


DATA NOTE

Whole genome and transcriptome maps of the entirely black native Korean chicken breed Yeonsan Ogye

Jang-il Sohn^{1,2,†}, Kyoungwoo Nam^{1,†}, Hyosun Hong^{1,†}, Jun-Mo Kim^{3,†}, Dajeong Lim⁴, Kyung-Tai Lee⁴, Yoon Jung Do⁴, Chang Yeon Cho⁵, Namshin Kim⁶, Han-Ha Chai^{4,7,*} and Jin-Wu Nam ^{1,2,*}

¹Department of Life Science, Hanyang University, Seoul, 133-791, Republic of Korea, ²Research Institute for Convergence of Basic Sciences, Hanyang University, Seoul, 133-791, Republic of Korea, ³Department of Animal Science and Technology, Chung-Ang University, Anseong, Gyeonggi-do, 17546, Republic of Korea, ⁴Department of Animal Biotechnology & Environment, National Institute of Animal Science, RDA, Wanju, 55365, Republic of Korea, ⁵Animal Genetic Resource Research Center, National Institute of Animal Science, RDA, Namwon, 55717, Republic of Korea, ⁶Personalized Genomic Medicine Research Center, KRIBB, Daejeon, 34141, Republic of Korea and ⁷College of Pharmacy, Chonnam National University, Kwangju, 61186, Republic of Korea

*Correspondence address: Jin-Wu Nam, E-mail: jwnam@hanyang.ac.kr  <http://orcid.org/0000-0003-0047-3687>; Han-Ha Chai, E-mail: hanha@korea.kr

[†]These authors contributed equally.

Abstract

Background: Yeonsan Ogye (YO), an indigenous Korean chicken breed (*Gallus gallus domesticus*), has entirely black external features and internal organs. In this study, the draft genome of YO was assembled using a hybrid *de novo* assembly method that takes advantage of high-depth Illumina short reads (376.6X) and low-depth Pacific Biosciences (PacBio) long reads (9.7X). **Findings:** The contig and scaffold NG50s of the hybrid *de novo* assembly were 362.3 Kbp and 16.8 Mbp, respectively. The completeness (97.6%) of the draft genome (Ogye.1.1) was evaluated with single-copy orthologous genes using Benchmarking Universal Single-Copy Orthologs and found to be comparable to the current chicken reference genome (galGal5; 97.4%; contigs were assembled with high-depth PacBio long reads (50X) and scaffolded with short reads) and superior to other avian genomes (92%–93%; assembled with short read-only or hybrid methods). Compared to galGal4 and galGal5, the draft genome included 551 structural variations including the fibromelanosis (FM) locus duplication, related to hyperpigmentation. To comprehensively reconstruct transcriptome maps, RNA sequencing and reduced representation bisulfite sequencing data were analyzed from 20 tissues, including 4 black tissues (skin, shank, comb, and fascia). The maps included 15,766 protein-coding and 6,900 long noncoding RNA genes, many of which were tissue-specifically expressed and displayed tissue-specific DNA methylation patterns in the promoter regions. **Conclusions:** We expect that the resulting genome sequence and transcriptome maps will be valuable resources for studying domestic chicken breeds, including black-skinned chickens, as well as for understanding genomic differences between breeds and the evolution of hyperpigmented chickens and functional elements related to hyperpigmentation.

Keywords: *Gallus gallus domesticus*; Yeonsan Ogye; whole genome *de novo* assembly; transcriptome maps; hyperpigmentation

Received: 15 January 2018; Revised: 19 May 2018; Accepted: 4 July 2018

© The Author(s) 2018. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Background

The *Yeosan Ogye* (YO), a designated natural monument of Korea (no. 265), is an indigenous Korean chicken breed that is notable for its entirely black plumage, skin, beak, comb, eyes, shank, claws, and internal organs [1]. In terms of its plumage and body color, as well as its number of toes, this unique chicken breed resembles the indigenous Indonesian chicken breed *Ayam cemani* [2–4]. YO also has some morphological features that are similar to those of the *Silkie* fowl, with the exception of the *Silkie*'s veiled black walnut comb and hair-like, fluffy plumage that is white or variably colored [5, 6]. Although the exact origin of the YO breed has not yet been clearly defined, its features and medicinal usages were recorded in *Dongui Bogam* [7], a traditional Korean medical encyclopedia compiled and edited by Heo Jun in 1613.

To date, a number of avian genomes from both domestic and wild species have been assembled and compared, revealing genomic signatures associated with the domestication process and genomic differences that provide an evolutionary perspective [8]. The chicken reference genome was first assembled using the red junglefowl [9], first domesticated at least 5,000 years ago in Asia; the latest version of the reference genome was released in 2015 (galGal5, GenBank Assembly ID GCA_000002315.3) [10]. However, because domesticated chickens exhibit diverse morphological features, including skin and plumage colors, the genome sequences of unique breeds are necessary for understanding their characteristic phenotypes through analyses of single nucleotide polymorphisms (SNPs), insertions and deletions (INDELs), structural variations (SVs), and coding and non-coding transcriptomes. Here, we provide the first version of the YO genome (Ogye.1.1), which includes annotations of large SVs, SNPs, INDELs, and repeats, as well as coding and noncoding transcriptome maps along with DNA methylation landscapes across 20 YO tissues.

Data Description

Sample collection

An 8-month-old YO chicken (object no. 02127), obtained from the Animal Genetic Resource Research Center of the National Institute of Animal Science (Namwon, Korea), was used in the study (Fig. 1A; [8–34]). All sequencing data in this study (including data from whole genome sequencing, RNA sequencing [RNA-seq], and reduced representation bisulfite sequencing [RRBS]) were obtained from this sample bird. The protocols for the care and experimental use of YO were reviewed and approved by the Institutional Animal Care and Use Committee of the National Institute of Animal Science (no. 2014-080). YO management, treatment, and sample collection took place at the National Institute of Animal Science.

Whole-genome sequencing

Genomic DNA was extracted from blood using the Wizard DNA extraction kit [35] and prepared for DNA sequencing library construction. According to the DNA fragment (insert) size, three different library types were constructed: paired-end libraries for small inserts (280 and 500 bp), mate-pair libraries for large inserts (3, 5, 8, and 10 Kbp), and FSMID libraries for very large inserts (40 Kbp) using Illumina's protocols (Illumina, San Diego, CA, USA) (Table 1). The constructed libraries were sequenced using Illumina's HiSeq2000 platform. In total, 376.6X raw Illumina short reads (100.2X from the small insert libraries and 276.4X

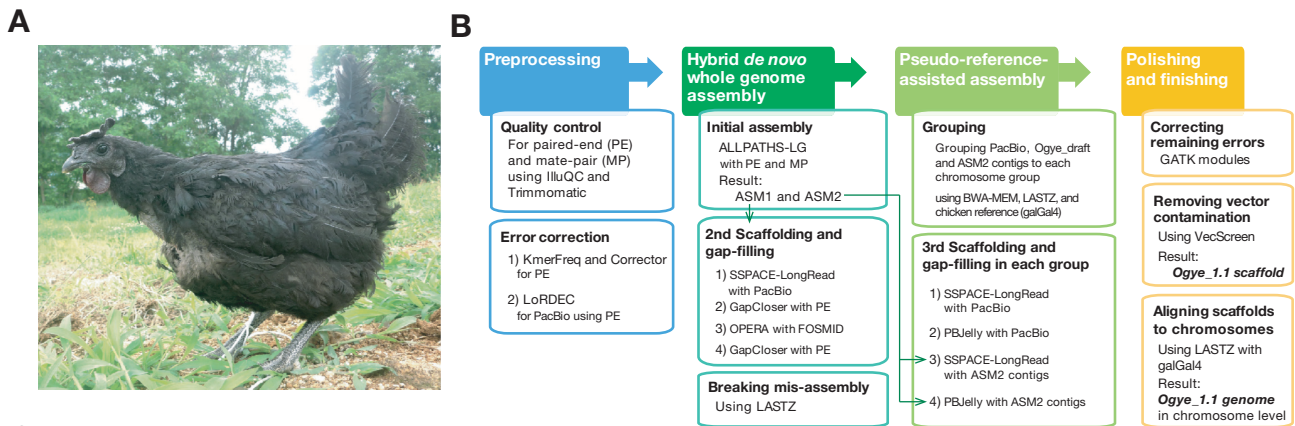
from the large insert libraries) were generated (Table 1 and Supplementary Table S1). To fill gaps and improve the scaffold N50, 9.7X Pacific Biosciences (PacBio) long reads were additionally sequenced using the PacBio RS II platform with P6C4 chemistry; the average length of the long reads was 6 Kbp (Table 1).

Whole transcriptome sequencing

Total RNAs were extracted from 20 tissues using 80% EtOH and TRIzol (Sigma-Aldrich, St. Louis, MO, USA). The RNA concentration was checked using Quant-IT RiboGreen (Invitrogen, Carlsbad, CA, USA). To assess the integrity of the total RNA, samples were run on the Agilent 2200 TapeStation system (Agilent Technologies, Waldbronn, Germany). Only high-quality RNA samples (RNA integrity number ≥ 7.0) were used for RNA-seq library construction. Each library was independently prepared with 300 ng of total RNA using an Illumina TruSeq Stranded Total RNA Sample Prep Kit (Illumina, San Diego, CA, USA). The rRNA in the total RNA was depleted using a Ribo-Zero kit. After rRNA depletion, the remaining RNA was purified, fragmented, and primed for cDNA synthesis. The cleaved RNA fragments were copied into the first cDNA strand using reverse transcriptase and random hexamers. This step was followed by second strand cDNA synthesis using DNA polymerase I, RNase H, and dUTP. The resulting cDNA fragments then underwent an end-repair process, the addition of a single "A" base, after which adapters were ligated. The products were purified and enriched with polymerase chain reaction (PCR) to create the final cDNA library. The libraries were quantified using qPCR according to the qPCR Quantification Protocol Guide (KAPA Library Quantification kits for Illumina sequencing platforms) and the integrity of the cDNA libraries was examined using the Agilent 2200 TapeStation system. In sum, about 1.5 billion RNA-seq reads were sequenced from the following 20 tissues from the same bird: breast, liver, bone marrow, fascia, cerebrum, gizzard, mature and immature eggs, comb, spleen, cerebellum, gallbladder, kidney, heart, uterus, pancreas, lung, skin, eye, and shank (Table 2).

Reduced representation bisulfite sequencing

RRBS libraries were prepared following Illumina's RRBS protocol. To prepare the libraries, 5 μ g of genomic DNA that had been digested with the restriction enzyme *MspI* and purified with a QIAquick PCR purification kit (QIAGEN, Hilden, Germany); a TruSeq Nano DNA Library Prep Kit (Illumina, San Diego, CA, USA) was used. Eluted DNA fragments were end-repaired, extended on the 3' end with an "A," and ligated with TruSeq adapters. The products, which ranged from 175 to 225 bp in length (insert DNA of 55–105 bp plus adaptors of 120 bp), were excised from 2% (w/v) Low Range Ultra Agarose gel (Biorad, Hercules, CA, USA) and purified using the QIAquick gel extraction protocol. The purified DNA underwent bisulfite conversion using the Epi-Tect Bisulfite Kit (Qiagen, 59 104). The bisulfite-converted DNA libraries were amplified by PCR (four cycles) using PfuTurbo Cx DNA polymerase (Agilent, 600 410). The quantity of the DNA libraries was then examined using qPCR, and the integrity was examined using the Agilent 2200 TapeStation system. The final product was sequenced using the HiSeq 2500 platform (Illumina, San Diego, CA, USA). Ultimately, 123 million RRBS reads were produced from 20 tissues from the same bird (see Table 3).



C

| Species | NCBI assembly name | Est. size (Gbp) | Assembly length (Gbp) | Pseudo-contig | | | Scaffold | | | Gaps in scaffold | | Assembly method | |
|---------------------------------|------------------------------|-----------------|-----------------------|---------------|-----------------|-------------|----------|-----------------|-------------|--------------------|--------------|-----------------|---------------------|
| | | | | Number | Ave. len. (Kbp) | NG50† (Kbp) | Number | Ave. len. (Kbp) | NG50† (Mbp) | Total length (Mbp) | Fraction (%) | Assembler | Sequencing platform |
| Chicken (Yeonsan Ogye) | Ogye_1.1 | 1.25 [10] | 1.00 | 8,241 | 119.8 | 362.3 | 1,906 | 517.8 | 16.8 | 8.5 | 0.85 | Our pipeline | I/P |
| Chicken (Red junglefowl) [8-10] | Gallus_gallus-4.0 | 1.25 [10] | 1.05 | 27,142 | 38.1 | 211.9 | 16,846 | 62.1 | 11.0 | 13.4 | 1.28 | Celara | S/4 |
| | Gallus_gallus-5.0 | 1.25 [10] | 1.22 | 24,701 | 49.3 | 2,718.7 | 23,870 | 51.2 | 6.3 | 2.8 | 0.23 | MHAP/PbCR | I/S/4/P |
| Zebra finch [11] | Taeniopygia_guttata-3.2.4 | 1.22 [12] | 1.23 | 124,806 | 9.8 | 38.8 | 37,422 | 32.9 | 8.5 | 8.7 | 0.71 | PCAP | S |
| Turkey [13] | Turkey_5.0 | 1.28 [14] | 1.13 | 296,315 | 3.7 | 26.7 | 233,806 | 4.8 | 3.0 | 35.0 | 3.11 | MaSuRCA | IS/4 |
| Hooded crow [15] | Hooded_Crow_genome | 1.26 [15] | 1.05 | 28,920 | 35.4 | 68.7 | 1,299 | 787.1 | 13.5 | 27.5 | 2.62 | ALLPATHS-LG | I |
| Golden eagle [16] | Aquila_chrysaetos-1.0.2 | 1.28 [16] | 1.19 | 17,032 | 69.3 | 156.4 | 1,142 | 1,033.3 | 8.7 | 12.7 | 1.07 | ALLPATHS-LG | I |
| Medium ground-finch [8, 17] | GeoFor_1.0 | 1.25 [16] | 1.07 | 95,828 | 10.9 | 24.0 | 27,239 | 38.2 | 3.7 | 24.0 | 2.25 | ALLPATHS-LG | I |
| Blue-crowned manakin [18] | Lepidothrix_coronata-1.0 | 1.16 [12] | 1.08 | 23,501 | 45.0 | 127.1 | 4,612 | 229.2 | 4.6 | 22.4 | 2.07 | ALLPATHS-LG | I |
| White-throated sparrow [19] | Zonotrichia_albicollis-1.0.1 | 1.30 [20] | 1.05 | 37,661 | 26.7 | 68.5 | 6,018 | 167.2 | 3.5 | 46.3 | 4.40 | ALLPATHS-LG | I |
| Silvereye [21] | ASM128173v1 | 1.35 [12] | 1.04 | 65,519 | 15.3 | 20.7 | 2,933 | 341.5 | 2.3 | 34.3 | 3.31 | ALLPATHS-LG | I |
| Tibetan ground-tit [22] | PseHum1.0 | 1.22 [22] | 1.04 | 27,052 | 38.1 | 132.7 | 5,406 | 190.5 | 11.8 | 13.0 | 1.24 | SOAPdenovo | I |
| Bald eagle [8, 23, 24] | Haliaeetus_leucocephalus-4.0 | 1.40 [12] | 1.18 | 31,786 | 36.5 | 82.3 | 1,023 | 1,133.2 | 7.4 | 19.2 | 1.63 | SOAPdenovo | I |
| American crow [8, 25] | ASM69197v1 | 1.24 [20] | 1.09 | 89,646 | 11.7 | 23.6 | 10,547 | 99.7 | 6.2 | 39.5 | 3.62 | SOAPdenovo | I |
| Saker falcon [26] | F_cherrug_v1.0 | 1.19 [26] | 1.17 | 75,898 | 15.2 | 30.2 | 5,863 | 196.3 | 4.1 | 23.8 | 2.03 | SOAPdenovo | I |
| Peregrine falcon [8, 26] | F_peregrinus_v1.0 | 1.22 [26] | 1.17 | 83,081 | 13.9 | 27.0 | 7,021 | 164.3 | 3.7 | 18.6 | 1.58 | SOAPdenovo | I |
| Rock pigeon [27] | Civ_1.0 | 1.30 [27] | 1.11 | 100,099 | 10.9 | 21.3 | 14,923 | 72.8 | 2.5 | 21.1 | 1.90 | SOAPdenovo | I |
| Budgerigar [28] | Melopsittacus_undulatus_6.3 | 1.19 [12, 29] | 1.12 | 70,891 | 15.3 | 49.2 | 25,212 | 43.1 | 9.3 | 30.8 | 2.75 | Celara | I/4 |
| Little egret [8, 23, 30] | ASM68718v1 | 1.39 * | 1.21 | 100,662 | 11.5 | 23.0 | 11,791 | 98.2 | 2.5 | 48.7 | 4.04 | SOAPdenovo | I |
| Hoatzin [8, 31] | ASM69207v1 | 1.68 * | 1.20 | 109,627 | 10.4 | 14.8 | 10,256 | 111.4 | 1.6 | 61.5 | 5.11 | SOAPdenovo | I |
| Golden-collared manakin [8, 32] | ASM171598v1 | 1.38 * | 1.21 | 29,998 | 38.9 | 137.7 | 15,315 | 76.3 | 13.1 | 45.5 | 3.75 | MaSuRCA | I/P |

* Estimated by k-mer counting method using KMC 2 [32] with 23-mer (SRR1144870-1 for Little egret, SRR947162-3 for Hoatzin, and SRR946955 for Golden-collared manakin).
 † In NG50 metric, estimated genome size is used rather than assembly length [34].

Figure 1: (A) A photograph of Yeonsan Ogye (YO) taken before sampling. (B) Hybrid genome assembly pipeline comprising four steps, each of which utilizes a different set of sequencing reads (see Table 1). Detailed methods for breaking misassembly and pseudo-reference-assisted assembly are depicted in Supplementary Figs. S2 and S3. (C) The NG50 and average length of pseudo contigs and scaffolds for the Ogye_1.1 and other avian genomes, generated using the indicated assembly methods (in the last column, sequencing platforms are designated as follows: I: Illumina, P: Pacific Biosciences, S: Sanger, 4: Roche454).

Hybrid Whole-Genome Assembly

The Ogye_1.1 genome was assembled using our hybrid genome assembly pipeline, employing the following four steps: 1) preprocessing, 2) hybrid *de novo* assembly, 3) pseudo-reference-assisted assembly, and 4) polishing and finishing (Fig. 1B and Supplementary Fig. S1). In the preprocessing step, reads in which $\geq 30\%$ of the nucleotides had a Phred score < 20 were excluded using the NGS QC Toolkit (IlluQC.PRL.L.pl) [36]; the adaptor sequences of the remaining reads were removed using Trimmomatic (Trimmomatic, RRID:SCR_011848) [37]; and three nucleotides at the 5' end and five nucleotides at the 3' end of the reads were trimmed using the NGS QC Toolkit (TrimmingReads.pl). After quality control, the sequencing errors in the Illumina paired-end short reads were corrected using KmerFreq and Corrector [38]. After these steps, 241.1X preprocessed reads were obtained for whole-genome assembly. In turn, using the corrected short reads, the sequencing errors in the PacBio long reads were corrected using LoRDEC [39].

In the hybrid *de novo* genome assembly, the initial assembly (ASM1) was done with 121.2X error-corrected short reads

from the paired-end and mate-pair libraries (see Table 1) using ALLPATHS-LG (ALLPATHS-LG, RRID:SCR_010742) [40] with the default option, producing contigs and scaffolds with N50 lengths of 53.6 Kbp and 10.7 Mbp, respectively (Fig. 1B; Supplementary Fig. S1). Additionally, another assembly (ASM2) was built with 109.2X paired-end and mate-pair reads that were unused in the initial assembly (see Table 1) using ALLPATHS-LG, resulting in 34,539 contigs with an N50 length of 59.2 Kbp. The resulting ASM2 contigs were then subjected to the pseudo-reference-assisted assembly step. In the second round of scaffolding and gap-filling (after the first scaffolding and gap-filling done during ASM1), the ASM1 scaffolds were connected with corrected PacBio long reads using SSPACE-LongRead [41], and gaps within and between scaffolds were examined with error-corrected short reads using GapCloser (GapCloser, RRID:SCR_015026) [38]. Then, the gap-filled scaffolds were connected again with FOSMID reads using OPERA [42], and the remaining gaps were re-examined with error-corrected short reads using GapCloser, resulting in scaffolds with an N50 length of 27.8 Mbp. However, some misassemblies (as illustrated in Supplementary Fig. S2A) were found by alignment of the resulting scaffolds with the galGal4 genome

Table 1: Summary of whole-genome sequencing data (estimated genome size 1.25 Gbp)

| Platform | Library type | Insert-size | Raw data | | | | Preprocessed data | | | | | | | | | |
|-------------------------|---------------------|-------------|-----------------------------------|------------------|--------------|---------------|-------------------|-----------------------------|-------|-------|------|------|-------|------|------|------|
| | | | Number of read (10 ⁶) | Total base (Gbp) | Coverage (X) | SRA accession | Coverage (X) | Usage of data (coverage, X) | | | | | | | | |
| | | | | | | | | SEC | ASM1 | ASM2 | SCF | GF | SV | SIC | | |
| Illumina HiSeq 2000 | Paired-end | 280 bp | 259.2 | 39.0 | 31.2 | SRR6189087 | 21.4 | 0 | 0 | | | | | 0 | 0 | |
| | | | 248.9 | 37.4 | 29.9 | SRR6189084 | 20.5 | 0 | 0 | | | 0 | 0 | 0 | | |
| | | 500 bp | 87.1 | 13.1 | 10.5 | SRR6189095 | 4.8 | 0 | 0 | | | | 0 | 0 | 0 | |
| | | | 94.4 | 14.2 | 11.4 | SRR6189097 | 5.2 | 0 | 0 | | | 0 | 0 | 0 | 0 | |
| | | | 28.1 | 4.2 | 3.4 | SRR6189096 | 1.3 | 0 | 0 | | | | 0 | 0 | 0 | |
| | | | 28.3 | 4.3 | 3.4 | SRR6189098 | 1.2 | 0 | 0 | | 0 | | 0 | 0 | 0 | |
| | | | 29.2 | 4.4 | 3.5 | SRR6189082 | 1.8 | 0 | 0 | | 0 | | 0 | 0 | 0 | |
| | | | 57.4 | 8.6 | 6.9 | SRR6189094 | 4.5 | 0 | 0 | | 0 | | 0 | 0 | 0 | |
| | | | Paired-end total | | 832.5 | 125.2 | 100.2 | | 60.7 | 60.7 | 37.2 | 23.5 | - | 31.4 | 60.7 | 60.7 |
| | | | Mate-pair | 3 Kbp | 293.1 | 43.6 | 34.9 | SRR6189093 | 23.6 | | | 0 | | | 0 | |
| | 270.0 | 40.2 | | | 32.1 | SRR6189083 | 21.6 | | 0 | | | | | | | |
| | 5 Kbp | 229.6 | | 34.2 | 27.4 | SRR6189081 | 16.9 | | 0 | | | | 0 | | | |
| | | 212.8 | | 31.7 | 25.4 | SRR6189088 | 15.7 | | | 0 | | | | | | |
| | 8 Kbp | 273.1 | | 40.7 | 32.6 | SRR6189085 | 20.2 | | 0 | | | | 0 | | | |
| | | 270.5 | | 40.4 | 32.3 | SRR6189086 | 19.7 | | | 0 | | | | | | |
| | 10 Kbp | 338.2 | | 50.4 | 40.3 | SRR6189091 | 26.7 | | | 0 | | | 0 | | | |
| | | 315.9 | | 47.1 | 37.7 | SRR6189092 | 25.3 | | 0 | | | | | | | |
| | 40 Kbp ^a | 169.9 | | 17.2 | 13.7 | SRR6189089 | 10.7 | | | | 0 | | | | | |
| | Mate-pair total | | | 2,373.2 | 345.5 | 276.4 | | 180.4 | - | 84.0 | 85.7 | 10.7 | - | 87.4 | - | |
| | PacBio RS II | Long-read | 6 Kbp ^b | 1.7 | 12.1 | 9.7 | SRR6189090 | 9.3 | | | 0 | 0 | | | | |
| Illumina total | | | 3,205.7 | 470.7 | 376.6 | | 241.1 | 60.7 | 121.2 | 109.2 | 20.0 | 40.7 | 148.1 | | | |
| Illumina + PacBio total | | | 3,207.4 | 482.8 | 386.3 | | 250.4 | 60.7 | 121.2 | 109.2 | 29.3 | 50.0 | 148.1 | | | |

^aFosmid^bAverage read length.

Abbreviations: ASM1: initial ALLPATHS-LG assembly; ASM2: additional ALLPATHS-LG assembly; GF: gap-filling; SCF: scaffolding; SEC: sequencing error correction; SIC: SNP/INDEL calling; SV: structural variation detection;.

Table 2: Sequencing and mapping summary of RNA-seq data

| Samples | Paired end | | | Single end | | |
|--------------|--------------|-----------------|---------------|--------------|-----------------|---------------|
| | No. of reads | Mapping rate, % | SRA accession | No. of reads | Mapping rate, % | SRA accession |
| Breast | 34,893,064 | 92.05 | SRX3223583 | 43,294,022 | 90.70 | SRX3223603 |
| Liver | 33,476,266 | 85.75 | SRX3223584 | 48,032,813 | 85.81 | SRX3223604 |
| Bone marrow | 30,975,506 | 85.00 | SRX3223585 | 40,286,974 | 87.99 | SRX3223605 |
| Fascia | 33,316,764 | 84.61 | SRX3223586 | 42,425,452 | 87.93 | SRX3223606 |
| Cerebrum | 30,887,821 | 89.95 | SRX3223587 | 46,455,658 | 92.32 | SRX3223607 |
| Gizzard | 31,537,118 | 84.00 | SRX3223588 | 38,689,871 | 85.82 | SRX3223608 |
| Immature egg | 32,009,437 | 87.73 | SRX3223589 | 32,048,703 | 87.80 | SRX3223609 |
| Comb | 31,936,332 | 85.34 | SRX3223590 | 37,985,049 | 87.76 | SRX3223610 |
| Spleen | 28,946,777 | 89.70 | SRX3223591 | 38,704,448 | 89.33 | SRX3223611 |
| Mature egg | 30,873,699 | 91.98 | SRX3223592 | 40,650,664 | 92.17 | SRX3223612 |
| Cerebellum | 30,798,145 | 93.53 | SRX3223593 | 39,940,946 | 93.34 | SRX3223613 |
| Gallbladder | 35,862,229 | 84.83 | SRX3223594 | 35,423,339 | 87.06 | SRX3223614 |
| Kidney | 29,953,007 | 87.25 | SRX3223595 | 39,894,009 | 89.99 | SRX3223615 |
| Heart | 30,986,431 | 94.14 | SRX3223596 | 45,951,338 | 91.49 | SRX3223616 |
| Uterus | 33,444,002 | 91.89 | SRX3223597 | 46,650,355 | 90.63 | SRX3223617 |
| Pancreas | 30,595,568 | 82.52 | SRX3223598 | 47,361,192 | 84.35 | SRX3223618 |
| Lung | 31,533,498 | 87.63 | SRX3223599 | 45,552,982 | 92.34 | SRX3223619 |
| Skin | 34,442,464 | 82.36 | SRX3223600 | 41,934,970 | 84.00 | SRX3223620 |
| Eye | 33,006,509 | 89.21 | SRX3223601 | 44,044,630 | 91.82 | SRX3223621 |
| Shank | 28,643,334 | 94.07 | SRX3223602 | 47,716,995 | 79.86 | SRX3223622 |

(GenBank assembly accession GCA.000002315.2) using LASTZ [43]. During an analysis of the resulting alignments, 30 misassemblies were detected and broken at each break point, as described in Supplementary Fig. S2. Breaking scaffolds at the break points resulted in a scaffold N50 length of 18.7 Mbp (Supplementary Fig. S1). For contigs, we considered a pseudo contig, broken at positions where two or more contiguous Ns appeared in scaffolds, resulting in a pseudo contig N50 of 108.6 Kbp.

In the pseudo-reference-assisted assembly step, error-corrected PacBio long reads and ASM2 contigs were utilized to reduce the topological complexity of the assembly graphs [44] (Fig. 1B). Because even scaffolding with long reads can be affected by repetitive sequences, the scaffolds mapped to each chromosome were transformed into a hierarchical bipartite graph to minimize the influence of repetitive sequences using TSRATOR [45] (Supplementary Fig. S3). In detail, error-corrected PacBio reads and ASM2 contigs were mapped to the scaffolds

Table 3: Sequencing and mapping summary of RRBS data

| Samples | No. of reads | Mapping rate, % | SRA accession |
|--------------|--------------|-----------------|---------------|
| Breast | 6,042,106 | 68.90 | SRX3223667 |
| Liver | 6,744,208 | 74.20 | SRX3223668 |
| Bone marrow | 5,736,011 | 72.00 | SRX3223669 |
| Fascia | 5,720,194 | 68.90 | SRX3223670 |
| Cerebrum | 6,078,989 | 70.00 | SRX3223671 |
| Gizzard | 5,731,878 | 69.40 | SRX3223672 |
| Immature egg | 6,741,258 | 67.70 | SRX3223673 |
| Comb | 5,948,687 | 72.90 | SRX3223674 |
| Spleen | 6,307,517 | 77.60 | SRX3223675 |
| Mature egg | 6,246,607 | 69.20 | SRX3223676 |
| Cerebellum | 6,291,610 | 68.20 | SRX3223677 |
| Gallbladder | 5,738,180 | 70.10 | SRX3223678 |
| Kidney | 5,470,502 | 68.60 | SRX3223679 |
| Heart | 5,462,739 | 69.40 | SRX3223680 |
| Uterus | 6,046,764 | 67.90 | SRX3223681 |
| Pancreas | 7,100,215 | 70.30 | SRX3223682 |
| Lung | 5,640,120 | 67.60 | SRX3223683 |
| Skin | 7,226,309 | 72.40 | SRX3223684 |
| Eye | 6,956,141 | 71.90 | SRX3223685 |
| Shank | 5,924,463 | 74.20 | SRX3223686 |

Table 4: Comparison of genome completeness using BUSCO

| Species | Assembly name | Complete | | | |
|-------------|---------------------------|----------------|----------------|-------------|------------|
| | | Single-copy, % | Duplication, % | Fragment, % | Missing, % |
| Chicken | Ogye.1.1 | 97.60 | 0.50 | 0.90 | 1.00 |
| | Gallus_gallus-4.0 | 96.90 | 0.90 | 1.10 | 1.10 |
| | Gallus_gallus-5.0 | 97.40 | 0.90 | 0.70 | 1.00 |
| Turkey | Turkey_5.0 | 93.70 | 0.50 | 4.10 | 1.70 |
| Duck | BGI.duck.1.0 | 92.60 | 0.40 | 4.80 | 2.20 |
| Zebra finch | Taeniopygia_guttata-3.2.4 | 93.60 | 2.20 | 2.70 | 1.50 |

using BWA-MEM and, in turn, the scaffolds were mapped to the galGal4 genome using LASTZ to build the hierarchical bipartite graph. Using the hierarchical bipartite graphs, all scaffolds, PacBio reads, and ASM2 contigs were finally grouped to each chromosome. Based on these results, a third round of scaffolding and gap-filling was performed with the long reads and the ASM2 contigs in each chromosome group using SSPACE-LongRead and PBjelly (PBjelly, [RRID:SCR.012091](#)) [46], respectively, resulting in a scaffold N50 of 21.2 Mbp with 0.85% gaps (Supplementary Fig. S1).

In the last step, nucleotide errors or ambiguities were corrected using the GATK (GATK, [RRID:SCR.001876](#)) pipeline [47] with paired-end reads. In turn, any vector contamination was removed using VecScreen with the UniVec database [48] (Fig. 1B), resulting in 506.3 Kbp and 21.2 Mbp contig and scaffold N50 lengths, respectively. The final assembly results (Ogye.1.1 scaffold) showed that the gap percentage and (pseudo-)contig N50 were significantly improved, from 1.87% and 53.6 Kbp in the initial assembly to 0.85% and 506.3 Kbp in the final assembly, respectively (Supplementary Fig. S1). Using the estimated chicken genome size (1.25 Gbp [10]), Ogye.1.1 scaffold's contig and scaffold NG50 lengths were estimated at 362.3 Kbp and 16.8 Mbp, respectively (Fig. 1C). The complete genome sequence at the chromosome level was built by connecting the final scaffolds in their order of appearance in each chromosome with the intro-

duction of 100 Kbp "N" gaps between them (Supplementary Fig. S4) (see [79]). To evaluate its completeness, the Ogye.1.1 genome was compared to the galGal4 (short-read-based assembly) and galGal5 (long-read-based assembly) genomes, with respect to 2,586 conserved vertebrate genes, using Benchmarking Universal Single-Copy Ortholog (BUSCO) (BUSCO, [RRID:SCR.015008](#)) [49] with OrthoDB v9 (OrthoDB, [RRID:SCR.011980](#)) [50]. The Ogye.1.1 genome contained more complete single-copy BUSCO genes (Table 4).

Large Structural Variations

When the Ogye.1.1 genome was compared to galGal4 and galGal5 using LASTZ [43], putative large SVs (>1 Kbp) were detected for each reference genome, and they were validated by four different SV prediction programs (Delly, Lumpy, FermiKit, and novoBreak) [51–54] (Supplementary Fig. S5 and Table S2). SVs, validated by at least one program, included 185 deletions, 180 insertions, 158 duplications, 23 inversions, and 5 intra- or inter-chromosomal translocations. A total of 290 and 447 distinct SVs were detected relative to galGal4 and galGal5, respectively, suggesting that either reference assembly could include misassemblies.

Although the fibromelanosis (FM) locus, which contains the hyperpigmentation-related *edn3* gene, is known to be duplicated

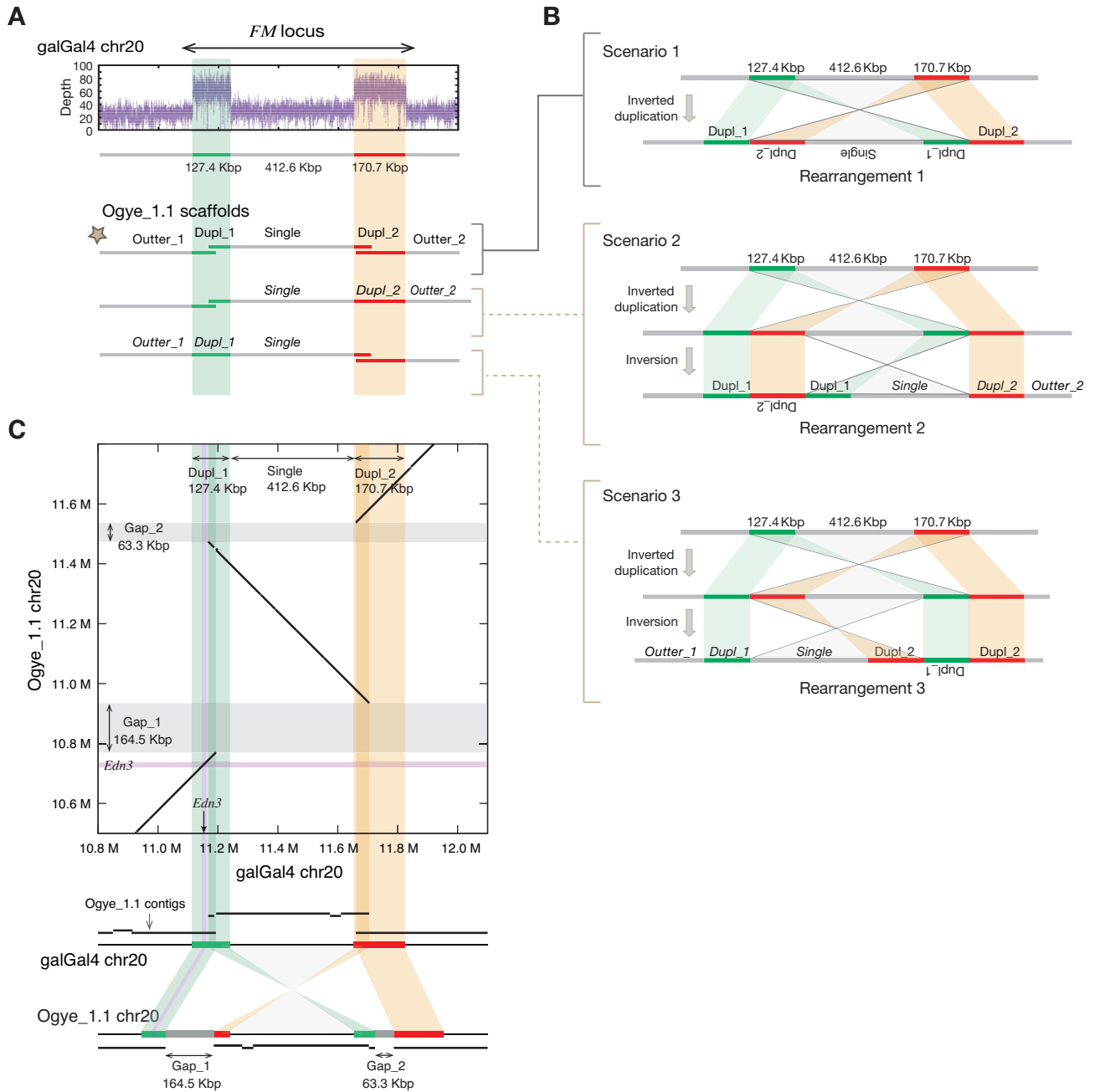


Figure 2: (A) The read depth of a locus on chromosome 20 is shown in the top panel, and the continuous/discontinuous patterns between mapped scaffolds are shown in the bottom panel. The star indicates the pattern that is discontinuous on both sides, validated in (C). (B) Three possible scenarios were developed based on the overlap patterns. The green and red lines indicated two duplicated genomic loci (Dupl.1 and Dupl.2, respectively) including the FM locus. Scenario 1 consists of a one-step rearrangement—an inverted duplication—whereas scenarios 2 and 3 consist of a simultaneous rearrangement of an inverted duplication and an inversion. The three rearrangements were suggested in a previous study [3]. (C) A comparison of the FM locus in galGal4 and the Ogye draft genome with aligned contigs (black lines) in each scaffold. The gray bands indicate the estimated gaps between contigs. The estimated sizes of Gap.1 and Gap.2 are 164.5 Kbp and 63.3 Kbp, respectively. The purple lines in the box indicate the *edn3* gene locus and the green and yellow shades indicate the duplicated regions (Dupl.1 and Dupl.2, respectively). The dark green and yellow shades indicate the discontinuous regions between scaffolds.

in the genomes of certain hyperpigmented chicken breeds, such as Silkie and Ayam cemani [3, 6], the exact structure of the duplicated FM locus in such breeds has not been completely resolved due to its large size (~1 Mbp). A previous study, using conventional PCR assays, suggested three possible rearrangements at the FM locus [3]. To understand more about the mechanism of FM locus rearrangement in the Ogye_1.1 genome, the FM loci

from YO and galGal4 were compared with mapped paired-end and mate-pair reads. A doubled read depth at two loci including the FM locus was detected in YO, indicating that the loci had been duplicated (Fig. 2A top). As previously reported [3, 6], our paired-end and mate-pair reads of YO's FM locus were discordantly mapped to the galGal4 FM locus (Supplementary Fig. S6). The intervening region between the two duplicated regions

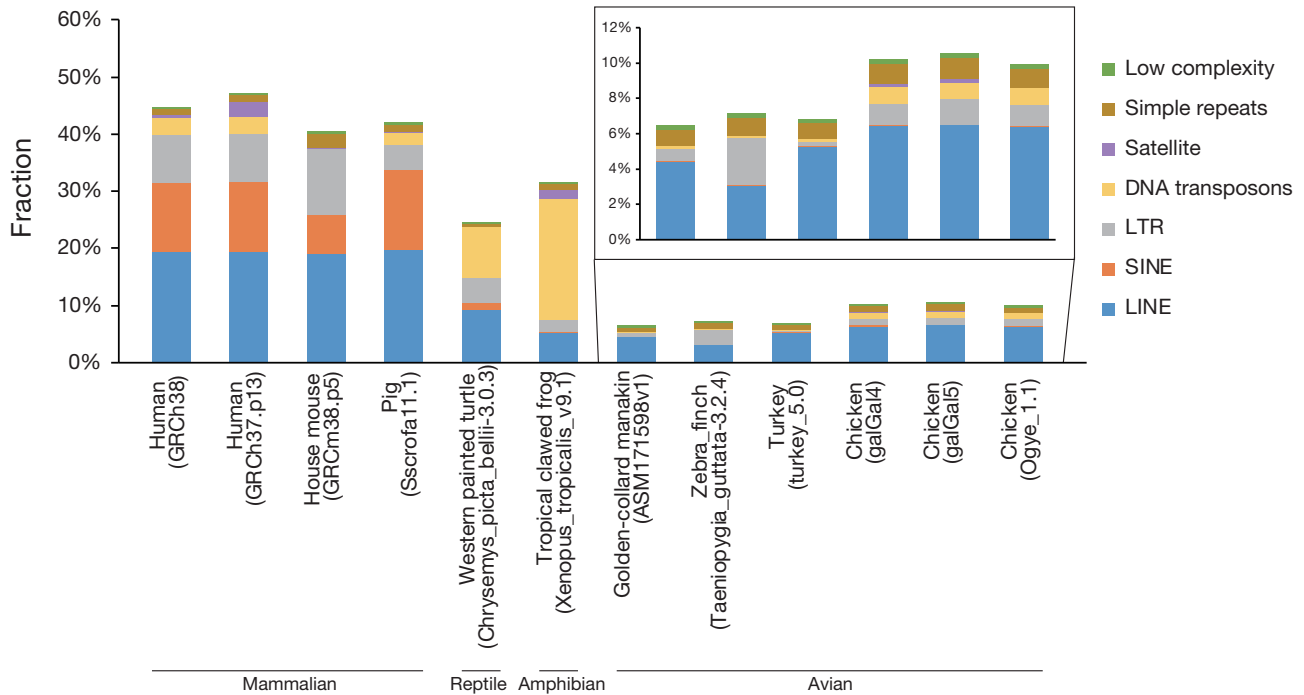


Figure 3: Composition of repeat elements in different assemblies of avian, amphibian, reptile, and mammalian genomes. The repeats in unplaced scaffolds were not considered.

was estimated to be 412.6 Kbp in length in the Ogye_1.1 genome. Based on these results, we propose three possible scenarios that might have produced the FM locus rearrangement (Fig. 2B). To discern which rearrangement best fits our results, the FM loci from galGal4 and the Ogye draft were compared with the resulting scaffolds, showing an inverted duplication with discontinued scaffolds at both duplicated regions (Fig. 2A, 2C). The results, with a discontinued scaffold on both sides, support rearrangement 1 rather than rearrangement 2 or 3, which have a discontinued scaffold on only one side. Although rearrangement 1 needs to be further validated, the FM locus in the Ogye_1.1 genome was updated according to the first rearrangement (Fig. 2C). Given the resulting alignment, the sizes of Gap_1 and Gap_2 were estimated to be 164.5 Kbp and 63.3 Kbp, respectively.

Annotations

Repeats

Repeat elements in the Ogye_1.1 and other genomes (human, mouse, pig, western painted turtle, tropical clawed frog, zebra finch, turkey, and chicken) were predicted by a reference-guided approach using RepeatMasker (RepeatMasker, [RRID:SCR_012954](https://doi.org/10.1093/bioinformatics/btt125)) [55] with Repbase libraries [56]. In the Ogye_1.1 genome, 205,684 retro-transposable elements (7.65%), including long interspersed nuclear elements (LINEs) (6.41%), short interspersed nuclear elements (SINEs) (0.04%), and long terminal repeat (LTR) elements (1.20%), 27,348 DNA transposons (0.94%), 7,721 simple repeats (0.12%), and 298 low-complexity repeats (0.01%) were annotated (Fig. 3 and Supplementary Table S3). Repeats are similarly distributed in the Ogye_1.1 and other avian genomes (Fig. 3 and Supplementary Table S4). Compared with other avian genomes, the Ogye_1.1 genome resembles galGal4 and galGal5 the most in terms of repeat composition except for that of simple repeats (0.12% for Ogye_1.1, 1.12% for galGal4, and 1.24%

for galGal5), low-complexity (0.01% for Ogye_1.1, 0.24% for galGal4, and 0.25% for galGal5), and satellite DNA repeats (0.01% for Ogye_1.1, 0.20% for galGal4, and 0.22% for galGal5). The distribution of transposable elements across all chromosomes is depicted in Supplementary Fig. S7.

SNPs and INDELS

To annotate SNPs and INDELS in the Ogye_1.1 genome, all paired-end libraries were mapped to the Ogye_1.1 genome using BWA-MEM and deduplicated using Picard modules [57]. We identified 3,206,794 SNPs and 302,463 INDELS across the genome using VarScan 2 with options `-min-coverage 8 -min-reads 2 -min-avg-qual 15 -min-var-freq 0.2 -p-value 1e-2` [58]. The densities of SNPs and INDELS across all chromosomes are depicted in Supplementary Fig. S7.

Protein-coding genes

To sensitively annotate protein-coding genes, all paired-end RNA-seq data were mapped on the Ogye_1.1 genome using STAR [59] for each tissue, and the mapping results were then assembled into potential transcripts using StringTie [60]. Assembled transcripts from each sample were merged using StringTie, and the resulting transcriptome was subjected to the prediction of coding DNA sequences (CDSs) using TransDecoder [61]. For high-confidence prediction, transcripts with intact gene structures (5'UTR, CDS, and 3'UTR) were selected. To verify their coding potential, the candidate sequences were examined using CPAT [62] and CPC [63]. Candidates with a high CPAT score (>0.99) were directly assigned to be protein-coding genes, and those with an intermediate score (0.8–0.99) were re-examined to determine whether the CPC score was >0 . Candidates with low coding potential or that were partially annotated were examined to determine if their loci overlapped with annotated protein-

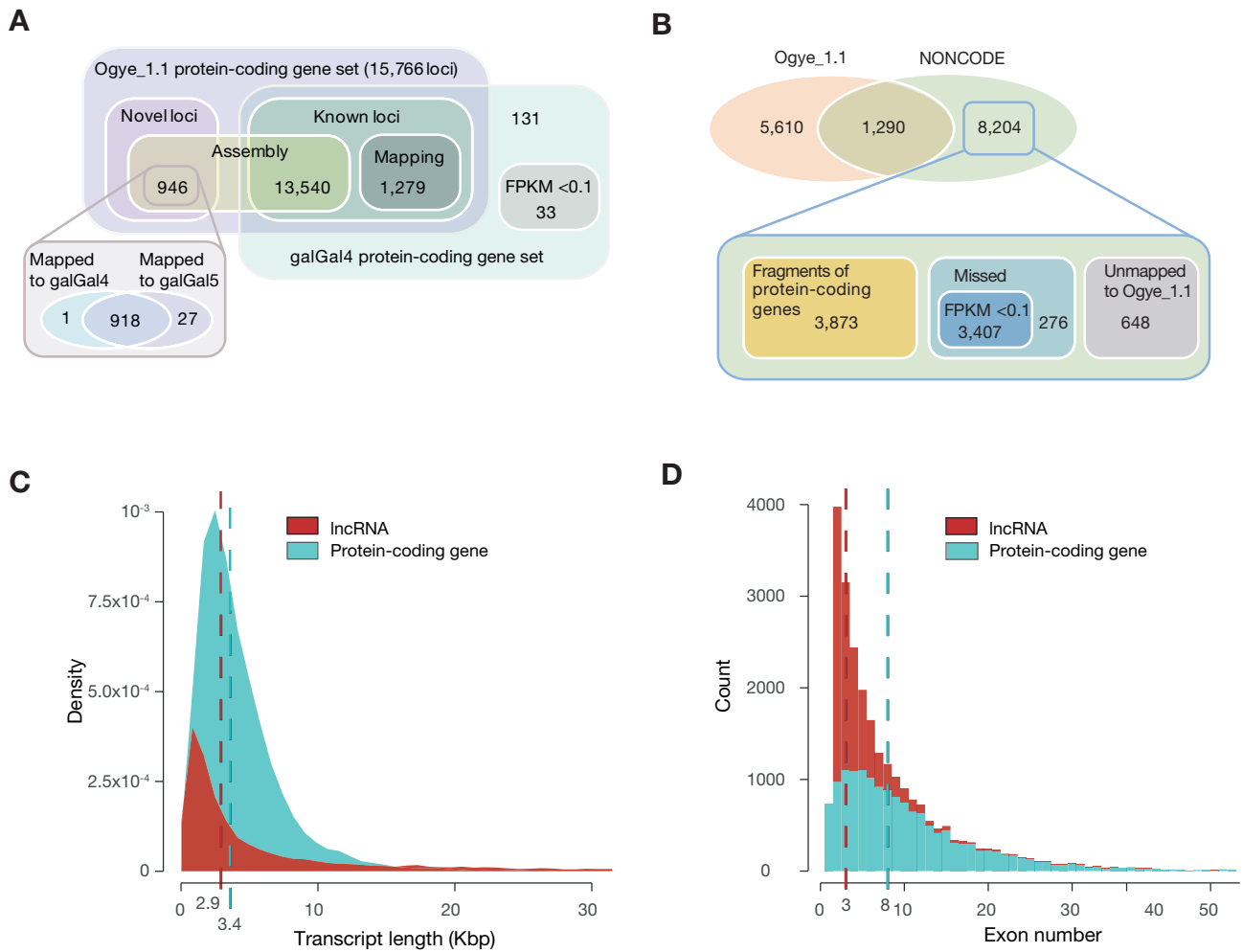


Figure 4: (A) A Venn diagram showing the number of protein-coding genes in the Ogye.1.1 genome. (B) A Venn diagram showing the number of Ogye.1.1 and galGal4 lncRNAs. (C) Distribution of transcript length (red for lncRNAs and cyan for protein-coding genes). The vertical dotted lines indicate the median length. (D) Distribution of the number of exons per transcript. Otherwise, as in (C).

coding genes from galGal4 (ENSEMBL cDNA release 85). Overlapping genes were added to the set of Ogye.1.1 protein-coding genes. Using this protein-coding gene annotation pipeline (Supplementary Fig. S8), 15,766 protein-coding genes were finally annotated in the Ogye.1.1 genome, including 946 novel genes and 14,819 known genes (Fig. 4A). However, 164 galGal4 protein-coding genes were not mapped to the Ogye.1.1 genome by GMAP (Supplementary Table S5), 131 of which were confirmed to be expressed in YO (≥ 0.1 FPKM) using all paired-end YO RNA-seq data. In contrast, the remaining 33 genes were not expressed in YO (< 0.1 FPKM) or were lost from the Ogye.1.1 genome. Of the 33 missing genes, 26 appeared to be located on unknown chromosomes and the remainder are on autosomes (six genes) or the W sex chromosome (one gene) in galGal4. The density of protein-coding genes across all chromosomes is depicted in Supplementary Fig. S7.

lncRNAs

To annotate and profile lncRNA genes, we used our lncRNA annotation pipeline (Supplementary Fig. S9), adopted from our previous study [64]. Pooled single- and paired-end RNA-seq reads from each tissue were mapped to the Ogye.1.1 genome (PR-

JNA412424) using STAR [59] and subjected to transcriptome assembly using Cufflinks (Cufflinks, [RRID:SCR_014597](#)) [65], leading to the construction of transcriptome maps for 20 tissues. The resulting maps were combined by Cuffmerge and, in total, 206,084 transcripts from 103,405 loci were reconstructed in the Ogye genome. We removed other RNA biotypes (the sequences of mRNAs, tRNAs, rRNAs, snoRNAs, miRNAs, and other small noncoding RNAs downloaded from ENSEMBL biomaart) and short transcripts (less than 200 nt in length). A total of 54,760 lncRNA candidate loci (60,257 transcripts) were retained and compared with a chicken lncRNA annotation from NONCODE (v2016) [66]. Of the candidates, 2,094 loci (5,215 transcripts) overlapped with previously annotated chicken lncRNAs. Then, 52,666 nonoverlapping loci (55,042 transcripts) were further examined to determine whether they had coding potential using CPC score [63]. Those with a score greater than -1 were filtered out, and the remainder (14,108 novel lncRNA candidate loci without coding potential) were subjected to the next step. Because many candidates still appeared to be fragmented, those with a single exon but with neighboring candidates within 36,873 bp, which is the length of introns in the 99th percentile, were re-examined using both exon-junction reads consistently presented over 20 tissues and the maximum entropy score [67], as done in our previous

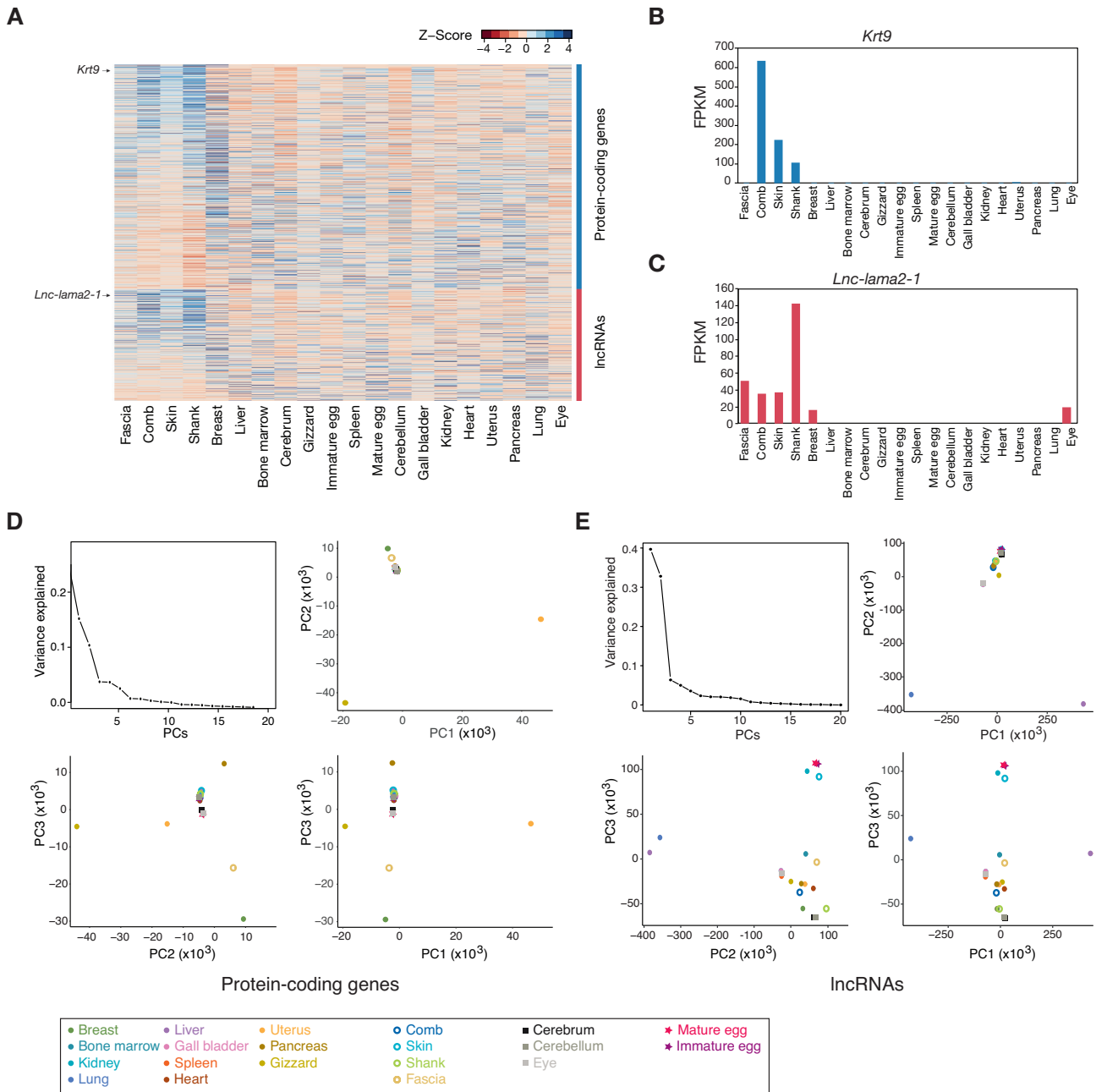


Figure 5: (A) The expression patterns of the genes expressed with ≥ 10 FPKM in black tissues. Expression levels are indicated with a color-coded Z-score (red for low and blue for high expression) as shown in the key. (B) Expression levels of *krt9* across 20 tissues. (C) Expression levels of *lnc-lama2-1* across 20 tissues. (D) Principal component analysis (PCA) using tissue-specific protein-coding genes. PCs explaining the variances are indicated with the amount of the contribution in the top-left plot. PCA plots are shown with PC1, PC2, and PC3 plotted in a pairwise manner. Each tissue is indicated on the PCA plot with a specific color. (E) PCA using tissue-specific lncRNAs. Otherwise, as in (A).

study [64]. If there were at least two junction reads spanning two neighboring transcripts or if the entropy score was greater than 4.66 in the interspace, the two candidates were reconnected, and those with a single exon were discarded. In the final version, 6,900 loci (5,610 novel and 1,290 known) were annotated as lncRNAs (see Fig. 4B), which included 6,170 (89.40%) intergenic lncRNAs and 730 (10.57%) anti-sense ncRNAs. Consistent with previous results [68–71], the median Ogye lncRNA transcript length and exon number were less than those of protein-coding genes (Fig. 4C and 4D).

Whereas 13,540 of 14,983 protein-coding genes (90.4%) were redetected in our protein-coding gene annotations (see Fig. 4A), only 1,290 (13.6%) of NONCODE lncRNAs were redetected in our Ogye.1.1 lncRNA annotations (Fig. 4B). The majority of the missing NONCODE lncRNAs were either fragments of protein-coding genes or not expressed in all 20 Ogye tissues (Fig. 4B). Only 276 were actually missing in the transcriptome assembly, and 648 were not mapped to the Ogye.1.1 genome.

Table 5: Summary of methylated CpG sites across 20 tissues

| | All genomic region | | | Promoter region | | |
|--------------|--------------------|----------------------|-------------|--------------------|----------------------|-------------|
| | Total no. of sites | Methylated CpG sites | | Total no. of sites | Methylated CpG sites | |
| | | No. of sites | Fraction, % | | No. of sites | Fraction, % |
| Breast | 994,326 | 621,751 | 62.53 | 228,673 | 91,704 | 40.10 |
| Liver | 1,641,060 | 505,775 | 30.82 | 522,590 | 97,597 | 18.68 |
| Bone marrow | 1,096,466 | 671,781 | 61.27 | 254,978 | 100,385 | 39.37 |
| Fascia | 1,146,350 | 670,181 | 58.46 | 278,618 | 99,802 | 35.82 |
| Cerebrum | 1,246,514 | 748,323 | 60.03 | 298,677 | 112,689 | 37.73 |
| Gizzard | 1,024,125 | 609,010 | 59.47 | 234,379 | 85,273 | 36.38 |
| Immature egg | 1,416,686 | 809,214 | 57.12 | 334,813 | 115,195 | 34.41 |
| Comb | 1,035,966 | 642,138 | 61.98 | 239,319 | 92,436 | 38.62 |
| Spleen | 995,639 | 401,080 | 40.28 | 298,833 | 74,473 | 24.92 |
| Mature egg | 1,144,589 | 695,258 | 60.74 | 269,124 | 102,282 | 38.01 |
| Cerebellum | 1,279,666 | 775,513 | 60.60 | 305,489 | 117,950 | 38.61 |
| Gallbladder | 953,630 | 595,681 | 62.46 | 225,122 | 89,174 | 39.61 |
| Kidney | 1,016,035 | 610,941 | 60.13 | 238,066 | 89,255 | 37.49 |
| Heart | 1,000,957 | 611,343 | 61.08 | 235,853 | 90,434 | 38.34 |
| Uterus | 893,101 | 543,931 | 60.90 | 203,102 | 77,365 | 38.09 |
| Pancreas | 1,119,795 | 647,577 | 57.83 | 267,036 | 94,371 | 35.34 |
| Lung | 985,824 | 594,046 | 60.26 | 229,316 | 87,140 | 38.00 |
| Skin | 868,368 | 565,815 | 65.16 | 198,275 | 85,094 | 42.92 |
| Eye | 1,051,332 | 663,413 | 63.10 | 252,991 | 105,539 | 41.72 |
| Shank | 862,931 | 512,853 | 59.43 | 210,905 | 76,512 | 36.28 |

Coding and noncoding transcriptome maps

Using paired-end YO RNA-seq data, the expression levels of protein-coding and lncRNA genes were calculated across 20 tissues (Supplementary Fig. S10). In the profiled transcriptomes, 1,814 protein-coding and 1,226 lncRNA genes were expressed with ≥ 10 FPKM in only one tissue, whereas 1,559 protein-coding and 351 lncRNA genes were expressed with ≥ 10 FPKM in all tissues. In black tissues (fascia, comb, skin, and shank), we have found that 6,702 protein-coding and 3,291 lncRNA genes were expressed with ≥ 10 FPKM, the majority of which appeared to be expressed in a tissue-specific manner (Fig. 5A). For instance, the protein-coding gene *krt9* and the lncRNA *lnc-lama2-1* are highly expressed in black tissues, particularly in comb and shank, respectively (Fig. 5B and 5C).

Because lncRNAs tend to be specifically expressed in a tissue or in related tissues, they could be more useful than protein-coding genes for defining genomic characteristics of tissues. To prove this idea, principle component analyses were performed with 9,153 tissue-specific protein-coding and 5,191 tissue-specific lncRNA genes using the reshape2 R package (Fig. 5D and 5E) [72]. Here, we classified a gene as tissue-specific if the maximum expression value was at least four-fold higher than the mean value over 20 tissues. As expected, the first, second, and third PCs of lncRNAs enabled us to predict the majority of variances and to better discern distantly related tissues and functionally and histologically related tissues (i.e., black tissues and brain tissues) (Fig. 5E) than those of protein-coding genes (Fig. 5C).

DNA Methylation Maps

After mapping RRBS reads to the Ogye.1.1 genome (Table 3), DNA methylation signals (C to T changes in CpGs) were calculated across chromosomes using Bismark [73]. Of all CpG sites in the genome, 31%–65% were methylated across tissues, whereas

only 19%–43% were methylated in gene promoters (the region 2 Kbp upstream of the transcription start site [TSS]) (Table 5), indicating that the promoters of expressed genes tended to be hypomethylated. The DNA methylation landscapes in the regions 2 Kbp upstream of the protein-coding and lncRNA gene TSSs are shown in Supplementary Fig. S11. Based on the CpG methylation pattern, hierarchical clustering was performed using the rsgcc R package, and clusters including adjacent or functionally related tissues, such as cerebrum and cerebellum, immature and mature eggs, and comb and skin, were identified (Fig. 6A).

We then examined the average methylation landscapes over protein-coding and lncRNA loci to check whether the CpG methylation profiles were properly processed. As previously shown [74–77], the average methylation levels in gene body regions were much higher than those in promoters across tissues (Fig. 6B and 6C). To investigate the association between CpG methylation in the promoter and target gene expression, the average methylation levels of tissue-specific genes (280 protein-coding and 392 lncRNA genes with expression ≥ 10 FPKM in at least one tissue and with a maximum expression value four-fold higher than the mean expression level in 20 tissues) were compared to those of others expressed in their specific tissues. The methylation levels of highly expressed genes appeared to be lower than those of others (Fig. 6D and 6E). We then searched for genes with tissue-specific expression that was significantly correlated to the promoter methylation level using the Spearman correlation method (Fig. 6F). To exclude stochastic noise, only tissues in which a certain position had a sufficient number of reads (at least five) were taken into account for measuring the correlation. We found that the expression levels of 454 protein-coding and 25 lncRNA genes displayed a negative correlation to promoter methylation levels, whereas 157 protein-coding and 20 lncRNA genes had a positive correlation (box plots in Fig. 6F).

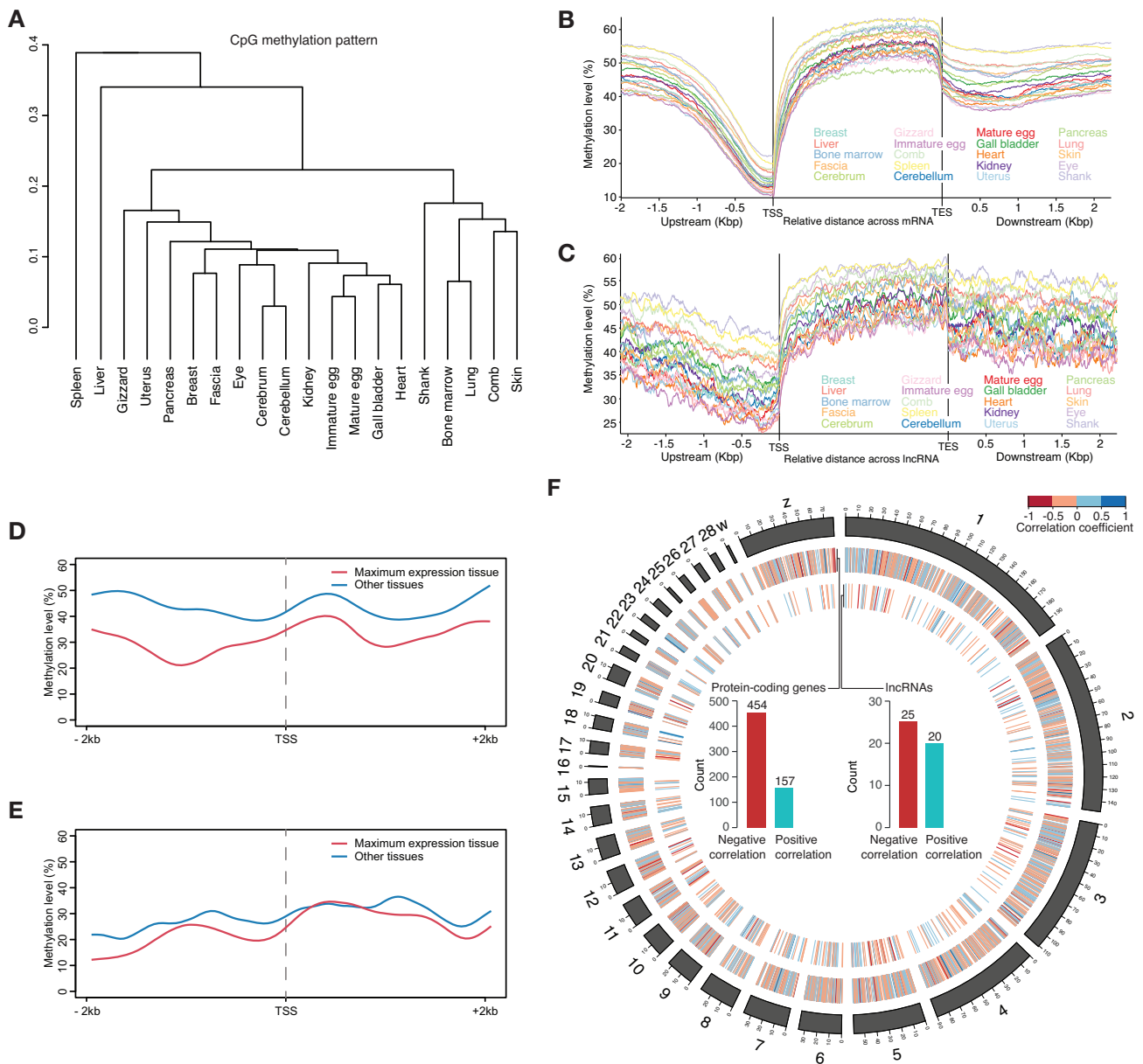


Figure 6: (A) Hierarchical clustering using Pearson correlation of DNA methylation patterns between tissues. (B and C) Average DNA methylation landscapes along protein-coding (B) and lncRNA (C) gene bodies and their flanking regions across 20 tissues. (D and E) Average DNA methylation levels of protein-coding (D) and lncRNA (E) genes in the tissue of maximum expression (red) and the other tissues (blue). (F) Spearman correlation coefficients between gene expression and promoter methylation levels are shown across chromosomes (heat maps) in a Circos plot. The bar charts indicate the number of genes (left for protein-coding genes and right for lncRNAs) with significant negative (red) and positive (cyan) correlations ($P < 0.05$) between their promoter methylation levels and their expression values.

Discussion

In this work, the first draft genome of YO, Ogye.1.1, was constructed with genomic variation, repeat, and protein-coding and noncoding gene maps. Compared with the chicken reference genome maps, many more novel coding and noncoding elements were identified from large-scale RNA-seq datasets across 20 tissues. Although the Ogye.1.1 genome is comparable with galGal5 with respect to genome completeness evaluated using BUSCO, Ogye.1.1 seems to lack simple and long repeats compared with galGal5, which was assembled from high-depth PacBio long reads (50X) that can capture simple and long repeats. Although PacBio long reads were also produced in our

study, they were only used for scaffolding and gap-filling because of their shallow depth (9.7X), probably resulting in some simple and satellite repeats being missed in Ogye.1.1. A similar tendency can be seen in the golden-collared manakin genome (ASM171598v1) [32] (Fig. 3) and the gray mouse lemur genome (Mmur3.0) [78], which were also assembled in a hybrid manner with high-depth Illumina short reads and low-depth PacBio long reads.

A total of 15,766 protein-coding and 6,900 lncRNA genes were annotated from 20 YO tissues. Also, 946 novel protein-coding genes were identified, while 164 *Gallus gallus red junglefowl* genes were missed in our annotations. In the case of lncRNAs, only about 13.6% of previously annotated chicken lncRNAs were re-

detected, and the remainder were mostly not expressed in YO or were false annotations, suggesting that the current chicken lncRNA annotations should be carefully examined. Our Ogye lncRNAs resembled previously annotated mammalian lncRNAs in their genomic characteristics, including transcript length, exon number, and tissue-specific expression patterns, providing evidence for the accuracy of the new annotations. Hence, our lncRNA catalogue may help us improve lncRNA annotations in the chicken reference genome.

Availability of supporting data

All of our sequence data and the genome sequence have been deposited in National Center for Biotechnology Information's Gene Expression Omnibus superseries GSE 104358 and BioProject PRJNA412408. All supporting data (genome and gene sequence files, the expression tables for protein-coding and lncRNA genes, and the RRBS, protein-coding, lncRNA, SNP, and INDEL annotation files) are available in the GigaScience repository GigaDB [79].

Additional files

Additional file 1: Supplementary Figures and Tables.

Additional file 2: Description (README) of available data in GigaDB.

Additional file 3: Command lines of programs and pipelines with run-time options used in this study.

Figure S1: Ogye.1.1 genome assembly statistics at each step.

Figure S2: A. An example of mis-assemblies in a scaffold. The x-axis represents the positions on chr1 or chr2 in galGal4 and the y-axis represents the position in scaffold_22 of the scaffold at the second step of the second stage (i.e., Opera scaffolder's result); **B.** In this example, there are two translocations: at P1 between L1 and L.2 and at P2 between L2.and L3. Since L.1, L.2 and L.3 are all >1Mbp, we broke the scaffold at P1 and P2. In this manner, we found 30 break points over all scaffolds in the breaking step of the second stage in Fig. 1B and Fig. S1.

Figure S3: Pseudo-reference-assisted assembly pipeline utilizing a hierarchical bipartite graph of PacBio long reads, scaffolds, and galGal4 chromosomes. The tools, used in grouping PacBio reads and scaffolds, are available in <https://github.com/sohnjangil/tsrator.git>.

Figure S4: Alignment of the Ogye.1.1 genome to galGal4/5 drawn by MUMmer.

Figure S5: Structural variation (SV) map of the Ogye.1.1 genome compared with galGal4 and galGal5. Insertions (red), deletions (blue), duplications (yellow), inversions (green), inter-chromosomal translocations (gray; Inter-translocation), and intra-chromosomal translocations (orange; Intra-translocation) are shown. SVs between the Ogye.1.1 genome and galGal4 or 5 are shown with Venn diagrams.

Figure S6: Mapping positions of mate-pair reads in the FM locus. The x- and y-axes indicate the positions of the first- and second-fragments, respectively, of a mate-pair read (insert size 3–10Kbp). The distance between the positions is the insert size of a mate-pair read.

Figure S7: Gene (protein-coding and lncRNA) annotation maps of the Ogye.1.1 genome with TE, SNV/INDEL, and GC ratio landscapes shown in a Circos plot. Color codes indicate coverage (%) of TE in a Mbp window, the number of protein-coding genes in a Mbp window, the number of lncRNAs in a Mbp window, SNP

and INDEL frequencies in a 100Kbp window, and the GC ratio in a 100Kbp window.

Figure S8: A schematic flow of our protein-coding gene annotation pipeline.

Figure S9: A computational pipeline for lncRNA annotations.

Figure S10: Circos plots illustrating the expression levels of protein-coding genes (bottom) and lncRNAs (top) across twenty tissues. The expression levels are indicated with a color-coded Z-score, described in the key.

Figure S11: Circos plots illustrating the CpG methylation levels in the promoters of protein-coding genes (bottom) and lncRNAs (top) across twenty tissues. The methylation levels are indicated with a color-coded Z-score, described in the key.

Table S1: Statistics of whole genome sequencing data (Illumina) after quality control.

Table S2: Structural variations in the Ogye.1.1 genome.

Table S3: Repeats in the Ogye.1.1 genome.

Table S4: Repeat composition in different assemblies.

Table S5: 164 galGal4 protein-coding genes missed in the Ogye.1.1 protein-coding gene annotations.

Abbreviations

BUSCO: Benchmarking Universal Single-Copy Orthologs; FM: fibromelanosis; CDS: coding DNA sequence; INDEL: insertions and deletions; PacBio: Pacific Biosciences; PC: principle component; PCA: principle component analyses; PCR: polymerase chain reaction; RNA-seq: RNA sequencing; RRBS: reduced representation bisulfite sequencing; SNP: single nucleotide polymorphism; SV: structural variation; TSS: transcription start site.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Cooperative Research Program for Agriculture Science and Technology Development (project title: National Agricultural Genome Program, Project No. PJ01045301 and PJ01045303).

Author contributions

K.T.L., N.S.K., H.H.C., and J.W.N. designed the study. K.T.L., Y.J.D., and C.Y.C. collected samples. D.J.L., H.H.C., and K.T.L. collected sequencing data. J.I.S., K.W.N., N.S.K., J.M.K., H.H.C., and J.M.N. performed the analysis and developed the methodology. J.I.S., K.W.N., J.M.K., H.S.H., and J.W.N. wrote the manuscript.

Acknowledgements

We thank all members of the BIG lab for helpful comments and discussions.

References

1. Domestic Animal Diversity Information System. <http://dad.fao.org/>.
2. Dorshorst B, Okimoto R, Ashwell C. Genomic regions associated with dermal hyperpigmentation, polydactyly and other morphological traits in the Silkie chicken. *J Hered* 2010;101(3):339–50.

3. Dorshorst B, Molin AM, Rubin CJ, et al. A complex genomic rearrangement involving the endothelin 3 locus causes dermal hyperpigmentation in the chicken. *PLoS Genet* 2011;7(12):e1002412.
4. Arora G, Mishra SK, Nautiyal B, et al. Genetics of hyperpigmentation associated with the fibromelanosis gene (Fm) and analysis of growth and meat quality traits in crosses of native Indian Kadaknath chickens and non-indigenous breeds. *Br Poult Sci* 2011;52(6):675–85.
5. Lukasiewicz M, Niemiec J, Wnuk A, et al. Meat quality and the histological structure of breast and leg muscles in Ayam Cemani chickens, Ayam Cemani × Sussex hybrids and slow-growing Hubbard JA 957 chickens. *J Sci Food Agric* 2015;95(8):1730–5.
6. Dharmayanthi AB, Terai Y, Sulandari S, et al. The origin and evolution of fibromelanosis in domesticated chickens: genomic comparison of Indonesian Cemani and Chinese Silkie breeds. *PLoS One* 2017;12(4):e0173147.
7. UNESCO's Memory of the World Programme . <http://www.unesco.org/new/en/communication-and-information/memory-of-the-world/>.
8. Zhang G, Li C, Li Q, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* 2014;346(6215):1311–20.
9. International Chicken Genome Sequencing C. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 2004;432(7018):695–716.
10. Warren WC, Hillier LW, Tomlinson C, et al. A new chicken genome assembly provides insight into avian genome structure. *G3 (Bethesda)* 2017;7(1):109–17.
11. Warren WC, Clayton DF, Ellegren H, et al. The genome of a songbird. *Nature* 2010;464(7289):757–62.
12. Animal genome size database (release 2.0). <http://www.genomesize.com/>.
13. Dalloul RA, Long JA, Zimin AV, et al. Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis. *PLoS Biol* 2010;8(9):e1000475.
14. Krishan A, Dandekar P, Nathan N, et al. DNA index, genome size, and electronic nuclear volume of vertebrates from the Miami Metro Zoo. *Cytometry A* 2005;65(1):26–34.
15. Poelstra JW, Vijay N, Bossu CM, et al. The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science* 2014;344(6190):1410–4.
16. Doyle JM, Katzner TE, Bloom PH, et al. The genome sequence of a widespread apex predator, the golden eagle (*Aquila chrysaetos*). *PLoS One* 2014;9(4):e95599.
17. Zhang G, Parker P, Li B, et al. The genome of Darwin's finch (*Geospiza fortis*) GigaScience Database 2012. <http://dx.doi.org/10.5524/100040>.
18. *Lepidothrix coronata* (blue-crowned manakin). <https://www.ncbi.nlm.nih.gov/genome/?term=Blue-crowned%20manakin>.
19. Tuttle EM, Bergland AO, Korody ML, et al. Divergence and functional degradation of a sex chromosome-like supergene. *Curr Biol* 2016;26(3):344–50.
20. Andrews CB, Mackenzie SA, Gregory TR. Genome size and wing parameters in passerine birds. *Proc Biol Sci* 2009;276(1654):55–61.
21. Cornetti L, Valente LM, Dunning LT, et al. The genome of the “great speciator” provides insights into bird diversification. *Genome Biology and Evolution* 2015;7(9):2680–91.
22. Qu Y, Zhao H, Han N, et al. Ground tit genome reveals avian adaptation to living at high altitudes in the Tibetan plateau. *Nature Communications* 2013;4:2071.
23. Li S, Li B, Cheng C, et al. Genomic signatures of near-extinction and rebirth of the crested ibis and other endangered bird species. *Genome Biol* 2014;15(12):557.
24. Warren W, Jarvis ED, Wilson RK, et al. Genomic data of the bald eagle (*Haliaeetus leucocephalus*) GigaScience Database 2014. <http://dx.doi.org/10.5524/101040>
25. Zhang G, Li B, Li C, et al. Genomic data of the American crow (*Corvus brachyrhynchos*) GigaScience Database 2014. <http://dx.doi.org/10.5524/101008>.
26. Zhan XJ, Pan SK, Wang JY, et al. Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle. *Nat Genet* 2013;45(5):563–U142.
27. Shapiro MD, Kronenberg Z, Li C, et al. Genomic diversity and evolution of the head crest in the rock pigeon. *Science* 2013;339(6123):1063–7.
28. Ganapathy G, Howard JT, Ward JM, et al. High-coverage sequencing and annotated assemblies of the budgerigar genome. *GigaScience* 2014;3(1):11.
29. Andrews CB, Gregory TR. Genome size is inversely correlated with relative brain size in parrots and cockatoos. *Genome* 2009;52(3):261–7.
30. Zhang G, Li B, Li C, et al. Genomic data of the little egret (*Egretta garzetta*) GigaScience Database 2014. <http://dx.doi.org/10.5524/101002>.
31. Zhang G, Li B, Li C, et al. Genomic data of the hoatzin (*Opisthocomus hoazin*) GigaScience Database 2014. <http://dx.doi.org/10.5524/101011>.
32. Zhang G, Li B, Li C, et al. Genomic data of the golden-collared manakin (*Manacus vitellinus*) GigaScience Database 2014. <http://dx.doi.org/10.5524/101010>.
33. Deorowicz S, Kokot M, Grabowski S, et al. KMC 2: fast and resource-frugal k-mer counting. *Bioinformatics* 2015;31(10):1569–76.
34. Earl D, Bradnam K, St John J, et al. Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res* 2011;21(12):2224–41.
35. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* 1988;16(3):1215.
36. Patel RK, Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* 2012;7(2):e30619.
37. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30(15):2114–20.
38. Luo R, Liu B, Xie Y, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* 2012;1(1):18.
39. Salmela L, Rivals E. LorDEC: accurate and efficient long read error correction. *Bioinformatics* 2014;30(24):3506–14.
40. Gnerre S, Maccallum I, Przybylski D, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A* 2011;108(4):1513–8.
41. Boetzer M, Pirovano W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* 2014;15(1):211.
42. Gao S, Sung WK, Nagarajan N. Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. *J Comput Biol* 2011;18(11):1681–91.
43. Harris R. Improved pairwise alignment of genomic DNA. PhD Thesis, The Pennsylvania State University, 2007.
44. Sohn JI, Nam JW. The present and future of de novo whole-

- genome assembly. *Brief Bioinform* 2018;**19**(1):23–40.
45. TSRATOR. <https://github.com/sohnjangil/tsrator.git>.
 46. English AC, Richards S, Han Y, et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* 2012;**7**(11):e47768.
 47. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;**20**(9):1297–303.
 48. VecScreen (<https://anonsvn.ncbi.nlm.nih.gov/repos/v1/trunk/c+/>) and UniVec database (<https://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/>).
 49. Simao FA, Waterhouse RM, Ioannidis P, et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;**31**(19):3210–2.
 50. OrthoDB. <http://www.orthodb.org/>.
 51. Rausch T, Zichner T, Schlattl A, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 2012;**28**(18):i333–i9.
 52. Layer RM, Chiang C, Quinlan AR, et al. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* 2014;**15**(6):R84.
 53. Li H. FermiKit: assembly-based variant calling for Illumina resequencing data. *Bioinformatics* 2015;**31**(22):3694–6.
 54. Chong Z, Ruan J, Gao M, et al. novoBreak: local assembly for breakpoint detection in cancer genomes. *Nat Methods* 2017;**14**(1):65–7.
 55. Tempel S. Using and understanding RepeatMasker. *Mobile Genetic Elements: Protocols and Genomic Applications* 2012:29–51.
 56. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* 2015;**6**(1):11.
 57. Picard Tools. <http://broadinstitute.github.io/picard/>.
 58. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 2012;**22**(3):568–76.
 59. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**(1):15–21.
 60. Pertea M, Pertea GM, Antonescu CM, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 2015;**33**(3):290–5.
 61. TransDecoder. <https://github.com/TransDecoder/TransDecoder/>.
 62. Wang L, Park HJ, Dasari S, et al. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res* 2013;**41**(6):e74.
 63. Kong L, Zhang Y, Ye ZQ, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* 2007;**35**(Web Server issue):W345–9.
 64. You BH, Yoon SH, Nam JW. High-confidence coding and non-coding transcriptome maps. *Genome Res* 2017;**27**(6):1050–62.
 65. Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010;**28**(5):511–5.
 66. Zhao Y, Li H, Fang S, et al. NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res* 2016;**44**(D1):D203–8.
 67. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* 2004;**11**(2–3):377–94.
 68. Pauli A, Valen E, Lin MF, et al. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res* 2012;**22**(3):577–91.
 69. Weikard R, Hadlich F, Kuehn C. Identification of novel transcripts and noncoding RNAs in bovine skin by deep next generation sequencing. *BMC Genomics* 2013;**14**(1):789.
 70. Billerey C, Boussaha M, Esquerre D, et al. Identification of large intergenic non-coding RNAs in bovine muscle using next-generation transcriptomic sequencing. *BMC Genomics* 2014;**15**(1):499.
 71. Al-Tobasei R, Paneru B, Salem M. Genome-wide discovery of long non-coding RNAs in rainbow trout. *PLoS One* 2016;**11**(2):e0148940.
 72. reshape2. <https://github.com/hadley/reshape>.
 73. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics* 2011;**27**(11):1571–2.
 74. Laurent L, Wong E, Li G, et al. Dynamic changes in the human methylome during differentiation. *Genome Res* 2010;**20**(3):320–31.
 75. Huang YZ, Sun JJ, Zhang LZ, et al. Genome-wide DNA methylation profiles and their relationships with mRNA and the microRNA transcriptome in bovine muscle tissue (*Bos taurus*). *Sci Rep* 2014;**4**:6546.
 76. Laine VN, Gossman TI, Schachtschneider KM, et al. Evolutionary signals of selection on cognition from the great tit genome and methylome. *Nat Commun* 2016;**7**:10474.
 77. Li A, Zhou ZY, Hei X, et al. Genome-wide discovery of long intergenic noncoding RNAs and their epigenetic signatures in the rat. *Sci Rep* 2017;**7**(1):14817.
 78. Larsen PA, Harris RA, Liu Y, et al. Hybrid de novo genome assembly and centromere characterization of the gray mouse lemur (*Microcebus murinus*). *BMC Biol* 2017;**15**(1):110.
 79. Sohn J, Nam K, Hong H, et al. Supporting data for “Whole genome and transcriptome maps of the entirely black native Korean chicken breed Yeonsan Ogye.” *GigaScience Database* 2018. <http://dx.doi.org/10.5524/100467>.