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 Lim[3], Kyung-Tai Lee[3], Yoon Jung Do[3], Chang Yeon Cho[4], Namshin Kim[5], Han-Ha Chai[3,§] and Jin-Wu Nam[1,2,‡]

[1]Department of Life Science, Hanyang University, Seoul 133-791, Republic of Korea
[2]Research Institute for Convergence of Basic Sciences, Hanyang University, Seoul 133-791, Republic of Korea
[3]Department of Animal Biotechnology & Environment, National Institute of Animal Science, RDA, Wanju 55365, Republic of Korea
[4]Animal Genetic Resource Research Center, National Institute of Animal Science, RDA, Namwon 55717, Republic of Korea
[5]Personalized Genomic Medicine Research Center, KRIBB, Daejeon 34141, Republic of Korea

[*] These authors contributed equally to this paper
[‡] jwnam@hanyang.ac.kr
[§] hanha@korea.kr

Abstract

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 (232.2X) and low-depth PacBio long-reads (11.5X). Although the contig and scaffold N50s (defined as the shortest contig or scaffold length at 50% of the entire assembly) of the initial *de novo* assembly were 53.6Kbp and 10.7Mbp, respectively, additional and pseudo-reference-assisted assemblies extended the assembly to 504.8Kbp for contig N50 (pseudo-contig) and 21.2Mbp for scaffold N50, which included 551 structural variations including the Fibromelanosis (*FM*) locus duplication, compared to galGal4 and 5. The completeness (97.6%) of the draft genome (Ogye_1) was evaluated with single copy orthologous genes using BUSCO, and found to be comparable to the current chicken reference genome (galGal5; 97.4%),

which was assembled with a long read-only method, and superior to other avian genomes (92~93%), assembled with short read-only and hybrid methods. To comprehensively reconstruct transcriptome maps, RNA sequencing (RNA-seq) and representation bisulfite sequencing (RRBS) data were analyzed from twenty different tissues, including black tissues. The maps included 15,766 protein-coding and 6,900 long non-coding RNA genes, many of which were expressed in the tissue-specific manner, closely related with the DNA methylation pattern in the promoter regions.

Keywords: Gallus gallus domesticus; *Yeonsan Ogye*; whole genome *de novo* assembly; Transcriptome maps; Hyperpigmentation

Background

The *Yeonsan 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, except for a 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 constructed 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 five thousand 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 *YO* genome (Ogye_1) and transcriptome maps, which include annotations of large SVs, SNPs, INDELs, and repeats, as well as coding and non-coding transcriptome maps along with DNA methylation landscapes across twenty different tissues of *YO*.

Results

Sample collection and data description

8-month-old *YO* chickens (object number: 02127), obtained from the Animal Genetic Resource Research Center of the National Institute of Animal Science (Namwon, Korea), were used in the study (**Figure 1A**). 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 (IACUC 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 Wizard DNA extraction kit [11] and prepared for DNA sequencing library construction. According to the DNA fragment (insert) size, three different library types were constructed: paired-end library for small inserts (280 and 500 bp) and mate-pair library for large inserts (3, 5, 8, and 10 Kbp), and fosmid libraries (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, 232.2 X Illumina short reads were obtained (59.6 X from the small insert libraries and 172.6 X from the large insert libraries) and, after filtering raw data with low quality (> 30% of the base-pairs in a read have a Phred score <20), 163.5X were used for genome assembly. To fill gaps and improve the scaffold quality, 11.5X PacBio long reads were additionally sequenced; the average length of the long reads was 6Kbp (**Table 1**).

Whole transcriptome sequencing

Total RNAs were extracted from twenty different tissues using 80% EtOH and TRIzol. The RNA concentration was checked by Quant-IT RiboGreen (Invitrogen, Carlsbad, USA). To assess the integrity of the total RNA, samples were run on a TapeStation RNA screentape (Agilent, Waldbronn, Germany). Only high quality RNA samples (RIN ≥R7.0) were used for RNA-seq library construction. Each library

was independently prepared with 300ng 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 PCR to create the final cDNA library. The libraries were quantified using qPCR according to the qPCR Quantification Protocol Guide (KAPA Library Quantificatoin kits for Illumina Sequecing platforms) and qualified using the TapeStation D1000 ScreenTape assay (Agilent Technologies, Waldbronn, Germany). As a result, we have sequenced about 1.5 billion RNA-seq reads from twenty different tissues, which are Breast, Liver, Bone marrow, Fascia, Cerebrum, Gizzard, Immature egg, Comb, Spleen, Mature egg, Cerebellum, Gall bladder, Kidney, Heart, Uterus, Pancreas, Lung, Skin, Eye, and Shank (**Table 2**).

Reduced representation bisulfite sequencing

Preparation of reduced representation bisulfite sequencing (RRBS) libraries was done following Illumina's RRBS protocol. 5µg of genomic DNA that had been digested with the restriction enzyme MspI and purified with a QIAquick PCR purification kit (QIAGEN, Hilden, Germany) was used for library preparation, which was done using a TruSeq Nano DNA Library Prep Kit (Illumina, San Diego, USA). Eluted DNA fragments were end-repaired, extended on the 3' end with an 'A', and ligated with Truseq adapters. After ligation had been assessed, the products, which ranged from 175 to 225bp in length (insert DNA of 55–105 bp plus adaptors of 120 bp), were excised from a 2%(w/v) Low Range Ultra Agarose gel (Biorad, Hercules, USA) and purified using the QIAquick gel extraction protocol. The purified DNA underwent bisulfite conversion using an EpiTect Bisulfite Kit (Qiagen, 59104). The bisulfite-converted DNA libraries were amplified by PCR (four cycles) using PfuTurbo Cx DNA polymerase (Agilent, 600410). The final product was then quantified using qPCR and qualified using the Agilent Technologies

2200 TapeStation assay (Agilent, Waldbronn, Germany). The final product was sequenced using the HiSeq™ 2500 platform (Illumina, San Diego, USA). As a result, we have produced 123 million RRBS reads (see **Table 3**) from twenty different tissues.

Hybrid whole genome assembly

The Ogye_1 genome was assembled using our hybrid genome assembly pipeline, employing the following four steps: preprocessing, hybrid *de novo* assembly, super-scaffolding, and polishing (**Figure 1B** and **Figure S1).** During the preprocessing step, the errors in the Illumina short-reads were corrected by KmerFreq and Corrector [12] using sequencing quality scores. In turn, using the corrected short-reads, the sequencing errors in the PacBio long reads were corrected by LoRDEC [13].

In hybrid *de novo* genome assembly, the initial assembly was done with the error-corrected short reads from the paired-end and mate-pair libraries using ALLPATHS-LG [14] with the default option, producing contigs and scaffolds. The resulting contigs and scaffolds showed 53.6 Kbp and 10.7 Mb of N50 (**Figure S1**), respectively. Next, the scaffolds were additionally connected using SSPACE-LongRead [15] and OPERA [16] with corrected PacBio long reads and FOSMID reads. The gaps within and between scaffolds were re-examined with GapCloser [12] with error-corrected short-reads. All resulting scaffolds were aligned to the galGal4 genome (GenBank assembly accession: GCA_000002315.2) by LASTZ [17] to detect putative mis-assemblies, verified by paired-end and mate-pair reads mapped to the scaffolds using BWA-MEM [18]. Comparison with results of LASTZ and BWA-MEM detected 30 mis-assemblies, break points of which were detected (**Figure S2.**) using Integrative Genomics Viewer (IGV) [19] and in-house programs. Breaking scaffolds at the break points resulted in pseudo-contig N50 of 108.6 Kbp and scaffold N50 of 18.7 Mb (**Figure S1**). A pseudo-contig is defined as a sequence broken by gaps of >1bp or a single N, which are assumed to be gaps or errors.

In the super-scaffolding stage, pseudo-reference-assisted assembly was done using LASTZ, BWA-MEM, PBJelly [20], and SSPACE-LongRead to enhance the quality of assembly using error-

corrected PacBio long-reads to reduce the topological complexity of the assembly graphs [21]. Because even scaffolding with long-reads can be affected by repetitive sequences, the results of mapping scaffolds to each chromosome were transformed into a hierarchical bipartite graph (**Figure S3**) to minimize the influence of repetitive sequences. The hierarchical bipartite graph was built by mapping PacBio (error-corrected) reads to scaffolds using BWA-MEM and, in turn, mapping scaffolds to the galGal4 genome (GCA_000002315.2) using LASTZ. Using the hierarchical bipartite graphs, all scaffolds and PacBio reads were finally assigned to each chromosome. Based on the results, super-scaffolding and additional gap-filling was performed by SSPACE-LongRead and PBJelly, respectively, resulting in scaffold N50 of 21.2Mbp (**Figure 1C** and **Figure S1**). In the last stage, nucleotide errors or ambiguities were corrected by the GATK pipeline [22] with paired-end reads, and in turn, any vector contamination was removed using VecScreen with UniVec database [23]. The final assembly results 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 504.8 Kbp in the final assembly, respectively. Among avian genome assemblies, these results are second best and the scaffold N50 is the best (**Figure 1C**). The complete genome sequence at the chromosome level (**Figure S4**) was built by connecting final scaffolds in the order of appearance in each chromosome with the introduction of 100 Kbp 'N' gaps between them. To evaluate the completeness of the genome, the *YO* draft 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 BUSCO [24]. The results indicated that the Ogye_1 genome contains more complete single-copy BUSCO genes, suggesting that the Ogye_1 genome is slightly more complete than the others (**Table 4**).

Large structural variations

When the Ogye_1 genome assembly was compared to two versions of the chicken reference genome assembly, galGal4 and 5, using LASTZ [17] and in-house programs/scripts, 551 common large (≥1 Kbp) structural variations (SVs) evident in both assembly versions were detected by at least two different SV prediction programs (Delly, Lumpy, FermiKit, and novoBreak) [25-28] (**Figure 1A**; **Table S1**). SVs

included 185 deletions (DELs), 180 insertions (INSs), 158 duplications (DUPs), 23 inversions (INVs), and 5 intra or inter-chromosomal translocations (TRs). 290 and 447 distinct SVs were detected relative to galGal 4 and galGal5, respectively (**Figure 2A**), suggesting that the two reference assemblies could include mis-assembly.

Although the Fibromelanosis (*FM*) locus, which contains the hyperpigmentation-related *edn3* gene, is known to be duplicated 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 (~1Mbp). A previous study suggested that the inverted duplication of the *FM* locus could be explained by three possible mechanistic scenarios (**Figure 2B**) [3]. To understand more about the mechanism of *FM* locus SV in the Ogye_1 genome, we compared it to the same locus in the galGal4 genome. Aligning paired-end reads of the Ogye_1 genome to the galGal4 genome, we found higher read depth at the *FM* locus in *YO*, indicating a gain of copy number at the locus (**Figure 2C** top). Also, we have identified same mate-pair information using paired-end and mate-pair reads (see Supplementary **Figure S5**). The intervening region between the two duplicated regions was estimated to be 412.6 Kbp in length in Ogye_1. Regarding possible mechanistic scenarios, mate-pair reads (3-10 Kbp, and FOSMID) mapped to the locus supported all three suggested scenarios, but an alignment of chromosome 20 from Ogye_1 and galGal4 showed that the intervening regions, including inner-partial regions in both duplicated regions, were inverted at the same time, which supports the first mechanistic scenario in **Figure 2B**. Given the resulting alignments, the *FM* locus of the Ogye_1 genome was updated according to the first scenario. The size of Gap_1 and Gap_2 were estimated at 164.5 Kbp and 63.3 Kbp, respectively.

Additionally, a large inversion was detected near a locus including the *tyrp1* gene (**Figure 2D**), which is known to be related to melanogenesis [29, 30]. However, when resequencing data from *White Leghorn* (white skin and plumage), *Korean Black* (white skin and black plumage), and white *Silkie* (black skin and white plumage) were compared to the galGal4 or 5 genome assemblies, the same break points

related to the inversion were detected (see supplementary **Figure S6** and **Figure S7**), suggesting that the inversion including *tyrp1* is not specifically related to skin hyperpigmentation.

Annotations

Repeats

Repeat elements in the Ogye_1 and other genomes were predicted by a reference-guided approach, RepeatMasker [31], which utilizes Repbase libraries [32]. In the Ogye_1 genome, 205,684 retro-transposable elements (7.65%), including LINEs (6.41%), SINEs (0.04%) and 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 (**Figure 3**, Supplementary **Table S2** and **Table S3**). It turns out that the composition of repeats in the Ogye_1 genome is similar to that in other avian genomes (**Figure 3**). Compared with other avian genomes, the Ogye_1 genome is more similar to galGal4 and 5 in terms of repeat composition including transposable elements (TEs) except for the fractions of simple repeats (0.12% for Ogye_1, 1.12% for galGal4 and 1.24% for galGal5), low-complexity (0.01% for Ogye_1, 0.24% for galGal4 and 0.25% for galGal5) and satellite DNA repeats (0.01% for Ogye_1, 0.20% for galGal4 and 0.22% for galGal5). The density of TEs across all chromosomes was depicted in **Figure 4A**.

SNPs/INDELs

To annotate SNPs and INDELs in the Ogye_1 genome, we mapped all paired-end libraries (SRR6189087-SRR6189094) to the Ogye_1 genome using BWA-MEM, and performed a series of post-processes including deduplication by Picard modules [33]. As a result, we have annotated 895,988 SNPs and 82,392 insertions/deletions (INDELs) across the genome using Genome Analysis Toolkit (GATK) modules. In performing GATK, we used HaplotypeCaller, combineGVCF, GenotypeGVCFs and VariantFiltration (with options "`QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0`") [22]. The densities of SNPs and INDELs across all chromosomes are depicted in **Figure 4A**.

Protein-coding genes

By analyzing large-scale of RNA-seq data from twenty different tissues through our protein-coding gene annotation pipeline (**Figure 4B**), 15,766 protein-coding genes were annotated in the Ogye_1 genome (**Figure 4C**), including 946 novel genes and 14,820 known genes. 164 protein-coding genes annotated in galGal4 were missing from the Ogye_1 assembly. The density of protein-coding genes across all chromosomes was depicted in **Figure 4A**.

To sensitively annotate protein-coding genes, all paired-end RNA-seq data were mapped on the Ogye_1 genome by STAR [34] for each tissue and the mapping results were then assembled into potential transcripts using StringTie [35]. 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 [36]. For high-confidence prediction, transcripts with intact gene structures (5'UTR, CDS, and 3'UTR) were selected. To verify the coding potential, the candidate sequences were examined using CPAT [37] and CPC [38]. 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 is >0. Candidates with low coding potential or that were partially annotated were examined to determine if their loci overlapped with annotated protein-coding genes from galGal4 (ENSEMBL cDNA release 85). Overlapping genes were added to the set of Ogye_1 protein-coding genes. Finally, 164 genes were not mapped to the Ogye_1 genome by GMAP, 131 of which were confirmed to be expressed in *YO* (≥ 0.1 FPKM) using all paired-end *YO* RNA-seq data. However, expression of the remaining 33 genes was not confirmed, suggesting that they are not expressed in *YO* (< 0.1 FPKM) or have been lost from the Ogye_1 genome. The missing genes are listed in **Table S4**. In contrast, the 946 newly annotated Ogye_1 genes appeared to be mapped to the galGal4 or galGal5 genomes (**Figure 4C**).

lncRNAs

In total, 6,900 *YO* lncRNA genes, including 5,610 novel loci and 1,290 known loci, were confidently annotated from RNA-seq data using our lncRNA annotation pipeline, adopted from our

previous study [39] (**Figure 5A**).

To annotated and profile lncRNA genes, pooled single- and paired-end RNA-seq reads of each tissue were mapped to the Ogye_1 genome (PRJNA412424) using STAR [34], and subjected to transcriptome assembly using Cufflinks [44], leading to the construction of transcriptome maps for twenty tissues. The resulting maps were combined using Cuffmerge and, in total, 206,084 transcripts from 103,405 loci were reconstructed in the Ogye genome. We removed other biotypes of RNAs (the sequences of mRNAs, tRNAs, rRNAs, snoRNAs, miRNAs, and other small non-coding RNAs downloaded from ENSEMBL biomart) and short transcripts (less than 200nt in length). 54,760 lncRNA candidate loci (60,257 transcripts) were retained, and which were compared with a chicken lncRNA annotation of NONCODE (v2016) [45]. Of the candidates, 2,094 loci (5,215 transcripts) overlapped with previously annotated chicken lncRNAs. 52,666 non-overlapping loci (55,042 transcripts) were further examined to determine whether they had coding potential using CPC score [38]. 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 36,873bp, which is the intron length of the 99th percentile, were re-examined using both exon-junction reads consistently presented over twenty tissues and the maximum entropy score [46], as done in our previous study [39]. 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, two candidates were reconnected, and those with a single exon were discarded. In the final version, 9,529 transcripts from 6,900 lncRNA loci (5,610 novel and 1,290 known) were annotated as lncRNAs (see **Figure 5B**), which included 6,170 (89.40 %) intergenic lncRNAs and 730 (10.57 %) anti-sense ncRNAs. Consistent with other species [40-43], the median gene length and the median exon number of Ogye lncRNAs were less than those of protein-coding genes (**Figure 5C** and **D**).

Although 13,540 of 15,766 protein-coding genes (92%) were redetected by our transcriptome assembly and protein-coding gene annotations (see **Figure 4C**), 8,204 of NONCODE lncRNAs were

missed in our Ogye lncRNA annotations (**Figure 5B**), the majority of which were either fragments of protein-coding genes or not expressed in all twenty Ogye tissues (**Figure 5B**). Only 276 were missed in our transcriptome assembly, and 648 were missed in our draft genome.

Coding and non-coding transcriptome maps

Using paired-end *YO* RNA-seq data, the expression levels of protein-coding and lncRNA genes were calculated across twenty tissues (**Figure 6A**), which were dynamically changed. The profiled transcriptomes included 1,814 protein-coding genes and 1,226 lncRNA genes, expressed ≥ 10 FPKM in only one tissue as well as 1,559 and 351 expressed ≥ 10 FPKM in all tissues. The Ogye black tissues (fascia, comb, skin, and shank) expressed 6,702 protein-coding and 3,291 lncRNA genes ≥ 10 FPKM, the majority of which appeared to be expressed in tissue-specific manner (**Figure 6B**). For instance, two neighboring genes, *krt9* and *lnc-lama2-1* are highly expressed in comb and shank, and (**Figure 6C** and **D**).

As lncRNAs tend to be specifically expressed in a tissue or in related tissues, they could be better factors for defining genomic characteristics of tissues than protein-coding genes. To prove this idea, principle component analyses (PCA) were performed with tissue-specific lncRNAs and protein-coding genes using reshape2 R package (**Figure 7**) [47]. As expected, the 1st, 2nd, and 3rd PCs of lncRNAs enabled us to predict the majority of variances, and better discerned distantly-related tissues and functionally and histologically-related tissues (i.e., black tissues and brain tissues) (**Figure 7B**) than those of protein-coding genes (**Figure 7A**).

DNA methylation maps

RRBS reads were mapped to Ogye draft genome (**Table 3**), and calculated DNA methylation signals (C to T changes in CpGs) across chromosomes using Bismark in each sample [48]. The DNA methylation landscape in protein-coding and lncRNA genes were shown in **Figure 8A**. Based on CpG methylation pattern in the promoter of genes, hierarchical clustering was performed using rsgcc R package, and clusters including adjacent or functionally similar tissues, such as cerebrum and cerebellum, immature

and mature eggs, or comb and skin were identified (**Figure 8B**). Of all CpG sites in genomes, 31~65% were methylated across tissues while only 19~43% were methylated in the promoter (**Table 5**), indicating hypomethylation status in the promoters of expressed genes. Exceptionally, the methylation levels of spleen (40.28% in all genomic regions; 24.92% in promoter region) and liver (30.82% in all genomic region; 18.68% in promoter regions) displayed much less methylation signal, compared to those of others. In fact, examining the averaged methylation landscapes over protein-coding and lncRNA loci showed that the methylation levels in gene body regions were much higher than those in 2 Kbp upstream regions from transcript start sites (TSS) (**Figure 8C** and **D**). To confirm that CpG methylation represses the expression, 280 protein-coding and 392 lncRNA genes of $Max[FPKM] \geq 10$ with highly tissue-specific expression pattern were selected. The methylation level of highly expressed genes was much lower than others (**Figure 8E** and **F**).

To correlate the expression of genes with their methylation levels, only tissues in which a certain position had sufficient read coverage (at least five) were considered for measuring the correlation. The Spearman correlation coefficients between the expression and methylations levels were observed over twenty tissues (**Figure 9**). 454 protein-coding and 25 lncRNA genes displayed a negative correlation to methylation levels in promoter regions, while 157 protein-coding and 20 lncRNA genes have a positive correlation (box plots in **Figure 9**)

Discussions

In this work, the first draft genome from a Korean native chicken breed, *YO*, was constructed with genomic variation, repeat, and protein-coding and non-coding gene maps. Compared with the chicken reference genome maps, many more novel coding and non-coding elements were identified from large-scale RNA-seq datasets across twenty different tissues. Although the Ogye_1 genome is comparable with galGal5 with respect to genome completeness evaluated by BUSCO, the Ogye_1 seems to lack simple and long repeats compared with galGal5, which was assembled from high-depth PacBio long-reads (50X)

[10] that can capture simple and long repeats. Although we also used PacBio long reads, some simple and satellite repeats would be missed during assembly, because we used the PacBio data of shallow depth (11.5X) for scaffolding and gap filling. A similar tendency can be seen in the Golden-collared manakin genome (ASM171598v1) (**Figure 3**), which was also assembled in a hybrid manner using MaSuRCA assembler with high-depth Illumina short-reads and low-depth PacBio long-reads.

15,766 protein-coding 6,900 lncRNA genes were annotated from twenty tissues of *YO*. 946 novel protein-coding genes were identified while 164 genes of *Galllus gallus red junglefowl* were missed in our annotations. In the case of lncRNAs, only about 19% of previously annotated chicken lncRNAs were redetected, and the remainders 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 lncRNAs in mammals in their genomic characteristics, including transcript length, exon number, and tissue-specific expression pattern, providing evidence for the accuracy of the new annotations. Hence, our lncRNA catalogue may help us to improve lncRNA annotations in the chicken reference genome.

Annotated genomic variations and comparative analysis of coding and non-coding genes will provide a resource for understanding genomic differences and evolution of *YO* as well as identifying functional elements related to its unique phenotypes, such as fibromelanosis. Additionally, such analyses will be useful for future genome-based animal genetics.

Availability of data

All of our sequencing data and the genome sequence have been deposited in NCBI's BIOPROJECT under the accession number PRJNA412424. The raw sequence data have been deposited in the Short Read Archive (SRA) under accession numbers SRR6189081-SRR6189098 (**Table 1**), SRX3223583-SRX3223622 (**Table 2**), and SRX3223667-SRX3223686 (**Table 3**).

Additional files

The supplementary Figures and tables have been included in a supplementary file:

Figure S1. Assembly statistics of Ogye_1 gnome assembly at each step.

Figure S2. Filtration of noise and mis-assembly detection using Lastz

Figure S3. Hierarchical mapping information in the reference-assisted additional assembly pipeline.

Figure S4. Alignment of the Ogye_1 genome to galGal4/5 drawn by MUMmer.

Figure S5. Mate-pair information in *FM* locus.

Figure S6. Mate-pair information near tyrp1 in chrZ of galGal4 and galGal5.

Figure S7. Break points near *tyrp1* in galGal4 chrZ.

Table S1. Structural variations in the Ogye_1 genome

Table S2. Repeats in the Ogye_1 genome

Table S3. Repeat contents in different assemblies.

Table S4. 164 unmapped genes among galGal4 protein-coding genes.

Acknowledgements

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

KTL, NSK, HHC, and JWN designed the study, KTL, YJD and CYC collected samples, DJL, HHC and KTL collected sequencing data, and JIS, KWN, NSK, JMK, HHC and JMN performed the analysis and developed the methodology. JIS, KWN and JMK wrote the manuscript.

Competing interests

The authors declare that they have no competing interests.

Tables

Table 1. Whole genome sequencing data

Platform	Library type	Insert-size	No. of read (10^6)	Total base (Gbp)	Coverage (X)	SRA accession
Illumina HiSeq 2000	Paired-end	280 bp	129.6	19.5	18.6	SRR6189087
			124.5	18.7	17.8	SRR6189084
		500 bp	43.6	6.6	6.2	SRR6189095
			47.3	7.1	6.8	SRR6189097
			14.0	2.1	2.0	SRR6189096
			14.1	2.1	2.0	SRR6189098
			14.6	2.2	2.1	SRR6189082
			28.7	4.3	4.1	SRR6189094
	Mate-pair	3Kbp	146.5	21.8	20.8	SRR6189093
			135.0	20.1	19.1	SRR6189083
		5Kbp	114.8	17.1	16.3	SRR6189081
			106.4	15.6	15.1	SRR6189088
		8Kbp	136.6	20.4	19.4	SRR6189085
			135.3	20.2	19.2	SRR6189086
		10Kbp	169.1	25.2	24.0	SRR6189091
			157.9	23.5	22.4	SRR6189092
	FOSMID	40Kbp	169.9	17.6	16.3	SRR6189089
PacBio RS II	Long-read	6Kbp (ave. read length)	1.7	12.1	11.5	SRR6189090

Table 2. Sequencing and mapping summary of RNA-seq reads

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
Gall bladder	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

Table 3. Sequencing and mapping summary of RRBS reads

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
Gall bladder	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

Assembly	Complete		Fragment	Missing
	Single-copy	Duplication		
Ogye_1	97.60%	0.50%	0.90%	1.00%
galGal4	96.90%	0.90%	1.10%	1.10%
galGal5	97.40%	0.90%	0.70%	1.00%
Turkey_5.0	93.70%	0.50%	4.10%	1.70%
BGI_1.0	92.60%	0.40%	4.80%	2.20%
taeGut3.2.4	93.60%	2.20%	2.70%	1.50%

Table 5. Summary of methylated CpGs across twenty tissues

	All genomic region			Promoter region		
	Total No. of site	Methylated CpG sites		Total No. of site	Methylated CpG sites	
		No. of site	Fraction (%)		No. of site	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
Gall bladder	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

Figure legends

Figure 1. A. Picture of *Yeonsan Ogye*; **B.** Hybrid genome assembly pipeline; **C.** The N50 and average length of pseudo-contigs and scaffolds of the Ogye_1 and other avian genomes created using the indicated assembly methods (last column; here, sequencing platforms are designated as follows: "I" indicates Illumina, "P" is PacBio, "S" is Sanger, and "4" is Roche454).

Figure 2. A. Structural variation (SV) map of the Ogye_1 genome compared with galGal4 and galGal5. Insertions (red), deletions (green), duplications (yellow), inversions (blue), inter-chromosomal translocations (gray; inter-trans), and intra-chromosomal translocations (orange; intra-trans) are shown. Variations in common between the genomes are shown in the middle with Venn diagrams; **B.** Three possible scenarios that could have led to SV (inverted duplication) of the Fibromelanosis (*FM*) locus in the genomes of hyperpigmented chicken breeds; **C.** Copy gain of the *FM* locus, which includes the *end3* gene (indicated by the thin purple-shaded boxes), was identified on chromosome 20. The green- and yellow-shaded boxes indicate duplicated regions (Dupl_1 and Dupl_2, respectively) and the gray-shaded boxes indicate gaps (Gap_1 and Gap_2). The sizes of Gap_1 and Gap_2 were estimated to be 164.5 Kbp and 63.3 Kbp, respectively; **D.** Inversion of a genomic region on chromosome Z that includes *tyrp1*. The purple-shaded boxes indicate the *tyrp1* locus.

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

Figure 4. A. Gene (protein-coding and lncRNA) annotation maps of the Ogye_1 genome with TE, SNV/INDEL, and GC ratio landscapes are shown in a Circos plot. Note that TE and SNV denote transposable element and single nucleotide variation, respectively; **B.** A schematic flow of our protein-coding gene annotation pipeline and a Venn diagram showing the number of protein-coding genes in the Ogye_1 genome. **C.** We have annotated 946 novel genes, and found 13,541 known genes by transcript assembly. 1,279 known genes were annotated by mapping using GMAP. 164 annotated genes were not

included in our *YO* protein-coding gene set, among which 33 were not expressed (<0.1 FPKM). All of the 946 newly annotated genes are mapped to the galGal4 or galGal5 genomes.

Figure 5. A. A computational pipeline for lncRNA annotations. **B.** The number of Ogye_1 and galGal4 lncRNAs are shown in the Venn diagram. About 51% (3,873) of non-overlapping galGal4 lncRNAs are predicted to be fragments of protein-coding genes. 45% (3,407) were not expressed in any tissue. Only 4% (276) of the lncRNAs are actually missing from the set of Ogye lncRNAs (bottom). **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 exon number per transcript. Otherwise, as in **C**.

Figure 6. A. The circus plot illustrates the expression levels of protein-coding genes (left) and lncRNA (right) across twenty tissues. The expression levels are indicated with a color-coded z-score, described in the key; **B.** The expression patterns of black tissues-specific genes. Expression levels are indicated with a color-coded Z-score (red for low and blue for high expression) as shown in the key; **C**. Expression levels of *krt9* across twenty tissues; **D.** Expression levels of *lnc-lama2-1* across twenty tissues.

Figure 7. A. Principal component analysis (PCA) using tissue-specific lncRNAs. PCs explaining the variances are indicated with the amount of the contribution in the left-top plot. PCA plots with PC1, PC2, and PC3 were demonstrated in a pairwise manner. Each tissue is indicated on the PCA plot with a specific color; **B.** PCA using tissue-specific protein-coding genes. Otherwise, as in **A**.

Figure 8. A. The circus plot illustrates the CpG methylation levels in the promoters of protein-coding genes (left) and lncRNA (right) across twenty tissues. The methylation levels are indicated with a color-coded z-score, described in the key; **B.** Hierarchical clustering using Pearson correlation of DNA methylation patterns between tissues; **C.** Average DNA methylation landscape across gene bodies of protein-coding genes and their flanking regions; **D.** Average DNA methylation landscape across gene bodies of lncRNAs and their flanking regions; **E.** Average DNA methylation level of the protein-coding gene in tissues where the gene is highest expressed and lowest expressed, respectively. **F.** Average DNA methylation level of the lncRNA gene in tissues where the gene is highest expressed and lowest expressed, respectively. The methylation level is indicated with a color-coded z-score, described in the key.

Figure 9. A circos plot illustrating the Spearman correlation coefficients between expression levels and methylation levels of genes across chromosomes (heatmaps). The bar chart indicates the count of genes (left for protein-coding genes and right for lncRNAs) with a significant negative (red) and positive (cyan) correlation between the methylation level in their promoters and their expression values.

References

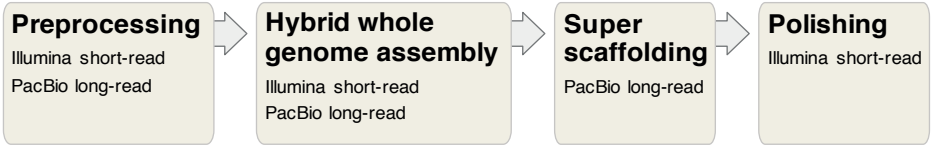
1. Domestic Animal Diversity Information System. http://dad.fao.org/.
2. Dorshorst B, Okimoto R and Ashwell C. Genomic regions associated with dermal hyperpigmentation, polydactyly and other morphological traits in the Silkie chicken. J Hered. 2010;101 3:339-50. doi:10.1093/jhered/esp120.
3. Dorshorst B, Molin AM, Rubin CJ, Johansson AM, Stromstedt L, Pham MH, et al. A complex genomic rearrangement involving the endothelin 3 locus causes dermal hyperpigmentation in the chicken. PLoS Genet. 2011;7 12:e1002412. doi:10.1371/journal.pgen.1002412.
4. Arora G, Mishra SK, Nautiyal B, Pratap SO, Gupta A, Beura CK, 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. doi:10.1080/00071668.2011.635637.
5. Łukasiewicz M, Niemiec J, Wnuk A and Mroczek-Sosnowska N. 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. Journal of the Science of Food and Agriculture. 2015;95 8:1730-5.
6. Dharmayanthi AB, Terai Y, Sulandari S, Zein MS, Akiyama T and Satta Y. The origin and evolution of fibromelanosis in domesticated chickens: Genomic comparison of Indonesian Cemani and Chinese Silkie breeds. PLoS One. 2017;12 4:e0173147. doi:10.1371/journal.pone.0173147.
7. It has been registered on UNESCO's Memory of the World Programme in 2009. http://www.unesco.org/new/en/communication-and-information/memory-of-the-world/.
8. Zhang GJ, Li C, Li QY, Li B, Larkin DM, Lee C, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. Science. 2014;346 6215:1311-20. doi:10.1126/science.1251385.
9. Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, et al. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 2004;432 7018:695-716. doi:10.1038/nature03154.
10. Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, et al. A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. G3-Genes Genomes Genetics. 2017;7 1:109-17. doi:10.1534/g3.116.035923.
11. Miller SA, Dykes DD and Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. 1988;16 3:1215.
12. 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;1 1:18. doi:10.1186/2047-217X-1-18.
13. Salmela L and Rivals E. LoRDEC: accurate and efficient long read error correction. Bioinformatics. 2014;30 24:3506-14. doi:10.1093/bioinformatics/btu538.
14. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, 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. doi:10.1073/pnas.1017351108.
15. Boetzer M and Pirovano W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. Bmc Bioinformatics. 2014;15 1:211. doi:10.1186/1471-2105-15-211.

16. Gao S, Sung WK and Nagarajan N. Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. J Comput Biol. 2011;18 11:1681-91. doi:10.1089/cmb.2011.0170.
17. Harris R. *Improved pairwise alignment of genomic DNA.* PhD Thesis, 2007.
18. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997. 2013.
19. Thorvaldsdottir H, Robinson JT and Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14 2:178-92. doi:10.1093/bib/bbs017.
20. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One. 2012;7 11:e47768. doi:10.1371/journal.pone.0047768.
21. Sohn JI and Nam JW. The present and future of de novo whole-genome assembly. Brief Bioinform. 2016:bbw096. doi:10.1093/bib/bbw096.
22. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20 9:1297-303. doi:10.1101/gr.107524.110.
23. VecScreen https://anonsvn.ncbi.nlm.nih.gov/repos/v1/trunk/c++/ and UniVec database https://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/.
24. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV and Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31 19:3210-2. doi:10.1093/bioinformatics/btv351.
25. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V and Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28 18:i333-i9. doi:10.1093/bioinformatics/bts378.
26. Layer RM, Chiang C, Quinlan AR and Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014;15 6:R84. doi:10.1186/gb-2014-15-6-r84.
27. Li H. FermiKit: assembly-based variant calling for Illumina resequencing data. Bioinformatics. 2015;31 22:3694-6. doi:10.1093/bioinformatics/btv440.
28. Chong Z, Ruan J, Gao M, Zhou W, Chen T, Fan X, et al. novoBreak: local assembly for breakpoint detection in cancer genomes. Nat Methods. 2017;14 1:65-7. doi:10.1038/nmeth.4084.
29. Xu Y, Zhang XH and Pang YZ. Association of Tyrosinase (TYR) and Tyrosinase-related Protein 1 (TYRP1) with Melanic Plumage Color in Korean Quails (Coturnix coturnix). Asian-Australas J Anim Sci. 2013;26 11:1518-22. doi:10.5713/ajas.2013.13162.
30. Yu S, Liao J, Tang M, Wang Y, Wei X, Mao L, et al. A functional single nucleotide polymorphism in the tyrosinase gene promoter affects skin color and transcription activity in the black-boned chicken. Poult Sci. 2017;96 11:4061-7. doi:10.3382/ps/pex217.
31. Tempel S. Using and understanding RepeatMasker. Mobile Genetic Elements: Protocols and Genomic Applications. 2012:29-51.
32. Bao W, Kojima KK and Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6 1:11. doi:10.1186/s13100-015-0041-9.
33. Picard Tools. http://broadinstitute.github.io/picard/.

34. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29 1:15-21. doi:10.1093/bioinformatics/bts635.
35. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT and Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33 3:290-5. doi:10.1038/nbt.3122.
36. TransDecoder. https://github.com/TransDecoder/TransDecoder/.
37. Wang L, Park HJ, Dasari S, Wang SQ, Kocher JP and Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Research. 2013;41 6:e74-e. doi:10.1093/nar/gkt006.
38. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Research. 2007;35 suppl_2:W345-W9. doi:10.1093/nar/gkm391.
39. You B-H, Yoon S-H and Nam J-W. High-confidence coding and noncoding transcriptome maps. Genome research. 2017;27 6:1050-62.
40. Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, et al. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome research. 2012;22 3:577-91.
41. Weikard R, Hadlich F and Kuehn C. Identification of novel transcripts and noncoding RNAs in bovine skin by deep next generation sequencing. BMC genomics. 2013;14 1:789.
42. Billerey C, Boussaha M, Esquerré D, Rebours E, Djari A, Meersseman C, et al. Identification of large intergenic non-coding RNAs in bovine muscle using next-generation transcriptomic sequencing. BMC genomics. 2014;15 1:499.
43. Al-Tobasei R, Paneru B and Salem M. Genome-wide discovery of long non-coding RNAs in rainbow trout. PLoS One. 2016;11 2:e0148940.
44. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology. 2010;28 5:511-5.
45. Zhao Y, Li H, Fang S, Kang Y, Hao Y, Li Z, et al. NONCODE 2016: an informative and valuable data source of long non-coding RNAs. Nucleic acids research. 2016;44 D1:D203-D8.
46. Yeo G and Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. Journal of computational biology. 2004;11 2-3:377-94.
47. reshape2. https://github.com/hadley/reshape.
48. Krueger F and Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27 11:1571-2. doi:10.1093/bioinformatics/btr167.

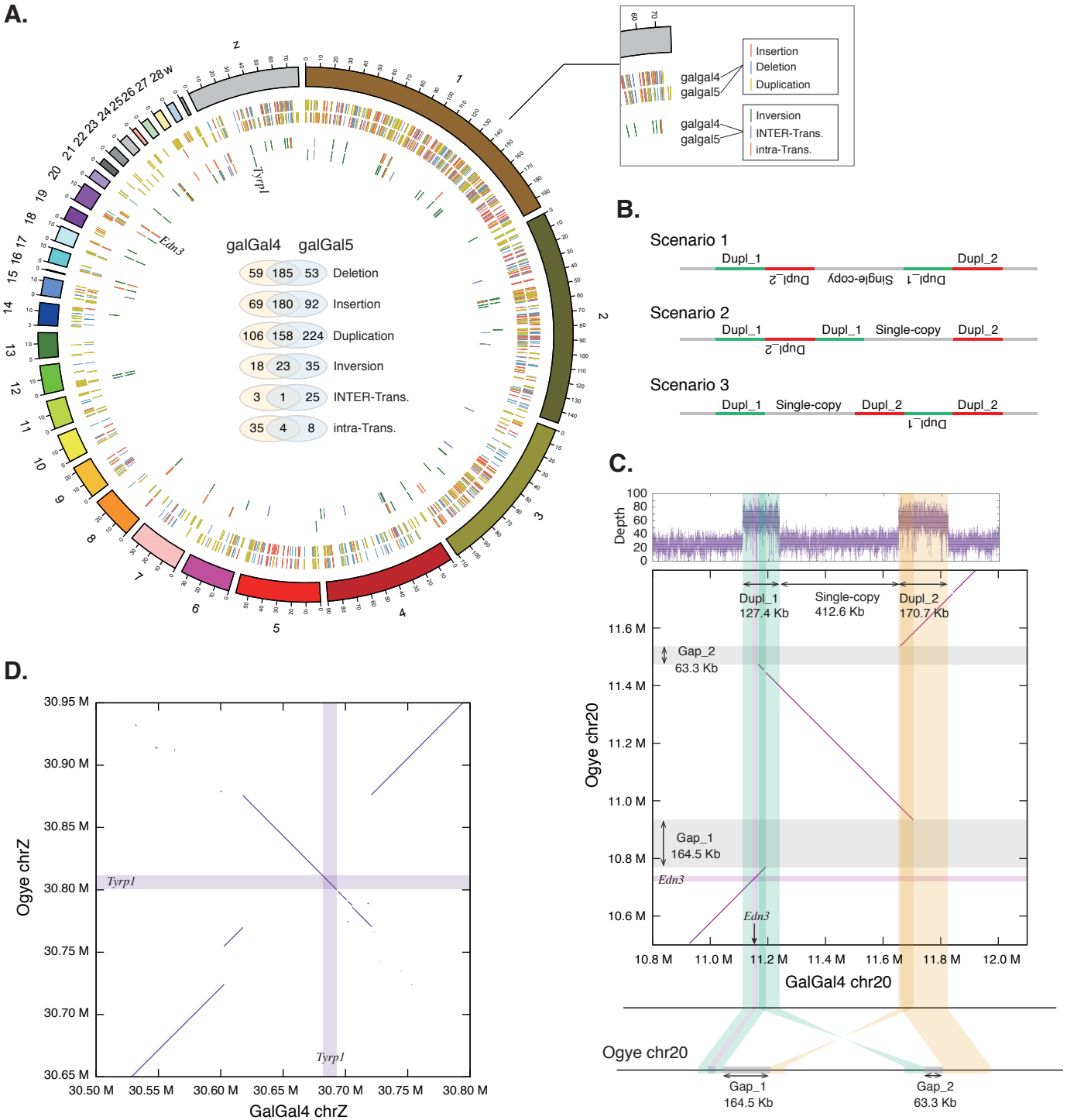
A.

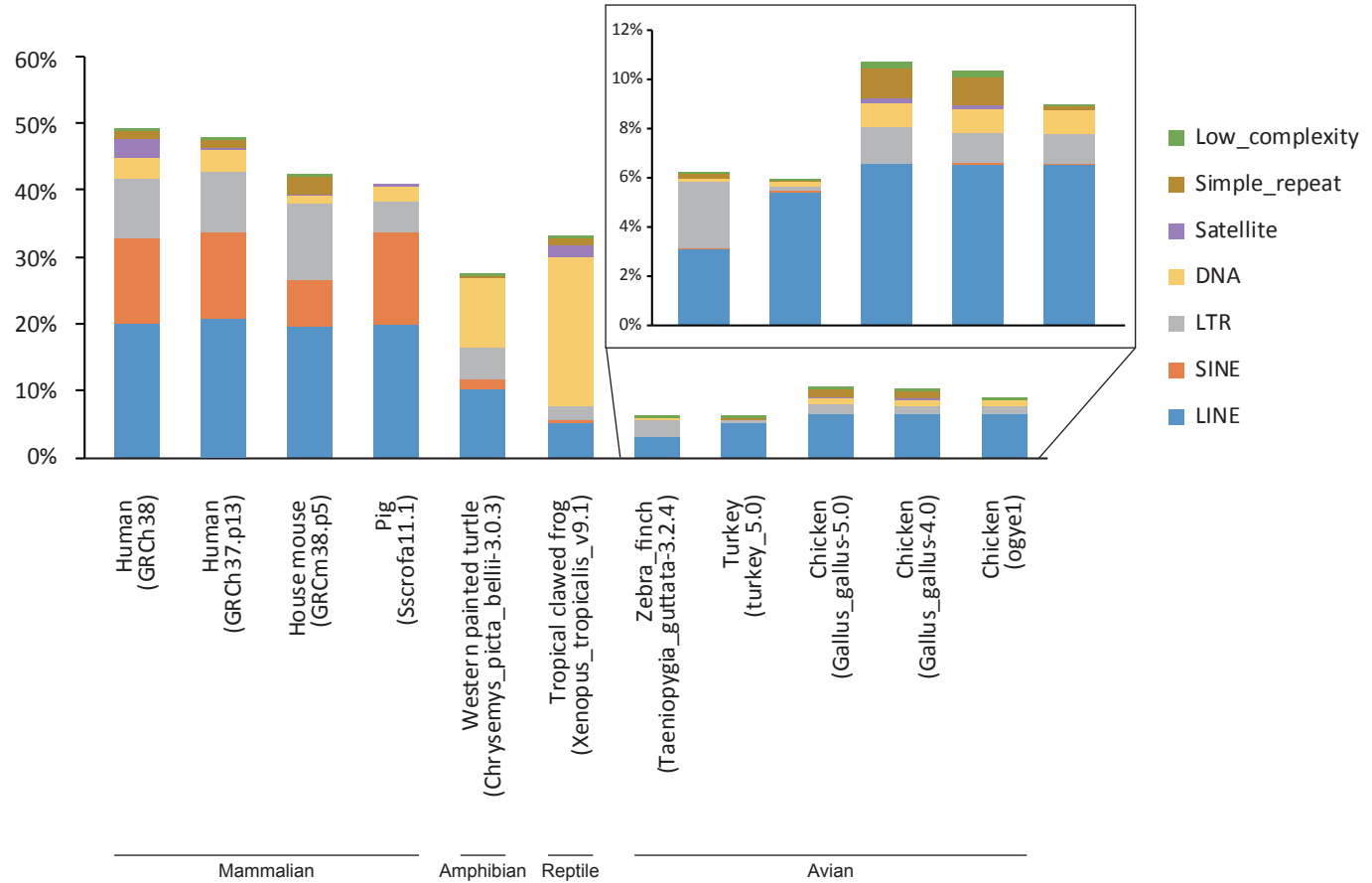
B.

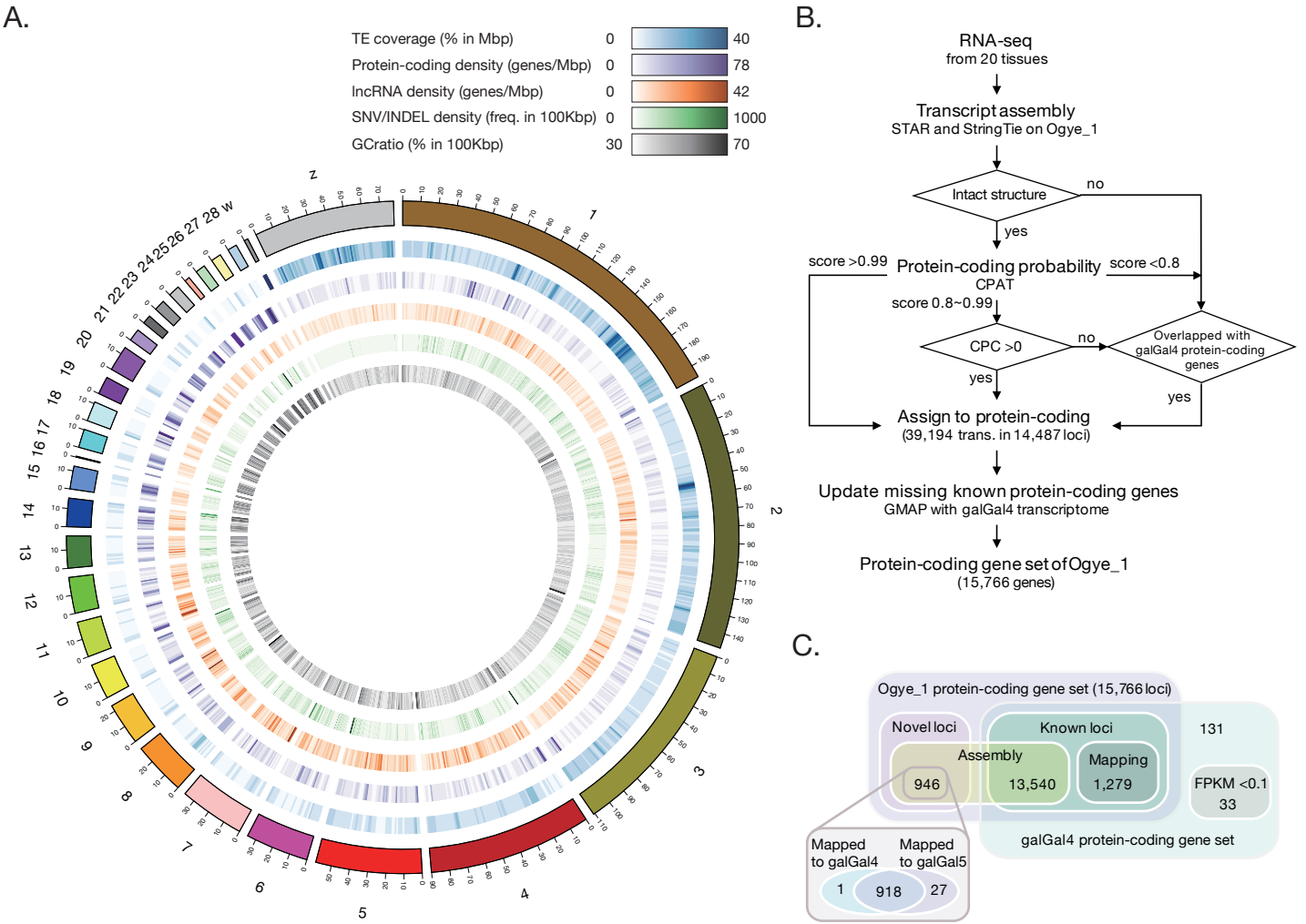


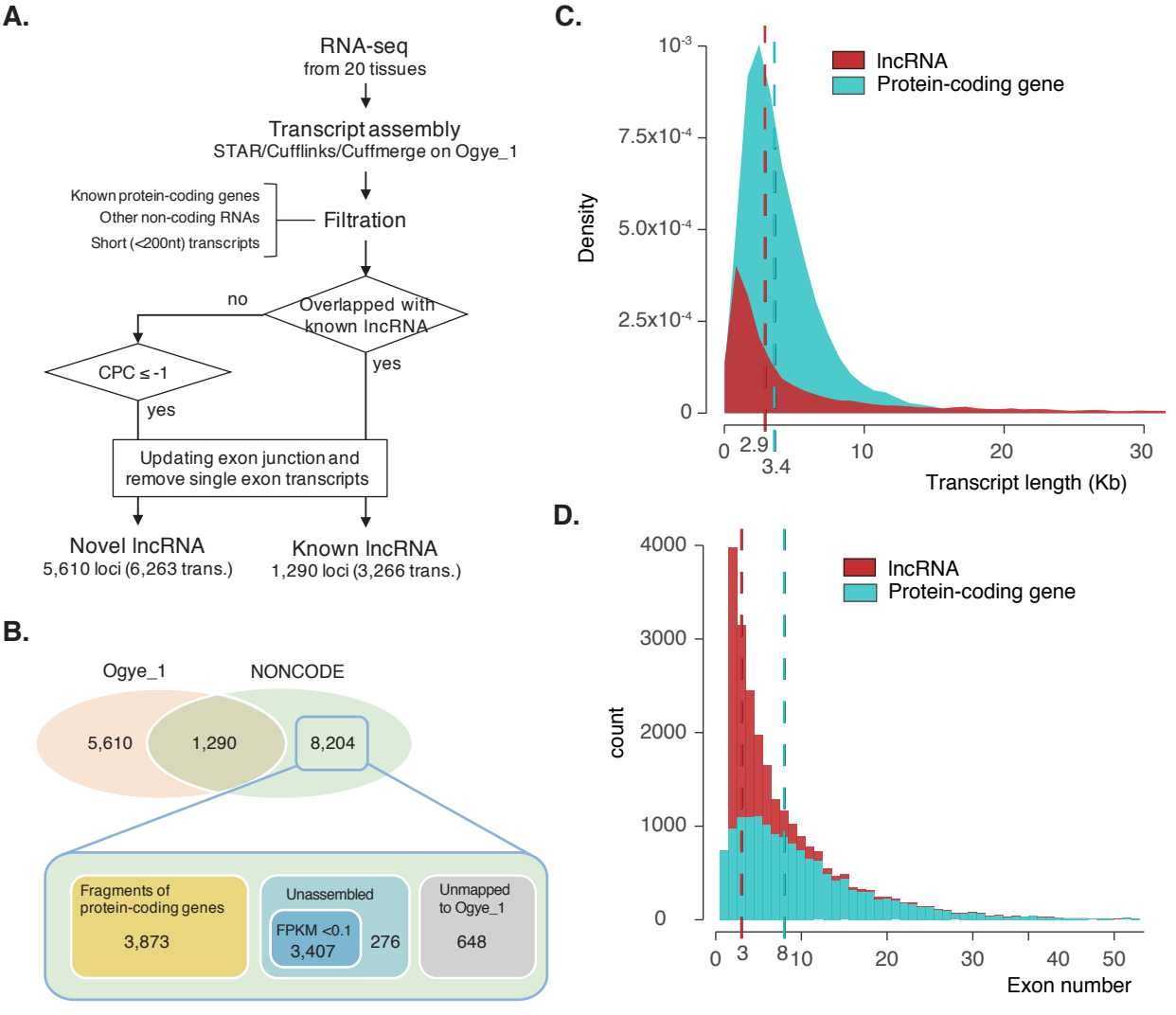
C.

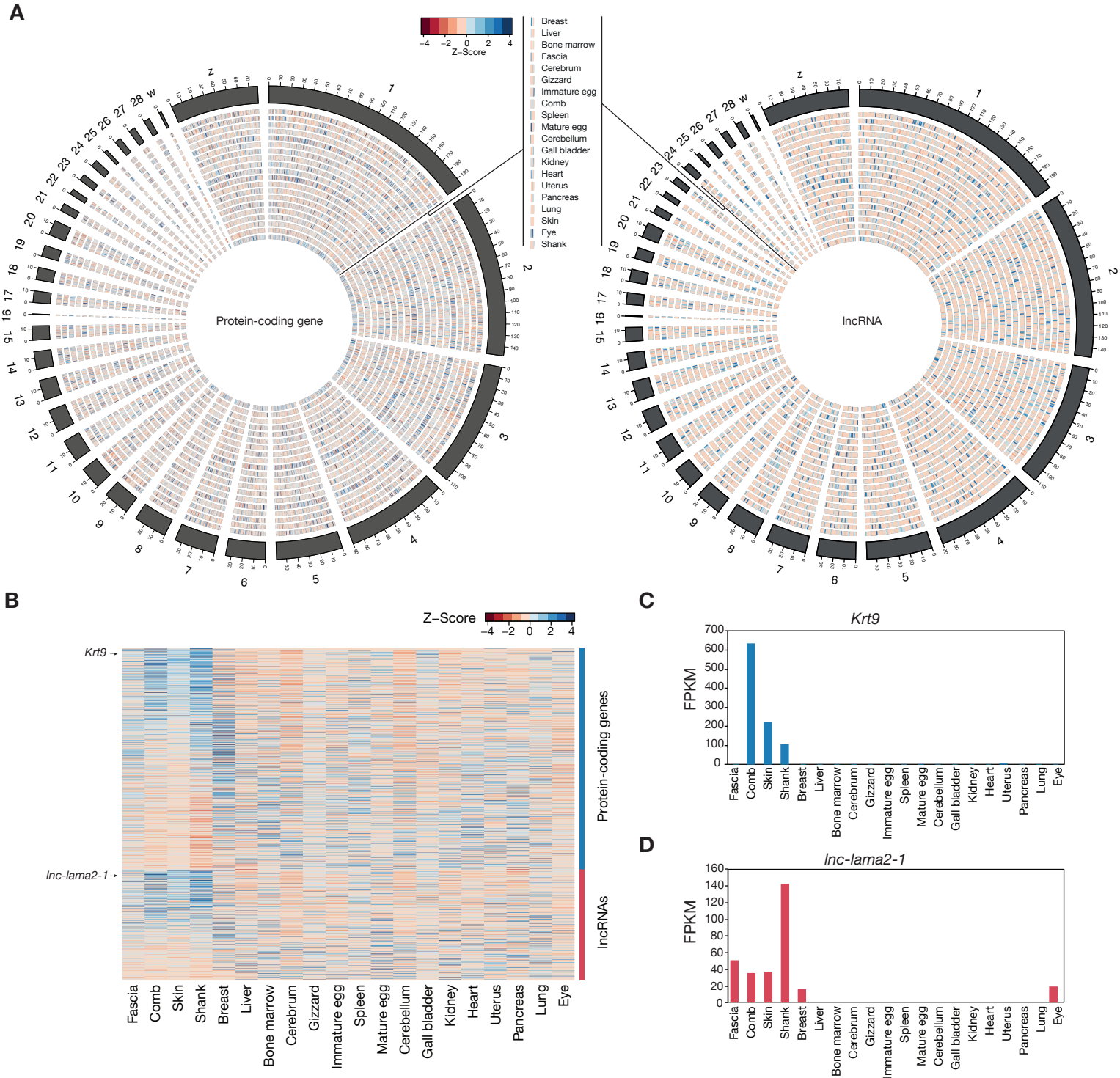
Species	Assembly	Total length (Gbp)	Pseudo-contig			Scaffold			Gaps in scaffold		Assembly method	
			Number	Average length (Kbp)	N50 (Kbp)	Number	Average length (Kbp)	N50 (Mbp)	Total length (Mbp)	Fraction (%)	Assembler	Sequeing platform
Chicken (*Yoensan Ogye*)	Ogye_1	1.00	***8,448***	***118.6***	504.8	1,906	517.8	***21.2***	8.5	0.85	Our pipeline	I/P
Chichen	Gallus_gallus-4.0	1.05	27,143	38.1	279.0	***915***	***1,128.8***	12.9	14.1	1.34	Celara	S/4
	Gallus_gallus-5.0	1.23	24,698	49.3	***2,894.8***	23,870	51.0	6.4	11.8	0.96	MHAP/PBcR	I/S/4/P
Hoodedcrow	Hooded_Crow_genome	1.05	28,920	35.4	94.4	1,299	787.1	16.4	27.5	2.62	ALLPATHS-LG	I
Golden eagle	Aquila_chrysaetos-1.0.2	1.19	17,032	69.3	172.3	1,142	1,033.3	9.2	12.7	1.07	ALLPATHS-LG	I
Medium ground-finch	GeoFor_1.0	1.07	95,828	10.9	30.5	27,239	38.2	5.3	24.0	2.25	ALLPATHS-LG	I
Blue-crowned manakin	Lepidothrix_coronata-1.0	1.08	23,501	45.0	141.8	4,612	229.2	5.0	22.4	2.07	ALLPATHS-LG	I
White-throated sparrow	Zonotrichia_albicollis-1.0.1	1.05	37,661	26.7	112.7	6,018	167.2	4.9	46.3	4.40	ALLPATHS-LG	I
Silver-eye	ASM128173v1	1.04	65,519	15.3	32.2	2,933	341.5	3.6	34.3	3.31	ALLPATHS-LG	I
Tibetan ground-tit	PseHum1.0	1.04	27,052	38.1	165.3	5,406	190.5	16.3	13.0	1.24	SOAPdenovo	I
Bald eagle	Haliaeetus_leucocephalus-4.0	1.18	31,786	36.5	105.5	1,023	1,133.2	9.1	19.2	1.63	SOAPdenovo	I
American crow	ASM69197v1	1.09	89,646	11.7	29.1	10,547	99.7	7.0	39.5	3.62	SOAPdenovo	I
Saker falcon	F_cherrug_v1.0	1.17	75,898	15.2	31.3	5,863	196.3	4.2	23.8	2.03	SOAPdenovo	I
Peregrine falcon	F_peregrinus_v1.0	1.17	83,081	13.9	28.6	7,021	164.3	3.9	18.6	1.58	SOAPdenovo	I
Rock pigeon	Cliv_1.0	1.11	100,099	10.9	26.6	14,923	72.8	3.1	21.1	1.90	SOAPdenovo	I
Little egret	ASM68718v1	1.21	100,662	11.5	29.0	11,791	98.2	3.1	48.7	4.04	SOAPdenovo	I
Hoatzin	ASM69207v1	1.20	109,627	10.4	28.2	10,256	111.4	2.9	61.5	5.11	SOAPdenovo	I
Golden-collared manakin	ASM171598v1	1.21	29,998	38.9	185.6	15,315	76.3	16.6	45.5	3.75	MaSuRCA	I/P
Turkey	Turkey_5.0	1.13	296,315	3.7	26.7	233,806	4.7	3.8	35.3	3.13	MaSuRCA	IS/4
Parrot	Melopsittacus_undulatus_6.3	1.12	70,891	15.3	55.6	25,212	43.1	10.6	30.8	2.75	Celara	I/4
Zebra finch	Taeniopygia_guttata-3.2.4	1.23	124,806	9.8	38.6	37,422	32.7	8.2	9.3	***0.75***	PCAP	S

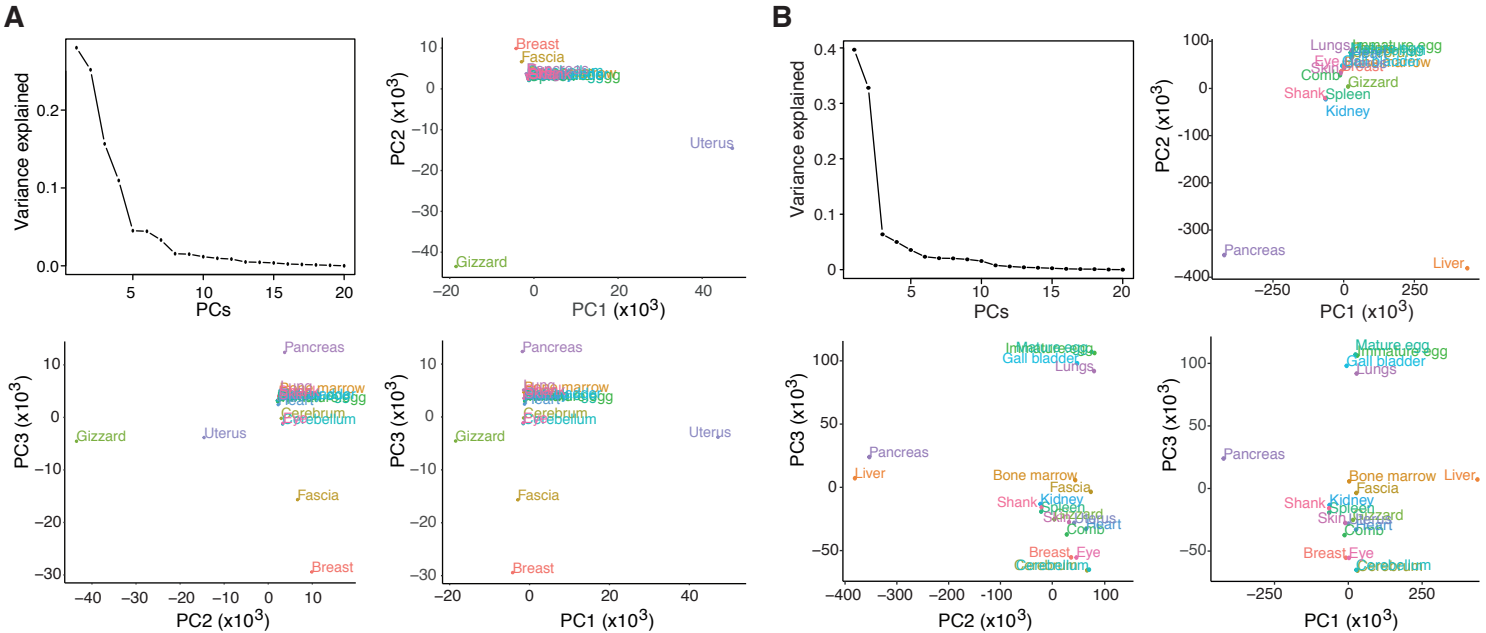

A.
galGal4 galGal5
59 185 53 Deletion
69 180 92 Insertion
106 158 224 Duplication
18 23 35 Inversion
3 1 25 INTER-Trans.
35 4 8 intra-Trans.
Tyrp1
Edn3
galgal4
galgal5
Insertion
Deletion
Duplication
Inversion
INTER-Trans.
intra-Trans.
B.
Scenario 1
Dupl_1
Dupl_2
Single-copy
Scenario 2
Scenario 3
C.
Depth
Dupl_1
127.4 Kb
Single-copy
412.6 Kb
Dupl_2
170.7 Kb
Gap_2
63.3 Kb
Gap_1
164.5 Kb
Ogye chr20
GalGal4 chr20
D.
Ogye chrZ
GalGal4 chrZ

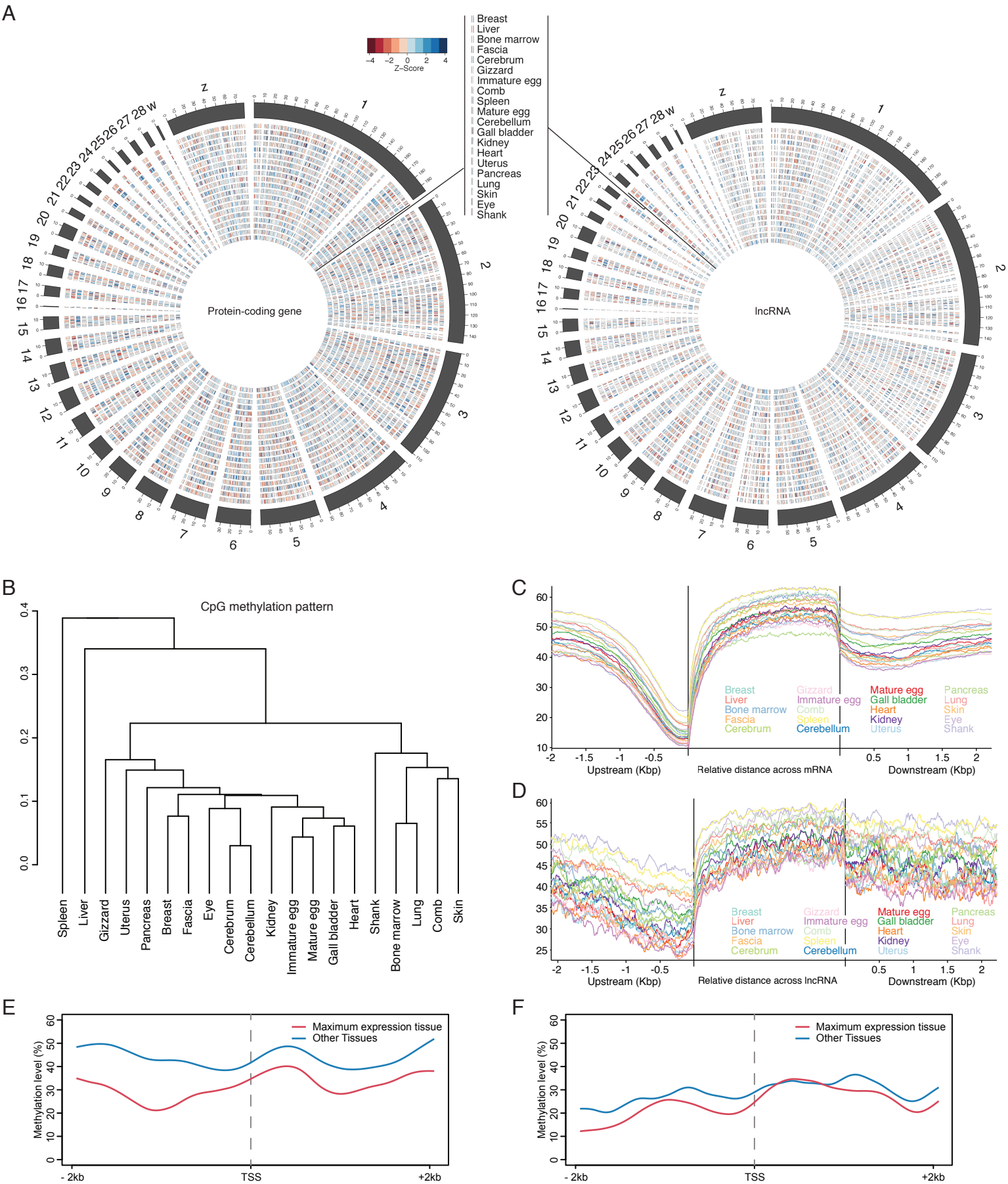

60%
50%
40%
30%
20%
10%
0%
12%
10%
8%
6%
4%
2%
0%
Low_complexity
Simple_repeat
Satellite
DNA
LTR
SINE
LINE
Human (GRCh38)
Human (GRCh37.p13)
House mouse (GRCm38.p5)
Pig (Sscrofa11.1)
Western painted turtle (Chrysemys_picta_bellii-3.0.3)
Tropical clawed frog (Xenopus_tropicalis_v9.1)
Zebra_finch (Taeniopygia_guttata-3.2.4)
Turkey (turkey_5.0)
Chicken (Gallus_gallus-5.0)
Chicken (Gallus_gallus-4.0)
Chicken (ogye1)
Mammalian
Amphibian
Reptile
Avian

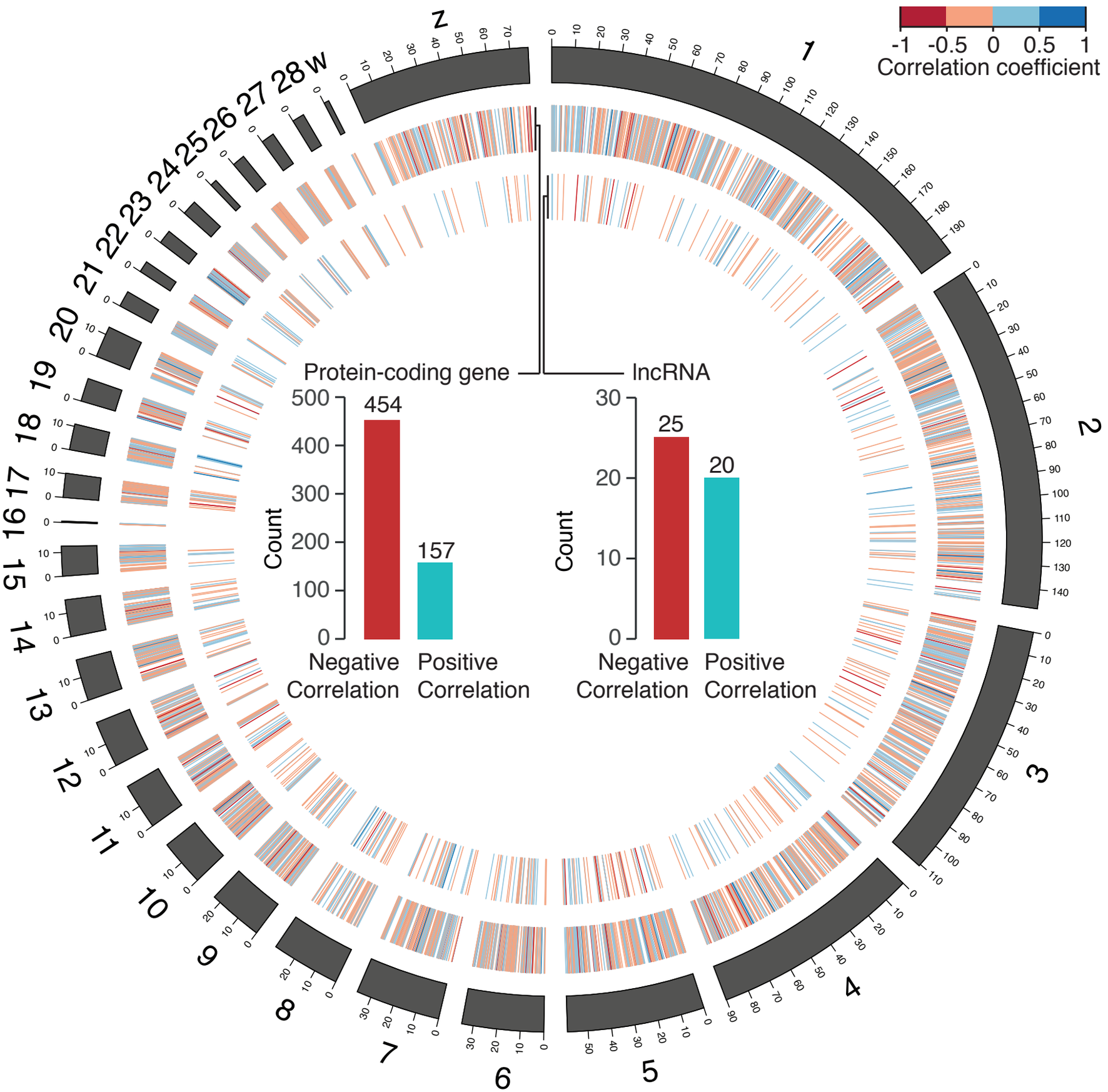

A.
TE coverage (% in Mbp) 0 40
Protein-coding density (genes/Mbp) 0 78
lncRNA density (genes/Mbp) 0 42
SNV/INDEL density (freq. in 100Kbp) 0 1000
GCratio (% in 100Kbp) 30 70
B.
RNA-seq
from 20 tissues
Transcript assembly
STAR and StringTie on Ogye_1
Intact structure
no
yes
score >0.99
Protein-coding probability
CPAT
score <0.8
score 0.8~0.99
CPC >0
no
Overlapped with galGal4 protein-coding genes
yes
yes
Assign to protein-coding
(39,194 trans. in 14,487 loci)
Update missing known protein-coding genes
GMAP with galGal4 transcriptome
Protein-coding gene set of Ogye_1
(15,766 genes)
C.
Ogye_1 protein-coding gene set (15,766 loci)
Novel loci
Known loci
131
Assembly
946
13,540
Mapping
1,279
FPKM <0.1
33
galGal4 protein-coding gene set
Mapped to galGal4
Mapped to galGal5
1
918
27


A.
RNA-seq
from 20 tissues
Transcript assembly
STAR/Cufflinks/Cuffmerge on Ogye_1
Known protein-coding genes
Other non-coding RNAs
Short (<200nt) transcripts
Filtration
Overlapped with known lncRNA
no
yes
CPC ≤ -1
yes
Updating exon junction and remove single exon transcripts
Novel lncRNA
5,610 loci (6,263 trans.)
Known lncRNA
1,290 loci (3,266 trans.)
B.
Ogye_1
NONCODE
5,610
1,290
8,204
Fragments of protein-coding genes
3,873
Unassembled
FPKM <0.1
3,407
276
Unmapped to Ogye_1
648
C.
lncRNA
Protein-coding gene
Density
10-3
7.5x10-4
5.0x10-4
2.5x10-4
0
0
2.9
3.4
10
20
30
Transcript length (Kb)
D.
lncRNA
Protein-coding gene
count
4000
3000
2000
1000
0
0
3
8
10
20
30
40
50
Exon number


A
Z-Score
−4 −2 0 2 4
Breast
Liver
Bone marrow
Fascia
Cerebrum
Gizzard
Immature egg
Comb
Spleen
Mature egg
Cerebellum
Gall bladder
Kidney
Heart
Uterus
Pancreas
Lung
Skin
Eye
Shank
Protein-coding gene
lncRNA
B
Z-Score
−4 −2 0 2 4
Krt9
lnc-lama2-1
Protein-coding genes
lncRNAs
Fascia
Comb
Skin
Shank
Breast
Liver
Bone marrow
Cerebrum
Gizzard
Immature egg
Spleen
Mature egg
Cerebellum
Gall bladder
Kidney
Heart
Uterus
Pancreas
Lung
Eye
C
Krt9
FPKM
D
lnc-lama2-1
FPKM


A
B
Variance explained
PCs
PC1 (x10^3)
PC2 (x10^3)
PC3 (x10^3)
Breast
Fascia
Uterus
Gizzard
Pancreas
Cerebrum
Cerebellum
Liver
Lungs
Immature egg
Gall bladder
Bone marrow
Shank
Spleen
Kidney
Skin
Comb
Breast
Eye
Protein-coding genes
lncRNAs


A
Z-Score
Breast
Liver
Bone marrow
Fascia
Cerebrum
Gizzard
Immature egg
Comb
Spleen
Mature egg
Cerebellum
Gall bladder
Kidney
Heart
Uterus
Pancreas
Lung
Skin
Eye
Shank
Protein-coding gene
lncRNA
B
CpG methylation pattern
Spleen
Liver
Gizzard
Uterus
Pancreas
Breast
Fascia
Eye
Cerebrum
Cerebellum
Kidney
Immature egg
Mature egg
Gall bladder
Heart
Shank
Bone marrow
Lung
Comb
Skin
C
Upstream (Kbp)
Relative distance across mRNA
Downstream (Kbp)
D
Upstream (Kbp)
Relative distance across lncRNA
Downstream (Kbp)
E
Methylation level (%)
Maximum expression tissue
Other Tissues
- 2kb
TSS
+2kb
F
Methylation level (%)
Maximum expression tissue
Other Tissues
- 2kb
TSS
+2kb


Correlation coefficient
-1 -0.5 0 0.5 1
Protein-coding gene
lncRNA
Count
454
157
25
20
Negative Correlation
Positive Correlation
Negative Correlation
Positive Correlation