G3 Genes | Genomes | Genetics

GENOME REPORT

A Highly Contiguous Genome for the Golden-Fronted Woodpecker (*Melanerpes aurifrons*) via Hybrid Oxford Nanopore and Short Read Assembly

Graham Wiley*[,1] and Matthew J. Miller[†,1,2]

*Clinical Genomics Center, Oklahoma Medical Research Foundation, Oklahoma City, Oklahoma and †Sam Noble Oklahoma Museum of Natural History and Department of Biology, University of Oklahoma, Norman, Oklahoma

ORCID IDs: 0000-0003-1757-6578 (G.W.); 0000-0002-2939-0239 (M.J.M.)

ABSTRACT Woodpeckers are found in nearly every part of the world and have been important for studies of biogeography, phylogeography, and macroecology. Woodpecker hybrid zones are often studied to understand the dynamics of introgression between bird species. Notably, woodpeckers are gaining attention for their enriched levels of transposable elements (TEs) relative to most other birds. This enrichment of TEs may have substantial effects on molecular evolution. However, comparative studies of woodpecker genomes are hindered by the fact that no high-contiguity genome exists for any woodpecker species. Using hybrid assembly methods combining long-read Oxford Nanopore and short-read Illumina sequencing data, we generated a highly contiguous genome assembly for the Golden-fronted Woodpecker (*Melanerpes aurifrons*). The final assembly is 1.31 Gb and comprises 441 contigs plus a full mitochondrial genome. Half of the assembly is represented by 28 contigs (contig L50), each of these contigs is at least 16 Mb in size (contig N50). High recovery (92.6%) of bird-specific BUSCO genes suggests our assembly is both relatively complete and relatively accurate. Over a quarter (25.8%) of the genome consists of repetitive elements, with 287 Mb (21.9%) of those elements assignable to the CR1 superfamily of transposable elements, the highest proportion of CR1 repeats reported for any bird genome to date. Our assembly should improve comparative studies of molecular evolution and genomics in woodpeckers and allies. Additionally, the sequencing and bioinformatic resources used to generate this assembly were relatively low-cost and should provide a direction for development of high-quality genomes for studies of animal biodiversity.

KEYWORDS
Piciformes
hybrid assembly
repetitive elements

Because of their near-global distribution, woodpeckers are often used as model systems in biogeographic and phylogeographic studies (Zink *et al.* 2002; Fuchs *et al.* 2006, 2007; Klicka *et al.* 2011; Navarro-Sigüenza *et al.* 2017), as well as macroecological studies (Bock and Bock 1974; Blackburn *et al.* 1998; Husak and Husak 2003; Ilsøe *et al.* 2017). In North America, four woodpecker hybrid zones have also been studied for insights into avian speciation. These include flickers (Giller 1959; Smith 1987; Moore 1987; Moore and Price 1993; Seneviratne *et al.* 2012, 2016), sapsuckers (*Sphyrapicus*; Giller 1959; Smith 1987; Moore 1987; Moore and Price 1993; Seneviratne *et al.* 2012, 2016), Nuttall's/Ladder-back woodpeckers (Manthey *et al.* 2019) and Red-bellied/Golden-fronted woodpeckers (Giller 1959; Smith 1987; Moore 1987; Moore and Price 1993; Seneviratne *et al.* 2012, 2016). More recently, woodpeckers have gained attention for the high amount of repetitive DNA found in their genomes relative to other bird taxa (Sotero-Caio *et al.* 2017; de Oliveira *et al.* 2017; Bertocchi *et al.* 2018), the result of high levels of genome-wide transposable elements (TEs), which are scarce in most bird genomes (Gao *et al.* 2017). Manthey *et al.* (2018) surveyed several woodpecker genomes, and found that TEs make up 17–31%

doi: https://doi.org/10.1534/g3.120.401059
Manuscript received January 9, 2020; accepted for publication April 17, 2020; published Early Online April 21, 2020.

Supplemental material available at figshare: https://doi.org/10.25387/g3.11952324.

[1] The authors contributed equally to the work.

[2] Corresponding author: Sam Noble Oklahoma Museum, 2401 Chautauqua Ave, Norman, OK 73072. E-mail: biodiversity@onehealth-genomics.org

of woodpecker genomes, compared to <10% for other bird species. Increasingly, researchers have suggested that TEs may play a critical role in driving avian evolution (Gao *et al.* 2017; Suh *et al.* 2018). However, only two woodpecker genomes (*Picoides* [*Dryocopus*] *pubescens*, Zhang *et al.* 2014; and *Dendrocopos noguchii*, GCA_004320165.1) have been published, limiting our ability to undertake comparative genomics and to understand the architecture of TE evolution in woodpeckers.

The genus *Melanperpes* represents the largest radiation of New World woodpeckers (Aves: Picidae; Navarro-Sigüenza *et al.* 2017). *Melanerpes* woodpeckers are found almost everywhere where forest occurs in the Americas. Various species range continuously from southern Canada to Argentina, with three species occurring in the West Indies. Members of the Red-bellied/Golden-fronted Woodpecker species complex (*Melanerpes aurifrons/carolinus*) are notable for the discord between plumage variation and phylogenetic structure, especially in Mexico and northern Mesoamerica, where various races show considerable plumage variation despite a lack of phylogeographic variation (García-Trejo *et al.* 2009; Navarro-Sigüenza *et al.* 2017; but see Barrowclough *et al.* (2018).

Here, we provide a high-quality, highly contiguous, reference genome for the Golden-fronted Woodpecker (*Melanerpes aurifrons*; Figure 1), which we expect to be foundational for research in avian hybrid zone dynamics, the evolution of reproductive isolation, and the role of TEs and other repetitive elements in driving molecular evolution and speciation in woodpeckers and other birds. Additionally, we believe our workflow will be of general interest to researchers looking to develop high quality reference genomes for non-model birds and other vertebrates. Our genome assembly was accomplished with only ∼46X coverage of long-read Oxford Nanopore and ∼52X coverage of short-read Illumina sequence. Furthermore, after troubleshooting the various steps in our workflow, we were able to go from raw data to completed assembly in less than two weeks, using a compute cluster with 72 CPU cores, 384 GB of RAM, and two NVIDIA Tesla P100 GPU accelerators. These sequencing and computational requirements should be fairly accessible to most research groups with modest budgets.

METHODS

Specimen collection and DNA extraction

We collected an adult female Golden-fronted Woodpecker (*Melanerpes aurifrons*) in March 2019 in Dickens, Texas. This specimen and all associated genomic resources have been archived as a museum voucher at the Sam Noble Oklahoma Museum of Natural History (specimen number 24340, tissue number SMB682). This specimen is registered with NCBI as Biosample SAMN13719207. Scientific collecting was done under the following permits: Texas Scientific Collecting Permit SPR-0916-229, US Migratory Bird Treaty Act permit MB02276C-0, and University of Oklahoma IACUC permit R15-016A. Approximately ten milliliters of whole blood were transferred immediately to blood tubes coated with EDTA as an anticoagulant and kept cool on wet ice for 6 hr. Upon return to the lab, blood was aliquoted to several tubes and stored at -20°. We extracted high molecular weight DNA from a single aliquot using the Nanobind CBB Big DNA Kit (Circulomics, Inc). We followed the manufacturer's high molecular weight (50–300+ kb) DNA protocol for cultured mammalian cells, with the exception that we diluted 10 uL of whole woodpecker blood into 190 uL of fresh PBS as sample input. This was done per recommendation of the manufacturer (personal communication) given that avian red blood cells are nucleated. We performed three rounds of HMW DNA extraction, with each extraction being used as input for an Oxford Nanopore sequencing library. For the initial MinION nanopore sequencing run, we treated the HMW DNA with Short Read Eliminator Kit (Circulomics, Inc.), following the manufacturer's instructions. For the PromethION sequencing runs, this kit was not used in order to maximize DNA concentration for downstream library construction. For Illumina sequencing, we performed a standard genomic DNA extraction using the Quick-DNA miniprep kit (Zymo Research). Again, 10 uL of blood was diluted into 190 uL of PBS before following the manufacturer's protocol.

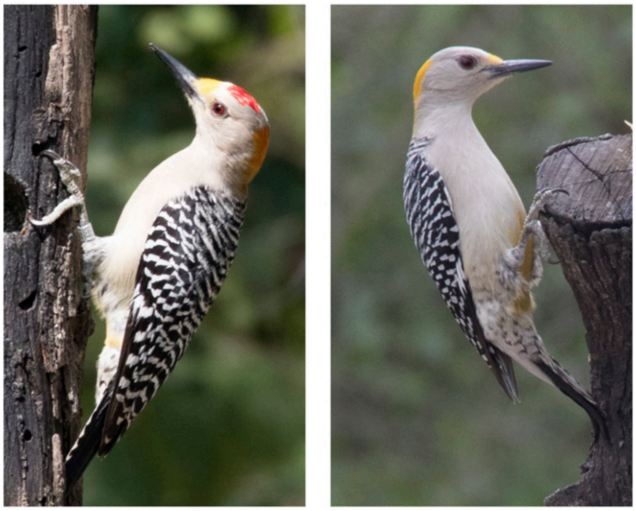

Figure 1 Male (left) and female (right) Golden-fronted Woodpecker (*Melanerpes aurifrons*). Photos by Bettina Arrigoni, cropped, and used under CC BY 2.0 license. Original photos available at: https://flickr.com/photos/69683857@N05/39849351035 and https://flickr.com/photos/69683857@N05/26752528708.

Whole genome library preparation and sequencing

We first generated Oxford Nanopore long reads in-house on the MinION device using the SQK-LSK109 library preparation kit followed by sequencing on a R9.4.1 flow cell per the manufacturer's instructions (Oxford Nanopore Technologies). After recovering less than 3 Gb of sequence data, we sent our remaining two HMW extracts to the UC Davis Core Lab to run on two PromethION flow cells. For the first PromethION run, a single library was prepared, loaded, and run for 48 hr. In an attempt to recover more sequence data, for the second PromethION flow cell, the HMW DNA was sheared using a Megaruptor (Diagenode, Inc.) set to a 50 kb target prior to a double library preparation allowing for a nuclease flush and a fresh library reload at the 24 hr mark of a 48 hr run. To generate the Illumina sequencing library, we used the NEBNext Ultra II FS DNA library kit (New England Biolabs, Inc); the initial enzymatic shearing step was accomplished via 10 min of incubation at 37°, after which we followed the manufacturer's instructions. The library was indexed using NEBNext Multiplex Oligos for Illumina unique dual index kit (New England Biolabs, Inc). This library was run on one lane of a NovaSeq S-Prime flow cell at the Oklahoma Medical Research Foundation Clinical Genomics Center.

Read QC and trimming

Raw Nanopore reads (fast5) were converted into fastq format using Oxford Nanopore's proprietary base-calling software guppy v.3.4

Table 1 Run statistics for three Oxford Nanopore sequencing runs

Sequencing Run	Reads	Median Read Length	Read Length N50	Median Read Qual	Total Bases
MinION Run	101,989	15,105	39,203	10.5	2.29 X 10^9
PromethION Run 1	1.77 X 10^6	9,478	30.742	10.6	27.43 X 10^9
PromethION Run 2	2.09 X 10^6	9,178	34,314	10.5	34.27 X 10^9

(https://community.nanoporetech.com). We evaluated Nanopore read quality in NanoPlot, and filtered reads ($Q > 7$, length $>$ 10,000) using NanoFilt; both are part of the NanoPack distribution (De Coster *et al.* 2018). We filtered and trimmed Illumina reads to remove adapters and low-quality bases using the standard settings in the bbduk program, which is part of BBMap v38.00 (http://sourceforge.net/projects/bbmap).

Genome size estimation

We used jellyfish v2.2.3 (Marçais and Kingsford 2011) to count the frequency of three distinct k-mers (17-mers, 26-mers, and 31-mers) in our trimmed Illumina sequencing reads. We plotted the resulting histogram file to establish the peak coverage at each k-mer. We then used the following formula from Liu *et al.* (2013) to estimate genome size:

$$G = N / D * L/(L - K + 1)$$

where G equals genome size, N equals the total number of base pairs in all reads, D equals the expected k-mer coverage depth, L equals the average read length, and K equals the k-mer size.

Genome assembly

We assembled an initial draft assembly using flye v2.6 (Lin *et al.* 2016; Kolmogorov *et al.* 2019). We corrected errors in that assembly using three iterations of racon v1.4.7 (Vaser *et al.* 2017), which uses the original nanopore long reads to correct the assembly consensus output. Subsequently, we ran two iterations of short-read polishing in pilon v1.23 (Walker *et al.* 2014). Pilon uses read alignment analysis to identify inconsistencies between a draft genome and an alignment of reads to that genome and attempts to correct base call errors, small and large indels and block substitutions, and to identify local mis-assemblies. We used minimap2 v2.17 (Li 2018) to map the Illumina short reads to our assembly to generate input alignments for pilon polishing. Finally, we removed redundant haploid and low coverage contigs from the polished assembly using purge_haplotigs v1.1 (Roach *et al.* 2018). That workflow begins by generating a read depth histogram to establish coverage cutoff levels for low coverage, high coverage, and the midpoint in coverage between haploid and diploid coverage. Subsequent steps remove contigs if they fall outside this range. To prevent interference from repeat elements, a bed file containing repetitive element coordinates was provided to purge_haplotigs (see below). At each step, we evaluated the completeness of the genome assembly via BUSCO v3 benchmarking (Simão *et al.* 2015) using the aves_odb9 dataset, which contains 4915 near-universal single-copy orthologs.

Identifying repetitive elements and annotating the mitochondrial genome

We estimated the portion of our genome comprised of repetitive elements using RepeatMasker v4.0.9p2. (http://www.repeatmasker.org/). To ensure avian-specific annotations, we utilized the 'chicken' species entry in the RepeatMasker database. The output of this was converted to a bed file for use in purge_haplotigs (see above). Using the Zebra Finch (*Taeniopygia guttata*), rather than chicken, resulted in less than 0.1% difference in the estimated repetitive elements frequencies reported by RepeatMasker.

To identify contigs corresponding to the mitochondrial genome, we performed BLASTn using an existing sequence of *M. aurifrons* NADH-dehydrogenase subunit II (MF766655) against our draft assembly, which recovered a single contig. Using that contig as a draft mitochondrial genome, we annotated all mitochondrial genes using the MitoAnnotator web server (http://mitofish.aori.u-tokyo.ac.jp/; Iwasaki *et al.* 2013). As an additional metric to evaluate the robustness of our assembly, we assessed the quality of the mtDNA assembly by comparing all protein coding gene sequences to an annotated mitochondrial genome of the Great Spotted Woodpecker (*Dendrocopos major*, NC028174.1) and we confirmed proper secondary structure for all tRNAs using the tRNAscan-SE 2.0 web server (http://lowelab.ucsc.edu/tRNAscan-SE/index.html; Lowe and Chan 2016). Finally, we performed a Google Scholar search to look for other avian genomes published in 2019 to compare our results to other high-quality wild bird reference genomes.

Reference assisted pseudomolecule scaffolding

We generated pseudomolecule chromosomal scaffolds using a reference genome assisted approach in RaGOO v1.1 (Alonge *et al.* 2019). RaGOO attempts to cluster, order, and orient assembly contigs based on a Minimap2 (Li 2018) alignment of those contigs to a reference genome. There is no chromosomal-level genome available for woodpeckers and allies (Order: Piciformes) and all bird species with chromosomal-level genomes are roughly similar in phylogenetic distance (Zhang *et al.* 2014). Therefore, we chose to use the Anna's Hummingbird chromosome-assembled genome (*Calypte anna* bCalAnn1_v1.p; GCA_003957555.2; Korlach *et al.* 2017) as our reference. To evaluate the validity of our reference-guided scaffold we measured genome synteny by aligning the RaGOO-scaffolded assembly to the Anna's Hummingbird reference genome as well as to two additional chromosomal-scale bird genomes: Zebra Finch (*Taeniopygia guttata*, GCA_008822105.1) and Kakapo (*Strigops habroptila*, *GCA_004027225.1*). We used the nucmer module in MUMMER v4.0.0b2 (Kurtz *et al.* 2004) to perform the alignments which were subsequently filtered using MUMMER's delta_filter module with many-to-many alignments allowed, minimum alignment length equal to 500, and minimum alignment identity equal to 80%. We filtered this output to only select mapped clusters equaling at least 3000 base pairs in the RaGOO (query) assembly. To visualize genome synteny we generated circos plots in Circa (http://omgenomics.com/circa); to improve visual clarity, only chromosomes 1 through 14 (including 4A, 4B and 5A) and the Z and W chromosomes were used.

Table 2 Results of genome size estimate using *k*-mer analysis

k-mer size	*k*-mer coverage	Estimated genome size
17	46	1.377 Gb
26	42	1.404 Gb
31	41	1.379 Gb

Table 3 Assembly contiguity statistics for the initial assembly and after multiple iterations of consensus correction

Stage	# of contigs	Assembly size	Max contig size	contig L50	contig L90	contig N50	contig N90
flye (no correction)	1519	1,353 Mb	42.8 Mb	33 fragments	136 contigs	14.5 Mb	1.4 Mb
racon 1	847	1344 Mb	46.8 Mb	29 contigs	117 contigs	15.9 Mb	1.5 Mb
racon 2	823	1343 Mb	46.8 Mb	29 contigs	117 contigs	15.9 Mb	1.5 Mb
racon 3	808	1343 Mb	46.8 Mb	29 contigs	117 contigs	15.9 Mb	1.5 Mb
pilon 2	808	1346 Mb	46.8 Mb	29 contigs	117 contigs	15.9 Mb	1.5 Mb
purge_haplotigs	441	1309 Mb	46.8 Mb	28 contigs	100 contigs	16 Mb	2.3 Mb

Data availability

The genome assembly and raw reads have been deposited to NCBI (project PRJNA598863; WWNC00000000). Output from RaGOO-scaffolded assembly is available from figshare. Code to replicate our analyses is available at: https://github.com/wileygenomics/Nanopore_Avian_Genome_Assembly. Supplemental material available at figshare: https://doi.org/10.25387/g3.11952324.

RESULTS AND DISCUSSION

Sequencing run statistics and genome assembly results

We generated a total of 3.9 M Oxford Nanopore reads for a total of 63.99 Gb of long read sequence data (Table 1). The initial MinION sequencing run produced only limited data but had longer reads than either the unsheared or sheared HMW runs on the PromethION (Table 1). Megaruptor shearing resulted in slightly lower median read length but increased read length N50. Comparing our two PromethION runs, HMW shearing, along with reloading a fresh library after the initial 24-hour runtime, increased the number of reads recovered by 18% and total base pairs yielded by 40%. These results, along with the slightly higher read length N50, suggest that users should consider HMW DNA shearing and a nuclease flush prior to Oxford Nanopore library preparation for metazoan whole genome assemblies.

Our initial flye assembly resulted in 1519 contigs which were organized into 1456 scaffolds with 41x mean coverage. The total length of the assembly was 1.35 Gb with a contig L50 of 14.5 Mb and a scaffold L50 of 15.9 Mb. The largest contig was 4.28 Mb. Fifty percent of the assembly was recovered in just 33 contigs (L50) – those contigs were at least 14.5 Mb in size (N50). Ninety percent of the assembly was recovered in 136 contigs (L90) all of which were at least 1.4 Mb in size (N90). In contrast, our final genome assembly after polishing and haplotig purging (see below) resulted in 441 unscaffolded contigs with an estimated genome size of 1.309 Gb, plus a complete circularized mitochondrial genome (N = 16,844). The largest contig was 46.8 Mb. Contig L50 was 28 contigs and contig N50 was 16 Mb.

Our Illumina sequencing generated a total of 490.7 M paired end reads for a total of 74.1 Gb of short read sequence data. After bbduk trimming, we retained 71.2 Gb of Illumina data from 487.8 M reads. Histograms for k-mer coverage (Supplementary Figure S1) suggest an average k-mer depth ranging from 46 (*k*-mer = 17 bp) to 41 (*k*-mer = 31 bp). This results in an estimated genome size of 1.38 – 1.40 Gb, depending on the *k*-mer value (Table 2). For resulting analyses, we consider 1.38 Gb to be the best estimate of genome size. This suggests that our assembly recovered 94.2% of the true Golden-fronted Woodpecker genome, whereas a recent review suggests that on-average, bird genomes generated with short reads alone often recover only 71–89% of the full genome (Peona *et al.* 2018). Given our genome size estimate, our Oxford Nanopore reads were at an average depth of 43.4, while our trimmed Illumina reads were at an average sequencing depth of 51.6.

Racon correction eliminates gaps and merges small contigs

Consensus correction using racon eliminated gaps in scaffolds, meaning that all recovered contigs were ungapped. Racon greatly reduced the number of contigs, but this was apparently mostly achieved by merging small contigs with overlapping regions from larger ones. Thus, racon had the effect of slightly reducing assembly size with only slight increases in contig L50, L90, N50, and N90 (Table 3). The largest effect of racon correction was observed in the first iteration, by the third round, little change was observed.

Polishing and haplotig purging reduce genome assembly errors

Whereas racon improved the contiguity of the assembly, it achieved only modest improvement in BUSCO benchmarking (Table 4). On the other hand, BUSCO scores greatly improved (76.5–92.6%) after pilon polishing. A second iteration of pilon resulted in only negligible improvement in BUSCO benchmarking. The improvement in BUSCO scores after polishing is likely due to correction of base-calling, indel, and local assembly errors that obscured identification of orthologs during AUGUSTUS annotation in the BUSCO pipeline. Haplotig purging had negligible impact on BUSCO scores, with a 0.2% reduction in duplications and a 0.1% increase in missing orthologs. Instead, purge_haplotigs identified 366 contigs as redundant, resulting in a more accurate genome representation while maintaining the overall genome assembly. Our final assembly recovered 92.6% of the BUSCO orthologs as complete.

Repetitive elements in the assembly

RepeatMasker estimated that 305 Mb (25.8%) of our assembly is comprised of repetitive elements. Nearly all of these elements (93.9%) – a total of 287 Mb (21.9% of the genome) – are part of the CR1 transposable element (TE) superfamily.

Birds have smaller genomes than most other tetrapods (Gregory *et al.* 2007), likely due to the metabolic cost of flight (Wright *et al.* 2014). The streamlining of avian genomes appears to be driven by a substantial reduction in the frequency of TEs in the genome (Gao *et al.* 2017). For most tetrapods, TEs represent a sizeable portion of the genome. For example, in mammals TEs typically account for 1/3

Table 4 BUSCO summarized benchmarking at various stages of genome assembly

Stage	Complete (Single, Duplicate)	Fragmented	Missing
flye (no correction)	76.2% (74.7%, 1.5%)	10.6%	13.2%
racon 3	76.5% (75.1%, 1.4%)	11.0%	12.5%
pilon 1	92.6% (90.8%, 1.8%)	4.5%	2.9%
pilon 2	92.7% (90.8%, 1.9%)	4.5%	2.8%
purge_haplotigs	92.6% (90.9%, 1.7%)	4.5%	2.9%

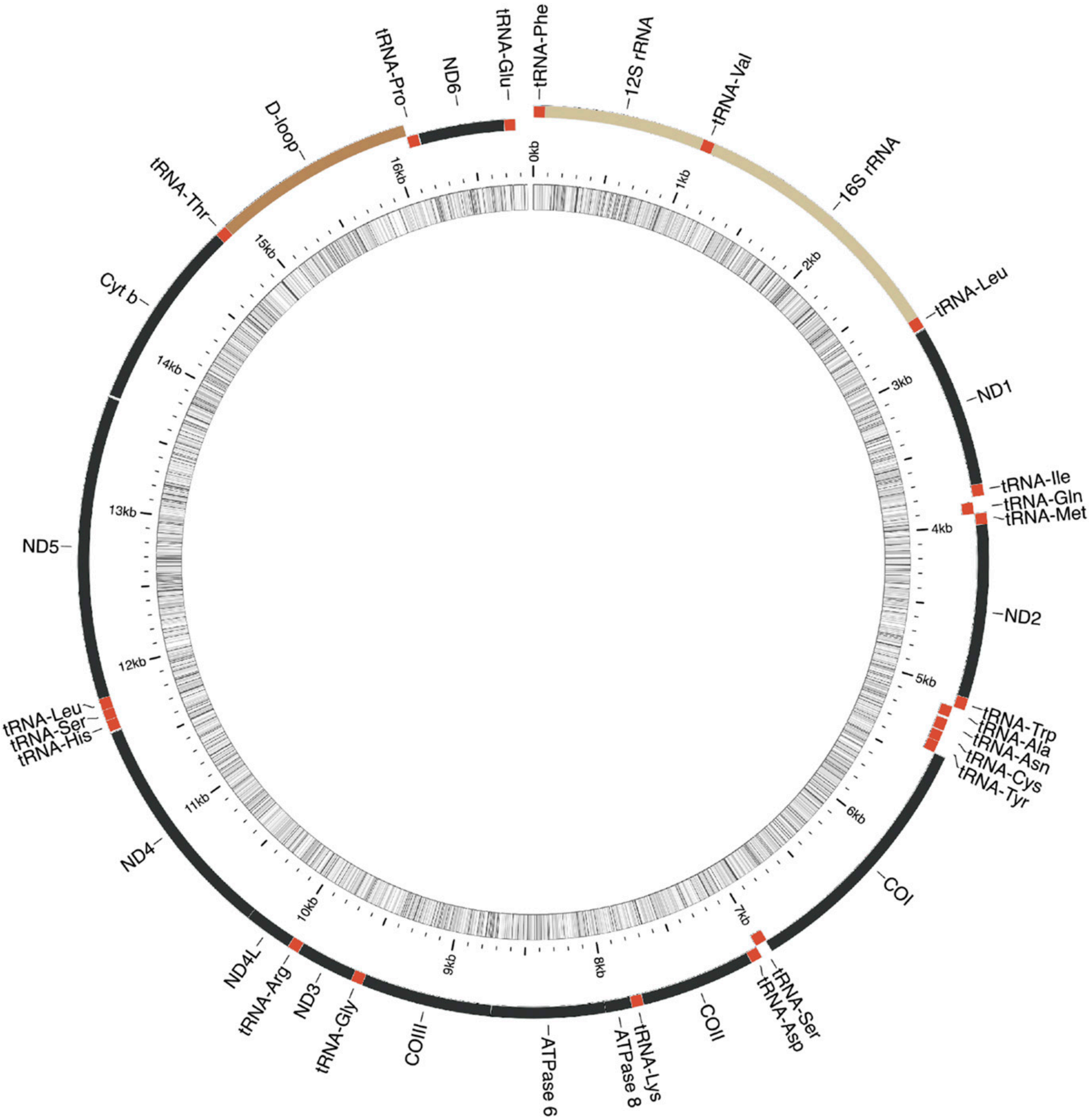

Figure 2 Circularized annotated mitochondrial genome assembly for the Golden-fronted Woodpecker (*Melanerpes aurifrons*). Figure generated by MitoAnnotator [36].

to 1/2 of the entire genome (Platt *et al.* 2018), while in birds' closest living relatives, the crocodiles, TEs represent about 1/3 of the genome (Green *et al.* 2014). In contrast, birds have substantially lower frequency of TEs – typically ranging from 4–10% of the total genome size (Sotero-Caio *et al.* 2017). Woodpeckers and allies are the exception. Manthey *et al.* (2018) demonstrated that TEs occurred at higher frequency across several genera of woodpeckers (17–30%), due principally to expansion of CR1 TEs (15–17% of total TEs). Here, CR1 TEs represent nearly 22% of the Golden-fronted Woodpecker genome. Manthey *et al.* (2018) also observed that their TE identification protocol may underestimate the true TE frequency by 2–5%, so a direct comparison should be considered tentative. However, our result does represent the highest CR1 proportion reported for any avian genome to date.

Mitogenome assembly and annotation results

The mitochondrial genome assembly (16,844 base pairs; Figure 2) recovered all 13 protein-coding genes, 12S and 16S ribosomal RNAs, and 22 tRNAs, all in the standard order observed in other woodpeckers and allies (Mindell *et al.* 1998; Eo 2017). Pairwise alignment of each of these features shows all features starting and ending coincident with homologous features in the Great Spotted Woodpecker mitochondrial

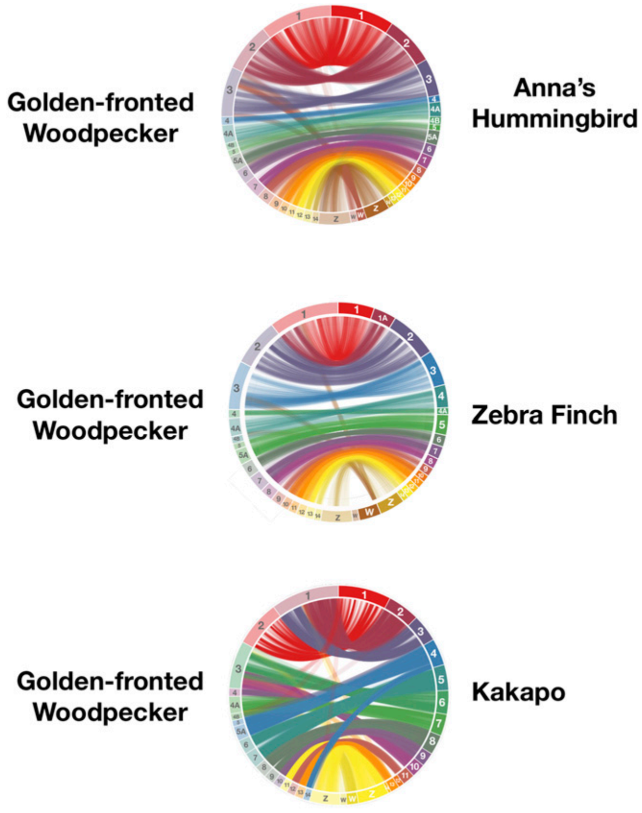
Golden-fronted Woodpecker
Anna's Hummingbird
Golden-fronted Woodpecker
Zebra Finch
Golden-fronted Woodpecker
Kakapo

Figure 3 Ideograms showing synteny between Golden-fronted Woodpecker pseudo-chromosomes and chromosomes from three different bird species.

genome, except in a small number of instances where our initial annotation failed to properly assess protein-coding genes with incomplete (-T) stop codons, which are common in vertebrate mitochondrial genomes (Ojala *et al.* 1981). In nearly all cases where this occurred, the adjacent 3` base would result in a valid stop codon, so this "error" seems to be a limitation of the MitoAnnotator algorithm rather than a measure of error in our mitogenome assembly. Our initial flye assembly recovered the mitochondrion as a single contig, with a sequence length of 16,858 bases, and 99.6% pairwise identity to the final polished mitochondrion sequence. Differences were due to homopolymer indel errors that resulted in frameshift errors. This comparison provides independent evidence to suggest that racon consensus correction plus pilon short read polishing is able to correct most errors in long read genome assemblies.

Synteny with other avian genomes and the usefulness of reference-based scaffolding

In general, synteny plots show that most of our Anna's Hummingbird guided reference-based scaffolded pseudo-chromosomes had largely one-to-one synteny plots with chromosomes of all three bird genomes evaluated (Figure 3). In some instances, the chromosomal numbers do not agree. For example, our pseudo-chromosome 5A maps almost exclusively to the Kakapo chromosome 4. This is likely to be simply the result of differences between chromosome numbering conventions. Both the Zebra Finch and Anna's Hummingbird synteny plot show that a portion of those birds' W chromosomes map to our Golden-fronted Woodpecker pseudo-chromosome 3, suggesting that this pseudo-molecule may be mis-assembled. We agree with Alonge *et al.* (2019) that reference-guided

Table 5 Comparison of various genome assembly statistics for recently published wild bird genomes. * N50 scaffold not reported for assemblies comprised only of ungapped contigs. ** Statistics only reported for scaffolded contigs

Species	Reference	Method	Assembly size	# of contigs/# of scaffolds	N50 contig	N50 scaffold	BUSCO complete	% Repeats
Pavo cristatus (Galliformes)	(Dhar et al. 2019)	Nanopore (low coverage) + short read	932 Mb	685,241/ 15,025	14.7 kb	0.23 Mb	Not reported	7.3%
Melospiza melodia (Passeriformes)	(Louha et al. 2019)	Paired end + HiC reads	978 Mb	Not reported	31.7 kb	5.6 Mb	87.5%	7.4%
Rhegmatorhina melanosticta (Passeriformes)	(Coelho et al. 2019)	Paired end + 10× linked reads	1.03 Gb	Not reported / 715	137 kb	3.3 Mb	89.2%	Not reported
Numida meleagris (Galliformes)	(Vignal et al. 2019)	Paired end + Mate pair	1.04Gb	Not reported / 2,739	234 kb	7.8 Mb	90.7%	19.5%
Colinus virginanus (Galliformes)	(Salter et al. 2019)	Paired end + HiC reads	866 Mb	Not reported **/ 1,512	Not reported **	66.8 Mb	90.8%	Not reported
Eremophila alpestris (Passeriformes)	(Mason et al. 2019)	Paired end + mate pair	1.04 Gb	Not reported ** / 2,708	Not reported**	10.6 Mb	94.5%	Not reported
Grus nigricollis (Gruiformes)	(Zhou et al. 2019)	Nanopore + paired end	1.33 Gb	1,837/ NA*	17.9 Mb	NA*	97.7%	8.1%
Melanerpes aurifrons (Piciformes)	This study	Nanopore + paired end	1.30 Gb/	441 / NA*	16.0 Mb	NA*	92.6%	25.8%

assemblies will be generally useful for comparative genomics studies of animal biodiversity, such as to map genome-wide phylogenetic markers, or to map the locations of structural variants and/or F_{st} outliers, etc. However, given the relatively low number of contigs in our assembly, researchers may prefer to do those analyses on the 441 unscaffolded contigs rather than on the pseudo-chromosomes.

Comparisons to recently published avian genomes

There is increasing interest in obtaining reference genomes for wild bird research. While a full chromosome level assembly would be ideal, many research questions would benefit from a highly contiguous but less-than-chromosome assembly. How should research groups with limited resources go about generating such genomes? Comparing assembly statistics for some recently published avian genomes (Table 5), we find that hybrid (PacBio or Oxford Nanopore long reads + Illumina short reads) assemblies are generally more contiguous than assemblies generated with only Illumina sequencing, including those with third generation Illumina libraries (*e.g.*, 10X and Hi-C) with our genome possibly being more complete than many other published genomes (Peona *et al.* 2018). As *de novo* genome assembly algorithms such as flye (Kolmogorov *et al.* 2019), RedBean (Ruan and Li 2020), Shasta (https://chanzuckerberg.github.io/shasta/), and Canu (Koren *et al.* 2017) that are capable with dealing with noisy Oxford Nanopore data continue to be developed, we expect to see more researchers pursuing similar, low overhead, genome assemblies for their own studies of animal biodiversity genomics.

ACKNOWLEDGMENTS

Funding for this research was provided by the Sam Noble Oklahoma Museum of Natural History. We also received a subsidy through the University of Oklahoma Consolidated Core Lab for Illumina sequencing which was undertaken at the Oklahoma Medical Research Foundation (OMRF) Clinical Genomics Center. Promethion library preparation and sequencing was carried out by Ruta Sahasrabudhe at the DNA Technologies and Expression Analysis Cores at the UC Davis Genome Center, supported by NIH Shared Instrumentation Grant 1S10OD010786-01. We thank OMRF for providing access to their computation cluster for data analysis. Scientific collecting was authorized by the Texas Parks and Wildlife Department and the United States Fish and Wildlife Service and supported by the Ornithology Department at the Sam Noble Museum. We thank them for their support of scientific collecting. We thank Joe Manthey for discussions of woodpecker genomics. We thank Brant Faircloth and an anonymous reviewer for comments that improved our manuscript.

LITERATURE CITED

Alonge, M., S. Soyk, S. Ramakrishnan, X. Wang, S. Goodwin *et al.*, 2019 RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biol. 20: 224. https://doi.org/10.1186/s13059-019-1829-6

Barrowclough, G. F., J. G. Groth, E. K. Bramlett, J. E. Lai, and W. M. Mauck, 2018 Phylogeography and geographic variation in the Red-bellied Woodpecker (*Melanerpes carolinus*): characterization of mtDNA and plumage hybrid zones. Wilson J Ornithol. 130: 671–683. https://doi.org/10.1676/17-070.1

Bertocchi, N. A., T. D. de Oliveira, A. Del Valle Garnero, R. L. B. Coan, R. J. Gunski *et al.*, 2018 Distribution of CR1-like transposable element in woodpeckers (Aves Piciformes): Z sex chromosomes can act as a refuge for transposable elements. Chromosome Res. 26: 333–343. https://doi.org/10.1007/s10577-018-9592-1

Blackburn, T. M., J. H. Lawton, and K. J. Gaston, 1998 Patterns in the geographic ranges of the world's woodpeckers. Ibis 140: 626–638. https://doi.org/10.1111/j.1474-919X.1998.tb04708.x

Bock, C. E., and J. H. Bock, 1974 On the geographical ecology and evolution of the Three-toed Woodpeckers, *Picoides tridactylus* and *P. arcticus*. Am. Midl. Nat. 92: 397–405. https://doi.org/10.2307/2424304

Coelho, L. A., L. J. Musher, and J. Cracraft, 2019 A multireference-based whole genome assembly for the obligate ant-following antbird, *Rhegmatorhina melanosticta* (Thamnophilidae). Diversity (Basel) 11: 144. https://doi.org/10.3390/d11090144

De Coster, W., S. D'Hert, D. T. Schultz, M. Cruts, and C. Van Broeckhoven, 2018 NanoPack: visualizing and processing long-read sequencing data. Bioinformatics 34: 2666–2669. https://doi.org/10.1093/bioinformatics/bty149

Dhar, R., A. Seethy, K. Pethusamy, S. Singh, V. Rohil *et al.*, 2019 De novo assembly of the Indian blue peacock (*Pavo cristatus*) genome using Oxford Nanopore technology and Illumina sequencing. Gigascience 8: giz038. https://doi.org/10.1093/gigascience/giz038

Eo, S. H., 2017 Complete mitochondrial genome of white-backed woodpecker *Dendrocopos leucotos* (Piciformes: Picidae) and its phylogenetic position. Mitochondrial DNA B Resour. 2: 451–452. https://doi.org/10.1080/23802359.2017.1357454

Fuchs, J., J. I. Ohlson, P. G. P. Ericson, and E. Pasquet, 2006 Molecular phylogeny and biogeographic history of the piculets (Piciformes: Picumninae). J. Avian Biol. 37: 487–496. https://doi.org/10.1111/j.0908-8857.2006.03768.x

Fuchs, J., J. I. Ohlson, P. G. P. Ericson, and E. Pasquet, 2007 Synchronous intercontinental splits between assemblages of woodpeckers suggested by molecular data. Zool. Scr. 36: 11–25. https://doi.org/10.1111/j.1463-6409.2006.00267.x

Gao, B., S. Wang, Y. Wang, D. Shen, S. Xue *et al.*, 2017 Low diversity, activity, and density of transposable elements in five avian genomes. Funct. Integr. Genomics 17: 427–439. https://doi.org/10.1007/s10142-017-0545-0

García-Trejo, E. A., A. E. De Los Monteros, M. D. C. Arizmendi, and A. G. Navarro-Siüenza, 2009 Molecular systematics of the Red-Bellied and Golden-Fronted Woodpeckers. Condor 111: 442–452. https://doi.org/10.1525/cond.2009.080017

Giller, B. S., 1959 Interspecific relations of woodpeckers in Texas. Wilson Bull. 71: 107–124.

Green, R. E., E. L. Braun, J. Armstrong, D. Earl, N. Nguyen *et al.*, 2014 Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. Science 346: 1254449. https://doi.org/10.1126/science.1254449

Gregory, T. R., J. A. Nicol, H. Tamm, B. Kullman, K. Kullman *et al.*, 2007 Eukaryotic genome size databases. Nucleic Acids Res. 35: D332–D338. https://doi.org/10.1093/nar/gkl828

Husak, M. S., and A. L. Husak, 2003 Latitudinal patterns in range sizes of New World woodpeckers. Southwest. Nat. 48: 61–69. https://doi.org/10.1894/0038-4909(2003)048<0061:LPIRSO>2.0.CO;2

Ilsøe, S. K., W. D. Kissling, J. Fjeldså, B. Sandel, and J.-C. Svenning, 2017 Global variation in woodpecker species richness shaped by tree availability. J. Biogeogr. 44: 1824–1835. https://doi.org/10.1111/jbi.13009

Iwasaki, W., T. Fukunaga, R. Isagozawa, K. Yamada, Y. Maeda *et al.*, 2013 MitoFish and MitoAnnotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline. Mol. Biol. Evol. 30: 2531–2540. https://doi.org/10.1093/molbev/mst141

Klicka, J., G. M. Spellman, K. Winker, V. Chua, and B. T. Smith, 2011 A phylogeographic and population genetic analysis of a widespread, sedentary North American Bird: the Hairy Woodpecker (*Picoides villosus*). Auk 128: 346–362. https://doi.org/10.1525/auk.2011.10264

Kolmogorov, M., J. Yuan, Y. Lin, and P. A. Pevzner, 2019 Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. 37: 540–546. https://doi.org/10.1038/s41587-019-0072-8

Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman *et al.*, 2017 Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27: 722–736. https://doi.org/10.1101/gr.215087.116

Korlach, J., G. Gedman, S. B. Kingan, C.-S. Chin, J. T. Howard *et al.*, 2017 De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. Gigascience 6: 1–16. https://doi.org/10.1093/gigascience/gix085

Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway *et al.*, 2004 Versatile and open software for comparing large genomes. Genome Biol. 5: R12. https://doi.org/10.1186/gb-2004-5-2-r12

Li, H., 2018 Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Lin, Y., J. Yuan, M. Kolmogorov, M. W. Shen, M. Chaisson *et al.*, 2016 Assembly of long error-prone reads using de Bruijn graphs. Proc. Natl. Acad. Sci. USA 113: E8396–E8405. https://doi.org/10.1073/pnas.1604560113

Liu, B., Y. Shi, J. Yuan, X. Hu, H. Zhang *et al.*, 2013 Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. arXiv. 1308.2012.

Louha, S., D. A. Ray, K. Winker, and T. Glenn, 2019 A high-quality genome assembly of the North American Song Sparrow, *Melospiza melodia*. G3 (Bethesda). 10: 1159–1166. https://doi.org/10.1534/g3.119.400929

Lowe, T. M., and P. P. Chan, 2016 tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res. 44: W54–W57. https://doi.org/10.1093/nar/gkw413

Manthey, J., S. Boissinot, and R. G. Moyle, 2019 Biodiversity genomics of North American *Dryobates* woodpeckers reveals little gene flow across the *D. nuttallii* x *D. scalaris* contact zone. Auk 136: ukz015. https://doi.org/10.1093/auk/ukz015

Manthey, J. D., R. G. Moyle, and S. Boissinot, 2018 Multiple and independent phases of transposable element amplification in the genomes of Piciformes (Woodpeckers and Allies). Genome Biol. Evol. 10: 1445–1456. https://doi.org/10.1093/gbe/evy105

Marçais, G., and C. Kingsford, 2011 A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27: 764–770. https://doi.org/10.1093/bioinformatics/btr011

Mason, N. A., P. Pulgarin, C. D. Cadena, and I. J. Lovette, 2019 Genome Report: De novo assembly of a high-quality reference genome for the Horned Lark (*Eremophila alpestris*). G3 (Bethesda) 10: 475–478. https://doi.org/10.1534/g3.119.400846

Mindell, D. P., M. D. Sorenson, and D. E. Dimcheff, 1998 Multiple independent origins of mitochondrial gene order in birds. Proc. Natl. Acad. Sci. USA 95: 10693–10697. https://doi.org/10.1073/pnas.95.18.10693

Moore, W. S., 1987 Random mating in the Northern Flicker hybrid zone: implications for the evolution of bright and contrasting plumage patterns in birds. Evolution 41: 539–546. https://doi.org/10.1111/j.1558-5646.1987.tb05824.x

Moore, W. S., and J. T. Price, 1993 Nature of selection in the Northern Flicker hybrid zone and its implications for speciation theory. pp. 196–225 in Hybrid zones and the evolutionary process. edited by Harrison R. G. Oxford: Oxford University Press.

Navarro-Sigüenza, A. G., H. Vázquez-Miranda, G. Hernández-Alonso, E. A. García-Trejo, and L. A. Sánchez-González, 2017 Complex biogeographic scenarios revealed in the diversification of the largest woodpecker radiation in the New World. Mol. Phylogenet. Evol. 112: 53–67. https://doi.org/10.1016/j.ympev.2017.04.013

Ojala, D., J. Montoya, and G. Attardi, 1981 tRNA punctuation model of RNA processing in human mitochondria. Nature 290: 470–474. https://doi.org/10.1038/290470a0

de Oliveira, T. D., R. Kretschmer, N. A. Bertocchi, T. M. Degrandi, E. H. C. de Oliveira *et al.*, 2017 Genomic organization of repetitive DNA in woodpeckers (Aves, Piciformes): Implications for karyotype and ZW sex chromosome differentiation. PLoS One 12: e0169987. https://doi.org/10.1371/journal.pone.0169987

Peona, V., M. H. Weissensteiner, and A. Suh, 2018 How complete are "complete" genome assemblies? -An avian perspective. Mol. Ecol. Resour. 18: 1188–1195. https://doi.org/10.1111/1755-0998.12933

Platt, 2nd, R. N., M. W. Vandewege, and D. A. Ray, 2018 Mammalian transposable elements and their impacts on genome evolution. Chromosome Res. 26: 25–43. https://doi.org/10.1007/s10577-017-9570-z

Roach, M. J., S. A. Schmidt, and A. R. Borneman, 2018 Purge haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics 19: 460. https://doi.org/10.1186/s12859-018-2485-7

Ruan, J., and H. Li, 2020 Fast and accurate long-read assembly with wtdbg2. Nat. Methods 17: 155–158. https://doi.org/10.1038/s41592-019-0669-3

Salter, J. F., O. Johnson, N. J. Stafford, W. F. Herrin, D. Schilling *et al.*, 2019 A highly contiguous reference genome for Northern Bobwhite (*Colinus virginianus*). G3 (Bethesda). 9: 3929–3932. https://doi.org/10.1534/g3.119.400609

Seneviratne, S. S., P. Davidson, K. Martin, and D. E. Irwin, 2016 Low levels of hybridization across two contact zones among three species of woodpeckers (*Sphyrapicus* sapsuckers). J. Avian Biol. 47: 887–898. https://doi.org/10.1111/jav.00946

Seneviratne, S. S., D. P. L. Toews, A. Brelsford, and D. E. Irwin, 2012 Concordance of genetic and phenotypic characters across a sapsucker hybrid zone. J. Avian Biol. 43: 119–130. https://doi.org/10.1111/j.1600-048X.2012.05516.x

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Smith, J. I., 1987 Evidence of hybridization between Red-Bellied and Golden-Fronted Woodpeckers. Condor 89: 377–386. https://doi.org/10.2307/1368491

Sotero-Caio, C. G., R. N. Platt, 2nd, A. Suh, and D. A. Ray, 2017 Evolution and diversity of transposable elements in vertebrate genomes. Genome Biol. Evol. 9: 161–177. https://doi.org/10.1093/gbe/evw264

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

Vaser, R., I. Sović, N. Nagarajan, and M. Šikić, 2017 Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27: 737–746. https://doi.org/10.1101/gr.214270.116

Vignal, A., S. Boitard, N. Thébault, G.-K. Dayo, V. Yapi-Gnaore *et al.*, 2019 A guinea fowl genome assembly provides new evidence on evolution following domestication and selection in galliformes. Mol. Ecol. Resour. 19: 997–1014. https://doi.org/10.1111/1755-0998.13017

Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel *et al.*, 2014 Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9: e112963. https://doi.org/10.1371/journal.pone.0112963

Wright, N. A., T. R. Gregory, and C. C. Witt, 2014 Metabolic "engines" of flight drive genome size reduction in birds. Proc. Biol. Sci. 281: 20132780. https://doi.org/10.1098/rspb.2013.2780

Zhang, G., C. Li, Q. Li, B. Li, D. M. Larkin *et al.*, 2014 Comparative genomics reveals insights into avian genome evolution and adaptation. Science 346: 1311–1320. https://doi.org/10.1126/science.1251385

Zhou, C., H. Yu, Y. Geng, W. Liu, S. Zheng *et al.*, 2019 A high-quality draft genome assembly of the Black-Necked Crane (*Grus nigricollis*) based on nanopore sequencing. Genome Biol. Evol. 11: 3332–3340.

Zink, R. M., S. V. Drovetski, and S. Rohwer, 2002 Phylogeographic patterns in the great spotted woodpecker *Dendrocopos major* across Eurasia. J. Avian Biol. 33: 175–178. https://doi.org/10.1034/j.1600-048X.2002.330208.x

Communicating editor: A. Sethuraman