species at these timescales (Ponting, 2008; Rands et al., 2014). However, if the genomes of many species that are closely related to the species of interest can be aligned, reasonable branch lengths for the prediction of CEs should also be achievable at relatively recent evolutionary timescales (Lindblad-Toh et al., 2011). For two reasons, passerines (and songbirds in particular) may represent a particularly attractive group to further address this issue in the future. First, the passerine lineage (order Passeriformes) diversified ~33–39 mya and contains ~60% of all extant avian species (Jarvis et al., 2014; Moyle et al., 2016); songbirds (oscine passerines) constitute a single lineage within the passerines and alone comprise the vast majority of these species, representing the largest radiation of birds (Barker, Cibois, Schikler, Feinsein, & Cracraft, 2004). With such extensive species richness, the songbirds represent an exceptional opportunity to achieve a relatively large total branch length at a recent evolutionary timescale. Second, passerine birds are studied extensively to address numerous questions in evolutionary biology and ecology, and gaining further insight into lineage-specific functional innovations should be expected to aid and develop many of these research areas.

The ability to predict lineage-specific elements in this study was derived from the flexibility of the PROGRESSIVE CACTUS 23 sauropsid whole-genome alignment produced previously by Green et al. (2014). Approximately 1.8 Mb of conserved sequence was identified in the collared flycatcher genome that is not present in the chicken genome, revealing regions putatively involved in functional innovations on the collared flycatcher lineage, or functional losses on the chicken lineage. Such loss events for CEs have recently enabled regulatory sequence to be mapped to putative phenotypes in mammals (Marcovitz et al., 2016), exemplifying the potential uses of identifying such candidate regions. The whole-genome alignment used in this study could be adapted to predict CEs explicitly in any of the other 22 included species, and lineage-specific CEs could be identified in any of the clades containing several species, including the Psittacidae (parrots), Crocoddilia and Testudines; three lineages that to our knowledge have not been scored for CEs previously. Such flexibility is therefore of great potential use in exploring evolutionary innovations in previously understudied groups, and we expect that PROGRESSIVE CACTUS whole-genome alignments will be frequently used in the future to extend the prediction of CEs to a wide range of nonmodel species, especially in well-studied groups such as mammals and birds where increasingly species-rich alignments are anticipated.

### 5 Conclusions

PROGRESSIVE CACTUS whole-genome alignments can be used to efficiently identify conserved elements in any organism included within
the alignment. Such an approach avoids biases introduced by reference-based alignments and enables the identification of elements specific to previously unstudied lineages. Our results, based on sequence conservation across avian species, as well as reduced diversity within a population, reinforce the finding that the use of coding or exonic sequences alone will fail to accurately capture the density and distribution of sites evolving under purifying selection within eukaryotic genomes. With >70% of conserved sites being nonexonic, these data will help to reveal the full genomic landscape of putatively functional sequence including identification of elements involved with lineage-specific adaptations.

**ACKNOWLEDGEMENTS**

We thank Severin Uebbing, Dominik R. Laetsch and Judith E. Risse for advice on genome annotation, and Carina Farah Mugal, Ludovic Dutoit, Linnéa Smeds and Douglas Scofield for providing scripts and advice on various analyses. This work was supported by the Swedish Research Council (grant number 2013-8271) and the Knut and Alice Wallenberg Foundation (Project Grant). Computational analyses were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) through the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

**DATA ACCESSIBILITY**

Genomic coordinates for all conserved elements are available in the supporting information FicAlb1.5_conservd_elements.tar.gz, and CDS/UTR and long noncoding RNA annotations are available in FicAlb1.5_genome_annotation.tar.gz.

**AUTHORS’ CONTRIBUTIONS**

H.E. conceived the study, R.J.C. performed bioinformatics analyses. A.S. participated in the analysis of transposon data and M.W. participated in the analysis of transcriptomic data. R.J.C. wrote the manuscript together with H.E.

**REFERENCES**

Abe, H., & Gemmell, N. J. (2016). Evolutionary footprints of short tandem repeats in avian promoters. Scientific Reports, 6, 19421. https://doi.org/10.1038/srep19421

Alder, J., Di Palma, F., Grabherr, M., Williams, C., Kong, L., Maucci, E., & Lindblad-Toh, K. (2011). The genome of the green anole lizard and a comparative analysis with birds and mammals. Nature, 477(7366), 587–591. https://doi.org/10.1038/nature10390

Bao, W., Kojima, K. K., & Kohany, O. (2015). Rebase update, a database of repetitive elements in eukaryotic genomes. Mobile DNA, 6, 11. https://doi.org/10.1186/s13100-015-0041-9

Barker, F. K., Cibois, A., Schikler, P., Feinstein, J., & Cracraft, J. (2004). Phylogeny and diversification of the largest avian radiation. Proceedings of the National Academy of Sciences of the United States of America, 101(30), 11040–11045. https://doi.org/10.1073/pnas.0401892101

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., & Haussler, D. (2004). Ultraconserved elements in the human genome. Science, 304(5675), 1321–1325. https://doi.org/10.1126/science.1098119

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., & Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature, 456 (7218), 53–59. https://doi.org/10.1038/nature07517

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F., Roskin, K. M., & Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. Genome Research, 14(4), 708–715. https://doi.org/10.1101/gr.1933104

Braasch, I., Gehrké, A. R., Smith, J. J., Kawasaki, K., Manousakis, T., Pasquer, J., & Postlethwait, J. H. (2016). The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nature Genetics, 48(4), 427–437. https://doi.org/10.1038/ng.3526

Burri, R., Nater, A., Kawakami, T., Mugal, C. F., Olason, P. I., Smeds, L., & Ellegren, H. (2015). Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. Genome Research, 25(11), 1656–1665. https://doi.org/10.1101/gr.196485.115

Carroll, S. B. (2005). Evolution at two levels: On genes and form. PLoS Biology, 3(7), e245. https://doi.org/10.1371/journal.pbio.0030245

Carroll, S. B. (2008). Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. Cell, 134(1), 25–36. https://doi.org/10.1016/j.cell.2008.06.030

Chaisson, M. J., Wilson, R. K., & Eichler, E. E. (2015). Genetic variation and the de novo assembly of human genomes. Nature Reviews Genetics, 16(11), 627–640. https://doi.org/10.1038/nrg3933

Chamary, J. V., Parmley, J. L., & Hurst, L. D. (2006). Hearing silence: Non-neutral evolution at synonymous sites in mammals. Nature Reviews Genetics, 7(2), 98–108. https://doi.org/10.1038/nrg1770

Chan, Y. F., Marks, M. E., Jones, F. C., Villarreal, G. Jr, Shapiro, M. D., Brady, S. D., & Kingsley, D. M. (2010). Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science, 327(5963), 302–305. https://doi.org/10.1126/science.1182213

Chuong, E. B., Elde, N. C., & Feschotte, C. (2017). Regulatory activities of transposable elements: From conflicts to benefits. Nature Reviews Genetics, 18, 71–86. https://doi.org/10.1038/nrg.2016.139

Cromer, J. M. (2014). Background selection as baseline for nucleotide variation across the Drosophila genome. PLoS Genetics, 10(6), e1004434. https://doi.org/10.1371/journal.pgen.1004434

Cooper, G. M., Brudno, M., Program, N. C. S., Green, E. D., Batzoglou, S., & Sidow, A. (2003). Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. Genome Research, 13(5), 813–820. https://doi.org/10.1101/gr.1064503

Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Molecular Ecology, 23(13), 3133–3157. https://doi.org/10.1111/mec.12796

Cutter, A. D., & Payseur, B. A. (2013). Genomic signatures of selection at linked sites: Unifying the disparity among species. Nature Reviews Genetics, 14(4), 262–274. https://doi.org/10.1038/nrg3425

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a high fraction of the human genome to
