

Draft assemblies for 177 bird species enhance genus-level coverage

Guangji Chen^{1,*}, Shuang Wang^{1,*}, Daniel Bilyeli Øksnebjerg^{2,*}, Sascha Dreyer Nielsen^{2,*}, Wei Dai^{3,*}, Wei Jiang⁴, Jing Liang¹, Wei Han¹, Chengran Zhou³, Qiye Li^{4,5}, Bent Petersen^{2,6}, Ara Monadjem^{7,8}, Zamekile D. Bhembe⁷, Machawe Maphalala⁷, Diego Ocampo^{9,10}, Luis Sandoval^{11,12}, Jörns Fickel^{13,14}, Alex D. Greenwood^{15,16}, Claudia A. Szentiks¹⁵, Marco Roller^{17,18,19}, Sharon M. Birks²⁰, Adam D. Leaché^{20,21}, Alejandro Rico-Guevara^{20,21}, Jérôme Fuchs²², Nguyen Tran Vy²³, Christina Hvilsom²⁴, Juliana Andrea Berner²⁴, Jan Terje Lifjeld²⁵, Arild Johnsen²⁵, Lars Erik Johannessen²⁵, Kim Labuschagne²⁶, Knud Andreas Jønsson^{27,28}, Martin Irestedt²⁸, Andrew Hart Reeve^{27,28}, Leo Joseph²⁹, Olof Hellgren³⁰, Robb T. Brumfield^{31,32}, Theresa M. Burg³³, Juan Carlos Illera³⁴, Alexandre Aleixo^{35,36}, Ben Smit³⁷, Frank E. Rheindt³⁸, Jessica Lee³⁹, Isao Nishiumi⁴⁰, Javier Quesada⁴¹, John P. Dumbacher⁴², Manuel Schweizer^{43,44}, Michael J. Andersen⁴⁵, Christopher C. Witt⁴⁵, Richard A. Phillips⁴⁶, Richard Prum⁴⁷, Kristof Zyskowski⁴⁸, Steven M. Goodman⁴⁹, Marie Jeanne Raherilalao^{50,51}, Ulf Ottosson⁵², Yahkat Barshep^{52,53}, Sam Ivande^{52,54}, Vojtěch Brlík⁵⁵, Emmanuel Okposio⁵⁶, Bonny Koane⁵⁷, Tri Haryoko⁵⁸, Erich D. Jarvis⁵⁹, Carsten Rahbek^{60,61,62}, Fumin Lei⁶³, Gary R. Graves^{60,64}, Shaohong Feng^{1,65,#}, Peter A. Hosner^{27,60,61,#}, M. Thomas P. Gilbert^{2,66,#}, Guojie Zhang^{1,#}

Affiliations:

1. Center for Evolutionary & Organismal Biology, Liangzhu Laboratory & Women's Hospital, Zhejiang University School of Medicine, Hangzhou, China
2. Center for Evolutionary Hologenomics, Globe Institute, University of Copenhagen, Copenhagen, Denmark
3. BGI Research, Wuhan, China
4. State Key Laboratory of Genome and Multi-omics Technologies, & Shenzhen Key Laboratory of Forensics, BGI Research, Shenzhen 518083, China
5. College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China
6. Senckenberg Natural History Collections Dresden, Dresden, Germany
7. Department of Zoology and Entomology, Mammal Research Institute, University of Pretoria, Hatfield, Pretoria, South Africa
8. Department of Biological Sciences, University of Eswatini, Kwaluseni, Eswatini
9. Museum of Vertebrate Zoology, University of California Berkeley, Berkeley, CA 94720 USA
10. Centro de Investigación en Biodiversidad y Ecología Tropical, Universidad de Costa Rica, San José, Costa Rica
11. Laboratorio de Ecología Urbana y Comunicación Animal, Escuela de Biología, Universidad de Costa Rica, San José, Costa Rica
12. Centro de Investigación en Biodiversidad y Ecología Tropical, Universidad de Costa Rica, San José, Costa Rica
13. Department of Evolutionary Genetics, Leibniz Institute for Zoo and Wildlife Research (IZW), Alfred-Kowalke-Strasse 17, 10315 Berlin, Germany
14. Institute of Biochemistry and Biology, Potsdam University, Karl-Liebknecht-Straße 24-25, 14476 Potsdam, Germany

15. Department Wildlife Diseases, Leibniz Institute for Zoo and Wildlife Research, Alfred-Kowalke-Strasse 17, 10315 Berlin, Germany
16. Department of Veterinary Medicine, Freie Universität Berlin, Berlin, Germany
17. Wilhelma Zoological-Botanical Gardens Stuttgart, Wilhelma 13, D-70376 Stuttgart, Germany
18. Department of Animal Sciences, Division of Microbiology and Animal Hygiene, Faculty of Agricultural Science, Georg-August-University, Burckhardtweg 2, D-37077 Göttingen, Germany
19. Institute for Microbiology, University of Veterinary Medicine Hannover, Foundation, Bischofsholer Damm 15, D-30173 Hannover, Germany
20. Burke Museum of Natural History and Culture, University of Washington, Seattle, Washington, USA
21. Department of Biology, University of Washington, Seattle, WA 98105, USA
22. Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, Paris, France
23. Institute of Tropical Biology, VAST, Ho Chi Minh City, Vietnam
24. Copenhagen Zoo, Copenhagen, Denmark
25. Natural History Museum, University of Oslo, Norway
26. South African National Biodiversity Institute (SANBI), Pretoria, South Africa
27. Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark
28. Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm, Sweden
29. Australian National Wildlife Collection, National Research Collections Australia, CSIRO, Canberra, Australia
30. Department of Biology, Lund University, Lund, Sweden
31. Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803, USA.
32. Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA.
33. Department of Biological Sciences, University of Lethbridge, Lethbridge, AB, T1K 3M4, Canada
34. Biodiversity Research Institute (CSIC-Oviedo University-Principality of Asturias), University of Oviedo, Campus of Mieres, E-33006 Mieres, Asturias, Spain
35. Museu Paraense Emílio Goeldi, Belém, PA, Brazil
36. Vale Institute of Technology, Belém, PA, Brazil
37. Department of Zoology and Entomology, Rhodes University, Makhanda, South Africa
38. Department of Biological Sciences, National University of Singapore, Singapore 117558, Singapore
39. Mandai Nature, Mandai, Singapore
40. Department of Zoology, National Museum of Nature and Science, Tokyo, Japan
41. Department of Vertebrates, Natural Sciences Museum of Barcelona, Barcelona, Spain
42. Ornithology & Mammalogy Department, California Academy of Sciences, San Francisco, CA 94118, United States
43. Natural History Museum Bern, Bern, Switzerland
44. Division of Population Genetics, Institute of Ecology and Evolution, University of Bern, Bern, Switzerland
45. Museum of Southwestern Biology, University of New Mexico, New Mexico, United States of America

46. British Antarctic Survey (BAS), Natural Environment Research Council (NERC), High Cross, Madingley Road, CB3 0ET Cambridge, United Kingdom
47. Department of Ecology and Evolutionary Biology, and Peabody Museum of Natural History, Yale University, New Haven, CT, USA
48. Division of Vertebrate Zoology, Peabody Museum of Natural History, Yale University, New Haven, CT USA
49. Field Museum of Natural History, Chicago, IL 60605, USA
50. Association Vahatra, Antananarivo, Madagascar
51. Mention Zoologie et Biodiversité Animale, Université d'Antananarivo, Madagascar
52. A.P. Leventis Ornithological Research Institute (APLORI), Centre of Excellence, University of Jos Biological Conservatory, Jos, Nigeria
53. Department of Zoology, University of Jos, Jos, Nigeria
54. Global Center for Species Survival, Indianapolis Zoo, 1200 West Washington St. Indianapolis, IN 46222, United States
55. Department of Ecology, Charles University, Prague, Czech Republic
56. Department of Biology, California State University, Fresno, California, USA
57. New Guinea Binatang Research Centre, Madang, Papua New Guinea.
58. Museum Zoologicum Bogoriense, Research Center for Biosystematics and Evolution, National Research and Innovation Agency (BRIN), Cibinong, 16911, Indonesia
59. The Vertebrate Genome Laboratory, The Rockefeller University, New York, USA
60. Center for Macroecology, Evolution and Climate, Globe Institute, University of Copenhagen, Copenhagen, Denmark
61. Center for Global Mountain Biodiversity, Globe Institute, University of Copenhagen, Copenhagen, Denmark
62. Department of Biology, University of Southern Denmark, 5230 Odense M, Denmark
63. Institute of Zoology, Chinese Academy of Sciences, Beijing, China
64. Department of Vertebrate Zoology, National Museum of Natural History, Smithsonian Institution, Washington DC, USA
65. Department of General Surgery of Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China
66. University Museum, NTNU, Trondheim, Norway

* contributed equally

Corresponding author

ORCID: Guangji Chen [0000-0002-9441-1155]; Shuang Wang [0009-0008-9451-5684]; Sascha Dreyer Nielsen [0000-0003-2485-9314]; Wei Dai [0000-0001-9286-759X]; Chengran Zhou [0000-0002-9468-5973]; Qiye Li [0000-0002-5993-0312]; Bent Petersen [0000-0002-2472-8317]; Ara Monadjem [0000-0003-1906-4023]; Zamekile D. Bhembe [0009-0006-5354-5751]; Machawe Maphalala [0000-0002-9304-1391]; Luis Sandoval [0000-0002-0793-6747]; Jörns Fickel [0000-0002-0593-5820]; Alex D. Greenwood [0000-0002-8249-1565]; Adam D. Leaché [0000-0001-8929-6300]; Alejandro Rico-Guevaera [0000-0003-4067-5312]; Christina Hvilsom [0000-0001-

7870-6888]; Arild Johnsen [0000-0003-4864-6284]; Lars Erik Johannessen [0000-0001-5981-9190]; Kim Labuschagne [0000-0003-2784-4767]; Knud Andreas Jønsson [0000-0002-1875-9504]; Martin Irestedt [0000-0003-1680-6861]; Leo Joseph [0000-0001-7564-1978]; Robb T. Brumfield [0000-0003-2307-0688]; Theresa M. Burg [0000-0001-5096-3479]; Juan Carlos Illera [0000-0002-4389-0264]; Alexandre Aleixo [0000-0002-7816-9725]; Frank E. Rheindt [0000-0001-8946-7085]; Jessica Lee [0000-0003-0757-4237]; Javier Quesada [0000-0002-6010-8473]; Manuel Schweizer [0000-0002-7555-8450]; Michael J. Andersen [0000-0002-7220-5588]; Christopher C. Witt [0000-0003-2781-1543]; Richard A. Phillips [0000-0002-0208-1444]; Richard Prum [0000-0002-4741-7132]; Kristof Zyskowski [0000-0002-5680-6412]; Marie Jeanne Raheirilalao [0000-0002-8618-7157]; Ulf Ottosson [0000-0001-7914-0484]; Sam Ivande [0000-0003-4949-1376]; Vojtěch Brlík [0000-0002-7902-8123]; Bonny Koane [0000-0001-6770-5126]; Tri Haryoko [0000-0002-8549-3662]; Erich D. Jarvis [0000-0001-8931-5049]; Carsten Rahbek [0000-0003-4585-0300]; Fumin Lei [0000-0001-9920-8167]; Gary R. Graves [0000-0003-1406-5246]; Shaohong Feng [0000-0002-2462-7348]; Peter A. Hosner [0000-0001-7499-6224]; M. Thomas P. Gilbert [0000-0002-5805-7195]; Guojie Zhang [0000-0001-6860-1521]

Abstract

Background: With over 10,000 recognized species, birds constitute one of the most diverse and widely distributed vertebrate groups. Although avian genomics has advanced rapidly over the past decade, substantial gaps remain across the global avifauna. Filling these gaps is essential for understanding macroevolutionary patterns, population structure, and the molecular basis of ecological and behavioral diversity. Worldwide museum collections represent invaluable resources for filling these gaps, yet the typically degraded DNA and limited quantities from historical specimens have posed significant challenges for generating high-quality genome assemblies. **Results:** Here, the Bird Genome 10K (B10K) Project adopted low-input sequencing strategies that reduce costs while improving assembly quality compared with earlier order- and family-level genomes. Using mainly stLFR, complemented by 10X Genomics and standard next-generation sequencing, we assembled 177 avian genomes from museum specimens and tissue collections representing 161 genera, including 102 newly sequenced at the genomic level. The assemblies average ~1.2 Gb in size, with scaffold N50 = 8.03 Mb, contig N50 = 120 kb, 93% BUSCO completeness, and Merqury QV score of 56. **Conclusions:** These genomes greatly expand avian taxonomic coverage and demonstrate the efficiency of low-input sequencing for generating high-quality assemblies from limited and often degraded material sourced from museum specimens. This resource provides a foundation for comparative genomics, conservation genetics, and evolutionary studies across the avian tree of life.

Keywords: birds; genome sequencing; biodiversity.

Context

Birds, derived from theropod dinosaurs, constitute one of the most species-rich and widely distributed vertebrate radiations, encompassing more than 10,000 species and over 2,000 genera that inhabit nearly every ecosystem on Earth[1,2]. Over the past decade, facilitated by international efforts such as the Bird 10,000 Genomes (B10K) Project[3,4] and the Vertebrate Genomes Project, avian genome resources have expanded rapidly (**Figure 1**), providing unprecedented opportunities for comparative and functional genomics. This growth has been enabled not only by advances in sequencing and assembly technologies but also by the invaluable contribution of natural history museums and institutional biobanks, whose curated specimens have become indispensable for large-scale genome initiatives.

The B10K Project has achieved significant milestones over the past decade in generating genomic data for species representing order- and family-level diversity across the avian tree of life[3–6]. However, genus- and species-level coverage remains highly uneven. Several genera, such as *Aphelocoma*, *Falco*, *Gallus*, *Anas*, and *Haemorrhous* are represented by dozens of assemblies, whereas many lineages remain underrepresented due to limited access to fresh tissues or vouchered samples. For numerous species, especially those inhabiting remote regions or represented primarily by historical specimens, natural history collections remain the only feasible sources of DNA. However, these collections are often geographically dispersed and difficult to access for genomic research due to limitations in both the quality (high-molecular-weight genomic DNA) and quantity of available genetic material. As a result, molecular-level understanding of these taxa has long been constrained by the scarcity of genomic resources.

Recent advances in low-input sequencing and museomics technologies (e.g., stLFR[7]) are gradually bridging this gap, allowing for the assembly of genomes from previously inaccessible samples. B10K Project has strategically adopted these low-input sequencing methods, reducing costs while improving genome assembly quality compared to earlier order- and family-level studies. Here, using stLFR technologies, we report draft genome assemblies for 177 bird species representing 161 genera, including 102 genera newly covered at the genomic level (**Figure 2**). These newly sequenced genera increased by 4.43% the genus-level coverage of extant bird species, providing a more continuous phylogenetic framework for comparative genomics across the avian tree of life. These assemblies substantially enrich the genomic resources available for avian research, offering new insights into the genetic diversity and evolutionary history of this remarkable vertebrate group.

Materials and Methods

Sampling, sequencing, assembly, and annotation

DNA extraction

Genomic DNA was extracted from 177 avian species, comprising both frozen materials and historical museum specimens, as detailed in **Supplementary Table 1**. Specifically, 80.8% of the samples (n = 143) were collected before 2020 with many challenging historical sources that necessitated the low-input sequencing strategy. All DNA extractions were performed using the 'MagMAX DNA Multi-Sample Ultra 2.0' kit (Thermo Fisher) following the manufacturer's guidelines. For post-lysis, RNA degradation was performed by incubating the lysate with RNase A at 37°C for 5 minutes. Subsequently, genomic DNA was isolated using a bead-based purification method, utilizing 440 µL of DNA Binding Bead Mix per sample. The automated isolation process was executed on the Kingfisher Duo Prime system (Thermo Fisher Scientific), following the standard MMX_Ultra2_Cell_Tissue_96_Duo protocol. Quality control was performed on the 2200 TapeStation System, and all DNA extractions were stored at -80°C until library construction.

Library construction and sequencing

Linked-Reads Library

The selection of sequencing technologies was strategically determined by the interplay between DNA integrity and the progressive availability of sequencing platforms throughout the project. The majority of the DNA samples (n = 161) were prepared using the stLFR (Single-Tube Long Fragment Reads) technology, which served as our primary pipeline due to its ability to generate high-quality linked-read data with 10 ng of input DNA with fragment sizes exceeding 10 kb. During the early phase of this study, 13 DNA samples which exhibited exceptional DNA integrity (main fragment sizes > 40 kb) were prepared using the 10X Genomics Chromium library system.

For the 13 samples prepared using the 10X Genomics Chromium library system, high-molecular-weight genomic DNA was processed with the Chromium Genome Library Kit (10X Genomics, Pleasanton, USA) to generate barcoded linked-read libraries. To ensure compatibility and sequencing efficiency across platforms, the libraries were adapted for sequencing on the BGISEQ-500 platform (BGI-Shenzhen, China) following in-house optimization to produce 150 bp paired-end reads in total 1.6 Tb raw sequencing reads (on average ~ 103x).

For the 161 samples prepared using stLFR (Single-Tube Long Fragment Reads) technology (MGI), high-molecular-weight genomic DNA was processed following the manufacturer's barcoding protocol, which enables the physical linkage of short reads derived from the same long DNA fragment. The barcoded libraries were sequenced on the DNBSEQ-T7 platform (MGI, Shenzhen, China) with PE100+42 mode, which included two 100 bp reads for the genomic inserts and a dedicated 42 bp read to capture the co-barcoding sequences essential for the stLFR method[7], and produced in total 26.3 Tb raw sequencing reads (on average ~ 136x).

Standard Next-Generation Sequencing

For the remaining 3 samples with relative DNA degradation (fragment sizes < 10 kb) that failed to meet the requirements for linked-read library construction, standard short-read next-

generation technology with approximately 500 ng input DNA was employed to ensure sufficient genomic coverage despite suboptimal sample quality. Sequencing libraries with an average insert size of approximately 350 bp were constructed according to the manufacturer's protocol and sequenced on the DNBSEQ-T1 platform (BGI-Shenzhen, China) to produce 150 bp paired-end reads in total 226 Gb raw sequencing reads (on average 63x).

Assembly and Statistics of genomes

stLFR and 10X Genomics

For the 161 samples sequenced using stLFR technology, raw data were initially transformed into 10X Genomics linked-reads format following the stlfr2supernova pipeline [8]. Subsequently, these 161 samples, together with 13 samples sequenced by 10X Genomics linked-reads technology, were preprocessed with SOAPfilter (v2.2, RRID:SCR_000689) using parameters ``-i $insertsize -p -z -g 1 -M 2 -Q 10 lane.lst stat_file`` to remove adapters and low-quality reads. The cleaned reads were then introduced into the Supernova software (v2.0.1; RRID:SCR_016756) [9] to assemble the genomes under the "pseudohap" mode. The assembly was executed with the command: ``supernova run --id=ID --fastqs=./ --localcores=32 --localmem=300G --nopreflight``, followed by ``supernova mkoutput --style=pseudohap --asmdir=./ID/outs/assembly --outprefix=output`` to generate the final genomic sequences.

To ensure assembly quality, scaffolds containing more than 80% "N" bases were removed using housed Perl script, which typically resulted in the exclusion of approximately 2 scaffolds per assembly. Subsequently, gap filling was applied with Gapcloser (v1.12, RRID:SCR_015026) [10] by using the default parameters and the cleaned paired-end reads.

Standard Next-Generation technology

For the 3 samples sequenced by the standard Next-Generation technology, reads were cleaned by SOAPfilter (v2.2, RRID:SCR_000689) with previous parameters and assembled using the SOAPdenovo (v2.04, RRID:SCR_010752) [10] with a K-mer size of 23-mer following the B10K family-phased assembly strategy [4]. After removing scaffolds with "N" > 80%, GapCloser (v1.12, RRID:SCR_015026) [10] was used to close the intra-scaffold gaps with default parameters based on the cleaned paired-end reads.

Gene structures annotation

Annotation of protein-coding genes was conducted with a homology-based method for the 177 bird species following the pipeline implemented by the Bird 10,000 Genomes (B10K) consortium[4]. The reference gene set consists of the primary reference gene set (20,194 genes), the supplemental human gene set (20,169 genes), and the supplemental transcriptome gene set (5,257 transcripts), which could be accessed [11]. The protein sequences in the primary reference gene set were first aligned to each genome using TBLASTN (v2.2.26; RRID:SCR_011822) [12] with an e-value cut-off 1e-5, and multiple adjacent hits of the same query were connected by genBlastA (v1.0.4; RRID:SCR_020951) [13]. Homologous blocks with a length greater than 30% of the query protein length were retained. Each connected hit region was later extended to include its 2 kb upstream and downstream flanking regions, on which gene structure was predicted by

Genewise (v2.4.1; RRID:SCR_015054) [14]. MUSCLE (v3.8.31; RRID:SCR_011812) [15] was then used to align the annotated protein against its reference protein. Predicted proteins with length ≥ 30 amino acids and an identity value $\geq 40\%$ were retained. Pseudogenes (annotated genes containing > 2 frameshifts or > 1 premature stop codon) and retrogenes were further removed. Next, gene models that overlapped in $> 40\%$ of their coding sequence were clustered into one group and the one with the highest identity to the reference proteins was retained to form a non-redundant gene set for each species. Two supplemental gene sets were also used for homology-based gene prediction for these newly released assemblies as above, but only the newly annotated loci from these supplemental sets were kept. Finally, all candidate annotated genes that had > 10 duplications were removed. Among the protein-coding annotations of these 177 avian species, an average of 89.53% were annotated using the core avian gene set, 8.43% were supported by the transcriptome gene set, and 2.03% by the human gene set.

Phylogenetic information

Phylogenetic information for the 177 newly released avian species was derived from the recently large-scale avian phylogenies[16], and the tree was pruned to match the dataset using the R package ape[17], and visualized using the R packages ggtree[18].

Data Validation and quality control

Sample taxonomic verification

Voucher specimens were verified against reference collections. Mitogenomes were BLAST searched against references in the NCBI database. If no reference mitogenome was available, we checked NCBI for commonly sequenced mitochondrial genes (*COI*, *ND2*, *CYB*) as a reference to verify sequence authenticity. For genomes that lacked a mitochondrial assembly, which was frequent when blood was the source tissue, we used a collection of nuclear markers well-represented on NCBI for birds (*FGB*, *GAPDH*, *MB*, *MUSK*, *ODC*, *RAG1*, *TGFb2*).

Contamination screening

Genome assemblies were screened for contamination using the NCBI Foreign Contamination Screen (FCS-GX) pipeline [19]. Adapter and contaminant sequences were detected using the FCS adaptor module, and contaminant scaffolds were removed or trimmed using the FCS cleaning module with the default parameters.

Quality control on sequencing reads

Quality control steps for raw reads before assembly using the SOAPfilter2 package (v2.2, RRID:SCR_000689) were: Remove reads with more than 10% of N bases; Remove reads with more than 40% low-quality bases (Phred score ≤ 10); Remove reads with undersized insert size; Filter out the PCR duplicates.

Statistics of genome assembly

Assembly quality of 177 assemblies was assessed by contig N50, scaffold N50, and total assembly length. Genome completeness was measured with Compleasm (v0.2.7, RRID:SCR_026370)[20] using aves_odb12 as the reference gene set for these species. Three standard categories of

BUSCO results were assessed as follows: Complete and single-copy BUSCOs (S), Complete and duplicated BUSCOs (D), Fragmented BUSCOs (F), The Complete and single-copy BUSCOs (S) and duplicated BUSCOs (D) hits were combined to assess the degree of genome completeness. Furthermore, to estimate the base-level accuracy and consensus quality, Merqury (v1.3, RRID:SCR_022964)[21] was utilized to calculate the Consensus Quality Value (QV) and k-mer completeness by comparing the assemblies against the original high-quality sequencing reads in a k-mer-based, reference-independent manner.

Genomic quality

The genome assemblies of the 177 avian genomes have an average genome size of ~ 1.2 Gb, with average scaffold N50 = 8.03 Mb, and contig N50 = 120 kb (**Figure 3**). Overall, the assembly quality shows an improvement compared with previous datasets generated at the order and family levels, largely due to the adoption of low-input sequencing strategies. Only 2.5% of the core genes in aves_odb12 could not be predicted on these newly released genomes (ranging from 0.2% to 10.2%), suggesting that the completeness of these genomes was suitable for most comparative genome analyses. An average of 93% complete core genes in these 177 avian genomes was comparable to that of the previously published genomes (**Figure 4**). On average, only 4.4% of the core genes were partly annotated for the 177 avian genomes. Furthermore, k-mer-based evaluation yielded an average Merqury QV score of 56 (**Supplementary Table 1**), indicating a relatively high consensus accuracy at the base level.

Using the homologous annotation method, the 177 genomes were predicted to contain an average of 16,461 protein-coding genes, similar to previously published bird genomes[4]. The average gene length and coding sequence length are ~15 and 1.26 kb, respectively. Genes contained on average ~7 exons, with an average length of 172 bp and an average intron length of 2.17 kb (**Supplementary Table 1**). Across most annotation and assembly metrics, there was no significant difference between the two Linked-Reads technologies (10X Genomics and stLFR), both of which outperformed the standard short-read next-generation sequencing method (**Figure 5**). Investigations across different avian lineages showed that Piciformes have a distinctly higher GC content with larger genome size variation (**Supplementary Table 1** and **Figure 6**), which is consistent with prior research [22].

Re-use potential

We created this dataset to support future research on avian biology. While it will be useful for broad avian evolutionary studies, we expect its most significant value will be in conservation genetics and the practical management of the threatened species. Specifically, these genomic resources enable the assessments of genetic diversity in endangered birds, providing essential data to identify inbreeding depression or genetic bottlenecks. The dataset facilitates precise population genetic structure analysis and the mapping of gene flow, which are critical for designing effective translocation or captive breeding programs. Furthermore, we expect its most significant value will be in studying how global change affects the loss of genomic diversity in

avian species, which is a necessary input to cope with the current challenges of the Anthropocene[23,24].

However, since our gene prediction primarily relies on a homology-based pipeline, there may be reduced sensitivity in identifying lineage-specific genes that lack close homologs in reference databases. Additionally, potential gene fragmentation effects may exist in regions where the assembly is less contiguous, which is a common trade-off in homology-based structural annotations. It should be noted that while this study provides the protein-coding gene annotations, the identification and characterization of non-coding RNAs (ncRNAs), such as miRNAs and lncRNAs, were not conducted in the current study. Future studies leveraging small RNA sequencing and specialized bioinformatic pipelines will be necessary to fully capture the regulatory landscape of these birds' genomes.

Despite these, the potential uses include population genomics, mapping genetic structure and gene flow, identifying hybrid zones, studying local adaptation, and investigating disease ecology. The quality of the genomes released here represents a level that was feasible at the time of this project, given the financial, technical, and biosample constraints available to the project. Nevertheless, we believe the quality will be suitable for many of the above possible end uses.

Supplementary data

Supplementary Table is available online.

Data Availability

Genome assemblies, and annotations of the 177 species generated in this study have been deposited under the BioProject accession number PRJEB102986 (ENA). Sample information for each genome and the genome statistics can also be viewed online at B10K website[25]. Code to run the genome assembly pipeline can be found at [8] and the corresponding genome annotation pipeline [11]. All additional supporting data are available in the GigaScience repository, GigaDB [26].

Abbreviations

AviList: The Global Avian Checklist; NCBI: National Center for Biotechnology Information; stLFR: Single-Tube Long Fragment Reads; B10K: The Bird 10,000 Genomes (B10K) Project; BUSCOs: Benchmarking Universal Single-Copy Orthologs; NGDC: National Genomics Data Center.

Declarations

All field sampling efforts followed appropriate local permission structures, laws, and regulations, including the Nagoya Protocol and Institutional Review Boards where they apply. All tissue samples were received and processed at Globe Institute, University of Copenhagen, which is certified to receive and process biological tissues under a permit for Animal By-Products DK-123-oth-905637 issued by The Danish Veterinary and Food Administration. CITES samples received from outside the European Union are under the CITES Scientific Exchange Exemption, registration number DK 014 issued by the Danish Environmental Protection Agency.

Competing interests

The authors declare no competing interests.

Funding

This work was supported by the National Key Research and Development Program of China (no. 2024YFA1802500) to G.Z.; the National Natural Science Foundation of China grant (no. 32422009) and the National Key Research and Development Program of China grant (no. 2023YFA1800500) to S.F.; the Postdoctoral Fellowship Program of CPSF Grant (no. GZB2025059) to G.C.; Danish National Research Foundation (DNRF143) to M.T.P.G.; Independent Research Fund Denmark 1054-00039B to P.A.H., who was also supported by a research grant (25925) from VILLUM FONDEN; J.C.I. was funded by two research grants from the Spanish Ministry of Science, Innovation and Universities and the European Regional Development Fund (PGC2018-097575-B-I00; PID2022-140091NB-I00); K.A.J acknowledges support from the Villum Foundation (Young Investigator Programme, project No. 15560, K.A.J.), the Carlsberg Foundation (CF15-0079, K.A.J.).

Authors' contributions

G.Z., M.T.P.G., E.D.J., P.A.H., C.R., F.L., G.R.G. and S.F. conceived of the study. A.M., Z.D.B., M.M, D.O., L.S., J.Fi., N.T.V., A.D.G., C.A.S., M.R., S.M.B., A.D.L., A.R., J.Fu., C.H., J.A.B., J.T.L., A.J., L.E.J., K.L., K.A.J., A.H.R., M.I., L.J., O.H., R.B., T.M.B., J.C.I., A.A., B.S., F.R., Je.L., I.N., J.Q., J.P.D., M.S., M.A., C.C.W., R.A.Ph., R.Pr., K.Z., S.M.G., M.J.R., U.O., Y.B., S.I., V.B., E.O., B.K., T.H., F.L., G.R.G. and P.A.H. provided genetic samples from fieldwork. G.C., S.W., D.B.O., S.D.N., W.D., W.J., Ji.L., W.H., C.Z., Q.L. and B.P. performed sample acquisition, lab work, and bioinformatics. G.C., S.F., P.A.H., M.T.P.G. and G.Z. drafted the manuscript, which was edited and approved by all authors.

Acknowledgements

The authors would like to thank the various field workers, museums, universities and other genetic resource archives, permitting officials, and other support and administrative staff without whom sample collection programs could not function: Natural History Museum Denmark

(NHMD), University of Copenhagen, Smithsonian National Museum of Natural History (USNM), Natural History Museum of Oslo (NHMO), Senckenberg Natural History Collections Dresden (SNSD), University of Eswatini, University of Pretoria, CIBET-Universidad de Costa Rica, Leibniz Institute for Zoo and Wildlife Research (IZW), Freie Universität Berlin (FUB), Wilhelma Zoological-Botanical Gardens Stuttgart, University of Veterinary Medicine Hannover (TiHo Hannover), Burke Museum of Natural History and Culture (UWBM), University of Washington (UW), Muséum national d'Histoire naturelle (MNHN), Copenhagen Zoo, South African National Biodiversity Institute (SANBI), Swedish Museum of Natural History (NRM), Australian National Wildlife Collection (CSIRO ANWC), Lund University, Louisiana State University (LSU), University of Lethbridge (ULeth), Museu Paraense Emílio Goeldi (MPEG), Finnish Museum of Natural History (LUOMUS), University of Helsinki, Rhodes University, National University of Singapore (NUS), University of Oviedo (UOV), Mandai Nature, National Museum of Nature and Science Japan (NSMT), Natural Sciences Museum of Barcelona (MZB), California Academy of Sciences (CAS), Natural History Museum Bern (NMBE), University of Bern, Museum of Southwestern Biology (MSB), University of New Mexico (UNM), British Antarctic Survey (BAS), Yale Peabody Museum (YPM), Field Museum of Natural History (FMNH), Université d'Antananarivo, Association Vahatra, A.P. Leventis Ornithological Research Institute (APLORI), University of Jos, Indianapolis Zoo, Charles University, Institute of Zoology, Chinese Academy of Sciences (IOZ), and NTNU University Museum. We also thank the Zhejiang Lab, the Information Technology Center of Zhejiang University and China Mobile Zhejiang Co., Ltd (Hangzhou Branch) for providing computational resources.

Reference

1. Howard R, Moore A, Dickinson EC, Remsen JV. The Howard and Moore Complete Checklist of the Birds of the World. Aves Press;
2. Rheindt FE, Donald PF, Donsker DB, Gerbracht JA, Iliiff MJ, Lepage D, et al. AvIList: a unified global bird checklist. *Biodivers Conserv*. Springer Science and Business Media LLC; 2025; doi: 10.1007/s10531-025-03120-y.
3. Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*. American Association for the Advancement of Science (AAAS); 2014; doi: 10.1126/science.1251385.
4. Feng S, Stiller J, Deng Y, Armstrong J, Fang Q, Reeve AH, et al. Dense sampling of bird diversity increases power of comparative genomics. *Nature*. 2020; doi: 10.1038/s41586-020-2873-9.
5. Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*. 2014; doi: 10.1126/science.1253451.
6. Stiller J, Feng S, Chowdhury A-A, Rivas-González I, Duchêne D, Fang Q, et al. Complexity of avian evolution revealed by family-level genomes. *Nature*. Nature Publishing Group UK; 2024; doi: 10.1038/s41586-024-07323-1.

7. Wang O, Cheng X, Drmanac R, Peters BA. A simple cost-effective method for whole-genome sequencing, haplotyping, and assembly. *Methods Mol Biol*. New York, NY: Springer US; 2023; doi: 10.1007/978-1-0716-2819-5_7.
8. A pipeline to de novo assemble the stLFR reads using Supernova Assembler. https://github.com/BGI-Qingdao/stlfr2supernova_pipeline Accessed 2026 Apr 13.
9. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid genome sequences. *Genome Res*. 2017; doi: 10.1101/gr.214874.116.
10. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*. 2012; doi: 10.1186/2047-217X-1-18.
11. Scripts used in the annotation of B10K genomes. <https://github.com/B10KGenomes/annotation>.
12. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997; doi: 10.1093/nar/25.17.3389.
13. She R, Chu JS-C, Wang K, Pei J, Chen N. GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res*. 2009; doi: 10.1101/gr.082081.108.
14. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. *Genome Res*. 2004; doi: 10.1101/gr.1865504.
15. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004; doi: 10.1093/nar/gkh340.
16. Claramunt S, Sheard C, Brown JW, Cortés-Ramírez G, Cracraft J, Su MM, et al.. A new time tree of birds reveals the interplay between dispersal, geographic range size, and diversification. *Curr Biol*. Elsevier BV; 2025; doi: 10.1016/j.cub.2025.07.004.
17. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*. Oxford University Press (OUP); 2004; doi: 10.1093/bioinformatics/btg412.
18. Yu G. Using ggtree to Visualize Data on Tree-Like Structures. *Curr Protoc Bioinformatics*. 2020; doi: 10.1002/cpbi.96.
19. Foreign Contamination Screening caller scripts and documentation. <https://github.com/ncki/fcs>.
20. Huang N, Li H. compleasm: a faster and more accurate reimplement of BUSCO. *Bioinformatics*. 2023; doi: 10.1093/bioinformatics/btad595.
21. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. 2020; doi: 10.1186/s13059-020-02134-9.
22. Kapusta A, Suh A. Evolution of bird genomes—a transposon’s-eye view. *Ann N Y Acad Sci*. Wiley;

2017; doi: 10.1111/nyas.13295.

23. Cassin-Sackett L, Welch AJ, Venkatraman MX, Callicrate TE, Fleischer RC. The contribution of genomics to bird conservation. *Avian Genomics in Ecology and Evolution*. Cham: Springer International Publishing; doi: 10.1007/978-3-030-16477-5_10.

24. Bernatchez L, Ferchaud A-L, Berger CS, Venney CJ, Xuereb A. Genomics for monitoring and understanding species responses to global climate change. *Nat Rev Genet*. 2024; doi: 10.1038/s41576-023-00657-y.

25. B10K Website. <https://b10k.com/>.

26. Chen G, Wang S, Øksnebjerg D B, Nielsen S D, Dai W, Jiang W, et al. Supporting data for "Draft assemblies for 177 bird species enhance genus-level coverage." GigaScience Database. 2026; doi:10.5524/102772.

27. Bird species reference genomes in the NCBI database.

<https://www.ncbi.nlm.nih.gov/genome/browse/#!/aves/> Accessed 2025 Oct 11.

28. The IUCN Red List of Threatened Species. <https://www.iucnredlist.org>.

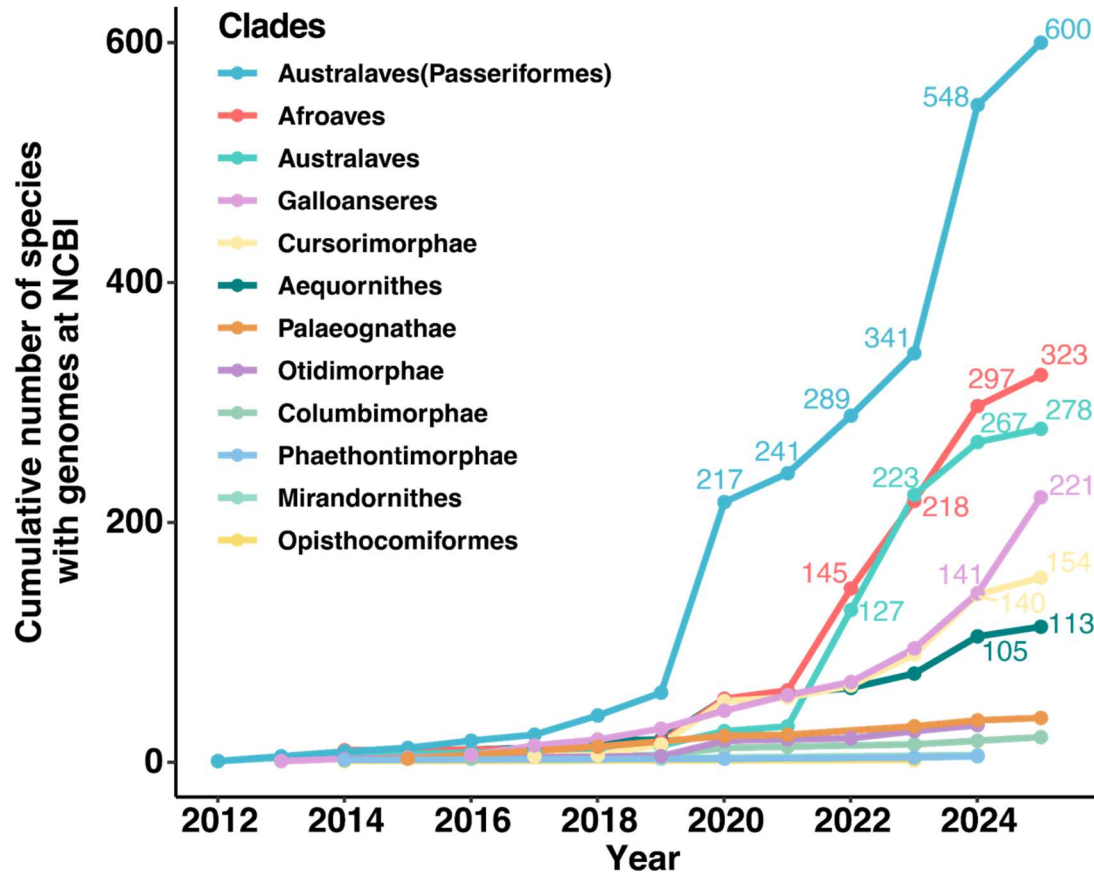


Figure 1. Temporal accumulation of avian reference genomes in the NCBI database. Cumulative number of bird species with reference genomes in the NCBI database (obtained from [27] on 11 October 2025). Data points are color-coded by taxonomic clades to illustrate the expanding genomic representation across different avian lineages.

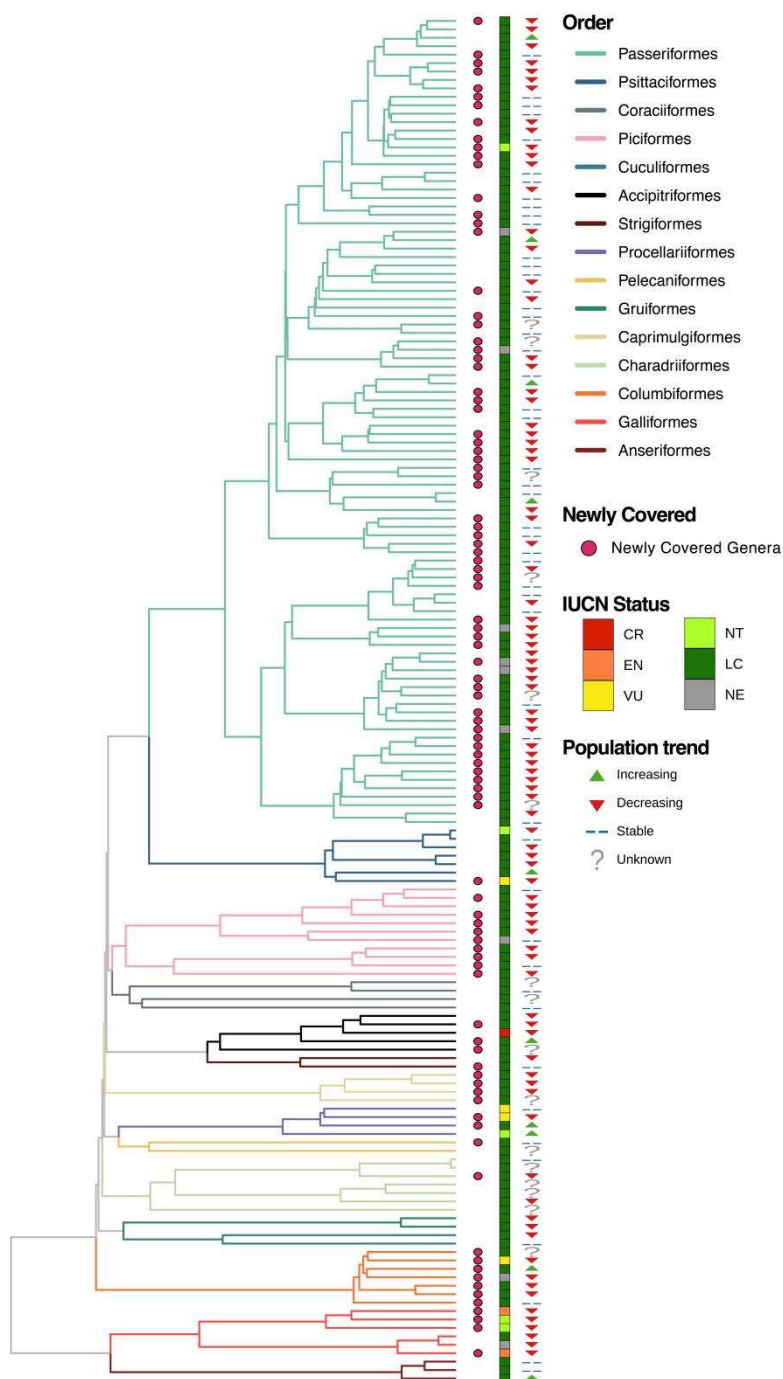


Figure 2. Phylogenetic distribution, taxonomic representation, and conservation profiles of the 177 newly sequenced avian genomes. Branches and internal clades are color-coded according to taxonomic orders. Red circles at the tips indicate newly covered genera. The color-coded strip represents the IUCN Red List categories (CR: Critically Endangered; EN: Endangered; VU: Vulnerable; NT: Near Threatened; LC: Least Concern; NE: Not Evaluated; obtained from [28]). And the symbols indicate the population trend, including increasing, decreasing, stable, and unknown.

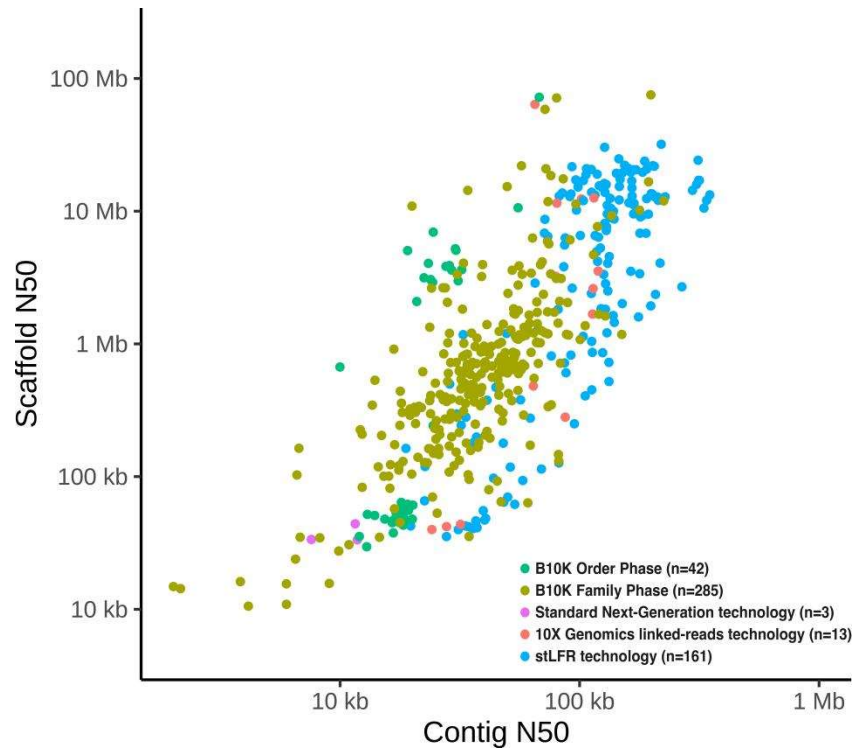


Figure 3. Evaluation of assembly continuity for 504 avian genomes. Scatter plot of Contig N50 versus Scaffold N50 metrics. Each dot represents a species. The dataset comprises 177 genomes from this study and 327 legacy genomes from the B10K project (Order Phase and Family Phase).

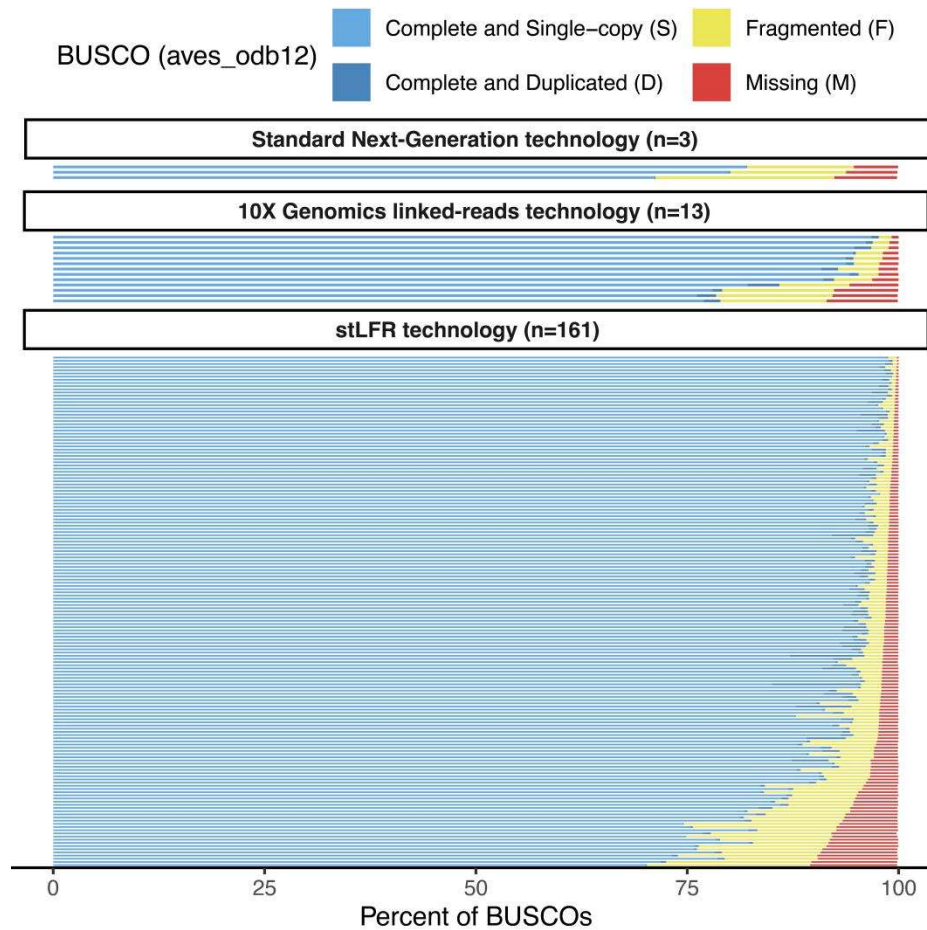


Figure 4. Assessment of genome assembly completeness for 177 newly sequenced avian genomes. Distribution of BUSCO scores (aves_odb12) for the 177 newly assembled genomes. Each bar represents an individual species, color-coded by gene category: complete single-copy, complete duplicated, fragmented, and missing.

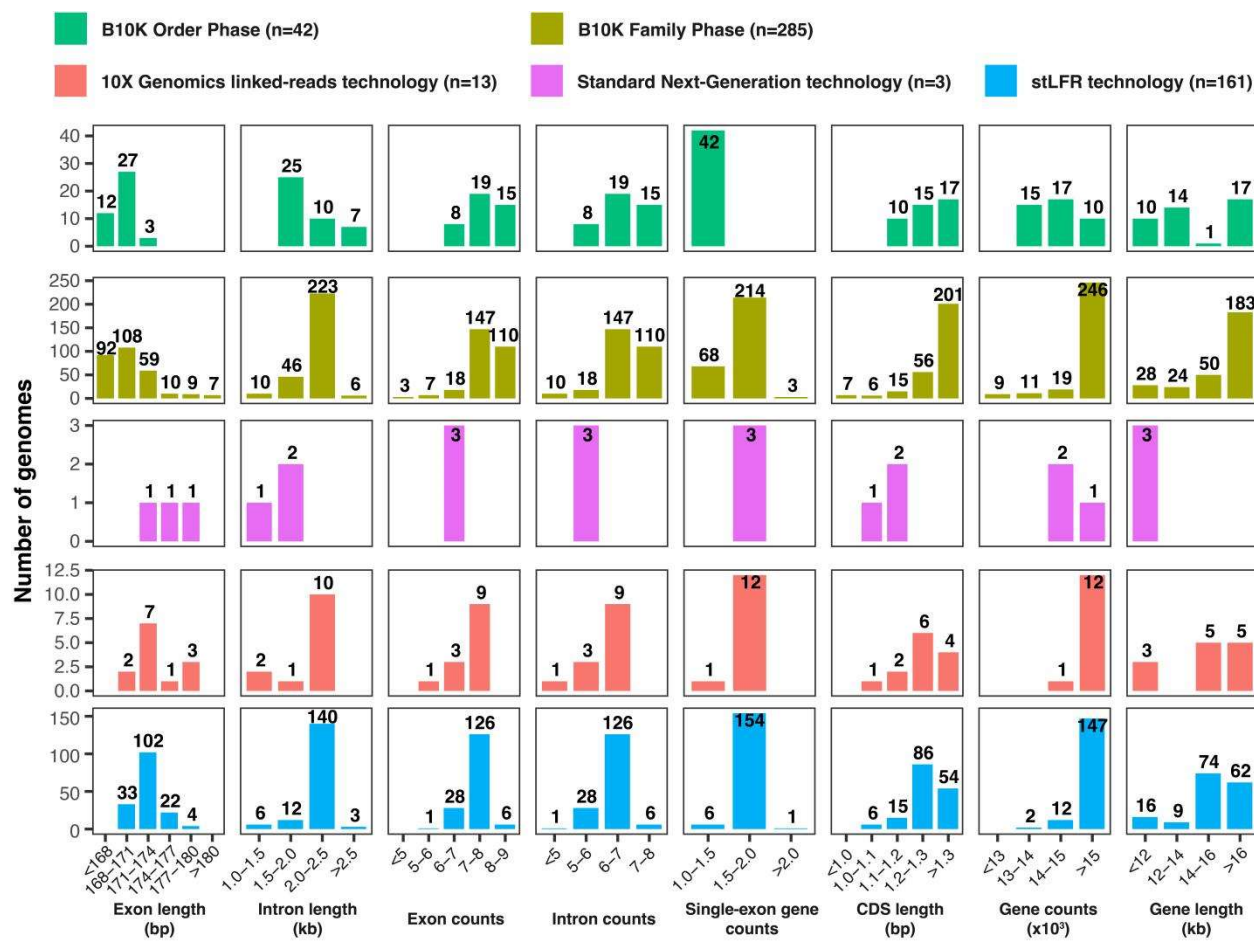


Figure 5. Evaluation of gene annotation quality and consistency. Distribution of eight primary annotation indicators for 504 avian genomes, stratified by B10K project phase and sequencing technology (stLFR, 10X Genomics, and NGS). From left to right, the panels display the frequency distributions of mean exon and intron lengths, mean numbers of exons and introns, single-exon gene counts, mean CDS lengths, total gene numbers, and mean gene lengths.

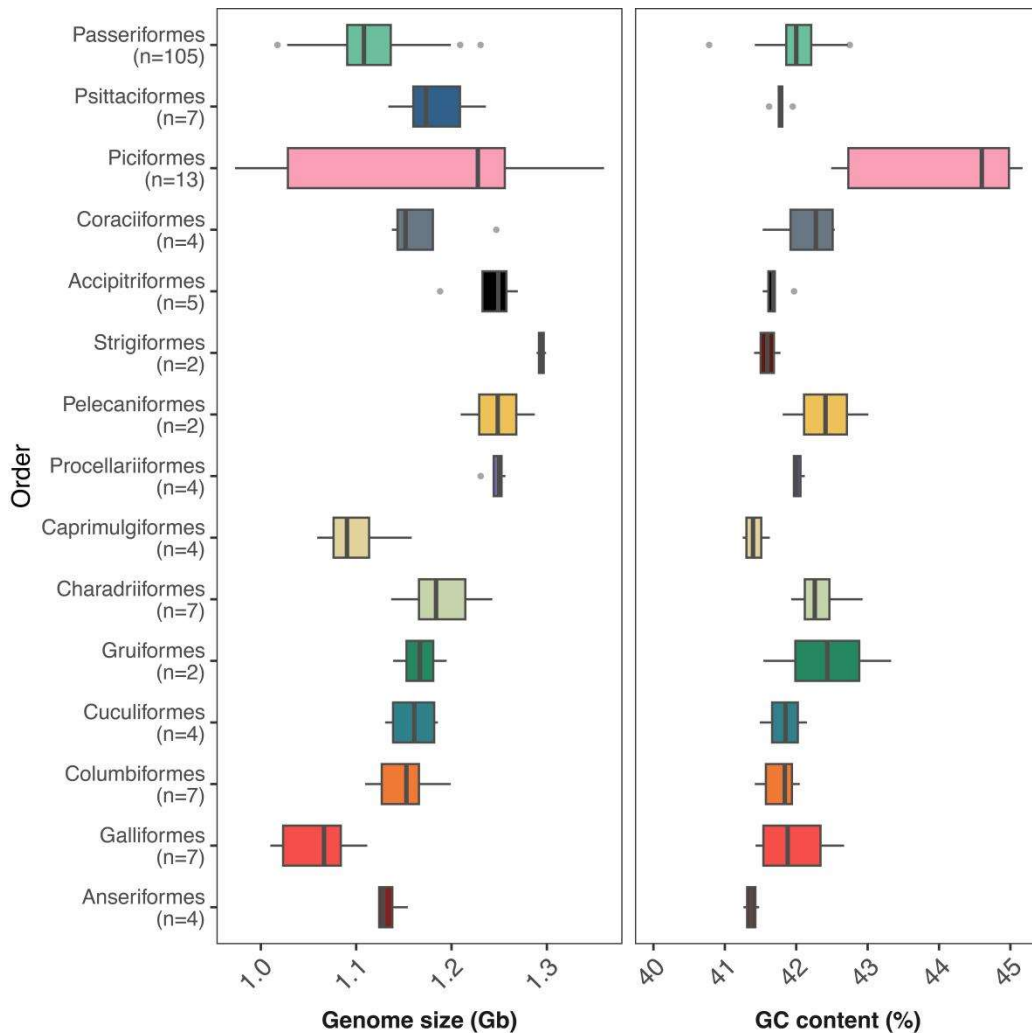


Figure 6. Comparative analysis of genome size and GC content across major avian taxonomic orders. Boxplots illustrating the distribution of estimated genome sizes (Gb) and GC content (%) for the 177 sequenced avian species, grouped by taxonomic order. The number of species (n) within each order is indicated in parentheses along the y-axis. For each boxplot, the thick vertical line represents the median, the box boundaries indicate the first and third quartiles (interquartile range, IQR), and the whiskers extend to the rest of the distribution, except for points that are determined to be outliers, which are plotted as individual gray dots.