


The De Novo Genome Sequencing of Silver Pheasant (*Lophura nycthemera*)

Xue-Juan Li ¹, Xiao-Yang Wang², Chao Yang^{1,3}, Li-Liang Lin¹, Le Zhao⁴, Xiao-Ping Yu¹, Fu-Min Lei⁵, and Yuan Huang^{1,*}

¹College of Life Sciences, Shaanxi Normal University, Xi'an, China

²School of Biological and Environmental Engineering, Xi'an University, China

³Shaanxi Institute of Zoology, Xi'an, China

⁴School of Biological Sciences and Engineering, Shaanxi University of Technology, Hanzhong, China

⁵Key Laboratory of the Zoological Systematics and Evolution, Institute of Zoology, The Chinese Academy of Sciences, Beijing, China

*Corresponding author: E-mail: yuanh@snnu.edu.cn.

Accepted: 6 December 2021

Abstract

Silver pheasant (*Lophura nycthemera*) belongs to Phasianidae, Galliformes, which exhibits high subspecific differentiation. In this study, we assembled a novel genome based on 98.42 Gb of Illumina sequencing data and 30.20 Gb of PacBio sequencing data. The size of the final assembled genome was 1.01 Gb, with a contig N50 of 6.96 Mb. Illumina paired-end reads (94.96%) were remapped to the contigs. The assembled genome shows high completeness, with a complete BUSCO score of 92.35% using the avian data set. A total of 16,747 genes were predicted from the generated assembly, and 16,486 (98.44%) of the genes were annotated. The average length of genes, exons, and introns were 19,827.53, 233.69, and 1841.19 bp, respectively. Noncoding RNAs included 208 miRNAs, 40 rRNAs, and 264 tRNAs, and a total of 189 pseudogenes were identified; 116.31 Mb (11.47%) of the genome consisted of repeat sequences, with the greatest proportion of LINEs. This assembled genome provides a valuable reference genome for further studies on the evolutionary history and conversion genetics of *L. nycthemera* and the phylogenomics of the Galliformes lineage.

Key words: *Lophura nycthemera*, PacBio sequencing, genome assembly.

Significance

The silver pheasant (*Lophura nycthemera*) is one of the least known pheasants of the world with a highly subspecific divergence. The high-quality reference genome assembly and annotation of *L. nycthemera* revealed several evolutionary features. This article provided a basic reference genome for facilitating studies on genomic characteristics and genome-based population divergence of *L. nycthemera* and phylogenomics of all pheasants of the world.

Introduction

The development of high-throughput sequencing technology represented the beginning of a new era of genomic studies (Giordano et al., 2017), involving platforms such as Illumina, Pacific, and Nanopore sequencing. Genome sequences of birds, such as *Gallus gallus* (e.g., International Chicken Genome Sequencing Consortium, 2004), *Pseudopodoces humilis* (Qu et al., 2013), and *Zosterops lateralis* (Cornetti

et al., 2015), facilitated by sequencing technologies, provide important information on avian evolution. Through comparative genomic analyses among Galliformes, some significant features have been found, such as characteristics related to high-altitude adaptation (Wang et al., 2015; Lee et al., 2018; Cui et al., 2019) and the coloration and pigmentation of plumage (Gao et al., 2018; Dhar et al., 2019), in addition to genetic and evolutionary characteristics (Jiang et al., 2019; Zhou et al., 2019). Recently, several Galliformes genomes

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

have been described in GenBank database, which species covered all five families. Although long-read sequencing platform has been used to obtain Galliformes genomes, most of Galliformes genomes were also sequenced by using Illumina HiSeq sequencing technology.

Silver pheasant (*Lophura nycthemera*) (Phasianidae, Galliformes) is widely distributed in southern China, eastern Myanmar, northern Thailand, and the Indo–China Peninsula (Cheng et al., 1978), with a forest canopy coverage preference ranging from altitudes of 0–1,000 m (BirdLife International, Chen et al., 2019). It exhibits 15 subspecies, nine of which occur in China (Johnsgard 1999; Gill and Donsker, 2020). The plumage pattern of the upper parts of the males and topographic barriers are used to establish its taxa and relationships (Delacour 1948). Females are smaller than males, with polygamous lifestyles (Grimmett et al., 1999). For *L. nycthemera*, some subspecies with limited ranges exhibit potential conservation problems due to habitat loss and other influences (McGowan and Garson, 1995).

To study the basic genomics of *L. nycthemera* and explore the evolution of all Galliformes, we performed the de novo genome sequencing by combining the Illumina and PacBio platforms. In addition, based on the assembled results, we also studied its genomic features. This study provides a high-quality genome assembly of *L. nycthemera*, and will be helpful for further studying evolutionary features of Galliformes species.

Results and Discussion

Genome Assembly and Completeness Assessment

In this study, approximately 98.42 Gb of raw sequencing data were obtained from Illumina platform, with a sequencing depth of 93.73× (supplementary table S1, Supplementary Material online). The PacBio sequencing platform generated ~30.20 Gb of raw data. A total assembly of 1.01 Gb with a contig N50 of ~6.96 Mb was obtained. The genome size was similar to that of some other Galliformes species, such as 1.09 Gb of *Arborophila rufipectus* (Zhou et al., 2019). The contig number, contig length, contig N90, contig max, and GC content of the assembly genome were 1,553, 1,014,408,745 bp, 643,154 bp, 23,586,999 bp, and 41.38%, respectively.

To assess assembled results, Illumina paired-end reads were remapped to the assembled genome, and 94.96% reads could be mapped to the contigs (supplementary table S2, Supplementary Material online). In addition, a total of 7,700 complete BUSCOs (92.35%) were identified in the assembly (fig. 1A). These results showed that the assembled genome was complete and presented a low error ratio. The complete and single-copy BUSCOs (92.18%) was higher than that of *A. rufipectus* (86.5%) (Zhou et al., 2019). The high-quality reference genomes of *L. nycthemera* could be a useful tool to understand genomic evolution.

Gene Prediction and Functional Annotation

The consensus gene set contained 16,747 genes. The lengths of the genes, exons, CDS, and introns were shown in supplementary table S3, Supplementary Material online. The averages of gene length, exon length, and intron length were 19,827.53, 233.69, and 1841.19 bp, respectively. A total of 13,447 (80.29%) genes were supported by all three methods (fig. 1B), which represented a good gene prediction effect. A total of 16,486 (98.44%) predicted genes were successfully annotated by using nine databases (supplementary table S4, Supplementary Material online). The noncoding RNAs included 208 miRNAs, 40 rRNAs, and 264 tRNAs, which belonged to 100, 4, and 23 families, respectively. In addition, a total of 189 pseudogenes were identified.

Repeat Sequences Annotation

It was estimated that 116.31 Mb (11.47%) of the genome consisted of repeat sequences (fig. 1C and supplementary table S5, Supplementary Material online). The percentage of repeat sequences was larger than those of other Galliformes species, such as 9.02% in *A. ardens* (Zhou et al., 2019) and 9.82% *G. gallus* (Dhar et al., 2019). Within Class I, the lengths of LINEs and SINEs sequences were 74 and 0.2 Mb, with percentages of 7.30% and 0.02%, respectively (supplementary table S5, Supplementary Material online). The LINEs represented the greatest proportion of the genome, which was also found in other avian genomes, such as *Pavo cristatus* (Dhar et al., 2019). Within Class II, 11,656,729 bp (1.15%) of TIR sequences were identified (supplementary table S5, Supplementary Material online).

Materials and Methods

Sampling and Sequencing

A *L. nycthemera* female species was collected from captive breeders in Lantian, Xi'an, Shaanxi Province, China in 2016. The muscle tissues were used for sequencing. DNA was extracted by the CTAB method, with DNA concentrations and quality measured by a NanoDrop 2000 system and a Qubit Fluorometer. Total RNA was extracted using TRIzol, with RNA concentrations measured by a NanoDrop 2000 system and an Agilent 2100 Bioanalyzer.

The Illumina HiSeq X-Ten and PacBio Sequel pipelines were used for genome sequencing. For the Illumina platform, five short-fragment paired-end libraries, including three of 270 bp and two of 350 bp, were constructed via Illumina sequencing. The genomic DNA was randomly fragmented by using the ultrasonic method, and target fragments were then filtered. A small-fragment sequencing library was constructed through a series of steps, including end repair, the addition of A and adaptor sequences, target fragment selection, and PCR. The size and quality of the libraries were detected by using an

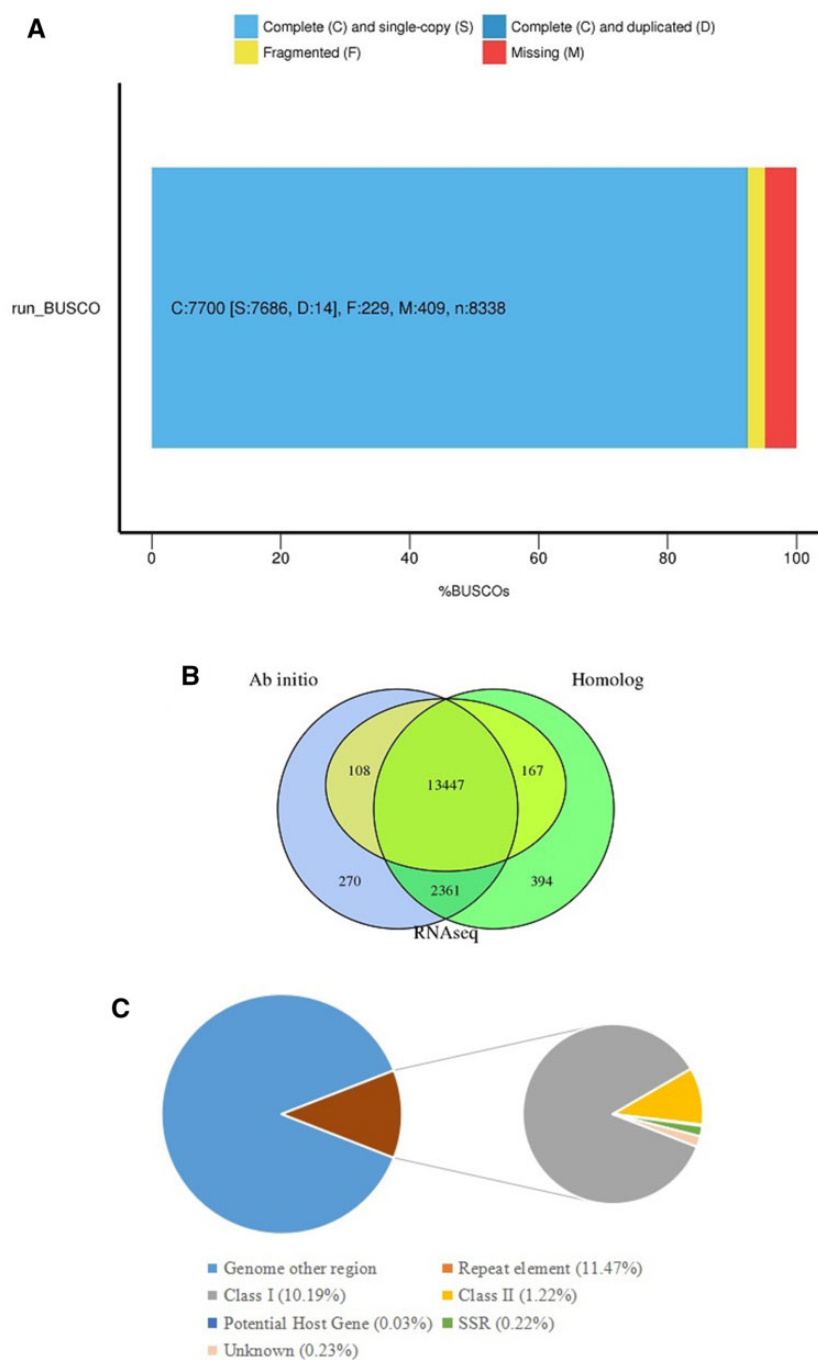


FIG. 1.—Statistics of genome assembly assessment, gene structure prediction and repeat element percentage of *Lophura nycthemera*. (A) BUSCO assessment results. (B) Predict results of gene structures using ab initio-based, homologue-based, and RNA-seq-based methods. (C) Percentage of repeat elements.

Agilent 2100 system and Q-PCR. Illumina double-ended sequencing with PE = 150 was applied. For the PacBio platform, long-fragment libraries were constructed. The DNA samples were sheared by using g-TUBE, and DNA damage was then repaired and end-repaired. Dumbbell-type adapters were

ligated using exonuclease digestion. For the sequencing libraries, target segment selection was performed using BluePippin.

The Illumina HiSeq X-Ten pipeline was also used to obtain RNA sequences. For RNA fragment libraries, rRNA was isolated from total RNAs, and then fragmented randomly. The

first-strand cDNA was synthesized using random hexamer primers by employing the fragmented rRNA-depleted RNA as a template. The second-strand cDNA was synthesized using DNA polymerase I and RNase H. After end-repair, A-tail, adaptor ligation, and purification, PCR amplification was conducted.

Genome Assembly and Assessment

After filtering low-quality and short length reads from the PacBio data, Wtdbg2 (Ruan and Li, 2020) was used for assembly. Pilon was used to correct this assembly results by using Illumina data with three times. Two methods were employed to assess assembled results, that is, Illumina paired-end reads remapped to the assembled genome, and BUSCO v4 databases (Waterhouse et al., 2018) with *aves_odb09* employed.

Repetitive Sequence Annotation

The database of repeat sequences was constructed using structure-based and ab initio-based strategies, employing LTR-FINDER v1.05 and RepeatScout v1.05. This database was classified by PASTEClassifier, and merged with the Repbase database into a final database of repeat sequences. RepeatMasker v4.0.6 (Tarailo-Graovac and Chen, 2009) was used to predict repeat sequences.

Gene Prediction and Function Annotation

Three strategies were employed to predict gene structures, including ab initio-based, homologue-based and RNA-seq-based methods. Genscan, Augustus v2.4, GlimmerHMM v3.0.4, GeneID v1.4, and SNAP (version 2006-07-28) were used for ab initio-based prediction. GeMoMa v1.3.1 was employed for homologue-based prediction, mainly employing six species (*G. Gallus*, *Meleagris gallopavo*, *Taeniopygia guttata*, *Ficedula albicollis*, *Parus major*, and *Coturnix japonica*). Hisat v2.0.4 and Stringtie v1.2.3 were used for assembly based on the referenced RNA-seq data. TransDecoder v2.0 and GeneMarkS-T v5.1 were used for predicting genes. PASA v2.0.2 was used to predict for assembled unigene sequences based on RNA-seq data without references. In addition, EVMv1.1.1 (Haas et al., 2008) was used to integrate the above prediction results, and PASA v2.0.2 was employed for modification.

For ncRNAs, microRNAs and rRNAs were predicted through genome alignment using Blastn, employing the Rfam database. TRNAscan-SE v1.3.1 was used to predict tRNAs. For pseudogenes, based on GenBlastA v1.0.4 alignment, homologous gene sequences were searched in the genome. GeneWise v2.4.1 was employed to search for premature stop codons and frame shifts, and to identify pseudogenes.

To assign gene functions, we aligned the genes to nine functional databases by using BLAST v2.2.3, with *E*-value = 1e-5. The databases included COG, GO, KEGG, KOG, Pfam, Swissprot, TrEMBL, eggNOG, and NR.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (31601846 and 31801993), Natural Science Foundation of Shaanxi Province, China (2020JM-270), Fundamental Research Funds for the Central Universities, China (GK202107011, GK202101003, and GK202103068), and Postdoctoral Science Foundation of Shaanxi Province, China (2017BSHEDZZ99). We are very grateful to Prof. Gang Li, Dr. Jie Yang, and Hao Yuan for their valuable suggestions.

Data Availability

The whole-genome project has been deposited at GenBank. The genome of *Lophura nycthemera* is accessible under accession number JADANK000000000. The National Center for Biotechnology Information (NCBI) BioProject and BioSample accession numbers are PRJNA354685 and SAMN06118903, respectively. The raw reads have been deposited in the NCBI SRA database with the accession numbers SRR12459247 (Illumina RNA-seq), SRR12459249 (Illumina DNA-seq), and SRR12459248 (PacBio).

Literature Cited

- Chen LJ, et al. 2019. Combined effects of habitat and interspecific interaction define co-occurrence patterns of sympatric Galliformes. *Avian Res.* 10(1):29.
- Cheng TS, et al. 1978. *Fauna Sinica, series Vertebrata, Aves, Vol. 4: Galliformes.* Beijing (China): Science Press.
- Cornetti L, et al. 2015. The genome of the “Great Speciator” provides insights into bird diversification. *Genome Biol Evol.* 7(9):2680–2691.
- Cui K, et al. 2019. The first draft genome of *Lophophorus*: a step forward for Phasianidae genomic diversity and conservation. *Genomics* 111(6):1209–1215.
- Delacour J. 1948. The subspecies of *Lophura nycthemera*. *Am Mus Novit.* 1377:1–12.
- Dhar R, et al. 2019. De novo assembly of the Indian blue peacock (*Pavo cristatus*) genome using Oxford Nanopore technology and Illumina sequencing. *Gigascience* 8(5):giz038.
- Gao G, et al. 2018. Comparative genomics and transcriptomics of *Chrysolophus* provide insights into the evolution of complex plumage coloration. *Gigascience* 7(10):gij113.
- Gill F, Donsker D, editors. 2020. IOC world bird list (v 10.1). Dataset. Available from: <http://www.worldbirdnames.org/>.
- Giordano F, et al. 2017. De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Sci Rep.* 7(1):3935.

- Grimmett R, Inskipp C, Inskipp T. 1999. Birds of India: Pakistan, Nepal, Bangladesh, Bhutan, Sri Lanka, and the Maldives, Princeton (NJ): Princeton University Press.
- Haas BJ, et al. 2008. Automated eukaryotic gene structure annotation using EVIDENCEModeler and the program to assemble spliced alignments. *Genome Biol.* 9(1):R7.
- International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432(7018):695–716.
- Jiang L, et al. 2019. Systematic identification and evolution analysis of *Sox* genes in *Coturnix japonica* based on comparative genomics. *Genes(Basel)* 10(4):314.
- Johnsgard P. 1999. Pheasants of the world. 2nd ed. Oxford: Oxford University Press.
- Lee CY, et al. 2018. Whole-genome de novo sequencing reveals unique genes that contributed to the adaptive evolution of the Mikado pheasant. *Gigascience* 7(5):gij044.
- McGowan PJK, Garson PJ. 1995. Pheasants: Status survey and conservation action plan 1995–1999. Gland (Switzerland): IUCN.
- Qu Y, et al. 2013. Ground tit genome reveals avian adaptation to living at high altitudes in the Tibetan plateau. *Nat Commun.* 4:2071.
- Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. *Nat Methods.* 17(2):155–158.
- Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics.* Chapter 4:Unit 4.10.
- Wang MS, et al. 2015. Genomic analyses reveal potential independent adaptation to high altitude in Tibetan chickens. *Mol Biol Evol.* 32(7):1880–1889.
- Waterhouse RM, et al. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol.* 35(3):543–548.
- Zhou C, et al. 2019. The draft genome of the endangered Sichuan Partridge (*Arborophila rufipectus*) with evolutionary implications. *Genes (Basel)* 10(9):677.

Associate editor: Bonnie Fraser