DATA NOTE

Whole genome hybrid assembly and protein-coding gene annotation of the entirely black native Korean chicken breed *Yeonsan Ogye*

Jang-il Sohn[1,2*], Kyoungwoo Nam[1,*], Jun-Mo Kim[3,*], Dajeong Lim[3], Kyung-Tai Lee[3], Yoon Jung Do[3], Chang Yeon Cho[4], Namshin Kim[5], Jin-Wu Nam[1,2,‡] and Han-Ha Chai[3,§]

[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 211,827 retrotransposons and 27,613 DNA transposons. The completeness (97.6%) of the *YO* genome was evaluated with 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) 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. By comparing to the *gallus gallus red junglefowl* chicken reference genome and re-sequencing data from additional *YO* chickens, 551 large structural variants, 895,988 single nucleotide polymorphisms, and 82,392 small insertions/deletions were detected across the *YO* genome.

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 (called *Ogolgye* in Korea), 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), and structural variations (SVs). Here, we provide the first version of the *YO* genome map, which includes annotations of large SVs, SNPs, INDELs, and repeat sequences, as well as transcriptome maps across twenty different tissues of Ogye.

Results

Sample collection and whole genome sequencing

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). Ogye management, treatment, and sample collection took place at the National Institute of Animal Science. 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 kb), and fosmid libraries (40 kb) 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 6kb (Table 1).

Whole transcriptome sequencing data from twenty tissues

To construct Ogye transcriptome maps, using Illumina's protocols we have sequenced 843 million single-end reads and 638 paired-end reads of 101bp from twenty different tissues: breast, liver, bone marrow, fascia, cerebrum, gizzard, mature and immature eggs, comb, spleen, cerebellum, gall bladder, kidney, heart, uterus, pancreas, lung, skin, eye, and shank. Detailed descriptions of the transcriptome data and related methods are included in our study of non-coding transcriptome maps (see ref. [12]).

Hybrid whole genome assembly

The *YO* genome was assembled using our hybrid genome assembly pipeline, employing the following four steps: preprocessing, hybrid *de novo* assembly, super-scaffolding, and polishing (Figures 1b and S1). During the preprocessing step, the errors in the Illumina short-reads were corrected by KmerFreq and Corrector [13] using sequencing quality scores. In turn, using the corrected short-reads, the sequencing errors in the PacBio long reads (SRR6189090) were corrected by LoRDEC [14].

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 [15] with the default option, producing contigs and scaffolds. The resulting contigs and scaffolds showed 53.6 Kb and 10.7 Mb of N50 (Table S1), respectively. Next, the scaffolds were additionally connected using SSPACE-LongRead [16] and OPERA [17] with corrected PacBio long reads and FOSMID reads. The gaps within and between scaffolds were re-examined with GapCloser [13] with error-corrected short-reads. All resulting scaffolds were aligned to the galGal4 genome (GenBank assembly accession: GCA_000002315.2) by LASTZ [18] to detect putative mis-assemblies, verified by paired-end and mate-pair reads mapped to the scaffolds using BWA-MEM [19]. 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) [20] and in-house programs. Breaking scaffolds at the break points resulted in pseudo-contig N50 of 108.6 Kb and scaffold N50 of 18.7 Mb (Table 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 [21], and SSPACE-LongRead to enhance the quality of assembly using error-corrected PacBio long-reads to reduce the topological complexity of the assembly graphs [22]. 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 Table S1). In the last stage, nucleotide errors or ambiguities were corrected by the GATK pipeline [23] with paired-end reads, and in turn, any vector contamination was removed using VecScreen with UniVec database [24]. The final assembly results showed that the gap percentage and (pseudo-)contig N50 were significantly improved, from 1.87% and 53.6 Kb 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 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 [25]. The results indicated that the *YO* genome contains more complete single-copy BUSCO genes, suggesting that the *YO* genome is slightly more complete than the others (Table 2).

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 [18] and in-house programs/scripts, 551 common large (≥1 Kb) structural variations (SVs) evident in both assembly versions were detected by at least two different SV prediction programs (Delly, Lumpy, FermiKit, and novoBreak) [26-29] (Figure 2A; 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 *YO* genome, we compared it to the same locus in the galGal4 genome. Aligning paired-end reads of the *YO* 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). The intervening region between the two duplicated regions was estimated to be 412.6 Kb in length. 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 Kb and 63.3 Kb, respectively.

Additionally, a large inversion was detected near a locus including the *tyrp1* gene (Figure 2D), which is known to be related to melanogenesis [30, 31]. 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, suggesting that the inversion including *tyrp1* is not specifically related to skin hyperpigmentation.

Annotations

Repeats

Repeat elements in the Ogye and other genomes were predicted by a reference-guided approach, RepeatMasker [32], which utilizes Repbase libraries [33]. In the Ogye_1 genome, 211,827 retro-transposable elements (8.01%), including LINEs (6.58%), SINEs (0.07%) and LTR elements (1.36%), 27,613 DNA transposons (1.02%), 8,131 simple repeats (0.97%), and 45,264 low-complexity repeats (0.21%) were annotated (Figure 3 and Supplementary Table S2). It turns out that the composition of repeats in the Ogye_1 genome is similar to that in other avian genomes (Figure 3), although the total percentage is slightly greater in Ogye_1 than in other avian genomes, with the exception of the galGal4 genome. Compared with other avian genomes, the Ogye_1 genome is more similar to galGal4 and 5 in terms of repeat composition except for the fractions of simple repeats (0.14% for Ogye_1 and 1.47% for galGal5) and satellite DNA repeats (0.01% for Ogye and 6% for galGal5).

SNPs/INDELs

To annotate SNPs and INDELs in the *YO* genome, resequencing data from 20 *YO* individuals were produced and mapped to the Ogye_1 genome using BWA-MEM, and a series of post-processes including deduplication were performed by Picard modules [34]. Using Genome Analysis Toolkit (GATK) modules, including HaplotypeCaller, combineGVCF, GenotypeGVCFs and VariantFiltration (with options "`QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0`") [23], 895,988 SNPs and 82,392 insertions/deletions (INDELs) were annotated across the genome (Figure 4A). 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, 15,766 protein-coding genes were annotated in the *YO* genome (see Figure 4B and C), including 946 novel genes and 14,820 known genes. 164 protein-coding genes annotated in galGal4 were missing from the Ogye_1 assembly. To sensitively annotate protein-coding genes, all paired-end RNA-seq data were mapped on the Ogye_1 genome by STAR [35] for each tissue and the mapping results were then assembled into potential transcripts using StringTie [36]. 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 [37]. 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 [38] and CPC [39]. 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 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 Ogye (≥ 0.1 FPKM) using all Ogye RNA-seq data. However, expression of the remaining 33 genes was not confirmed, suggesting that they are not expressed in Ogye (< 0.1 FPKM) or have been lost from the Ogye genome. The missing genes are listed in Table S3. In contrast, the 946 newly annotated Ogye genes appeared to be mapped to the galGal4 or galGal5 genomes (Figure 4C).

lncRNA

In total, 6,900 Ogye lncRNA genes, including 5,610 novel loci and 1,290 known loci, were identified in the Ogye_1 genome from RNA-seq data using an lncRNA annotation pipeline. 6,170 were intergenic lncRNA genes (89.40 %) and 730 were anti-sense ncRNA genes (10.57 %) (see ref. [12] for details about lncRNA of *Yeonsan Ogye*).

Discussion

In this work, the first draft genome from a Korean native chicken breed, *Yeonsan Ogye*, 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 genome completeness evaluated by BUSCO was better than the existing chicken reference genomes including galGal4 and 5, the Ogye_1 genome seems to lack simple and long repeats compared to 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, because their sequence depth is shallow (11.5X), they were only used for scaffolding and gap filling, so that some simple and satellite repeats would be missed during assembly. 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 Illumina short-reads and PacBio long-reads.

Annotated genomic variations and comparative analysis of coding and non-coding genes will provide a resource for understanding genomic differences and evolution of Ogye 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).

Additional files

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 genome to galGal4/5 drawn by MUMmer.

Table S1. Structural variations in the Ogye_1 genome

Table S2. Repeats in the Ogye genome

Table S3. 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. 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%

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 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 1. A. Structural variation (SV) map of the *YO* 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 2. Composition of repeat elements in different assemblies of avian, reptile, and mammalian genomes.

Figure 3. A. Gene annotation maps of the *YO* genome with repeats, SNVs (including SNPs and indels), and GC ratio landscapes are shown in a Circos plot; **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 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 Ogye 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.

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. Hong H, Chai H-H, Nam K, Lim D, Lee K-T, Do YJ, et al. Non-coding Transcriptome Maps Across Twenty Tissues of the Korean Black Chicken, Yeonsan Ogye. (In preparation).

13. 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.

14. Salmela L and Rivals E. LoRDEC: accurate and efficient long read error correction. Bioinformatics. 2014;30 24:3506-14. doi:10.1093/bioinformatics/btu538.

15. 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.

16. 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.

17. 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.

18. Harris R. *Improved pairwise alignment of genomic DNA.* PhD Thesis, 2007.

19. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997. 2013.

20. 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.

21. 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.

22. Sohn JI and Nam JW. The present and future of de novo whole-genome assembly. Brief Bioinform. 2016:bbw096. doi:10.1093/bib/bbw096.

23. 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.

24. VecScreen https://anonsvn.ncbi.nlm.nih.gov/repos/v1/trunk/c++/ and UniVec database https://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/.

25. 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.

26. 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.

27. 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.

28. Li H. FermiKit: assembly-based variant calling for Illumina resequencing data. Bioinformatics. 2015;31 22:3694-6. doi:10.1093/bioinformatics/btv440.

29. 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.

30. 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.

31. 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.

32. Tempel S. Using and understanding RepeatMasker. Mobile Genetic Elements: Protocols and Genomic Applications. 2012:29-51.

33. 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.

34. Picard Tools. http://broadinstitute.github.io/picard/.

35. 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.

36. 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.

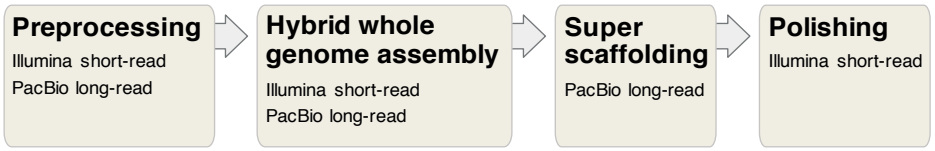
37. TransDecoder. https://github.com/TransDecoder/TransDecoder/.

38. 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.

39. 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.

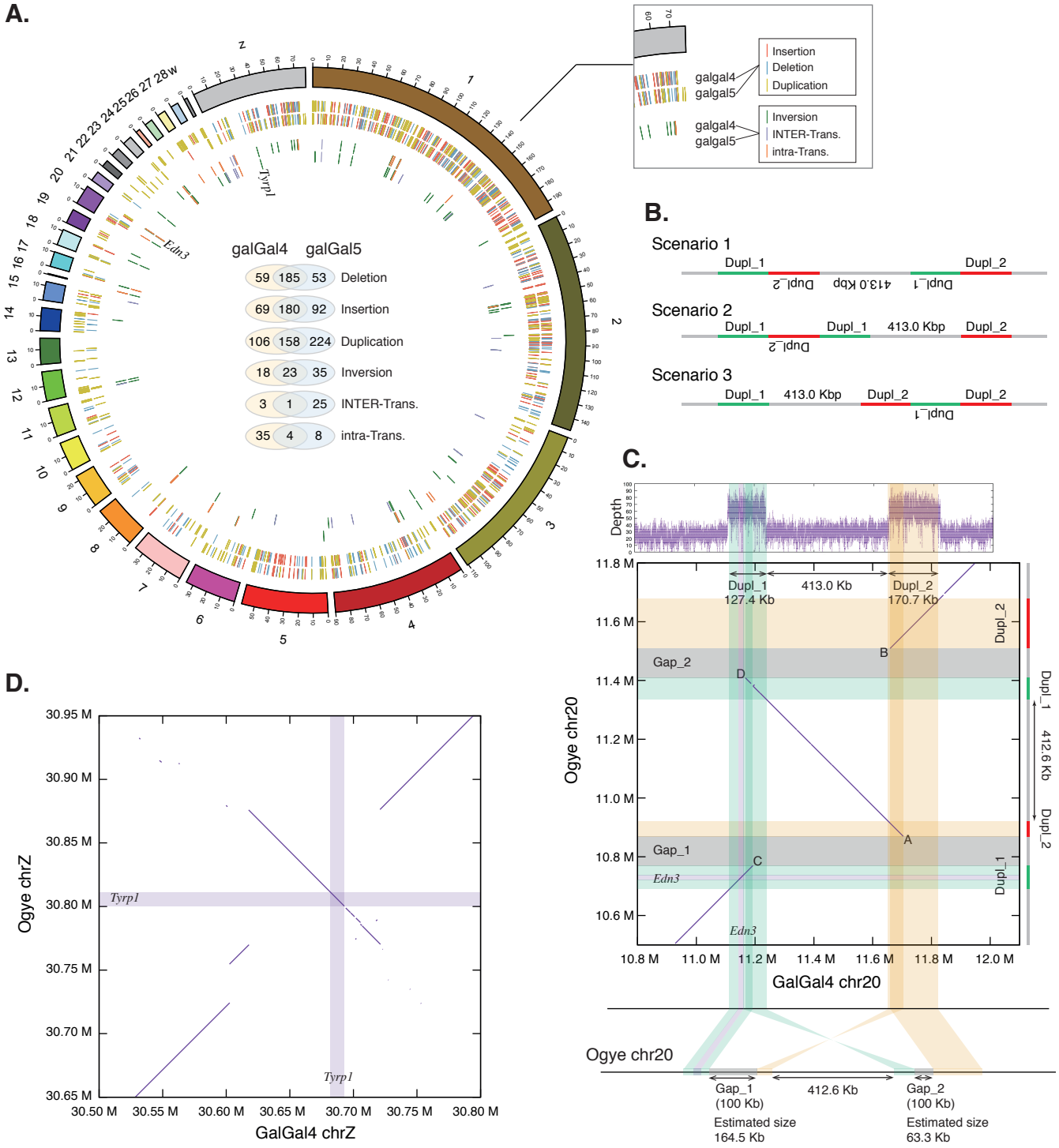
A.

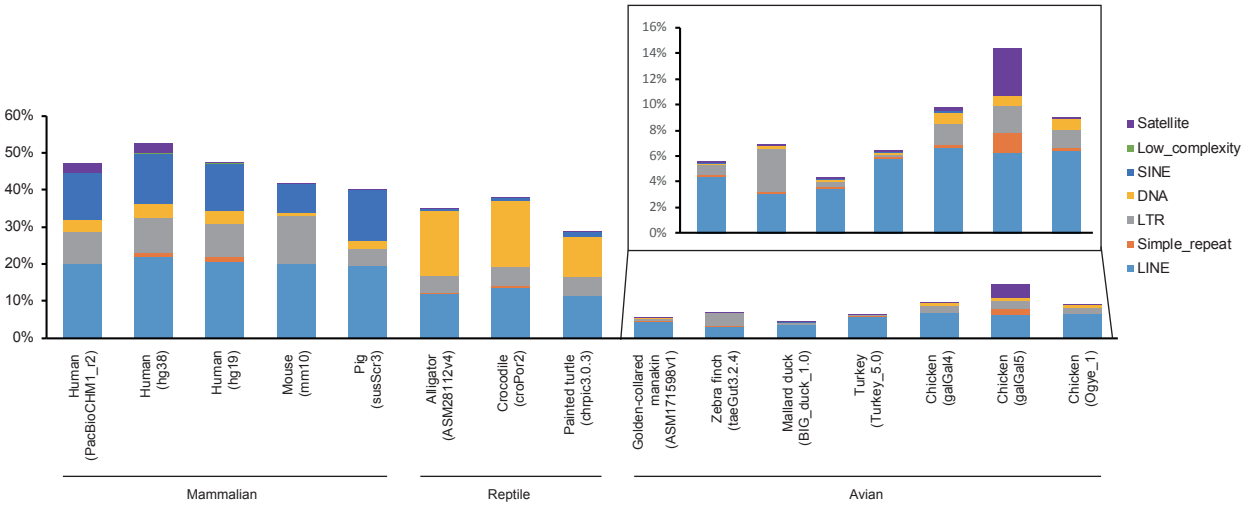
B.

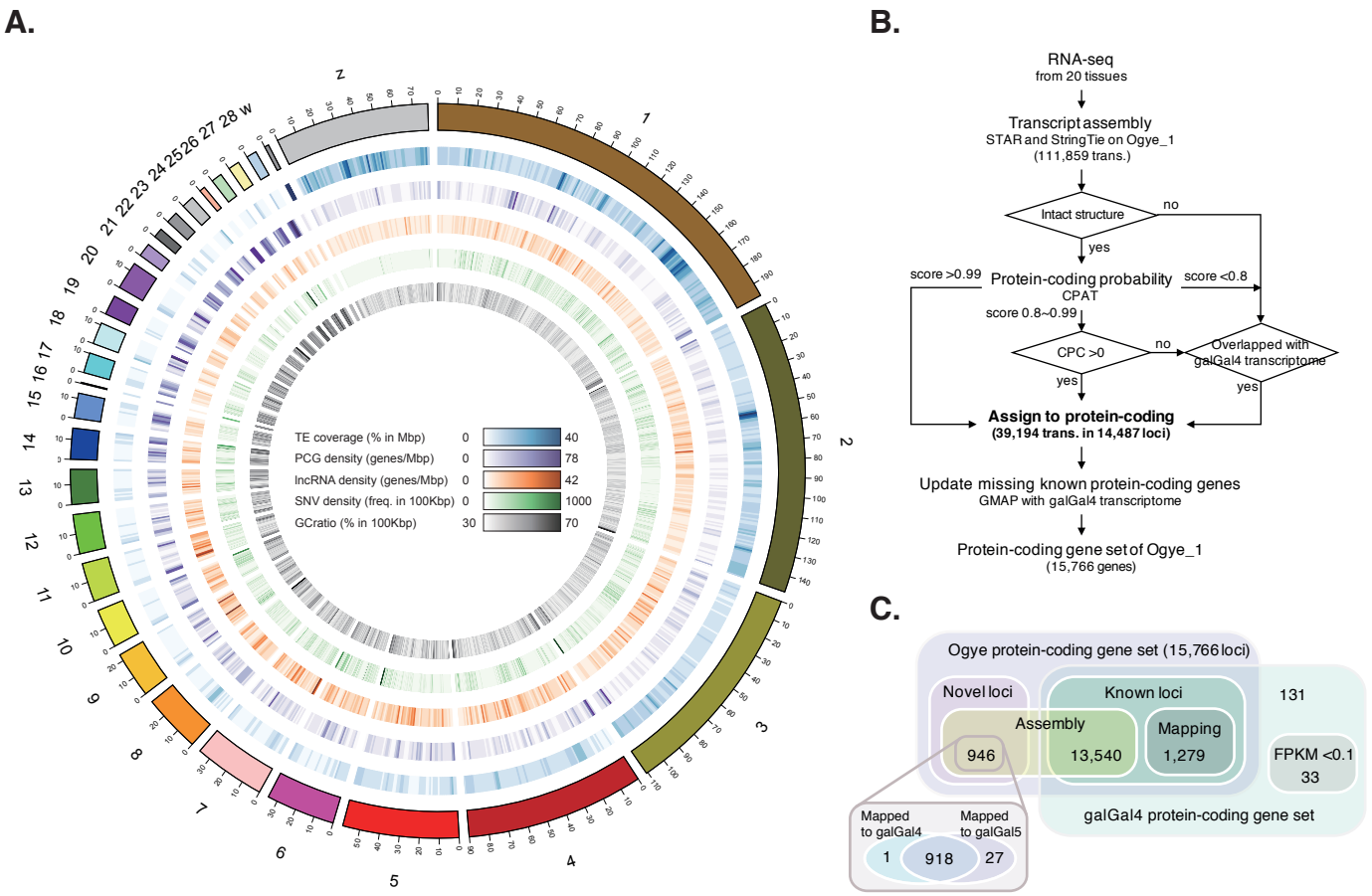


C.

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
Scenario 2
Dupl_1
413.0 Kbp
Dupl_2
Scenario 3
Dupl_1
413.0 Kbp
Dupl_2
C.
Depth
Dupl_1
127.4 Kb
413.0 Kb
Dupl_2
170.7 Kb
Gap_2
Gap_1
Edn3
Ogye chr20
GalGal4 chr20
412.6 Kb
Gap_1
(100 Kb)
Estimated size
164.5 Kb
Gap_2
(100 Kb)
Estimated size
63.3 Kb
D.
Ogye chrZ
GalGal4 chrZ
Tyrp1


60%
50%
40%
30%
20%
10%
0%
16%
14%
12%
10%
8%
6%
4%
2%
0%
Satellite
Low_complexity
SINE
DNA
LTR
Simple_repeat
LINE
Human (PacBioCHM1_r2)
Human (hg38)
Human (hg19)
Mouse (mm10)
Pig (susScr3)
Alligator (ASM28112v4)
Crocodile (croPor2)
Painted turtle (chrpic3.0.3)
Golden-collared manakin (ASM171598v1)
Zebra finch (taeGut3.2.4)
Mallard duck (BIG_duck_1.0)
Turkey (Turkey_5.0)
Chicken (galGal4)
Chicken (galGal5)
Chicken (Ogye_1)
Mammalian
Reptile
Avian


A.
TE coverage (% in Mbp) 0 40
PCG density (genes/Mbp) 0 78
lncRNA density (genes/Mbp) 0 42
SNV 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
(111,859 trans.)
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 transcriptome
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 protein-coding gene set (15,766 loci)
Novel loci
Known loci
131
Assembly
Mapping
946
13,540
1,279
FPKM <0.1
33
galGal4 protein-coding gene set
Mapped
to galGal4
Mapped
to galGal5
1
918
27