scientific data

Check for updates

OPEN

DATA DESCRIPTOR

# A complete reference genome assembly and annotation of the Black Redstart (*Phoenicurus ochruros*)

Prashant Ghimire[1,4], Nan Wang[2,4] ✉, Yiai Hu[2] & Sangeet Lamichhaney[1,3] ✉

**The Black Redstart (*Phoenicurus ochruros*) is one of the most widely distributed species, occupying diverse habitats and exhibiting altitudinal migration (0–3700 m), making it suitable model for studying altitudinal migration and high-altitude adaptation. In this study, we present the first reference genome of *Phoenicurus ochruros*, generated using PacBio HiFi long-read sequencing. The nuclear genome is 1.37 Gb in length, assembled into 296 contigs with a contig $N_{50}$ of 29.9 Mb. Multiple complementary genome validation approaches - including BUSCO analysis, mapping of raw PacBio HiFi reads, and comparative genomics analysis - confirmed the high contiguity and gene-space completeness of the genome. The Black Redstart genome contains one of the highest proportions of transposable elements among passerines (30.58%), second only to Bell's Sparrow (*Artemisiospiza belli*), 31.2%. We further integrated RNA-seq data, protein homology evidence, and *ab initio* gene prediction to build high-quality genome annotation of 18,609 genes in the *P. ochruros* genome. This genomic resource provides invaluable resources for evolutionary genomics studies of passerines and genetic studies of high-altitude adaptation.**

## Background & Summary

Altitudinal migrant birds are species that move up and down along the elevation gradients - usually within the same geographic region - in response to seasonal changes in climate, food availability, or breeding conditions[1]. Unlike long-distance migrants that travel between continents or across hemispheres, altitudinal migrants move vertically between lowland and highland areas, often over relatively short horizontal distances. The majority of birds residing in mountain ecosystems undergo seasonal altitudinal migration, experiencing rapid changes in climate, habitat, and oxygen availability, which can shape their patterns of local adaptation and drive their population divergence[2]. This form of migration allows them to exploit seasonal resources while avoiding harsh environmental conditions. Notably, more than ~65% of bird species in the Himalayas are altitudinal migrants, and they offer a unique model system for studying patterns of species divergence, mechanisms of rapid physiological acclimation, and long-term genetic adaptation to both short- and long-term environmental changes[3].

One well-known example of altitudinal migrants are the Redstarts (Family: Muscicapidae, genus: *Phoenicurus*), which exhibit seasonal elevational shifts in the Himalayas, moving to higher elevations in summer for breeding and descending to lower elevations in winter (Fig. 1a). Redstarts are a group of 14 species of small Old-World flycatchers, native to Asia, Europe and Africa[4], known for their striking plumage and vocalization diversity[5].

Among them, the Black Redstart (*Phoenicurus ochruros*) is distinguished by its dark plumage and orange lower belly. It is also one of the most widely distributed species ranging from western Europe and North Africa to Asia[5] (Fig. 1b). This species inhabits in a variety of habitats, including shrublands, grasslands, rocky slopes up to the snow line, and urban areas near human settlements[5,6]. It typically breeds at higher elevations above the tree line (>4000 m) and migrates down to lower elevation during the non-breeding winter season (0–2500 m)[5].

[1]Department of Biological Sciences, Kent State University, Kent, Ohio, USA. [2]School of Ecology and Nature Conservation, Beijing Forestry University, Beijing, China. [3]School of Biomedical Sciences, Kent State University, Kent, OH, 44240, USA. [4]These authors contributed equally: Prashant Ghimire, Nan Wang. ✉e-mail: wangnan761227@bjfu.edu.cn; slamichh@kent.edu

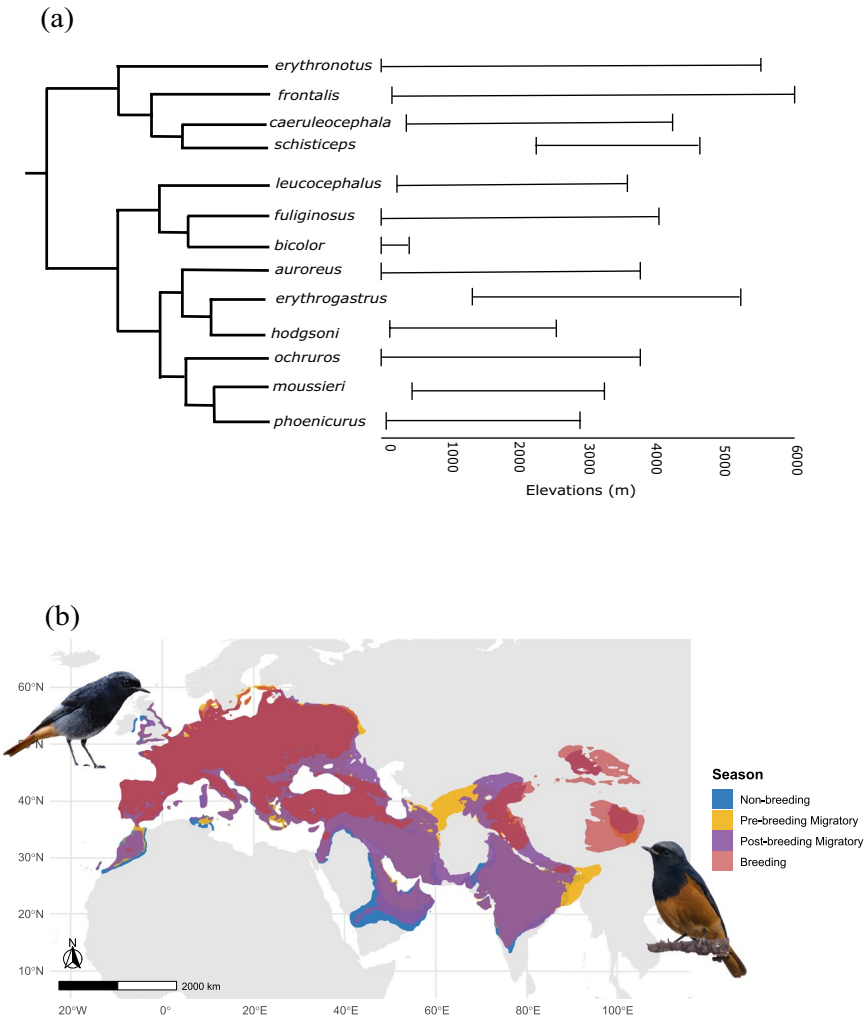


**Fig. 1** Redstarts (Genus: Phoenicurus) distribution across the altitudinal gradient (**a**). Phylogeny[44] and altitudinal range[45] show significant variation in the altitudinal distribution of redstarts **(b)** The seasonal distribution range of the Black Redstart *(Phoenicurus ochruros)* spans across Asia, Europe, and Africa[6]. The Eastern Black population (right; © Satdeep Gill, Wikimedia Commons) and Western Black population (left; © Rabiruivo comum, Wikimedia Commons) exhibit contrasting plumage patterns on their bellies, highlighting the geographical variation within the species[46].

Black Redstarts exhibit substantial morphological variation across its range, with eight subspecies currently recognized based on the variation in phenotypic traits[6] (Fig. 1b). Broadly, birds found in Europe and northern Africa with grey and white bellies are referred to as Western Black Redstarts, while those with orange bellies in Asia are known as Eastern Black Redstarts[6]. Although these birds show clear morphological differences, they are classified as a single species largely due to the presence of intermediate populations in the Middle East. These populations exhibit transitional trends in plumage, vocalizations, and mitochondrial DNA, suggesting gene flow across regions[7,8]. However, this taxonomic classification remains unresolved, due to the lack of comprehensive genomic studies on the species[6].

Genomic studies are essential to understand evolutionary patterns in Black Redstarts, as current subspecies designations rely on phenotypic traits such as plumage and vocalizations and limited molecular markers. These traits can be influenced by both genetic and environmental factors, making it difficult to infer evolutionary relationships without genome-wide data. The presence of intermediate populations suggests possible gene flow between Eastern and Western populations, and only genomic data can characterize the magnitude and direction of this gene flow, helping to determine whether these forms represent (a) distinct evolutionary lineages (potentially cryptic species), (b) recently diverged populations still connected by hybrid zones, or (c) subspecies undergoing incomplete lineage sorting. Moving beyond taxonomy, genomic analyses can identify patterns of selection, identify regions of the genome associated with phenotypic divergence between populations, and reveal how historical processes like glaciations or range shifts have shaped genetic structure across their distribution range. Most importantly, Black Redstarts offer a promising model system for investigating the genetic basis of altitudinal migration. Their seasonal movements are associated with rapid changes in temperature, oxygen availability, and resource distribution - conditions that are known to impose strong selective pressures on metabolic,

cardiovascular, and behavioral traits[9,10]. By studying the genomes of altitudinal migrants like Black redstarts, we can identify genetic variants and regulatory pathways linked to altitude-related physiological adaptations, such as hypoxia tolerance, energy metabolism and timing of migration. In addition, comparing genomes of Black redstarts with high altitude resident species such as Tibetan partridge *Perdix hodgsoniae*[11] will help us to assess whether molecular convergence underlie high-altitude adaptations or whether short term exposure shape evolution of distinct genetic pathways in species with diverse life history traits.

In this study, we present the first high-quality genome assembly and annotation of the Black Redstart, generated using a combination of long-read PacBio HiFi sequencing and Illumina short-read RNA sequencing. Our efforts resulted in the successful assembly of both the nuclear and mitochondrial genomes, marking the first time a reference genome has been produced for any species in the genus *Phoenicurus*. This reference genome provides a valuable genomic resource for future analyses of population demographic history and biogeographic patterns and facilitates studies on the genetic basis of high-altitude adaptation. The large distribution range of Black redstarts across various mountain ranges allow us to study how geography and climate shape population divergence. Furthermore, it lays the groundwork for comparative genomics studies across the Muscicapidae family, advancing our understanding of avian diversification, adaptation, and migration at both macro- and micro-evolutionary scales.

## Methods

**Sample collection.** An adult female Black Redstart *(Phoenicurus ochruros)* was captured using mist nets at an elevation of 3,000 meters in Qinghai Province, China, in the Qinghai-Tibetan Plateau (QTP). Blood sample was collected via brachial venipuncture and immediately stored in RNAlater until further processing for DNA extraction. Following blood sampling, the bird was humanely euthanized following institutional animal care guidelines. The individual was then dissected, and tissue samples were collected from the brain, liver, lungs, gonads, spleen, kidney, heart, and pectoral muscle. All tissues were preserved in RNAlater for subsequent RNA extraction. All procedures were conducted under approved animal care protocols and valid collection permits.

**DNA extraction and genome sequencing.** Genomic DNA was extracted from tissue samples using the DNeasy Blood & Tissue Kit (Qiagen, Valencia, CA, USA) following the manufacturer's protocol. DNA quality and quantity were assessed using spectrophotometry and gel electrophoresis to ensure high molecular weight suitable for long-read sequencing. Long-read whole-genome sequencing was performed on the PacBio Revio platform, which generated highly accurate HiFi (high-fidelity) reads, enabling robust assembly of complex genomic regions. We generated a total of ~93 Gb of raw PacBio HiFi long reads with an average read length of 17.39 Kb and read N50 of 17.41 kb (Fig. 2a).

**Genome assembly and quality assessment.** We assembled raw PacBio HiFi long reads using HiFiasm (v.0.24.0) - a fast and accurate assembler optimized for PacBio HiFi data[12]. We used the default parameters in HiFiasm with built-in duplication purging. The resulting *de-novo* genome assembly of *P. ochruros* was 1.37 Gb in total length, consisting of 296 contigs, with the Contig $N_{50}$ of 29.9 Mb and a longest contig of length of 113.3 Mb (Table 1). These results indicated a high level of genome contiguity. Furthermore, the completeness of genome assembly was evaluated using BUSCO (v.5.7.1)[13] against the passeriformes_odb10 (2019-11-20) database. The results revealed 96.80% of BUSCOs were successfully detected, including 96.10% complete and single copy, and 0.7% of duplicated BUSCOs (Table 1, Figs. 2b, 3a), indicating the high level of gene-space completeness. In addition, we mapped the raw HiFi reads back to the assembled genome using Minimap2 (v.2.21)[14], and found that 100% of the reads successfully mapped, further supporting the high completeness and accuracy of the genome assembly.

**Mitochondrial genome assembly and annotation.** We assembled the mitochondrial genome using the MitoHiFi v3.0.0 pipeline[15], which is optimized for extracting and assembling mitochondrial sequences from PacBio HiFi reads. We applied the parameter setting of "-p 90", which retained contigs with 90% or more of their length matching a reference mitochondrial sequence in BLAST searches. As a reference, we used the published mitochondrial genome of the closely related Daurian Redstart *(Phoenicurus auroreus)* (GenBank accession: MT366880.1[16]). We chose the mitochondrial genome of the Daurian Redstart *(Phoenicurus auroreus)* as a reference because it is the closest relative of the Black Redstart *(P. ochruros)* with a high-quality, publicly available mitochondrial genome in existing databases. Although a mitogenome for the Blue-fronted Redstart is also available, it has been reported to contain chimeric sequences, which led us use Daurian Redstart mtDNA genome[17] for this study. This approach enabled the successful assembly of a complete mitochondrial genome for *P. ochruros*. The final draft mitochondrial genome assembly of the *P. ochruros* was 18,613 bp in length and includes 13 protein-coding genes, 22 tRNA genes, and two ribosomal RNA genes, consistent with the typical structure of vertebrate mitogenomes (Fig. 3b).

**Repeat annotation and masking.** Before predicting protein-coding genes in the genome, repeat annotation was conducted using a combination of *de-novo* and homology based approaches to identify and mask repetitive elements in the genome. We first used RepeatModeler (v.2.0.5)[18] to build a *de-novo* custom-built species-specific repeat library generated for *P. ochruros* to increase the accuracy of detection and annotation of transposable elements. We used the NCBI BLAST engine for sequence comparisons during repeat classification. Generated custom repeat library was then merged with the reference repeat library downloaded from the Repbase database (release 20190301)[19] for identification and masking of transposable elements in the *P. ochruros* genome using RepeatMasker (v.4.1.6)[20]. Combining customs and known libraries improved the sensitivity and accuracy of repeat annotation across the genome.

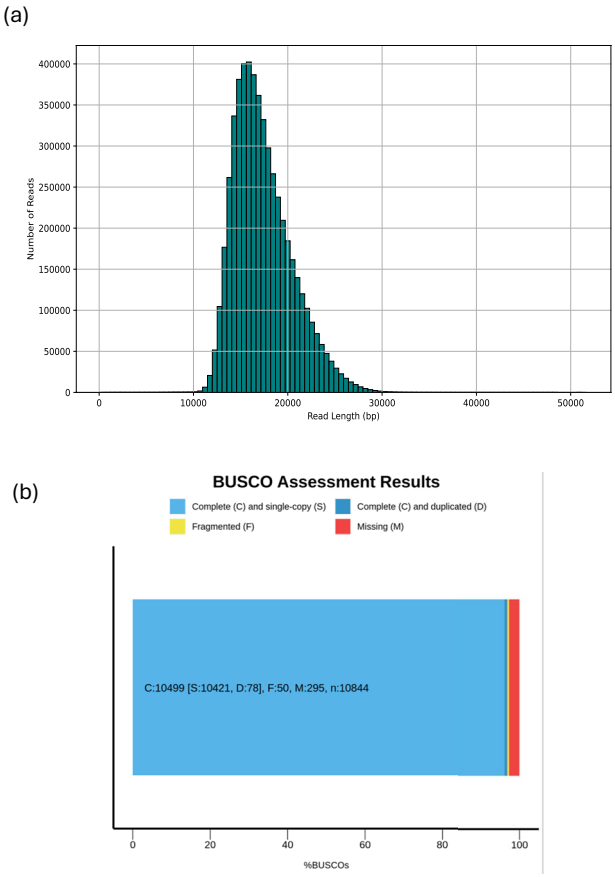


**Fig. 2** Statistics on raw PacBio HiFi long reads and de-novo genome assembly (**a**). Length distribution of PacBio HiFi raw reads **(b)** BUSCO (Benchmarking Universal Single-Copy Orthologs) assessment of genome completeness for *Phoenicurus ochruros*. The results indicate that 96.8% of BUSCO genes were identified as complete, including 96.1% complete and single-copy and 0.7% duplicated BUSCOs. Only 0.5% were fragmented and 2.7% were missing, demonstrating high completeness and gene-space coverage of the assembly.

| | |
|---|---|
| Total genome size | 1,371,345,233 bp |
| Minimum length of scaffold | 17,018 bp |
| Maximum length of scaffold | 113,313,014 bp |
| Number of scaffolds | 296 |
| Average length of scaffold | 4,632,923 bp |
| Scaffold N50 | 29,902,800 bp |
| **BUSCO Assessment Statistics (using *Passeriformes_odb10* dataset)** | |
| C (Complete) | 96.80% |
| S (Complete and Single-copy) | 96.10% |
| D (Complete and Duplicated) | 0.70% |
| F (Fragmented) | 0.50% |
| M (Missing) | 2.70% |
| Total BUSCO Groups Searched | 10,844 |
| Complete BUSCOs (C) | 10,499 |
| Complete and Single-copy BUSCOs (S) | 10,421 |
| Complete and Duplicated BUSCOs (D) | 78 |
| Fragmented BUSCOs (F) | 50 |
| Missing BUSCOs (M) | 295 |

**Table 1.** Genome assembly statistics.

We identified 30.58% of the Black Redstart genome as transposable elements. These included 0.13% Short Interspersed Nuclear Elements (SINEs), 5.44% Long Interspersed Nuclear Elements (LINEs), 7.75% Long Terminal Repeat (LTR) elements, 0.17% DNA transposons, and 0.10% rolling-circle elements (Table 2).

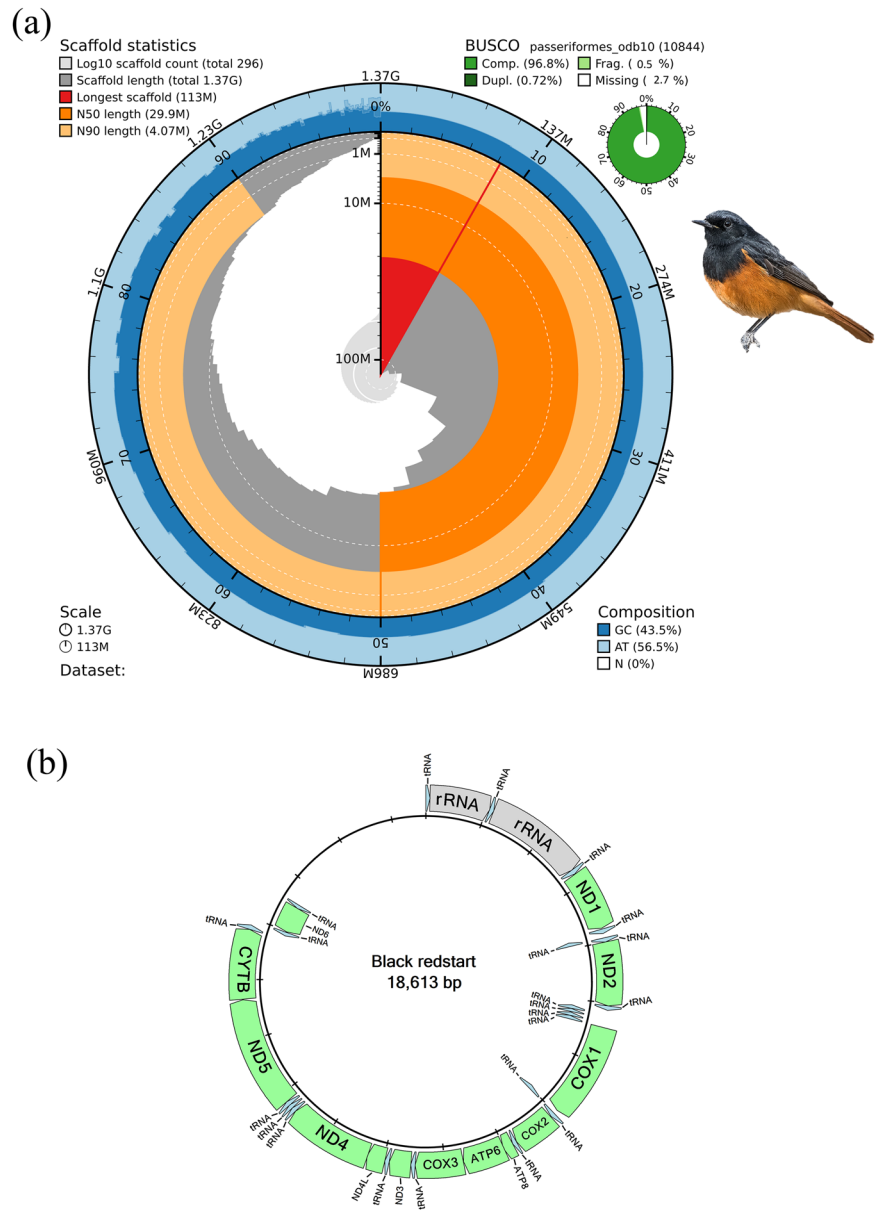


**Fig. 3** Summary of Genome assembly statistics (**a**). Statistics of the nuclear Genome. The distribution of sequence lengths is shown in dark grey with the plot radius scaled to the longest sequence present in the assembly (shown in red). Orange and pale-orange arcs show the N50 and N90 sequence lengths, respectively. The pale grey spiral shows the cumulative sequence count on a log scale with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT, and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated, and missing BUSCO genes in the passeriformes_odb10 set is shown in the top right **(b)** Details of the mitochondrial genome of the Black redstart showing position of respective genes.

Additionally, 4.37% of the genome consisted of unclassified repeats, and 10.77% was composed of satellite DNA. Such a high proportion of transposable elements is uncommon among birds[21]. Notably, only a few passerines - such as Bell's Sparrow *(Artemisiospiza belli)*, which exhibits a repeat content of 31.2%[22] display similarly elevated levels of repetitive elements. The genome size of the Black Redstart is comparable to that of Bell's Sparrow (Fig. 4a, Table 2), and since repeat content is generally correlated with genome size[21], the higher proportion of repetitive elements in the Black Redstart genome is expected. In contrast, a close relative within the Muscicapidae family, the Spotted Flycatcher *(Muscicapa striata)*, possesses a smaller genome size of 1.08 Gb and a significantly lower transposable element content of only 13.97%[23] (Fig. 4a, Table 2). Compared to Bell's Sparrow *(Artemisiospiza belli)*, which has 14.5% unclassified repeats, our repeat annotation for *Phoenicurus ochruros* appears more complete, with only 4.37% of repeats remaining unclassified. The high repeat content observed in *P. ochruros* suggests that its genome may serve as a valuable resource for investigating the evolutionary dynamics and regulatory roles of transposable elements in songbirds.

| Feature | Black Redstart | Bell's Sparrow | Chicken | Mexican Tetra Fish | Nelson's Sparrow | Paradise Crow | Spotted Flycatcher | White Wagtail | White-throated Sparrow |
|---|---|---|---|---|---|---|---|---|---|
| Genome size | 1371345233 | 1401818823 | 1053332251 | 1373169186 | 1185463352 | 1107771238 | 1083165776 | 1072670728 | 1052600561 |
| Log (genome_size) | 9.14 | 9.15 | 9.02 | 9.14 | 9.07 | 9.04 | 9.03 | 9.03 | 9.02 |
| Bases Masked (%) | 30.58 | 31.86 | 14.59 | 53.85 | 20.11 | 12.29 | 13.97 | 12.34 | 8.89 |
| SINE (%) | 0.13 | 0.07 | 0.09 | 0.08 | 0.08 | 0.09 | 0.08 | 0.08 | 0.08 |
| Retroelements (%) | 13.32 | 15.37 | 8.5 | 5.75 | 10.06 | 7.05 | 7.29 | 5.46 | 5.14 |
| LINEs (%) | 5.44 | 5.22 | 7.4 | 3.32 | 6.17 | 4.25 | 3.58 | 3.78 | 3.02 |
| LTR Elements (%) | 7.75 | 10.08 | 1.02 | 2.35 | 3.81 | 2.71 | 3.62 | 1.6 | 2.05 |
| DNA Transposons (%) | 0.17 | 0.36 | 0.98 | 20.97 | 0.26 | 0.31 | 0.24 | 0.28 | 0.27 |
| Rolling Circles (%) | 0.1 | 0.04 | 0 | 3.31 | 0.03 | 0.01 | 0.13 | 0.02 | 0.02 |
| Unclassified (%) | 4.37 | 14.55 | 3.38 | 19.91 | 8.07 | 3.16 | 3.14 | 5.1 | 5.2 |
| Total Interspersed Repeats (%) | 17.86 | 30.28 | 12.86 | 46.63 | 18.39 | 10.53 | 10.66 | 10.83 | 9.23 |
| Small RNA (%) | 0.22 | 0.03 | 0.06 | 0.36 | 0.07 | 0.28 | 0.09 | 0.02 | 0.01 |
| Satellites (%) | 10.77 | 0.27 | 0.04 | 0.77 | 0.2 | 0.49 | 1.34 | 0.22 | 0.18 |
| Simple Repeats (%) | 1.41 | 1.06 | 1.34 | 2.46 | 1.15 | 0.79 | 1.46 | 1.02 | 0.95 |
| Low Complexity (%) | 0.23 | 0.21 | 0.3 | 0.33 | 0.29 | 0.21 | 0.3 | 0.24 | 0.22 |

**Table 2.** Repeats statistics for Black redstart and other selected species.

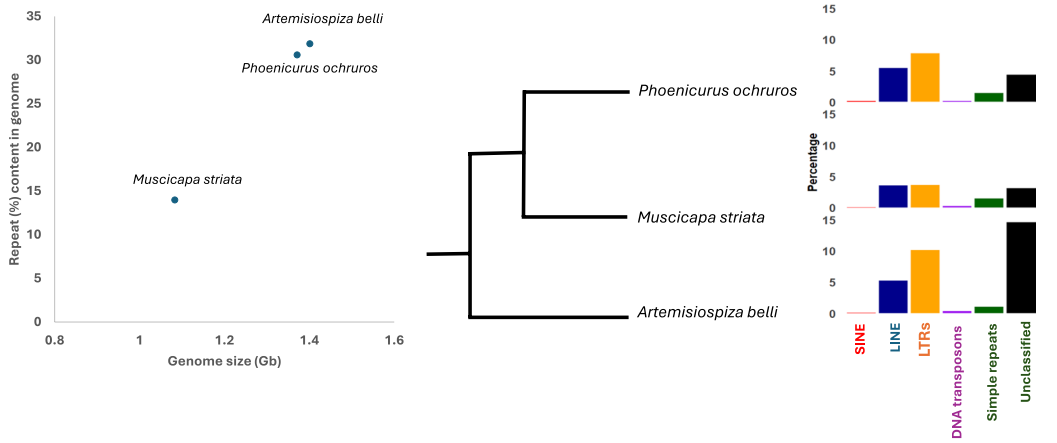


**Fig. 4** Landscape of Transposable Elements (TE) in *Phoenicurus ochruros*, *Artemisiospiza belli*, and *Muscicapa striata* genomes (**a**) Correlation of TE content and genome size **(b)** Comparison of various categories of TE across three species.

**RNA extraction and transcriptome sequencing.** Total RNA was extracted from eight tissue types (one sample each) - brain, liver, lungs, gonads, spleen, kidney, heart, and pectoral muscle - using the RNeasy Mini Kit (Qiagen, Valencia, CA, USA), following the manufacturer's protocol. These tissues were selected to maximize transcript diversity and support comprehensive genome annotation by capturing a wide range of expressed genes across major organ systems. RNA quality and integrity were assessed using a Bioanalyzer and spectrophotometry to ensure suitability for downstream sequencing. Each tissue sample was sequenced on the Illumina NovaSeq platform with 150 bp paired-end reads, generating approximately 45 million paired-end RNA-seq reads (total ~ 6.5 Gb of data) per sample (Table 3). Inclusion of diverse tissue will help to capture unique transcripts which may play important roles in key physiological processes related to high-altitude adaptation.

**Genome annotation.** We performed gene prediction and annotation using BRAKER3, an automated pipeline that integrates RNA-seq data, protein homology, and *ab initio* gene prediction to generate high-confidence gene models[24]. BRAKER3 runs the GeneMark-ETP pipeline[25], which executes a series of steps designed to accurately predict protein-coding genes. First, RNA-seq reads from eight tissues were aligned to the genome using HISAT2[26]. These alignments were used by StringTie2[27] to assemble transcript sequences, which were then analyzed using GeneMarkS-T[28] to identify putative protein-coding genes. The resulting protein sequences were searched against a reference vertebrate protein OrthoDB v11[29] database to determine similarity scores and select high-confidence gene predictions. The parameters of GeneMark.hmm were then trained on this high-confidence gene set, which was used to predict genes in the intermediate, less well-supported genomic regions. These intermediate predictions were matched against homologous proteins in the OrthoDB database, and the alignments were mapped back to the genome using ProtHint[30] to generate hints (e.g., intron-exon boundaries, start/stop codons).

| Sample | Tissue | Reads (bp) | Reads (Gb) |
|---|---|---|---|
| B-7-FE | Lungs | 44,465,278 | 6.67 |
| B-7-g | Liver | 44,008,762 | 6.60 |
| B-7-L | Gonads | 48,437,250 | 7.27 |
| B-7-N | Brain | 47,297,884 | 7.10 |
| B-7-P | Spleen | 45,808,536 | 6.90 |
| B-7-SH | Kidney | 41,073,742 | 6.16 |
| B-7-X | Heart | 44,556,880 | 6.68 |
| B-7-XJ | Pectoral muscle | 54,275,560 | 8.14 |

**Table 3.** Summary of RNAseq data generated from eight tissues.

Hints were classified into three groups: (a) Transcript and protein similarity support, (b) Transcript and *ab initio* support, and (c) Protein similarity support only. All three sets were used to expand the initial high-confidence gene set into a genome-wide gene annotation. Next, AUGUSTUS (v3.5.0)[31] was trained on the high-confidence gene models to generate a second set of gene predictions, integrating the RNA-seq and protein-derived hints. To further improve the accuracy of gene prediction, the *Passeriformes_odb10* lineage dataset from BUSCO[13,32] was used to optimize AUGUSTUS parameters by training on evolutionarily conserved genes. This ensured that gene models were informed by conserved genomic features specific to passerine birds, resulting in higher prediction quality. Finally, TSEBRA, a tool integrated within BRAKER3, was used to combine the gene predictions generated by GeneMark-ETP and AUGUSTUS. TSEBRA ensured the inclusion of high-confidence genes in the final annotation set and selected optimal gene models based on a scoring system that prioritizes evidence-supported predictions. This comprehensive approach resulted in a high-quality, evidence-supported annotation of 18,609 genes in the *P. ochruros* genome.

Functional annotation of protein-coding genes was carried out using BLASTp (v2.13.0 + )[33] against the Swiss-Prot database (UniProt release 2024_07)[34], applying an e-value threshold of $1 \times 10-5$. Protein domains were identified by querying the Conserved Domain Database (CDD) and Pfam[35] using InterProScan (v.5.67-99.0)[36]. A total of 18,609 gene models were identified, with an average gene length of 23,222 bp, approximately 20 exons per gene and with average exon length of 163 bp (Fig. 5). Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) annotation were obtained using eggNOG-mapper (v.2.0)[37] pipeline. We further evaluated the completeness of annotated protein sequences using BUSCO (v. 5.7.1)[13] against the *passeriformes_odb10* database. 97.8% of BUSCOs were successfully detected, including 96.6% complete and single copy, and 1.2% of duplicated BUSCOs. These results were consistent with the BUSCO assessment results based on the genome assembly (Table 1, Fig. 2b).

## Data Records

The PacBio HiFi long reads and the RNAseq data have been deposited in the Sequence Read Archive (SRA) database with accession number SRP584347[38] under BioProject accession PRJNA121710 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1217107). The genome assembly has been submitted to NCBI GenBank under the accession JBLKNL000000000 (https://www.ncbi.nlm.nih.gov/nuccore/JBLKNL000000000)[39]. The genome annotation files are available in Figshare (https://doi.org/10.6084/m9.figshare.28912433)[40]. The mitochondrial genome and its annotation have been submitted to NCBI GenBank under the accession ID PV606442 (https://www.ncbi.nlm.nih.gov/nucleotide/PV606442)[41].

## Technical Validation

DNA and RNA purity were assessed using a NanoDrop 2000 Spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). DNA and RNA concentration were quantified with a Qubit 2.0 Fluorometer (Thermo Fisher Scientific), and DNA/RNA integrity was evaluated using an Agilent 4200 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Only high-quality DNA and RNA samples were selected for genome and transcriptome sequencing. To validate the quality of the assembled genome, we employed multiple complementary approaches. First, we evaluated genome contiguity and gene-space completeness using standard assembly statistics and BUSCO[13]. The assembly achieved a contig $N_{50}$ of ~30 Mb and a BUSCO completeness score of 97.30% against the passeriformes_odb10 database, indicating a high contiguity and gene-space completeness.

Repeat-rich regions are often difficult to assemble accurately due to their repetitive nature, making them a common source of assembly errors or gaps. By evaluating the abundance and composition of transposable elements among representative species with published genomes, we can validate the quality of a *de novo* genome assembly. We compared the repeat landscape of *P. ochruros* with that of several other vertebrate species (Table 2); Blind cave fish *(Astyanax mexicanus)* (GenBank accession: GCA_023375975.1), Common frog (*Rana temporaria)* (GenBank accession: GCA_905171775.1), Chicken (*Gallus gallus)* (GenBank accession: GCA_016699485.1), Bell's sparrow (*Artemisiospiza belli)* (GenBank accession: PRJNA720569), Nelson's sparrow (*Ammospiza nelsoni)*[42], White-throated sparrow *(Zonotrichia albicollis)* (GenBank accession: PRJNA197293), Spotted flycatcher (*Muscicapa striata)*[23], Obi paradise-crow (*Lycocorax pyrrhopterus obiensis)*[43], and White wagtail (*Motacilla alba)* (GenBank accession: GCA_015832195.1). Our results indicated that the repeat content in *P. ochruros* is consistent with known evolutionary trends and genomic patterns observed in other passerines. We found a strong positive correlation between genome size and repeat content ($r = 0.87$), indicating that the

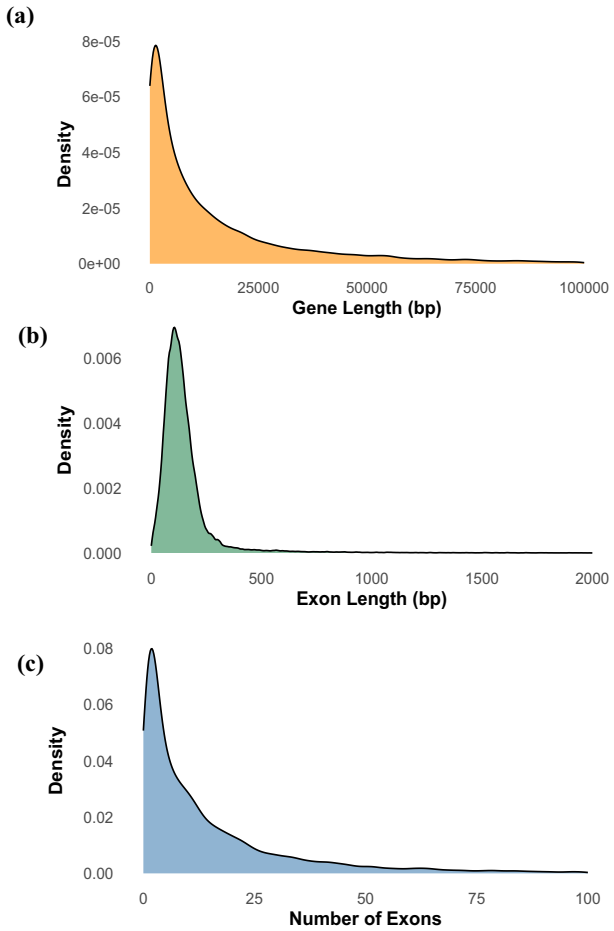


**Fig. 5** Gene structure statistics of *Phoenicurus ochruros* (**a**). The distribution of gene length **(b)** The distribution of exon length, and **(c)** The distribution of the number of exons per gene.

assembly has captured repetitive elements with high accuracy. This comparative approach not only aids in identifying potential assembly artifacts, such as collapsed or missing repeat regions, but also supports the completeness of repeat annotation. Together, these findings enhance confidence in the overall quality and biological validity of the *Phoenicurus ochruros* genome assembly.

## Data availability

PacBio HiFi long reads and the RNAseq data is available at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1217107. The genome assembly is available at. https://www.ncbi.nlm.nih.gov/nuccore/JBLKNL000000000. The genome annotation files are available at https://doi.org/10.6084/m9.figshare.28912433. The mitochondrial genome and its annotation are available at https://www.ncbi.nlm.nih.gov/nucleotide/PV606442.

## Code availability

All custom scripts and codes used in this study are available on GitHub at - https://github.com/Prashantevo/Black_redstart_genome_assembly.

Received: 15 May 2025; Accepted: 21 October 2025;
Published online: 28 November 2025

## References

1. Barçante, L., Vale, M. M. & Alves, M. A. S. Altitudinal migration by birds: a review of the literature and a comprehensive list of species. *J. Field Ornithol.* **88**, 321–335 (2017).
2. Williamson, J. L. & Witt, C. C. Elevational niche-shift migration: Why the degree of elevational change matters for the ecology, evolution, and physiology of migratory birds. *Ornithology* **138**, ukaa087 (2021).
3. Barve, S., Dhondt, A. A., Mathur, V. B. & Cheviron, Z. A. Life-history characteristics influence physiological strategies to cope with hypoxia in Himalayan birds. *Proc. R. Soc. B Biol. Sci.* **283**, 20162201 (2016).
4. Winkler, D. W., Billerman, S. M. & Lovette, I. J. Old World Flycatchers (Muscicapidae), version 1.0. in *Birds of the World* (eds Billerman, S. M., Keeney, B. K., Rodewald, P. G. & Schulenberg, T. S.) (Cornell Lab of Ornithology, Ithaca, NY, USA, 2020).

5. Birdlife International. *Species Factsheet: Black Redstart Phoenicurus Ochruros*. https://datazone.birdlife.org/species/factsheet/black-redstart-phoenicurus-ochruros (2018).
6. Collar, N. Black Redstart (Phoenicurus ochruros), version 1.1. *Birds World*, https://doi.org/10.2173/bow.blared1.01.1 (2024).
7. Sangster, G. Black Redstart: no evidence for multiple species or hybrid origin of. *Dutch Bird* **43**, 371–377 (2021).
8. Martinez, N. & Spek, V. van der. Geographical variation in Black Redstart Phoenicurus ochruros (S. G. Gmelin, 1774) calls. *Bull. Br. Ornithol. Club* **142**, 466–477 (2022).
9. Storz, J. F. High-Altitude Adaptation: Mechanistic Insights from Integrated Genomics and Physiology. *Mol. Biol. Evol.* **38**, 2677–2691 (2021).
10. Storz, J. F., Scott, G. R. & Cheviron, Z. A. Phenotypic plasticity and genetic adaptation to high-altitude hypoxia in vertebrates. *J. Exp. Biol.* **213**, 4125–4136 (2010).
11. Palacios, C. *et al*. Genomic Variation, Population History, and Long-Term Genetic Adaptation to High Altitudes in Tibetan Partridge (*Perdix hodgsoniae*). *Mol. Biol. Evol.* **40**, msad214 (2023).
12. Cheng, H., Asri, M., Lucas, J., Koren, S. & Li, H. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. *Nat. Methods* **21**, 967–970 (2024).
13. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
14. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
15. Uliano-Silva, M. *et al*. MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. *BMC Bioinformatics* **24**, 288 (2023).
16. Du, C., Liu, L., Liu, Y., Fu, Z. & Xu, Y. The complete mitochondrial genome of Daurian redstart Phoenicurus auroreus (Aves: Passeriformes: Muscicapidae). *Mitochondrial DNA Part B* **5**, 2231–2232 (2020).
17. Sangster, G. & Luksenburg, J. A. The published complete mitochondrial genome of Blue-fronted Redstart (Phoenicurus frontalis) is a chimera and includes DNA from Pink-rumped Rosefinch Carpodacus waltoni eos (Aves: Passeriformes). *Mitochondrial DNA Part B Resour* **9**, 861–864 (2024).
18. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* **117**, 9451–9457 (2020).
19. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
20. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0 http://www.repeatmasker.org. *RMDownload Html* (2013).
21. Kapusta, A., Suh, A. & Feschotte, C. Dynamics of genome size evolution in birds and mammals. *Proc. Natl. Acad. Sci.* **114**, E1460–E1469 (2017).
22. Benham, P. M. *et al*. Remarkably High Repeat Content in the Genomes of Sparrows: The Importance of Genome Assembly Completeness for Transposable Element Discovery. *Genome Biol. Evol.* **16**, evae067 (2024).
23. Baudrin, G. *et al*. A Reference Genome Assembly for the Spotted Flycatcher (Muscicapa striata). *Genome Biol. Evol.* **15**, evad140 (2023).
24. Gabriel, L. *et al*. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.* **34**, 769–777 (2024).
25. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes. *Genome Res.* **34**, 757–768 (2024).
26. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
27. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
28. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* **43**, e78 (2015).
29. Kuznetsov, D. *et al*. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res.* **51**, D445–D451 (2023).
30. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics Bioinforma.* **2**, lqaa026 (2020).
31. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
32. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinforma. Oxf. Engl.* **31**, 3210–3212 (2015).
33. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
34. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res* **43**, D204–D212 (2015).
35. Blum, M. *et al*. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**, D344–D354 (2021).
36. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
37. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
38. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP584347 (2025).
39. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBLKNL000000000 (2025).
40. Ghimire, P., Wang, N. & Lamichhaney, S. A complete reference genome assembly and annotation of the Black Redstart (Phoenicurus ochruros). https://doi.org/10.6084/m9.figshare.28912433 (2025).
41. *NCBI GenBank* https://identifiers.org/ncbi/insdc:PV606442 (2025).
42. Rhie, A. *et al*. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
43. Peona, V. *et al*. Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Mol. Ecol. Resour.* **21**, 263–286 (2021).
44. Voelker, G., Semenov, G., Fadeev, I. V., Blick, A. & Drovetski, S. V. The biogeographic history of Phoenicurus redstarts reveals an allopatric mode of speciation and an out-of-Himalayas colonization pattern. *Syst. Biodivers.* **13**, 296–305 (2015).
45. BirdLife International & Handbook of the Birds of the World. Bird species distribution maps of the world. Version 2022.1. http://datazone.birdlife.org/species/requestdis (2022).
46. Fink, D. *et al*. eBird Status and Trends, Data Version: 2023; 2025. Cornell Lab of Ornithology, Ithaca, New York. https://doi.org/10.2173/WZTW8903 (Released: 2025).


## Acknowledgements

The fieldwork was supported by Survey of National Key Protected Wildlife Resources and Habitat Assessment in Chengduo County, Qinghai, China, and Special investigation and monitoring of the Yarlung Zangbo Grand Canyon National Nature Reserve in Xizang, China. PG was supported by a graduate assistantship from Kent State University. Genome sequencing was supported from Kent State University to SL.

## Author contributions

S.L. designed the study, N.W. and Y.H. collected the samples, P.G. conducted the bioinformatics analysis under S.L. supervision. P.G. and S.L. drafted the manuscript. All authors read and approved the final version of the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to N.W. or S.L.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.