Genome Report

A chromosome-scale reference genome for Eastern Black-eared Wheatear (*Oenanthe melanoleuca*)

Valentina Peona [1,*], Octavio Manuel Palacios-Gimenez [1,2,3*], Dave Lutgen [4,2,5,*], Remi André Olsen [6], Niloofar Alaei Kakhki [2], Pavlos Andriopoulos [7], Vasileios Bontzorlos [8], Manuel Schweizer [9,4], Alexander Suh [1,10], Reto Burri [5,4,3]

[1] Department of Organismal Biology – Systematic Biology, Science for Life Laboratory, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

[2] Department of Population Ecology, Institute of Ecology and Evolution, Friedrich Schiller University Jena, Jena, Germany

[3] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany

[4] Institute of Ecology and Evolution, University of Bern, Bern, Switzerland

[5] Swiss Ornithological Institute, Sempach, Switzerland

[6] Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University, Solna, Sweden

[7] Section of Ecology and Systematics, Department of Biology, National and Kapodistrian University of Athens, Athens, Greece

[8] Department of Forestry, Wood Sciences and Design, University of Thessaly, Karditsa, Greece

[9] Natural History Museum Bern, Bern, Switzerland

[10] School of Biological Sciences, University of East Anglia, Norwich, United Kingdom

* These authors contributed equally to the work.

Correspondence
Reto Burri, Swiss Ornithological Institute, Seerose 1, CH-6204 Sempach.
Email: reto.burri@vogelwarte.ch

Abstract

Pervasive parallel phenotypic evolution and in part high incidences of hybridization distinguish wheatears (songbirds of the genus *Oenanthe*) as a versatile system to address questions at the forefront of research on the molecular bases of phenotypic and species diversification. To prepare the genomic resources for this venture, we here assembled a chromosome-scale assembly of the Eastern Black-eared Wheatear (*O. melanoleuca*). This species is part of the *O. hispanica*-complex that is characterized by parallel evolution of plumage coloration and high rates of hybridization. The long-read-based male nuclear genome assembly comprises 1.04 Gb in 32 autosomes, the Z chromosome, and the mitogenome. The assembly is highly contiguous (contig N50: 12.6 Mb; scaffold N50: 70 Mb), with 96% of the genome assembled at chromosome level and 96% BUSCO completeness. The chromosome-scale reference genome of Eastern Black-eared Wheatear provides a crucial resource for research into the genomics of adaptation and speciation in an intriguing group of songbirds.

Keywords

Birds, open-habitat chats, *Oenanthe hispanica*-complex, repeat content, transposable elements

Introduction

Wheatears of the genus *Oenanthe* and their relatives – together referred to as "open-habitat chats" – are a group of songbirds that display several remarkable characteristics distinguishing them as a versatile system to address key questions on the evolution of phenotypes and formation of species. Many phenotypes, including multiple conspicuous colour ornaments, seasonal migration, and sexual dimorphism appear independently in multiple branches within open-habitat chats, suggesting a high incidence of parallel evolution (Aliabadian et al. 2012; Schweizer et al. 2019b). Furthermore, hybridization is observed in several species complexes and occurs at notably high rates in the *O. hispanica*-complex that consists of four currently recognized taxa (Schweizer et al. 2019a): Western Black-eared Wheatear (*O. hispanica*), Pied Wheatear (*O. pleschanka*), Cyprus Wheatear (*O. cypriaca*), and Eastern Black-eared Wheatear (*O. melanoleuca*; **Fig. 1**). Pied and Eastern Black-eared Wheatear hybridize pervasively at the western shores of the Black Sea, in the Caucasus, and in the Elbourz mountains of northern Iran (Haffer 1977; Panov 2005). The resulting introgression reaches beyond the hybrid zones (Schweizer et al. 2019a), and hybrid zones themselves sport admixed phenotypes that display combinations of plumage color phenotypes divergent between species (mantle and neck-side coloration) (Haffer 1977; Panov 2005). Finally, a phenotype divergently expressed between many wheatear species, black-or-white throat coloration, segregates as polymorphisms in three species of the *O. hispanica*-complex. This polymorphism and the recombination of mantle and neck-side coloration in hybrids provide an excellent opportunity to map these phenotypes to the genome (Buerkle and Lexer 2008) and study their parallel evolution across open-habitat chats. Furthermore, hybridization in several geographic regions enables insights into common or idiosyncratic patterns of evolution under hybridization (Gompert et al. 2017).

Here, we describe the *de novo* assembly and annotation of a chromosome-scale reference genome for the Eastern Black-eared Wheatear (*O. melanoleuca*). The assembly includes 32

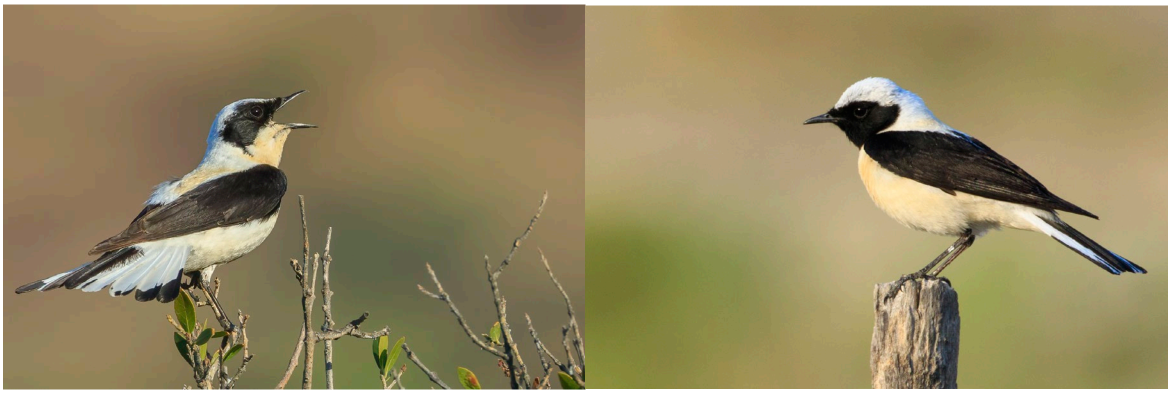

Figure 1. Eastern Black-eared Wheatear (*Oenanthe melanoleuca*). The species sports a white-throated (left; Agii Pantes, Greece, June 2022) and a black-throated phenotypes (right; Lesvos, Greece, May 2017) in males. © Reto Burri

autosomes, the Z chromosome, and the mitogenome that together cover 90 % of the k-mer-based genome size estimate (94 % with unplaced scaffolds included); it is highly contiguous with a scaffold N50 of 70 Mb and BUSCO completeness score of 96 %. This reference genome enables genomic research into the history of phenotypic and species diversification in wheatears and their close relatives.

Material and Methods

Sampling, tissue preservation, and nucleic acid extraction

To obtain optimal starting material for a reference individual, we freshly sampled a male Eastern Black-eared Wheatear (*Oenanthe melanoleuca*) in Galaxidi, Greece (sampling permit no. 181968/989, issued by the Ministry of Environment and Energy, General Secretariat of Environment, General Directorate of Forests and Forest Environment, Directorate of Forest Management, Department of Wildlife and Game Management; export permit no. 55980/1575, Regional CITES management authority Attika). For this purpose, we sampled about 100 µl of blood from the brachial vein, and, after euthanizing the bird, extracted all tissues possible. Tissues were immediately snap-frozen in liquid nitrogen. Throughout transportation and storage preceding DNA extraction, the samples were kept at a temperature below -80° C.

To obtain ultra-high molecular weight (UHMW) DNA from the reference individual, NGI Uppsala, Sweden, extracted DNA from the blood sample using the Bionano Prep™ Blood and Cell Culture DNA Isolation Kit (Binonano, San Diego, USA). Electrophoresis on a Femto Pulse instrument showed a mean DNA fragment length of about 200 kb, with fragments reaching up to 800 kb.

To prepare muscle tissue for Hi-C sequencing library preparation, we pulverized breast muscle tissue from the reference individual in a mortar. To avoid unfreezing of the tissue powder, the procedure was carried out in a climate chamber at 4°C under regular addition of liquid nitrogen.

De novo genome sequencing, and reference genome assembly

Assembly strategy and data acquisition

To obtain a chromosome-scale reference genome, our strategy largely followed the multiplatform approach recommended by Peona et al. (2021a). In brief, it consisted of (i) a phased primary assembly based on long reads, (ii) polishing and scaffolding of the primary assembly with linked-read sequencing data, and (iii) scaffolding of the secondary assembly with proximity ligation (Hi-C) information.

To this end, we obtained a total of 215 Gb (unique coverage 151 Gb) Pacific Biosciences (PacBio) long-read sequence data, 54 Gb linked-read sequence data, and 83 Gb Hi-C data. NGI Uppsala, Sweden, prepared a PacBio library from UHMW DNA and sequenced this library on 18 SMRT Cells 1M on a PacBio Sequel instrument. NGI Stockholm, Sweden, prepared a linked-read sequencing library using the 10X Genomics Chromium Genomic Kit (from the same DNA extraction as used for PacBio sequencing; 10X Genomics, Inc., Pleasanton, CA, USA; Cat No. 120215) and a Hi-C library the Dovetail Omni-C kit (Scotts Valley, CA, USA; Cat No. 21005). The linked-read and Hi-C libraries were prepared and sequenced on a NovaSeq 6000 instrument (S4 lane, 150 bp paired-end reads) at the facilities of NGI Stockholm, Sweden.

Genome size estimation

We estimated genome size by counting k-mer frequency of the quality-checked 10X Genomics linked reads. To this end, we first trimmed 22 bp from all 10X Genomics linked reads using fastp (Chen et al. 2018) to remove indices from R1 reads and keep symmetric read lengths for the R2 reads. We then counted k-mers of size 21 using jellyfish 2.2.10 (Marçais and Kingsford 2011) and used GenomeScope (Vurture et al. 2017) to estimate genome size from k-mer count histograms.

De novo genome assembly

We assembled the PacBio long reads into the phased primary assembly using the Falcon Unzip assembler (Chin et al. 2016), followed by polishing with Arrow. Before assembly polishing, we masked repeat regions of the phased primary assembly with RepeatMasker (Smit et al. 1996-2010) using a custom repeat library (Weissensteiner et al. 2020; Suh et al. 2018; Boman et al. 2019a; Peona et al. 2021c) to make accurate assembly corrections without overcorrecting large repeats. We then polished the masked assembly with two rounds in Pilon v1.22 (Walker et al. 2014) with the parameter "--fix indels" using the reference individual's linked-read data. To purge duplicate scaffolds from the assembly, we ran purge_dups (Guan et al. 2020) on the polished assembly. Prior to scaffolding with linked-read data, we split potential mis-assemblies with reference-individual linked-read data using Tigmint (Jackman et al. 2018). With the aim to scaffold the polished remaining contigs, we applied ARCS+LINKS using the reference individual's linked-read data using default parameters (Yeo et al. 2018; Warren et al. 2015).

To further scaffold the assembly, we applied the 3D-DNA pipeline (Dudchenko et al. 2017) to join the sequences into chromosomes. We first used Juicer v.1.6 (Durand et al. 2016) to map Hi-C data against the contigs and to filter reads, and then ran the asm-pipeline v.180922 to generate a draft scaffolding.

Finally, we corrected mis-assemblies based on the visual inspection of the proximity map using Juicebox (Robinson et al. 2018). The final chromosome-level assembly was polished with two additional rounds in Pilon as described above.

To assess homology of the assembled scaffolds with bird chromosomes, we aligned the final genome assembly to the genomes of collared flycatcher (*Ficedula albicollis*) (FicAlb1.5) (Kawakami et al. 2014), zebra finch (taeGut3.2.4) (Warren et al. 2010), and chicken (GRCg6a) (Bellott et al. 2017) using D-Genies (Cabanettes and Klopp 2018). Chromosomes were named according to homology with the latter three genomes. In cases, such as chicken chromosomes 1 and 4 that are split to multiple chromosomes in songbirds, the nomenclature in the wheatear genome was adapted to the species whose homologous chromosome matched closest.

Mitogenome assembly

To assemble the mitochondrial genome, we used the MitoFinder 1.4 (Allio et al. 2020) and mitoVGP 2.2 (Formenti et al. 2021) pipelines with the published *Oenanthe isabellina* mitochondrial genome (Genbank Accession Number: NC_040290.1) as a reference. We ran MitoFinder with the reference individual's short-read data (linked-read data but without making use of the linked-read haplotype information), and with mitoVGP we made joint use of the linked-read and long-read data. From MitoFinder we extracted the longest contig containing all 13 protein coding and two rRNA genes annotated by MitoFinder as mitogenome assembly. We annotated both assemblies using the MITOS WebServer (http://mitos2.bioinf.uni-leipzig.de/index.py).

We then aligned both resulting assemblies with the mitogenomes of Isabelline Wheatear (*O. isabellina*) (NC_040290.1) and Northern Wheatear (*O. oenanthe*) (MN356231.1) using MUSCLE (Edgar 2004) in MEGA X (Stecher et al. 2020) and generated a circular mitogenome map using CGView (Stothard and Wishart 2005).

Assembly quality evaluation

To evaluate assembly quality at each assembly step, we estimated basic assembly statistics using QUAST (Gurevich et al. 2013) and evaluated the completeness of expected gene content in the assembly based on benchmarking universal single-copy orthologs (BUSCO) (Simão et al. 2015) with the avian dataset aves_odb10 (8338 BUSCOs) in BUSCO 5.0.0.

Repeat annotation

The final version of the genome assembly was used to *de novo* characterize both interspersed and tandem repeats. For interspersed repeats, we used RepeatModeler2 (Flynn et al. 2020) with the option "-LTR_struct" to get an improved characterisation of LTR retrotransposons which are commonly found in avian genomes (Kapusta and Suh 2017; Peona et al. 2021b; Boman et al. 2019b). The resulting library of raw consensus sequences was filtered from consensus sequences of tandem repeats (for which we run a specific analysis; see below) and from protein-coding genes using the Snakemake pipeline repeatlib_filtering_workflow v0.1.0 (https://github.com/NBISweden/repeatlib_filtering_workflow)

For tandem repeats, we used RepeatExplorer2 (Novák et al. 2020) to search for satellite DNA (satDNA) sequences using the reference individual's 10X Genomics linked reads. Prior to RepeatExplorer2 graph-based clustering analysis, sequencing reads were preprocessed and checked by quality with FastQC (Babraham Bioinformatics: Cambridge 2012) using the public online platform at https://repeatexplorer.elixir-cerit-sc.cz. We processed the reads with the "quality trimming tool", "FASTQ interlacer on the paired end reads", "FASTQ to FASTQ converter", followed by "RepeatExplorer2 clustering" with default parameters. Each reference sequence assembled by RepeatExplorer2 consisted of a monomer of the satDNA consensus sequence. The relative genomic abundance and nucleotide divergence (Kimura-2-parameter distance) of each detected satDNA were estimated by sampling four million read pairs and aligning them to the satDNA library with RepeatMasker 4.1.2 (Smit et al. 1996-2010). The sampled reads were mapped to dimers of satDNA consensus sequences, and for smaller satDNAs, several monomers were concatenated until reaching roughly 150 bp array length. The resulting RepeatMasker.*align* file was then parsed to the script *calDivergenceFromAlign.pl* from RepeatMasker utils. The relative abundance of each satDNA sequence was then estimated as the proportion of nucleotides aligned with the reference sequence with respect to the total Illumina library size.

The RepeatModeler2 library was then merged with the satDNA library produced here and with known avian consensus sequences of transposable elements from Repbase (Bao et al. 2015), Dfam (Storer et al. 2021, 2021) and used to annotate the genome assembly with RepeatMasker 4.1.2 (Smit et al. 1996-2010). The annotation produced was processed with the script *calcDivergenceFromAlign.pl* from RepeatMasker utils to calculate the divergence between repeats and their consensus sequences using the Kimura 2-parameter distance.

Results and Discussion

Nuclear genome assembly

The polished, unzipped primary assembly contained a total of 1’681 contigs, of which all were >25 kb long and 1’610 >50 kb long (**Tab. 1**). Total assembly length was 1.29 Gb, with the longest contig spanning 45.3 Mb, contig N50 of 8.6 Mb and half of the assembly placed in 35 contigs. Avian BUSCOs were 96.9 % complete, with 90.6 % being single copy (**Tab. 1**).

Purging duplicated contigs resulted in an assembly constituted of 381 contigs with a total assembly length of 1.04 Gb, contig N50 of 13.5 Mb and half of the assembly placed in 23 contigs (**Tab. 1**). After this step, BUSCO completeness remained at 96.4 %, but an improvement to nearly 96 % single-copy BUSCOs was achieved (**Tab. 1**).

Starting from an already highly contiguous assembly, the linked-read data did not yield any scaffolding. Still, Tigmint detected several supposed mis-assemblies and split the assembly into 451 scaffolds. However, an alignment of the original contigs in D-Genies (Cabanettes and

Table 1. Assembly statistics.

		Falcon unzip, Arrow	+ Pilon, purge_dups	+ Tigmint	+ 3D DNA (all)	+ 3D DNA (chrom)
Basic stats	No. contigs/scaffolds*	1’681	381	383	588*	32*
	No. contigs/scaffolds* > 50 kb	1’610	347	348	143*	31*
	Assembly length (Gb)	1.29	1.04	1.04	1.04*	1.00*
	Contig/scaffold* N50 (Mb)	8.6	13.5	12.6	69.6*	69.7*
	Contig/scaffold* L50	35	23	24	6*	5*
	Largest contig/scaffold* (Mb)	45.3	45.3	45.3	148.4*	148.4*
BUSCO	Complete (%)	96.9	96.4	96.4	96.2	95.5
	Complete single-copy (%)	90.6	95.9	95.9	95.7	95.1
	Complete duplicated (%)	6.3	0.5	0.5	0.5	0.4
	Fragmented (%)	0.7	0.7	0.7	0.9	0.9
	Missing (%)	2.4	2.9	2.9	2.9	3.6

* Where numbers concern scaffolds instead of contigs, this is indicated by an asterisk.

Klopp 2018) showed that all but one of the original contigs (see below) were collinear with the collared flycatcher genome. Given this result and that the proximity ligation data would correct mis-assemblies in subsequent steps, we decided to keep the original contigs except for one aligning to flycatcher chromosomes 2 and 3. For the latter contig, we used the output of Tigmint that split the contig in line with the alignment. The two split parts covered all but 12’527 bp of the original contig. Visual inspection of the missing sequence showed that it almost entirely consisted of repeats. We left this sequence in the assembly as a separate contig.

The proximity ligation information obtained through Hi-C scaffolding corrected a number of scaffolds, resulting in a higher number of scaffolds (588) than the number of contigs it started from (383). However, the scaffolding yielded a highly contiguous chromosome-scale assembly (N50, 69.6 Mb; L50, 6) with BUSCO completeness of still >96 % and almost all BUSCOs in single copy (**Tab. 1**). This final assembly contained all macrochromosomes and the majority of microchromosomes usually found in the latest generation of avian genome assemblies (Peona et al. 2021a; Kapusta et al. 2017). 96 % of the assembly was placed into chromosome models, and the chromosomes-only assembly covered still 95.5 % of BUSCOs (**Tab. 1**).

The final assembly length closely matched the one of previous linked-read-based assemblies of the same species and closely related ones (Schweizer et al. 2019a; Lutgen et al. 2020). The genome size estimated from the k-mer distribution of linked reads sequence was between 1.105 and 1.106 Gb, with 0.925-0.926 Gb of unique sequence, 0.179-0.180 Gb (16 %) repeat sequence, and 0.75-0.76 % heterozygosity (GenomeScope model fit 98-99 %). The full final reference genome assembly thus covered 94 % of the k-mer-based genome size estimate, with 90 % of the estimated genome size placed in chromosomes. 96% of the assembly were placed in 33 chromosomes with homologs in flycatcher, zebra finch and chicken, according to which we adapted the chromosome nomenclature. The differences in genome size estimates based on the k-mer approach and the genome assembly length is likely the result of highly repetitive sequences (e.g. centromeres, telomeres, satDNAs) that collapsed during the assembly process (Peona et al. 2018).

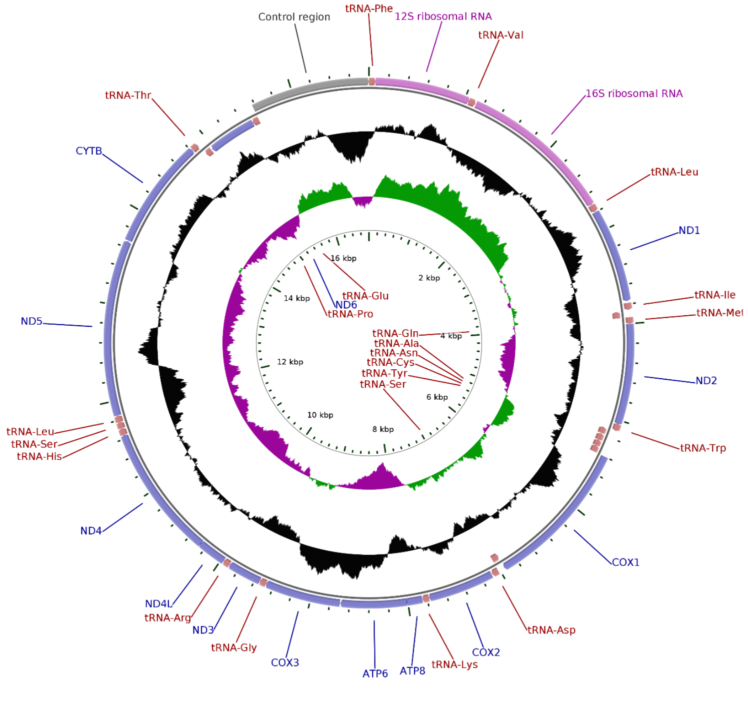

Figure 2. Mitogenome assembly.

Mitogenome assembly

MitoFinder and MitoVGP assembled mitogenomes of 16’937 bp and 18’631 bp length, respectively. The mitochondrial contigs assembled by the two pipelines were congruent, except for 9 single base pair mismatches, for a 1’827 bp long insert in the MitoVGP assembly and of a 141 bp long insert in the MitoFinder assembly, neither of which was observed either in the other assembly or in the mitogenomes of Isabelline and Northern Wheatear. Based on the specificity of the inserts in the respective assembly, on a strongly reduced short-read coverage in the insert regions (**Fig. S1**), and on the higher accuracy of short reads compared to long reads, we decided to remove the 147 bp insert from the MitoFinder assembly and included this curated mitogenome in the reference assembly.

The final assembled mitogenome (as also both original assemblies) contained all 13 protein-coding genes, 2 rRNAs, and 22 tRNAs (**Tab. S1**). All genes, except eight tRNAs and ND6, were located on the heavy DNA strand. Both gene order and strandedness were concordant with those observed in Northern Wheatear (*O. oenanthe*) (Wang et al. 2020).

Repetitive element annotation

The *de novo* identification of repetitive elements resulted in the characterisation of 572 raw consensus sequences from RepeatModeler2 and 16 satellite DNA consensus sequences from RepeatExplorer2. The consensus sequences from RepeatModeler2 were filtered from tandem repeats and protein-coding genes. This resulted in a final library of 462 consensus sequences (**Tab. S2**). Among these consensus sequences, RepeatModeler2 classified 226 sequences as LTR

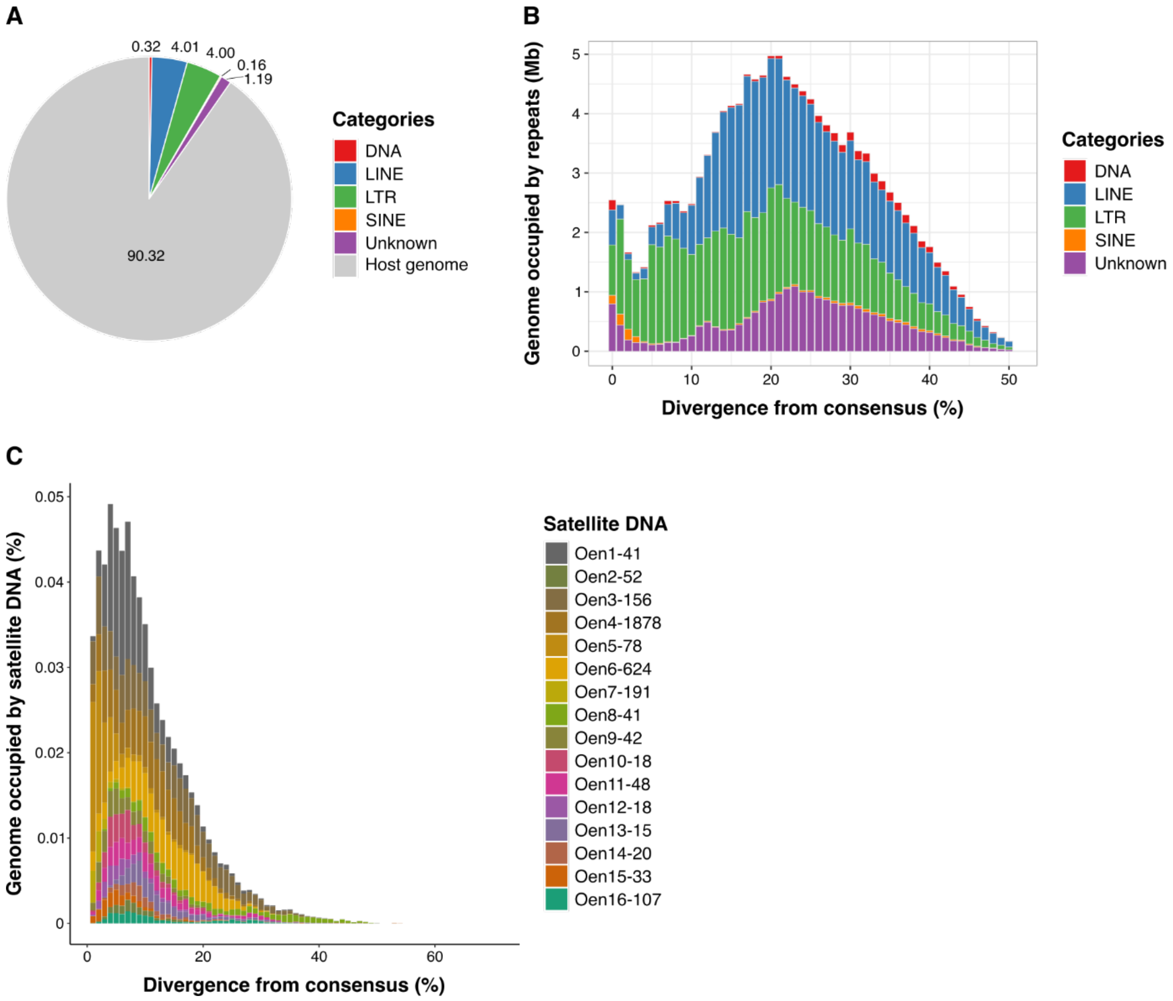

Figure 3. Repetitive element annotation landscapes. **A**) Pie-chart summarizing the transposable element content annotated in the genome assembly. **B**) Transposable element landscape. The divergence between interspersed repeat copies and their consensus sequences is shown on the X-axis as genetic distance calculated using the Kimura 2-parameter distance. The quantity of genome assembly occupied by transposable elements in Mb is shown on the Y-axis. **C**) Satellite DNA landscape. The divergence between the satellite DNA consensus sequences and sequences annotated in the short-read library is shown on the X-axis as genetic distance calculated using the Kimura 2-parameter distance. The percentage of genome (short reads) annotated as satellite DNA is shown on the Y-axis.

retrotransposons, 98 as LINE retrotransposons, 21 as DNA transposons, 5 as SINE retrotransposons, and 112 sequences remained unclassified (“unknown”).

The genome assembly annotation run with RepeatMasker using the repeat library produced here showed that ~10% of the assembled genome is repetitive (**Fig. 3A**, **Tab. S3** and **Tab. S4**). This finding indicates that many repeats collapsed during the genome assembly process. An example of this were satDNAs that represented ~0.8% of the sequenced reads but only < 0.3% of the genome assembly, suggesting that satDNA repeats (such as in (peri)centromeric and (sub)telomeric regions) are the most collapsed repeats. Most of the repeats annotated were LTR and LINE retrotransposons (**Fig. 3A**). While it is common to find abundant LINE retrotransposons in avian genomes (Manthey et al. 2018; Peona et al. 2021a; Kapusta and Suh 2017; Galbraith et al. 2021), it is less common to find so similar percentages of LINE and LTR retrotransposons. This is especially true for a male genome assembly such as the present one here that does not include the W chromosome which accumulates and acts as a refugium for most of the genomic LTR insertions in birds (Peona et al. 2021b; Warmuth et al. 2022). The transposable element landscape (**Fig. 3B**) suggests that LINE retrotransposons experienced a drop in their genomic accumulation in recent times (0-5% divergence; **Fig. 3B**), whereas LTR retrotransposons kept accumulating at the same rate. Such a recent replacement of LINE retrotransposon activity with a diversity of LTR retrotransposons has been noted in other songbirds and seems to have occurred independently in the so far analyzed passerine families, i.e., estrildid finches (Warren et al. 2010, Boman et al. 2019), flycatchers (Suh et al. 2018), crows (Weissensteiner et al. 2020), and birds-of-paradise (Peona et al. 2021a).

Data Availability

All data, including the assembly, its annotation, and the original sequencing data are available on the European Nucleotide Archive under project assession *XY (to be provided upon acceptance)*.

Acknowledgements

We warmly thank Marta Burri for preparing RNA, and Giulio Formenti, Remi Allio, and Lauren Coombe for support with computational questions. We are indebted to NGI Uppsala, namely Mai-Britt Mosbech and Olga Vinnere Pettersson for UHMW DNA extraction and long-read sequencing and Ignas Bunikis for running the primary assembly, as well as NGI Stockholm for the preparation of linked-read and Hi-C sequencing data. Finally, we thank the NGS platform at the University of Berne, namely Pamela Nicholson and Catia Coito for the preparation of IsoSeq data. Computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) through the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) and the High-Performance Computing Cluster EVE, a joint effort of the Helmholtz Centre for Environmental Research (UFZ) and the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig. We thank the administration and support staff of EVE: Thomas Schnicke and Ben Langenberg (UFZ), and Christian Krause (iDiv).

Conflict of Interest

The authors declare no conflict of interest.

Funder Information

The present project was supported by German Research Foundation (DFG), grant number BU3456/3-1 to RB and the National Research Fund (FNR) Luxembourg, grant number 14575729 to DL. OMPG was supported by the Swedish Research Council Vetenskapsrådet (grant number 2020-03866); and VP was supported via a grant to AS from the Swedish Research Council Vetenskapsrådet (grant number 2020-04436).

Publication bibliography

Aliabadian, Mansour; Kaboli, Mohammad; Förschler, Marc I.; Nijman, Vincent; Chamani, Atefeh; Tillier, Annie et al. (2012): Convergent evolution of morphological and ecological traits in the open-habitat chat complex (Aves, Muscicapidae: Saxicolinae). In *Molecular Phylogenetics and Evolution* 65 (2012), pp. 35–45.

Allio, Rémi; Schomaker-Bastos, Alex; Romiguier, Jonathan; Prosdocimi, Francisco; Nabholz, Benoit; Delsuc, Frédéric (2020): MitoFinder. Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. In *Molecular ecology resources* 20 (4), pp. 892–905. DOI: 10.1111/1755-0998.13160.

Babraham Bioinformatics: Cambridge (2012): FastQC; Version 0.10.1. A Quality Control Tool for High throughput Sequence Data; Babraham Bioinformatics. Cambridge, UK.

Bao, Weidong; Kojima, Kenji K.; Kohany, Oleksiy (2015): Repbase Update, a database of repetitive elements in eukaryotic genomes. In *Mobile DNA* 6, p. 11. DOI: 10.1186/s13100-015-0041-9.

Bellott, Daniel W.; Skaletsky, Helen; Cho, Ting-Jan; Brown, Laura; Locke, Devin; Chen, Nancy et al. (2017): Avian W and mammalian Y chromosomes convergently retained dosage-sensitive regulators. In *Nature genetics* 49 (3), pp. 387–394. DOI: 10.1038/ng.3778.

Boman, Jesper; Frankl-Vilches, Carolina; da Silva dos Santos, Michelly; de Oliveira, Edivaldo H. C.; Gahr, Manfred; Suh, Alexander (2019a): The Genome of Blue-Capped Cordon-Bleu Uncovers Hidden Diversity of LTR Retrotransposons in Zebra Finch. In *Genes* 10 (4), p. 301.

Boman, Jesper; Frankl-Vilches, Carolina; da Silva dos Santos, Michelly; Oliveira, Edivaldo H. C. de; Gahr, Manfred; Suh, Alexander (2019b): The Genome of Blue-Capped Cordon-Bleu Uncovers Hidden Diversity of LTR Retrotransposons in Zebra Finch. In *Genes* 10 (4). DOI: 10.3390/genes10040301.

Buerkle, C. Alex; Lexer, Christian (2008): Admixture as the basis for genetic mapping. In *Trends in Ecology & Evolution* 23 (12), pp. 686–694. DOI: 10.1016/j.tree.2008.07.008.

Cabanettes, Floréal; Klopp, Christophe (2018): D-GENIES. Dot plot large genomes in an interactive, efficient and simple way. In *PeerJ* 6, e4958. DOI: 10.7717/peerj.4958.

Chen, Shifu; Zhou, Yanqing; Chen, Yaru; Gu, Jia (2018): fastp. An ultra-fast all-in-one FASTQ preprocessor. In *Bioinformatics* 34 (17), i884-i890. DOI: 10.1093/bioinformatics/bty560.

Chin, Chen-Shan; Peluso, Paul; Sedlazeck, Fritz J.; Nattestad, Maria; Concepcion, Gregory T.; Clum, Alicia et al. (2016): Phased diploid genome assembly with single-molecule real-time sequencing. In *Nat Meth* 13 (12), pp. 1050–1054. DOI: 10.1038/nmeth.4035.

Dudchenko, Olga; Batra, Sanjit S.; Omer, Arina D.; Nyquist, Sarah K.; Hoeger, Marie; Durand, Neva C. et al. (2017): De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. In *Science (New York, N.Y.)* 356 (6333), pp. 92–95. DOI: 10.1126/science.aal3327.

Durand, Neva C.; Shamim, Muhammad S.; Machol, Ido; Rao, Suhas S. P.; Huntley, Miriam H.; Lander, Eric S.; Aiden, Erez Lieberman (2016): Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. In *Cell systems* 3 (1), pp. 95–98. DOI: 10.1016/j.cels.2016.07.002.

Edgar, Robert C. (2004): MUSCLE. multiple sequence alignment with high accuracy and high throughput. In *Nucl. Acids Res.* 32 (5), pp. 1792–1797. DOI: 10.1093/nar/gkh340.

Flynn, Jullien M.; Hubley, Robert; Goubert, Clément; Rosen, Jeb; Clark, Andrew G.; Feschotte, Cédric; Smit, Arian F. (2020): RepeatModeler2 for automated genomic discovery of transposable element families. In *Proceedings of the National Academy of Sciences of the United States of America* 117 (17), pp. 9451–9457. DOI: 10.1073/pnas.1921046117.

Formenti, Giulio; Rhie, Arang; Balacco, Jennifer; Haase, Bettina; Mountcastle, Jacquelyn; Fedrigo, Olivier et al. (2021): Complete vertebrate mitogenomes reveal widespread repeats and gene duplications. In *Genome Biol* 22 (1), p. 120. DOI: 10.1186/s13059-021-02336-9.

Galbraith, James D.; Kortschak, Robert Daniel; Suh, Alexander; Adelson, David L. (2021): Genome Stability Is in the Eye of the Beholder. CR1 Retrotransposon Activity Varies Significantly across Avian Diversity. In *Genome Biol Evol* 13 (12). DOI: 10.1093/gbe/evab259.

Gompert, Zachariah; Mandeville, Elizabeth G.; Buerkle, C. Alex (2017): Analysis of Population Genomic Data from Hybrid Zones. In *Annual Review of Ecology, Evolution, and Systematics* 48, pp. 207–229. DOI: 10.1146/annurev-ecolsys-110316-022652.

Guan, Dengfeng; McCarthy, Shane A.; Wood, Jonathan; Howe, Kerstin; Wang, Yadong; Durbin, Richard (2020): Identifying and removing haplotypic duplication in primary genome assemblies. In *Bioinformatics* 36 (9), pp. 2896–2898. DOI: 10.1093/bioinformatics/btaa025.

Gurevich, Alexey; Saveliev, Vladislav; Vyahhi, Nikolay; Tesler, Glenn (2013): QUAST. Quality assessment tool for genome assemblies. In *Bioinformatics* 29 (8), pp. 1072–1075. DOI: 10.1093/bioinformatics/btt086.

Haffer, Jürgen (1977): Secondary contact zones of birds in Northern Iran. In *Bonner Zoologische Monographien* (10), pp. 1–64.

Jackman, Shaun D.; Coombe, Lauren; Chu, Justin; Warren, Rene L.; Vandervalk, Benjamin P.; Yeo, Sarah et al. (2018): Tigmint. Correcting assembly errors using linked reads from large molecules. In *BMC bioinformatics* 19 (1), p. 393. DOI: 10.1038/nmeth.4035.

Kapusta, Aurélie; Suh, Alexander (2017): Evolution of bird genomes—a transposon's-eye view. In *Annals of the New York Academy of Sciences* 1389 (1), pp. 164–185. DOI: 10.1111/nyas.13295.

Kapusta, Aurélie; Suh, Alexander; Feschotte, Cédric (2017): Dynamics of genome size evolution in birds and mammals. In *Proceedings of the National Academy of Sciences of the United States of America* 114 (8), E1460-E1469.

Kawakami, Takeshi; Smeds, Linnéa; Backström, Niclas; Husby, Arild; Qvarnström, Anna; Mugal, Carina F. et al. (2014): A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. In *Molecular Ecology* 23 (16), pp. 4035–4058. DOI: 10.1111/mec.12810.

Lutgen, Dave; Ritter, Raphael; Olsen, Remi-André; Schielzeth, Holger; Gruselius, Joel; Ewels, Philip et al. (2020): Linked-read sequencing enables haplotype-resolved resequencing at population scale. In *Molecular Ecology Resources* 20 (5), pp. 1311–1322. DOI: 10.1111/1755-0998.13192.

Manthey, Joseph D.; Moyle, Robert G.; Boissinot, Stéphane (2018): Multiple and Independent Phases of Transposable Element Amplification in the Genomes of Piciformes (Woodpeckers and Allies). In *Genome Biol Evol* 10 (6), pp. 1445–1456. DOI: 10.1093/gbe/evy105.

Marçais, Guillaume; Kingsford, Carl (2011): A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. In *Bioinformatics* 27 (6), pp. 764–770. DOI: 10.1093/bioinformatics/btr011.

Novák, Petr; Neumann, Pavel; Macas, Jiří (2020): Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. In *Nature Protocols* 15 (11), pp. 3745–3776. DOI: 10.1038/s41596-020-0400-y.

Panov, E. N. (2005): Wheaters of the Palearctic. Ecology, Behaviour and Evolution of the genus *Oenanthe*. Sofia-Moscow: Pensoft (Series Faunistica).

Peona, Valentina; Blom, Mozes P. K.; Xu, Luohao; Burri, Reto; Sullivan, Shawn; Bunikis, Ignas et al. (2021a): Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. In *Molecular ecology resources* 21 (1), pp. 263–286. DOI: 10.1111/1755-0998.13252.

Peona, Valentina; Palacios-Gimenez, Octavio M.; Blommaert, Julie; Liu, Jing; Haryoko, Tri; Jønsson, Knud A. et al. (2021b): The avian W chromosome is a refugium for endogenous retroviruses with likely effects on female-biased mutational load and genetic incompatibilities. In *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 376 (1833), p. 20200186. DOI: 10.1098/rstb.2020.0186.

Peona, Valentina; Palacios-Gimenez, Octavio M.; Blommaert, Julie; Liu, Jing; Haryoko, Tri; Jønsson, Knud A. et al. (2021c): The avian W chromosome is a refugium for endogenous retroviruses with likely effects on female-biased mutational load and genetic incompatibilities. In *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 376 (1833), p. 20200186. DOI: 10.1098/rstb.2020.0186.

Peona, Valentina; Weissensteiner, Matthias H.; Suh, Alexander (2018): How complete are "complete" genome assemblies? —An avian perspective. In *Molecular Ecology Resources* 18 (6), pp. 1188–1195. DOI: 10.1111/1755-0998.12933.

Robinson, James T.; Turner, Douglass; Durand, Neva C.; Thorvaldsdóttir, Helga; Mesirov, Jill P.; Aiden, Erez Lieberman (2018): Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. In *Cell systems* 6 (2), 256-258.e1. DOI: 10.1016/j.cels.2018.01.001.

Schweizer, Manuel; Warmuth, Vera; Alaei Kakhki, Niloofar; Aliabadian, Mansour; Förschler, Marc I.; Shirihai, Hadoram et al. (2019a): Parallel plumage color evolution and pervasive hybridization in wheatears. In *Journal of Evolutionary Biology* 32 (1), pp. 100–110.

Schweizer, Manuel; Warmuth, Vera M.; Alaei Kakhki, Niloofar; Aliabadian, Mansour; Förschler, Marc; Shirihai, Hadoram et al. (2019b): Genome-wide evidence supports mitochondrial relationships and pervasive parallel phenotypic evolution in open-habitat chats. In *Molecular Phylogenetics and Evolution* 139, p. 106568.

Simão, Felipe A.; Waterhouse, Robert M.; Ioannidis, Panagiotis; Kriventseva, Evgenia V.; Zdobnov, Evgeny M. (2015): BUSCO. assessing genome assembly and annotation completeness with single-copy orthologs. In *Bioinformatics* 31 (19), pp. 3210–3212. DOI: 10.1093/bioinformatics/btv351.

Smit, A. F. A.; Hubley, R.; Green, P. (1996-2010): RepeatMasker Open 3.3.0.

Stecher, Glen; Tamura, Koichiro; Kumar, Sudhir (2020): Molecular Evolutionary Genetics Analysis (MEGA) for macOS. In *Mol Biol Evol* 37 (4), pp. 1237–1239. DOI: 10.1093/molbev/msz312.

Storer, Jessica; Hubley, Robert; Rosen, Jeb; Wheeler, Travis J.; Smit, Arian F. (2021): The Dfam community resource of transposable element families, sequence models, and genome annotations. In *Mobile DNA* 12 (1), p. 2. DOI: 10.1186/s13100-020-00230-y.

Stothard, Paul; Wishart, David S. (2005): Circular genome visualization and exploration using CGView. In *Bioinformatics* 21 (4), pp. 537–539. DOI: 10.1093/bioinformatics/bti054.

Suh, Alexander; Smeds, Linnéa; Ellegren, Hans (2018): Abundant recent activity of retrovirus-like retrotransposons within and among flycatcher species implies a rich source of structural variation in songbird genomes. In *Molecular Ecology*, in press. DOI: 10.1111/mec.14439.

Vurture, Gregory W.; Sedlazeck, Fritz J.; Nattestad, Maria; Underwood, Charles J.; Fang, Han; Gurtowski, James; Schatz, Michael C. (2017): GenomeScope. Fast reference-free genome profiling from short reads. In *Bioinformatics* 33 (14), pp. 2202–2204. DOI: 10.1093/bioinformatics/btx153.

Walker, Bruce J.; Abeel, Thomas; Shea, Terrance; Priest, Margaret; Abouelliel, Amr; Sakthikumar, Sharadha et al. (2014): Pilon. An integrated tool for comprehensive microbial variant detection and genome assembly improvement. In *PLoS ONE* 9 (11), e112963. DOI: 10.1371/journal.pone.0112963.

Wang, Erjia; Zhang, Dezhi; Braun, Markus Santhosh; Hotz-Wagenblatt, Agnes; Pärt, Tomas; Arlt, Debora et al. (2020): Can Mitogenomes of the Northern Wheatear (Oenanthe oenanthe) Reconstruct Its Phylogeography and Reveal the Origin of Migrant Birds? In *Sci Rep* 10 (1), p. 9290. DOI: 10.1038/s41598-020-66287-0.

Warmuth, Vera M.; Weissensteiner, Matthias H.; Wolf, Jochen B. W. (2022): Accumulation and ineffective silencing of transposable elements on an avian W Chromosome. In *Genome research* 32 (4), pp. 671–681. DOI: 10.1101/gr.275465.121.

Warren, René L.; Yang, Chen; Vandervalk, Benjamin P.; Behsaz, Bahar; Lagman, Albert; Jones, Steven J. M.; Birol, Inanç (2015): LINKS. Scalable, alignment-free scaffolding of draft genomes with long reads. In *GigaScience* 4, p. 35. DOI: 10.1186/s13742-015-0076-3.

Warren, W. C.; Clayton, D. F.; Ellegren, H.; Arnold, A. P.; Hillier, L. W.; Kunster, A. et al. (2010): The genome of a songbird. In *Nature* 464 (7289), pp. 757–762.

Weissensteiner, Matthias H.; Bunikis, Ignas; Catalán, Ana; Francoijs, Kees-Jan; Knief, Ulrich; Heim, Wieland et al. (2020): Discovery and population genomics of structural variation in a songbird genus. In *Nat Commun* 11 (1), p. 3403. DOI: 10.1038/s41467-020-17195-4.

Yeo, Sarah; Coombe, Lauren; Warren, René L.; Chu, Justin; Birol, Inanç (2018): ARCS. Scaffolding genome drafts with linked reads. In *Bioinformatics* 34 (5), pp. 725–731. DOI: 10.1093/bioinformatics/btx675.