DE NOVO ASSEMBLIES OF GENOMES OF FOUR INDIGENOUS CHICKENS AND A WHITE EARED PHEASANT REVEAL GENETIC BASIS OF CHICKEN DOMESTICATION AND ALTITUDE ADAPTATION

by

Siwen Wu

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics & Computational Biology

Charlotte

2023

Approved by:

_____
Dr. Zhengchang Su

_____
Dr. Jun-tao Guo

_____
Dr. Way Sung

_____
Dr. Bao-Hua Song

ABSTRACT

SIWEN WU. *De Novo* Assemblies of Genomes of Four Indigenous Chickens and a White Eared Pheasant Reveal Genetic Basis of Chicken Domestication and Altitude Adaptation (Under the direction of DR. ZHENGCHANG SU)

High-quality assembly and annotation of the genome of a species are critical in understanding the genetic basis of almost all aspects of the biology of the species. Although many genome assembly pipelines have been developed, they are either difficult to use or their assemblies are too fragmental. Moreover, although gene annotation pipelines have been developed at large genome centers, they are either too complicated for individual labs to use or not available to public. In this dissertation project, we have proposed a user-friendly pipeline that can assemble genome at chromosome-level with high-quality using PacBio/Nanopore long reads, Illumina paired-end short reads and Hi-C paired-end short reads. We have also developed an accompanying gene annotation pipeline using a combination of homology-based and RNA-based approached. The pipeline achieves high accuracy in protein-coding gene and pseudogene annotations.

Moreover, although multiple chicken genomes have been assembled, high-quality indigenous chicken genomes are still lacking, hampering the understanding of chicken domestication and evolution. Using the pipelines, we assembled and annotated the genomes of four indigenous chickens with distinct morphological traits at the chromosome-level. Our results challenge two earlier conclusions regarding chicken domestication and evolution. First, we found a total of 1,420 new protein-coding genes in the four chickens and recovered 51 of the 274 “missing” genes in birds in general and 36 of the 174 “missing” genes in chickens in particular. Most of these new genes are also

found in previously assembled GRCg6a and GRCg7b/w chicken genomes, and might play house-keeping roles. Counting these new genes, chicken genomes encode more genes than originally thought. Second, we identified a total of 2,015 non-processed pseudogenes in the seven genomes. Most pseudogenization mutations are fixed in their respective populations and preferentially occur at the two ends of genes. Purifying selection is relaxed on the pseudogenes, suggesting that they might lose their gene functions. Pseudogenization mutations segregate in the chickens as their phylogenetic tree does, which is based on more than 6,000 essential protein-coding genes. Thus, in contrast to the previous conclusion, loss-of-function mutations play a critical role in chicken domestication and evolution. Moreover, these assembled genomes are valuable resources for studying chicken domestication and evolution.

Furthermore, although many studies related to artificial selection signatures of commercial and indigenous chickens have been carried out, quite a small number of genes have been found to be under selection. To fill these gaps, we re-sequenced 85 individuals of five indigenous chicken breeds with distinct traits from Yunnan, a southwest province of China. By analyzing these indigenous chickens together with 116 individuals of commercial chickens (broilers and layers) and 35 individuals of red jungle fowl (RJF), we find a substantially large number of selective sweeps and affected genes for each chicken breed using a rigorous statistic model than previously reported. We confirm most of previously identified selective sweeps and affected genes. Meanwhile, the vast majority (~98.3%) of our identified selective sweeps overlap known chicken quantitative trait loci. Thus, our predictions are highly reliable. For each breed, we also identify candidate genes and selective sweeps that might be related to the unique traits of

the chickens. Most of these genes do not contain nonsense mutations, we therefore quantified the expression levels of eight genes in relevant tissues using RT-qPCR and found that most of them showed differential expression compared to their counterparts.

Finally, eared pheasant species are closely related but inhabit at highly varying altitudes from northeast to southwest China. To understand genetic bases of closely related species to adapt to different altitudes, we sequenced a population of 10 white eared pheasants (WT) (*Crossoptilon crossoptilon*) inhibiting in 3,000~4,300 m altitude niches in Yunnan, China, and assembled the genome of an individual at chromosome-level with a contig N50 of 19.63 Mb, a scaffold N50 of 29.59 Mb, and a total length of 1.02 Gb. This assembly with only few gaps is of higher quality than a previous one of a brown eared pheasant (BR) (*C. mantchuricum*) individual living at 20~1000 m altitude in northeast China. Interestingly, the WT genome encodes more protein genes than the BR genome (16,315 VS. 15,003), while the later contains more pseudogenes than the former (1,519 VS. 1,976). The two genomes shared 14,178 genes and 1,040 pseudogenes with 2,137 and 825 unique genes, and 479 and 936 unique pseudogenes, respectively. The unique genes and unique pseudogenes of both species are mainly involved in biological pathways of cardiovascular, energy metabolic, neuronal and immune functions, which are known to be related to adaptation to high altitude. Moreover, we compared the selective sweeps in the genomes of WT, BR and an additional species blue eared pheasant (BL) (*C. auritum*) inhibiting at 1,500~3,000 m altitude in central west China, using re-sequencing data of 10 WT, 12 BL and 41 BR individuals, respectively. Interestingly, genes under selection in each species converge on the same pathways of the aforementioned four functional categories. These results suggest that these species

adapted to highly varying altitudes by loss-of-function mutation and fine-tuning of genes in these common pathways. Our assembled WT genome and re-sequencing data can be valuable resources for studying the biology, evolution and developing conservation strategies of these endangered species.

ACKNOWLEDGEMENTS

I would sincerely thank my advisor Dr. Zhengchang Su for his professional and patient guidance in my PhD project. He gave many constructional ideas and instructions on the project so this thesis can be realized successfully. I would thank my committee members, Dr. Juntao Guo, Dr. Way Sung and Dr. Baohua Song for their insightful comments and suggestions. I would like to thank the members in my lab for their discussions on my project. I would thank my parents for their support. Finally, I would thank the Graduate Assistant Support Plan (GASP) for their financial support and the Yunnan Agriculture University for their sharing of data and collaboration on the project. I would also like to thank all staff members in the Department of Bioinformatics and Genomics and the international student and scholar office (ISSO) for their supportive help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

RJF	Red jungle fowl
VGP	Vertebrate Genomes Project
RT-qPCR	Real-time quantitative PCR
QTL	Quantitative trait loci
SNP	Single nucleotide polymorphism
SRA	Sequence Read Archive
PCA	Principal component analysis
DSS	Discrete selective sweep
WT	White eared pheasant
BR	Brown eared pheasant
BL	Blue eared pheasant

CHAPTER 1 Introduction

Assembling and annotating a vertebrate genome involve multiple piece-meal computational steps and are often done by an ad hoc manner, although many tools have been developed for each of these steps. Consequently, there is no easily used efficient pipeline to assemble vertebrates' genome at chromosome-level with high-quality and annotate genes accurately. Thus, we developed a user-friendly pipeline that can be used to assemble high-quality genomes at chromosome-level for vertebrates and annotate the genomes accurately using a combination of homology-based and RNA-based method. Since the first assembly of the red jungle fowl (RJF) (*Gallus gallus*) genome in 2004 [1], more completed versions (GRCg2~6a) of the RJF genome have been released using then newly available sequencing technologies [2, 3]. However, none of them reached a high-quality chromosome-level assembly, due largely to the difficulty to assemble the micro-chromosomes and sub-telomere regions. More recently, semi-haploid assemblies of a hybrid individual with a broiler mother (GRCg7b) and a white Leghorn layer father (GRCg7w) were sequenced and assembled at the chromosome-level by the Vertebrate Genome Project (VGP) consortium [4, 5]. However, no indigenous chicken breeds have been sequenced and assembled at a high-quality chromosome-level. To fill this gap, using our genome assembly pipeline, we have assembled all the chromosomes of four indigenous chicken breeds with distinct morphological traits from Yunnan province in southwestern China, one of geological places where chickens were first domesticated [6, 7]. Evaluations using multiple metrics proposed by the VGP consortium [4] indicate that we have achieved the most continuous chicken genome assemblies so far. Careful annotation of these high-quality assemblies allows us to uncover a large number of new

genes and pseudogenes in the four indigenous chicken genomes. These findings as well as the assembled genome well position us to understand the major genetic basis of chicken domestication and evolution. However, to fully reveal artificial selection footprints in the chicken domestication process, we also identified all types of genetic variations such as SNPs in larger populations of the RJF, the four indigenous chickens as well as commercial broilers and layers, and identified the selective sweeps on their genomes [8]. By analyzing the selective sweep of each chicken breed, we are able to find genes and genomic regions related to the specific traits of each chicken breed. In addition, to reveal genetic basis of adaptation to different altitude of eared pheasant, we sequenced and assembled the genome of a white eared pheasant inhabiting in high altitude and compared it with a brown eared pheasant individual. By the population genetic analysis of white, blue and brown eared pheasant, we are able to identify genes/pseudogenes and GO pathways related to the adaptation to different altitude.

In chapter 1, we introduce the whole aims and structure of the thesis.

In chapter 2, we develop a user-friendly genome assembly and annotation pipelines for vertebrates. A high-quality genome assembly and annotation are important to study vertebrates. However, many existing genome assembly tools can only assemble genomes at contig level, and the many existing gene annotation tools cannot annotate genes accurately. To fill the gaps, we have developed a pipeline that can be used to assemble high-quality genomes using a combination of short (Illumina), long (PacBio/Nanopore) and Hi-C reads and annotate the protein-coding genes accurately using a combination of homology-based and RNA-based methods.

In chapter 3, we reveal genetic basis of chicken domestication and evolution by *de novo* assembly and annotation of four indigenous chicken genomes. To understand the genetic basis of chicken domestication and evolution, high-quality indigenous chicken genomes and accurate gene annotations are needed. We assemble the genomes of four indigenous chickens with distinct traits at chromosome-level and accurately annotate the genes in them. By comparing the genomes and genes of the four chickens as well as the RJF (GRCg6a), the broiler (GRCg7b) and the layer (GRCg7w), we have found the major genetic basis in chicken domestication and evolution.

In chapter 4, we determine artificial selection signatures in chicken domestication process by population genetic analysis. Many complex or quantitative traits are not caused by protein-coding genes, but by variations such as SNPs in non-coding sequences. To determine artificial selection signatures in chicken domestication process, we call the SNPs in the population of the four indigenous chickens as well as RJF, broilers and layers. By analyzing the selective sweeps of these chicken breeds, we have found the artificial selection signature related to breed-specific traits of each chicken breed during the domestication process.

In chapter 5, we assemble the high-quality genome for a white eared pheasant individual. By annotating the genome of the white eared pheasant and comparing it with the genome of a brown eared pheasant individual, we have found the genetic region related to the different altitude adaptation of eared pheasants.

In chapter 6, we give the conclusions.

CHAPTER 2 A user-friendly genome assembly and annotation pipeline

2.1 Background

Complete and accurate genome assembly of an organism is the first step to fully understand the genetic bluebook of the organism. Subsequent annotation of all genes encoded or pseudogenized in the genome would further the understanding of all aspects of the organism including its evolution and physiology, etc. High-quality assembly and annotation of the genome would greatly facilitate the study of all aspects of the organism. The rapid development of sequencing technologies makes genome assembly more and more accurate and achievable in individual labs. Since 1977, Sanger sequencing [163] had been the most commonly used technology to sequence DNA before the advent of next-generation sequencing (NGS) technology, and a few genomes including the first human draft genome were sequenced by an automatic Sanger sequencing technology with ~1,000 bp reads length. However, due to its poor quality in the first 15~40 bases of the sequence and relatively high cost and low throughput, NGS [164] has replaced Sanger sequencing to become the working horse for genome sequencing since 2005. However, NGS, dominated by Illumina technology, is limited by its short reads length (150~300 bp), making it difficult to assemble repetitive regions. In 2014, the single molecule real time (SMRT) sequencing platform was developed by Pacific Biosciences (PacBio) [165], enabling the sequencing length up to 10 kbp [165, 166], allowing high quality *de novo* assembly of large vertebrate genomes. In addition, with the development of Oxford Nanopore Technology (ONT) [167, 168], sequencing length could reach to 2.3 Mbp, made *de novo* genome assembly more complete. However, both PacBio and ONT reads suffer high sequencing errors (10~20%), thus high quality NGS short read are often used

to correct these errors at different stages of assembly pipelines. Furthermore, with the introduction of Hi-C technology [169] that provides information about physical distances between genomic loci, genomes can be scaffolded into chromosomes.

Many contig assembly tools such as Wtdbg2 [15], Canu [170] and FALCON [171] have been developed using a combination of long and short sequencing reads. Moreover, tools such as SALSA [16, 17] have also been developed to scaffold the assembled contigs using Hi-C or optical mapping data. Numerous large vertebrate genomes have been assembled using sophisticated assembling pipelines that take advantages of short, long and Hi-C reads, mostly at large consortium such as Vertebrate Genome Consortium and genome. However, the genome assembly pipelines developed at these consortia and centers are usually not available to individual research groups, and also are difficult for most individual research groups to follow due to limited resources available.

At the same time, annotation of protein-coding genes in an assembled vertebrate genomes is a complex process, involving multiple tools for *de novo* gene prediction such as Augustus [172], GlimmerHMM [173], GeneID [174], Genscan [175], SNAP [176] and GeneMark-ES [177], homology-based method such as GeMoMa [178], and RNA-seq data based methods such as Braker2 [179]. Although annotation pipelines have been developed at genome centers such NCBI and ENSEMBL, they are not available to individual research groups and difficult to follow. To fill these gaps, we have developed a pipeline for assembling genomes of vertebrates with high-quality at chromosome-level using the PacBio/Nanopore long reads, Illumina short reads and Hi-C paired-end reads, and a pipeline for annotating protein-coding genes and at the same time identifying

pseudogenes using homology-based and RNA-seq data-based approaches. We have demonstrated the usefulness of the pipelines by assembling and annotating four indigenous chicken genomes (CHAPTER 3).

2.2 Methods and materials

2.2.1 Pipeline for genome assembly

2.2.1.1 Pipeline design

The length of PacBio/Nanopore long reads can reach up to 10 kbp, but their error rate can reach up to 10%. The length of Illumina short reads is usually only 150~300 bp, but the error rate is smaller than 0.1%. To assemble a high-quality genome, we need to use both long and short reads with appropriate methods. Besides, to assemble the genome at chromosome-level, Hi-C reads are also needed. In our genome assembly pipeline, the first step is to generate gapless contigs using long reads. To assemble the contigs into scaffolds with correct order, the next step is to bridge the contigs with gaps (Ns) into scaffolds using Hi-C reads. The third step is to fill the gaps introduced in the scaffolding step using long reads. Since the high error rate of long reads, the first three steps would introduce many sequence errors in the assembly. Thus, the fourth step and the fifth step are to polish the scaffolds using long reads and short reads, respectively. The final result is a highly continuous chromosome-level assembly with a few gaps. The flowchart of the genome assembly pipeline is shown in Figure 2-1.

2.2.1.2 Required software

- Wtdbg2 [15] software (https://github.com/ruanjue/wtdbg2)
- SALSA [16, 17] software (https://github.com/marbl/SALSA)
- PBJelly [18] software (https://github.com/esrice/PBJelly)

- Racon [180] software (https://github.com/isovic/racon)
- NextPolish [20] software (https://github.com/Nextomics/NextPolish)
- SAMtools [30] software (https://github.com/samtools/)
- Minimap2 [181] software (https://github.com/lh3/minimap2)
- BWA [29] software (https://github.com/lh3/bwa)
- BEDTools [182] software (https://github.com/arq5x/bedtools2)

2.2.1.3 Required raw data

Illumina paired-end short reads (150~300 bp), PacBio/Nanopore long reads and Hi-C paired-end reads with relatively high sequencing depths. We recommend 100X for short reads, at least 40X for long reads, and 100X for Hi-C reads.

2.2.1.4 Procedure

Step 1: Use Wtdbg2 [15] to generate contigs using long reads and polish the contigs using Illumina short reads.

Step 2: Use SALSA [16, 17] to bridge the contigs into scaffolds using Hi-C paired-end reads.

Step 3: Use PBjelly [18] to fill gaps introduced in scaffolds using long reads.

Step 4: Use Racon [180] to polish the scaffolds obtained in Step 3 using long reads.

Step 5: Use Nextpolish [20] to further polish the scaffolds obtained in Step 4 using short reads.

For the appropriate parameters of each step, we give the examples at: https://github.com/zhengchangsulab/A-genome-assebmly-and-annotation-pipeline

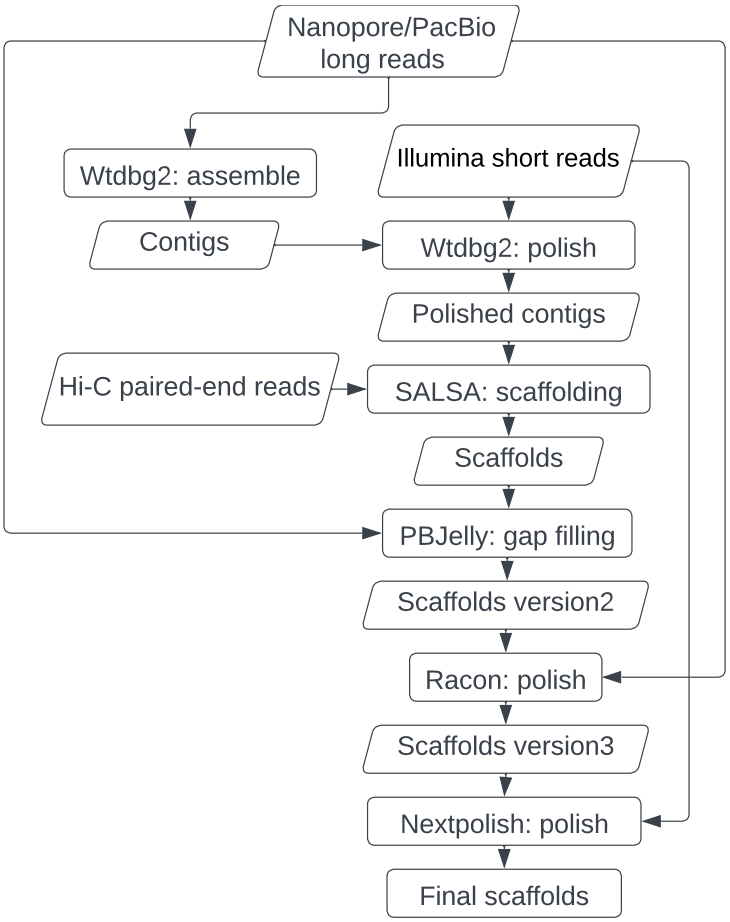

Figure 2-1 Flowchart of the genome assembly pipeline

2.2.2 Pipeline for gene annotation

2.2.2.1 Pipeline design

To annotate genes accurately in vertebrate genomes, we designed a pipeline with a combination of homology-based and RNA-based methods. For the homology-based annotation, the first step is to map the CDSs isoforms of the reference species which are closely related to the target species. For each reference gene whose CDSs could be mapped to the assembled genome, all the mapped CDSs are concatenated and checked to see whether the resulting sequence forms an ORF (the length is an integer time of three and contain no stop codon in the middle). If yes, the concatenated sequence is predicted

to be an intact gene. If the CDSs of a reference gene can be mapped to multiple loci in the assembled genome, the locus with the highest mapping identity is used. If the concatenated sequence did not form an ORF, i.e., it contains at least an earlier matured STOP codon, or an ORF shift mutation, we map the short reads to the sequence with no gaps allowed. If the sequence is completely covered by the short reads, the pseudogenization is fully supported by the short reads, and we predict the sequence to be a pseudogene; otherwise, the pseudogenization is not supported by the short reads, and we predicted the sequence to be a partially supported gene, because the pseudogenization might be artificially caused by errors of the long reads that could not be corrected by the short reads.

For the RNA-based annotation, the first step is to remove RNA-seq reads from rRNA genes by mapping all of the RNA-seq reads from various tissues of the target species to the rRNA database and filtering out the mapped reads. Then the unmapped reads are assembled into transcripts. Next, the assembled transcripts are mapped to the assembled genomes and those that at least partially overlap non-coding RNA genes (see below), protein-coding genes or pseudogenes predicted by the homology-based method are removed. For the remaining transcripts, if it contains an ORF with at least 300 bp, we predict it to be a protein-coding gene. If multiple ORFs were found in a transcript, the longest one is selected.

For the protein coding genes predicted by the homology-based and RNA-based methods, the CDS phase is corrected using GFF3toolkit [183]. The non-coding RNA genes are predicted using infernal (1.1.2) [38] with Rfam (v.14) database [39] as the

reference. The flowchart of the gene annotation pipeline is shown in Figure 2-2 (data preparations) and Figure 2-3 (our gene finding procedure).

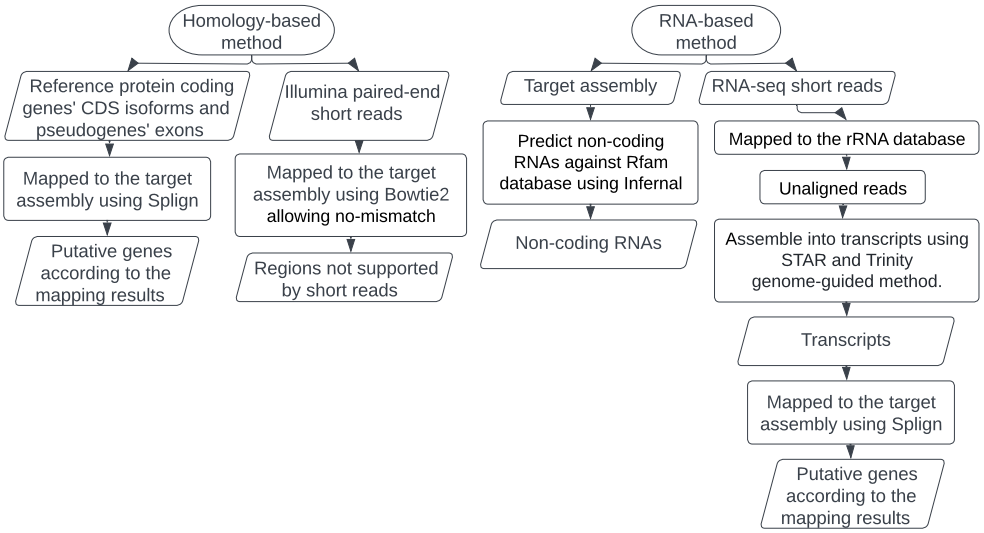

Figure 2-2 Flowchart of the data preparation of gene annotation pipeline

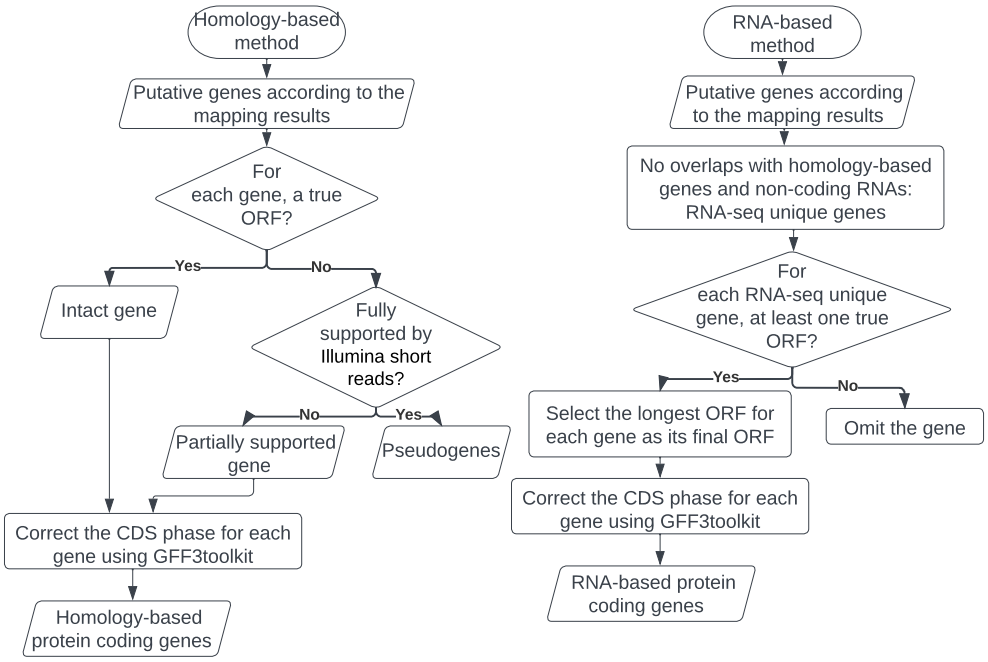

Figure 2-3 Flowchart of our gene finding procedure of gene annotation pipeline

2.2.2.2 Required software

- Splign [34] software (https://www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi)
- Bowtie2 [35, 184, 185] software (https://github.com/BenLangmead/bowtie2)

- SAMtools [30] software (https://github.com/samtools/)
- BEDTools [182] software (https://github.com/arq5x/bedtools2)
- Trinity [37] software (https://github.com/trinityrnaseq/trinityrnaseq)
- STAR [31] software (https://github.com/alexdobin/STAR)
- GFF3toolkit [183] software (https://github.com/NAL-i5K/GFF3toolkit)
- Infernal [38] software (https://github.com/EddyRivasLab/infernal)
- Our gene finding procedure (https://github.com/zhengchangsulab/A-genome-assebmly-and-annotation-pipeline)

2.2.2.3 Required raw data

Reference CDS isoforms from closely related species, RNA-seq short reads from as many as possible tissues of the target species and Illumina paired-end sequencing reads of the target species are required. Rfam database and rRNA database are required.

2.2.2.4 Procedure

2.2.2.4.1 Data preparations (Figure 2-2)

1) Homology-based method:

Step H1: Use Splign [34] to map the reference CDS isoforms to the target genome assembly.

Step H2: Use Bowtie2 [35, 184, 185] to map the Illumina paired-end sequencing reads to the target assembly allowing no-mismatch to get the region not supported by the short reads on the target assembly.

2) RNA-based method:

Step R1: Use Infernal [38] to predict non-coding RNAs against Rfam database.

Step R2: Use Bowtie2 [35, 184, 185] to map RNA-seq short reads to the rRNA database to get the unaligned reads, and assemble the unaligned reads into transcripts using STAR and Trinity genome-guided method.

Step R3: Use Splign [34] to map the transcripts obtained in step R2 to the target assembly.

2.2.2.4.2 Annotate genes and pseudogenes (Figure 2-3)

Step 1: Use our gene finding procedure (https://github.com/zhengchangsulab/A-genome-assebmly-and-annotation-pipeline) to get the primary gene annotation results. The inputs of the pipeline are the results from Step H1~H2 and Step R1~R3 of Data preparations.

Step 2: Use GFF3toolkit [183] to correct CDS phase of the protein coding genes from the primary gene annotation results.

2.3 Results

2.3.1 Results of genome assembly pipeline

Using the genome assembly pipeline, we have successfully assembled genomes of four indigenous chicken breeds at chromosome-level with high-quality (CHAPTER 3). We now use Daweishan chicken (raw sequencing data are shown in Supplementary Table 3-3) as an example to demonstrate how to use the pipeline to assemble its genome at chromosome-level with high continuity. As shown in Table 2-1, after step 1 (assemble contigs followed by polishing using short reads), we obtained 616 contigs with a Contig N50 of 23 Mbp and BUSCO [10] completeness of 94.20%. To evaluate the quality of the contigs, we mapped the short reads from the Daweishan chicken to these assembled contigs using Bowtie2 [35] allowing no mismatch, and found 3.56% of the assembled

nucleotides in the contigs could not be covered by the short reads. After step 2 (scaffolding), the number of contigs and BUSCO completeness decreased compared with the results of step 1, and the number of Ns, contig N50 and percentage of the assembly that cannot be covered by short reads increased compared with the results of step 1, these were caused by the fact that scaffolding made the assembly more continuous but at the same time it could introduce more gaps. After step 3 (fill gaps using long reads), the number of contigs and number of Ns decreased, and the contig N50 increased compared with the results of step 2. Meanwhile, the percentage of the assembly that cannot be covered by short reads increased compared with the results of step 2, because gap filling using long reads could reduce gaps but at the same time introduce sequence errors. After step 4 (polish using long reads), the number of contigs remained the same (368) but their total length increased and contig N50 increased, and the number of Ns decreased, compared with the results of step 3. Meanwhile, BUSCO completeness decreased, and percentage of the assembly that cannot be covered by short reads increased, because polishing using long reads can reduce gaps but at the same time introduce sequencing errors for the assembly. Thus, the assembly needs to be further polished using short reads in Step 5. After step 5 (polish using short reads), the number of contigs remained the same (368), the number of Ns increase slightly but was only 1,452 in 157 gaps, the contig N50 was 74 Mbp, the BUSCO completeness was 96.70%, and only 1.69% of assembled nucleotides could not be covered by short reads. The Hi-C interaction heatmap of the finally assembled scaffolds (Figure 2-4) shows that most of the scaffolds form a squared box along the main diagonal of the heatmap matrix, indicating the assembly is at

chromosome-level. Taken together, all these measures indicate the assembly is of high-quality.

Table 2-1 Genome assembly results of the Daweishan chicken

Steps	# Congits (bp) (#NS)	Contig N50 (bp) (L50)	BUSCO completeness	Percentage of assembly that cannot be covered by short reads allowing no mismatch
Step 1-Wtdbg2	616 (1,018,887,028) (0)	22,952,302 (13)	94.20%	3.56%
Step 2-SALSA	407 (1,018,887,028) (108,500)	73,917,222 (5)	94.10%	3.57%
Step 3-Pbjelly	368 (1,026,115,452) (14,224)	73,988,966 (5)	94.10%	4.22%
Step 4-Racon	368 (1,036,024,190) (1,306)	74,243,019 (5)	85.00%	14.95%
Step 5-NextPolish	368 (1,037,025,916) (1,452)	74,334,266 (5)	96.70%	1.69%

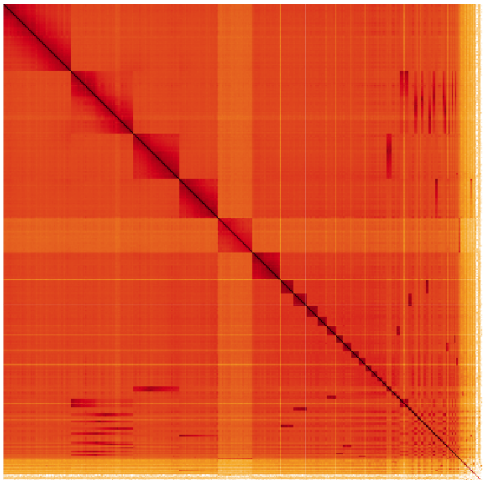

Figure 2-4 Hi-C interaction heatmap of the Daweishan chicken assembly

2.3.2 Results of gene annotation pipeline

Our gene annotation pipeline works seamlessly with our genome annotation pipeline for accurately identifying protein-coding genes, including new genes that are missed in the reference annotations as well as pseudogenes in the assembled genome as we have recently demonstrated in four indigenous chicken genomes (CHAPTER 3). We now use chr10 of our assembled Daweishan chicken genome as an example to demonstrate how to use the pipeline to annotate genes encoded in the chromosome. First, using the homology-based approach (Figures 2-2 and 2-3), we mapped all the CDS isoforms in closely related RJF (GRCg6a), and domesticated chickens (GRCg7b and GRCg7w) to the target chr10, found 421 intact genes, six partially supported genes and 13 pseudogenes from the homology-based method. Second, using the RNA-based

approach (Figures 2-2 and 2-3), we found 26 new genes in the chromosome (Table 2-2). Thus, we have identified a total of 453 protein-coding genes (including 26 new genes not annotated in the reference) and 13 pseudogenes on the chr10 of the Daweishan chicken. For the reference, 432, 433 and 445 protein-coding genes, and four, two and four pseudogenes, are annotated in GRCg6a, GRCg7b and GRCg7w, respectively (Table 2-2). By comparing our results with the reference, we have identified more pseudogenes and also found new genes that are not annotated in the reference, suggesting our pipeline is accurate and can find more putative genes.

Table 2-2 Gene annotation on chr10 of the four chickens

Chicken	Homology-based	RNA-based	# Total protein coding genes	# Total pseudogenes
Daweishan	# Intact genes: 421; # Partially supported genes: 6; # Pseudogenes: 13	# new genes: 26	453	13
GRCg6a	-	-	432	4
GRCg7b	-	-	433	2
GRCg7w	-	-	445	4

2.3.3 Comparison of our annotation pipeline with the pipeline of NCBI

To further demonstrate the accuracy of our gene annotation pipeline, we annotated the genes for a Tibetan ground-tit (an Aves) individual based on its assembly from NCBI (GCF_000331425.1) using our pipeline and compared our results with the gene annotation results from NCBI. By using the CDSs of 53 well-annotated Aves from NCBI (Supplementary Table 2-1), we annotated 14,659 intact genes, 1,165 partially supported genes and 765 pseudogenes for the individual (Table 2-3). By using the RNA-seq reads of the individual, we annotated 111 NT-supported genes and seven novel genes for the individual (Table 2-3). In total, we annotated 15,942 genes and 765 pseudogenes for the individual (Table 2-3). For the results from NCBI, there are 15,814 genes and 83 pseudogenes annotated for the individual (Table 2-3). When comparing the results from our pipeline and NCBI, we found that there are 15,058 genes shared by the results from the two pipelines, suggesting that our pipeline could annotate almost all the genes

(95.2%) predicted by the NCBI pipeline (Table 2-3). For the pseudogenes, there are 41 shared by the results of the two pipelines (Table 2-3). For the inconsistent genes and pseudogenes predicted by the two pipelines, we found that NCBI could mistakenly predict genes to pseudogenes, or mistakenly predict pseudogenes to genes. For example, gene *DIO*3 is annotated as a true gene by the NCBI pipeline, but it is annotated as a pseudogene by our pipeline since there is a stop codon (TGA) in the middle (position: 411) of the ORF of the gene (Figure 2-5a). Gene *LOC*102109485 is annotated as a pseudogene by the NCBI pipeline, but it is annotated as a true gene (named *LOC*107212260) by our pipeline since there is neither a stop codon nor a ORF shift mutation in the ORF of the gene (Figure 2-5b).

Table 2-3 Annotation of the Tibetan ground-tit genome using the two pipelines

Methods	Reference-based			RNA-based		# Total genes	# Total pseudogenes
	# Intact	# Partial	# Pseudogenes	# NT-supported	# Novel		
Our pipeline	14,659	1,165	765	111	7	15,942	765
NCBI pipeline	-	-	-	-	-	15,814	83
Common	-	-	-	-	-	15,058	41

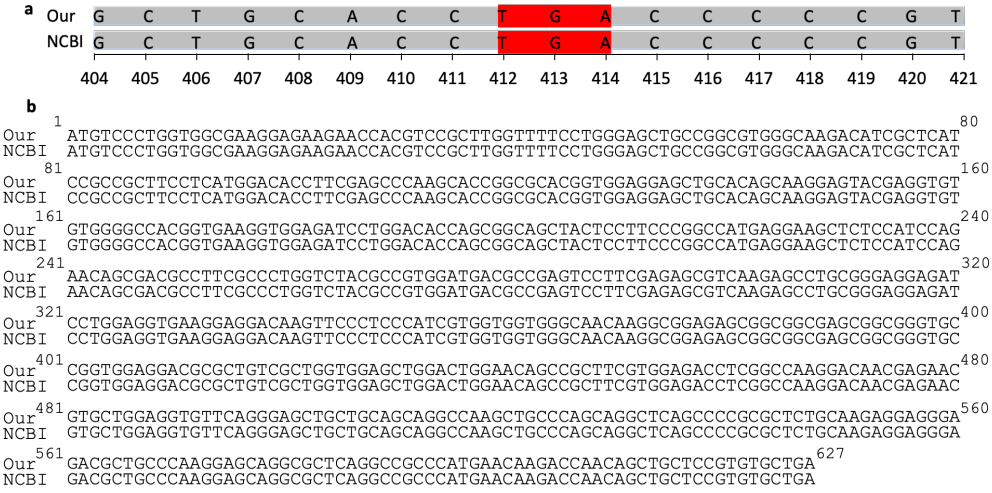

Figure 2-5 Examples of different annotations from the two pipelines. a. Part of ORFs of gene DIO3 annotated by the two pipelines, which contains a stop codon labeled in red. b. ORFs of gene LOC102109485 annotated by the two pipelines.

2.4 Discussion

When assembling a genome, most of the existing genome assembly tools can only achieve the contig-level assembly using the PacBio/Nanopore long reads and Illumina short reads, such as Wtdbg2 [15], Canu [170], and FALCON [171]. To get the chromosome-level assembly, more steps are needed based on the contigs, including scaffolding the contigs using Hi-C short reads, filling the gaps introduced in the scaffolding step using PacBio/Nanopore long reads and polishing the errors introduced in the gap filling step using PacBio/Nanopore long reads and Illumina short reads. Thus, we have proposed a genome assembly pipeline that could achieve the steps mentioned above to get the high-quality chromosome-level assembly for vertebrates. Compared with the existing genome assembly tools [15, 170, 171], our pipeline consists of several individual tools proposed before including Wtdbg2 [15] for assembling contigs, SALSA [16, 17] for scaffolding contigs, PBJelly [18] for gap filling, Racon [19] and NextPolish [20] for polishing, and could achieve high-quality genome assembly at chromosome level for vertebrates.

For the gene annotation process, most of the existing tools rely on either only *de novo* prediction such as Augustus [172], GlimmerHMM [173], GeneID [174], Genscan [175], SNAP [176] and GeneMark-ES [177], which often makes substantial errors that could hamper the analysis of important biological process, or only RNA-seq data such as Braker2 [179], which need a plenty of RNA-seq data and cannot identify pseudogenes, or only homology-based prediction such as GeMoMa [178], which can only predict protein coding genes based on the reference and cannot identify novel genes of the species. Although NCBI and ENSEMBL have proposed their own gene annotation pipeline, they

are not available to individual research groups and difficult to follow. Thus, we have proposed a user-friendly gene annotation pipeline using a combination of homology-based and RNA-seq data-based method. Compared with the existing gene annotation tools and the pipelines proposed by NCBI and ENSEMBL, our pipeline is an open source and easy to be followed by individual research lab. At the same time, our pipeline could identify protein-coding genes accurately that are shared among the homolog and the target species, pseudogenes that have open reading frame (ORF) shift or nonsense mutation compared with the homolog and new genes that are not existing or annotated in the homolog.

To run our genome assembly pipeline, Illumina paired-end short reads, PacBio/Nanopore long reads and Hi-C paired-end reads with enough sequencing depth are needed, so users without these raw data may not obtain a high-quality assembly. To run our gene annotation pipeline, CDS isoforms of reference from a near species and RNA-seq paired-end reads from as many as possible tissues with enough sequencing depth of the target species or near species are required, so users need to collect these raw data, otherwise they may not get the accurate gene annotation result.

2.5 Conclusion

High-quality genome assembly and accurate gene annotation provide good foundation to study a species. In our previous study (CHAPTER 3), we used our pipeline to assemble the genomes for four indigenous chicken individuals at chromosome-level and annotate the protein-coding genes and pseudogenes for them based on their assemblies. The result shows our pipeline could achieve high-quality assemblies at chromosome-level and accurate gene annotations for vertebrate.

CHAPTER 3 High-quality assemblies of four indigenous chicken genomes

3.1 Background

Evidence shows that red jungle fowl (RJF) (*Gallus gallus*) is the primary ancestor of domestic chickens (*Gallus gallus domesticus*) all over the world [6]. Since the release of the initial draft genome assembly (GRCg1) of an RJF individual [1], multiple improved assemblies (GRCg2~6a) have been made [2, 3]. More recently, the Vertebrate Genomes Project (VGP) also assembled pseudo-haplotype genomes (GRCg7b and GRCg7w) of a hybrid individual from a broiler mother and a layer father using long sequencing reads and multiple scaffolding data [4, 5]. Although these assemblies provide a good foundation to understand various aspects of chicken biology and guidance to poultry breeding, high-quality assemblies for indigenous chickens (traditionally domesticated village chickens) are still lacking, hampering the understanding of chicken domestication and evolution. Recently, Li, et al have assembled a pan-genome for chicken using short, long and Hi-C reads [9]. However, with a contig N50 of 5.89~16.72 Mb and a complete BUSCO [10] value of 92.4%~95.3%, this assembly does not allow the identification of the subtle differences in gene compositions of the chickens. Although they identified 1,335 new genes, more than half of them are micro-open reading frames (ORFs) with a coding DNA sequence (CDS) shorter than 300 bp. Thus, the existing assemblies are limited for revealing details of chicken domestication and evolution, and inconsistent conclusions have been drawn. For example, on one hand, it was reported that chicken genome has undergone a large number of segmental deletions [1], resulting in a large number of gene loss, thus chicken might have few genes than

other tetrapods [11]. On the other hand, it was concluded that selection for loss-of-function mutations had no prominent role in chicken domestication [12, 13].

To fill the gaps and clear contradictory conclusions, we assembled genomes of four individual indigenous chickens of Daweishan, Hu, Piao and Wuding breeds from Yunnan province in southwestern China, one of the major geographical places where the domesticated chickens originated [6]. These chicken breeds are formed by less-intensive traditional family-based artificial selection in villages in isolated mountainous areas in the province since 2000–6000 BC [14]. Each breed possesses distinct morphological traits: Daweishan chickens have a miniature body size (0.5~0.8kg for female and 0.8~1.2kg for male adults); Hu chickens have a large body size (3kg for female and 6kg for male adults) with extraordinarily stout legs; Piao chickens have a short tail (a rumpless phenotype); and Wuding chickens have a middle-sized body and are good at running. Using a combination of short, long and Hi-C reads, we assembled each genome at the chromosome-level with a contig N50 of 16.2~25.1 Mb and a complete BUSCO value of 96.5%~96.7%. By annotating these high-quality indigenous chicken genome assemblies, we identified numerous new protein-coding genes and non-processed pseudogenes. Most of these new genes are also found in the earlier assembled RJF (GRCg6a) and commercial chickens (GRCg7b/w) genomes. Counting these new genes, chickens have a similar number of protein-coding genes as other tetrapods do. Our analyses of the occurring patterns and evolutionary behaviors of the pseudogenes suggest that loss-of-function mutations plays a critical role in domestication and evolution of village chickens. These indigenous chicken genome assemblies will be valuable resources for studying chicken domestication and evolution.

3.2 Methods and materials

3.2.1 Chicken Populations

A total of 25 (nine males, 16 females) Daweishan chickens aged 10 months, 10 (five males, five females) Hu chickens aged seven months, 23 (11 males, 12 females) Piao chickens aged 10 months and 23 (11 males, 12 females) Wuding chickens aged 10 months were collected from the Experimental Breeding Chicken Farm of the Yunnan Agricultural University (Yunnan, China). One female individual chicken from each breed was randomly selected for genome assembly, and subject to short-reads DNA sequencing, PacBio or Oxford long-reads sequencing, and Hi-C sequencing. All other individuals were subject to short-reads DNA sequencing. We downloaded the re-sequencing short DNA reads of 35 RJFs, 60 broilers and 56 layers from the NCBI SRA database with the accession number of PRJEB15276, PRJEB30270 and PRJEB15189.

3.2.2 Short-reads DNA sequencing

Two milliliters of blood were drawn from the wing vein of each chicken in a centrifuge tube containing anticoagulant (EDTA-2K) and stored at -80℃ until use. Genomic DNA (10μg) in each blood sample was extracted using a DNA extraction kit (DP326, TIANGEN Biotech, Beijing, China) and fragmented using a Bioruptor Pico System (Diagenode, Belgium). DNA fragments around 350 bp were selected using SPRI beads (Beckman Coulter, IN, USA). DNA-sequencing libraries were prepared using Illumina TruSeq® DNA Library Prep Kits (Illumina, CA, USA) following the vendor's instructions. The libraries were subject to 150 cycles paired-end sequencing on an Illumina Novaseq 6000 platform (Illumina, CA, USA) at 100X coverage.

3.2.3 PacBio long-reads sequencing

Two milliliters of blood were drawn from the wing vein of a female Hu chicken (H3) in a centrifuge tube with anticoagulant (EDTA-2K) and stored at -80℃ until use. High molecular weight DNA was extracted from each blood sample using NANOBIND® DNA Extraction Kits (PacBio, CA, USA) following the vendor's instructions. DNA fragments of about 25 kb were size-selected using a BluePippin system (Sage Science, MA, USA). Sequencing libraries were prepared for the DNA fragments using SMRTbell® prep kits (PacBio, CA, USA) following the vendor's instructions, and subsequently sequenced on a PacBio Sequel II platform (PacBio, CA, USA) at 36X coverage.

3.2.4 Oxford long-reads sequencing

Two milliliters of blood were drawn from the wing vein of a female Daweishan (F025), Piao (P17) or Wuding (W17) chicken in a centrifuge tube with anticoagulant (EDTA-2K) and stored at -80℃ until use. High molecular weight DNA from each blood sample was prepared using Ultra-Long Sequencing Kits (Oxford Nanopore Technology (ONT), Oxford, UK) following the vendor's instructions. The integrity of DNA was determined using pulsed field electrophoresis. DNA fragments of about 20 kb were size-selected using a BluePippin system (Sage Science, MA, USA). Sequencing libraries were prepared for the DNA fragments using ONT Template prep kit (SQK-LSK109) and NEB Next FFPE DNA Repair Mix kit, following the vendors' instructions. The libraries were sequenced on a Nanopore PromethION P48 platform (ONT, Oxford, UK) at ~100X coverage using ONT sequencing kits (EXP-FLP001.PRO.6).

3.2.5 RNA-seq sequencing

One to two grams of various tissues were collected from the selected female individual chicken of each breed in a centrifuge tube and immediately frozen in liquid nitrogen, then stored at -80℃ until use. Total RNA from each tissue sample were extracted using TRlzol reagents (TIANGEN Biotech, Beijing China) according to the manufacturer's instructions. RNA-sequencing libraries for each tissue as well as for the mixture of all the tissues collected from a chicken were prepared using Illumina TruSeq® RNA Library Prep Kits (Illumina, San Diego) following the vendor's instructions. The libraries were subject to 150 cycles paired-end sequencing on an Illumina Novaseq 6000 platform at a sequencing depth of 80X. Additional individual chicken was randomly selected from each breed population and the same types of tissues were collected. Total RNA from each tissue sample were extracted in the same way as described above. Equal weight of total RNA of each tissue was mixed for preparing an RNA-seq library as described above.

3.2.6 Hi-C sequencing

Five milliliters of blood were drawn from the wing vein of the selected Daweishan (F025), Piao (P17), Hu chicken (H3) or Wuding (W17) chickens in a Streck Cell-free DNA BCT collecting vessel (Streck Corporate, USA), and stored at 4℃ and used in 24 hours. Hi-C libraries were constructed using Phase Genomics' Animal Hi-C kit following the vendor's instructions and subsequently sequenced on an Illumina's Novaseq 6000 platform at a sequencing depth of ~100X.

3.2.7 Real-time quantitative PCR (RT-qPCR) analysis

RT-qPCR was performed using the Bio-Rad CFX96 real-time PCR platform (Bio-Rad Laboratories. lnc, America) and SYBR Green master mix (iQTM SYBRGreen ® Supermix, Dalian TaKaRa Biotechnology Co. Ltd. Add). The primers of the 42 randomly selected putative new genes are listed in Supplementary Note 3-1. The β-actin gene was used as a reference. Primers were commercially synthesized (Shanghai Shenggong Biochemistry Company P.R.C). Each PCR reaction was performed in 25 μl volumes containing 12.5 μl of iQ™ SYBR Green Supermix, 0.5 μl (10 mM) of each primer, and 1 μl of cDNA. Amplification and detection of products was performed with the following cycle profile: one cycle of 95 °C for 2 min, and 40 cycles of 95 °C for 15 s, annealing temperature for 30 s, and 72 °C for 30 s, followed by a final cycle of 72 °C for 10 min. The specificity of the amplification product was verified by electrophoresis on a 0.8% agarose gel and DNA sequencing. The $2^{-\Delta Ct}$ method was used to analyze mRNA abundance. All samples were analyzed with at least three replicates, and the mean of these measurements was used to calculate mRNA expression.

3.2.8 Contig assembling and scaffolding

We filtered out PacBio/Nanopore long reads shorter than 5,000 bp in each library, and assembled contigs using Wtdbg (2.5) [15] with the remaining reads for each chicken. Then we bridged the contigs and obtained scaffolds using SALSA [16, 17] with the Hi-C data. We filled the gaps in the scaffolds using PBJelly [18] with long reads. We made two rounds of polishing on the resulting scaffolds, first by using Racon (1.4.21) [19] with the long reads, and second by using NextPolish (1.4.0) [20] with the paired-end short reads from the same individual chicken.

3.2.9 Chromosome-level genome assembling

In order to get the chromosome-level assembly for the four chickens, we mapped the scaffolds for each chicken to the GRCg7b assembly using blastn (2.11.0) [21]. We consider a scaffold is mapped to a chromosome if the ratio of the mapped length of a scaffold over minimum of (the length of the query scaffold, the length of the target chromosome) is greater than 0.5. We ordered and orientated the scaffolds based on the mappings using the GRCg7b chromosomes as templates. For the scaffolds mapped to the same chromosome, we concatenated them by 500 Ns according to their mapping orders. For the rest scaffolds that cannot be mapped to GRCg7b assembly, we consider them as unplaced scaffolds. In this way, we sorted the assembled scaffolds of the four chickens into 42 chromosomes (including two sexual chromosomes and one mitochondrial chromosome).

During the mapping, we found that some contigs were incorrectly concatenated into scaffolds by the scaffolding tool, we thus manually corrected the errors by splitting them from the scaffolds and reassembled them to their corresponding target chromosomes. In addition, there is no scaffold in Hu chicken's primary assembly, which can be mapped to the mitochondrial chromosome of GRCg7b. Thus, it appears that the assembly tools didn't assemble the mitochondrial chromosome for Hu chicken for unknown reason. We therefore mapped the short reads of Hu chicken to the mitochondrial chromosome of GRCg7b, and then assembled the mitochondrial chromosome for Hu chicken using Abyss (2.2.5) [22] with using the mapped short reads.

3.2.10 Correction of GRCg6a assembly

Since chromosomes chr29 and chr34~chr39 of the most recent RJF genome assembly (GRCg6a) are still missing, to facilitate our comparative analysis, we assembled these missing chromosomes by mapping all the assembled chromosomes and contigs of GRCg6a to the chromosomes of GRCg7b using blastn (2.11.0) [21]. If a chromosome of GRCg6a is exactly mapped to the corresponding chromosome of GRCg7b, then we kept it intact. For the chromosome that is not consistent between the two assemblies, we used GRCg7b as the template to resolve the inconsistence by appropriate reorienting, splitting and concatenation. Specifically, we first located the original contigs of inconsistent chromosomes of GRCg6a, and then mapped them as well as unplaced contigs of GRCg6a to the chromosomes in GRCg7b to find the target chromosomes. Finally, we concatenated the contigs of GRCg6a mapped to the same chromosome of GRCg7b by 500 Ns according to the mapping order.

3.2.11 Quality evaluation of assemblies

We masked the repeats of the four assemblies using WindowMasker (2.11.0) [23], and estimated the heterozygosity of each assembly using Jellyfish (2.3.0) [24] and GenomeScope [25]. To estimate the continuity of each assembly, we used QUAST (5.0.2) [26] to calculate the contig N50, scaffold N50 and chromosome N50. To estimate the structural accuracy, we used Asset [27] to calculate the reliable block N50 and used BUSCO (5.1.3) [10] to calculate the false duplications in each assembly. To estimate the base accuracy, we used Merqury (1.3) [28] to calculate the *k-mer* QV and *k-mer* completeness for the four assemblies, used BWA (0.7.17) [29] to map the short reads to the four assemblies, and used and SAMtools (1.10) [30] to analyze the mapping results.

To estimate the functional completeness, we used BUSCO (5.1.3) [10] to assess each assemblies' completeness against the avian gene set and used STAR (2.7.0) [31] to map the mRNA short reads to each assembly and calculate the mRNA completeness value. To plot the heatmap of the chromosomes of each assembly, we mapped the Hi-C paired-end reads to the assembly using BWA (0.7.17) [29], used SAMtools (1.10) [30] and Pairtools (0.3.0) [32] to analyze the mapping results, and used Higlass [33] to plot the heatmap for each assembly.

3.2.12 Protein-coding gene annotation

To annotate the protein-coding genes in the assembled indigenous chicken genomes, we masked the repeats in each genome using WindowMasker (2.11.0) [23], and then we annotated the protein-coding genes using a combination of homology-based and RNA-based methods. For homology-based annotation, we collected all the protein-coding genes, pseudogenes and their corresponding CDS isoforms or exons in GRCg6a, GRCg7b and GRCg7w as the templates (Supplementary Table 3-1). The protein-coding genes in these three assemblies were recently predicted by the NCBI eukaryotic genome annotation pipeline that uses a combination of mRNA- and protein-based homology methods and *ab initio* methods. We mapped all the CDS isoforms of protein-coding genes and exons of pseudogenes in GRCg6a, GRCg7b and GRCg7w to each of the assembled indigenous chicken genomes using Splign (2.0.0) [34]. For each template gene whose CDSs could be mapped to an assembled genome, we concatenated all the mapped CDSs, and checked whether the resulting sequence forms an intact ORF (the length is an integer time of three and contain no stop codon in the middle). If yes, we called it an intact gene. If the CDSs of a template gene can be mapped to multiple loci in an

assembled genome, we consider the locus with the highest mapping identity. If a concatenated sequence did not form an intact ORF, i.e., the length is not an integer time of three (ORF shift) or it contains a stop codon in the middle (nonsense mutation), we mapped the DNA short reads from the same individual to the CDSs of the gene using bowtie (2.4.1) [35] with no gaps and mismatches permitted. If the sequence could be completely covered by at least 10 short reads at each nucleotide position, we consider the pseudogenization is fully supported by the short reads, and called the sequence a pseudogene; otherwise, we consider the pseudogenization is not supported by the short reads, and called it a partially supported gene, because the pseudogenization might be artificially caused by errors of the long reads that could not be corrected by our assembly pipeline. For a few selenoprotein template genes where the "opal" stop-codon UGA encode selenocysteine, we manually checked the mapped loci in our assemblies, and annotate them accordingly.

For RNA-based annotation, we first mapped all of the RNA-seq reads from various tissues as well as the mixture of tissues of the four chicken breeds into the rRNA database SILVA_138 [36] and filtered out the mapped reads. We then mapped the unaligned reads to each of the four indigenous chicken genome assemblies using STAR (2.7.0c) [31]. Based on the mapping results, we assembled transcripts in each chicken using Trinity (2.8.5) [37] with its genome-guided option. Next, we mapped the assembled transcripts in each chicken to its assembled genomes using Splign (2.0.0) [34], and removed those that at least partially overlap non-coding RNA genes (see below), protein-coding genes or pseudogenes predicted by the homology-based method. For the remaining transcripts, if we could find a longest ORF with at least 300 pb, we called it a

protein-coding gene. If multiple ORFs were found in a transcript, we selected the longest one.

Additionally, to detect protein-coding genes that might be not included in our assemblies but expression in the tissues, we de novo assembled transcripts using Trinity (2.8.5) [37] with its de novo option using RNA-seq reads that could not be mapped to any of our four assembled genomes.

3.2.13 RNA-coding gene annotation

We annotated tRNA, rRNA, miRNA, snoRNA, telomerase RNA and SRP RNA using infernal (1.1.2) [38] with Rfam (v.14) database [39] as the reference. In addition, we predicted an assembled transcript longer than 1,000 bp but lacking an ORF as a lncRNA. The results are summarized in Supplementary Table 3-2.

3.2.14 Neighbor-joining tree construction

We mapped the 8,338 essential avian proteins from the BUSCO aves_odb10 database [10] to each of the seven chicken's CDSs as well as the coturnix japonica's CDSs using blastx (2.11.0) [21]. We selected the 6,744 genes with greater than 70% sequence identity with the essential avian proteins in each of the eight genomes to construct a neighbor-joining tree. Since it is hard to make multiple alignments for very long sequences, we evenly divided the genes in each bird into 68 groups (each contains about 100 genes) and concatenated CDSs in each group with fixed order. We then aligned the concatenated sequences of the same group in the eight birds using Clustal Omega (1.2.4) [40]. We finally concatenated the 68 multiple alignments and constructed a consensus neighbor-joining trees with 1,000 rounds of bootstrapping using Phylip (3.697) [41].

3.2.15 Prediction of miRNA binding sites

For each pair of pseudogene and its reference gene, we scanned their CDSs and 1,000 bp downstream sequences as putative 3'-UTRs for miRNA binding sites using RNAhybrid (2.1.2) [42]. The miRNAs predicted in the genome harboring the pseudogene are used as the database for the scanning. We consider the putative binding sites with a p-value<0.05.

3.2.16 Single nucleotide variants calling

To calculate the fixation rates of the pseudogenes in a chicken breed, we called SNVs in the pseudogenes using short DNA reads from the breed population (n= 25, 10, 23, 23, 35, 60 and 56 for Daweishan, Hu, Piao, Wuding, RJF, broiler and layer, respectively) using GATK (4.1.6) [43].

3.3 Results

3.3.1 High quality assemblies of four indigenous chicken genomes

Using a combination of Illumina short reads (89~143X) and PacBio (36X) or Nanopore (101~110X) long reads (Supplementary Table 3-3), we assembled the genomes of a female individual of the Daweishan, Hu, Piao and Wuding chicken in 462~1,364 contigs (Supplementary Table 3-4) with a contig N50 of 23.0, 16.2, 25.1 and 21.5 Mb, respectively. Hu chickens' smaller contig N50 (16.2 Mb) might be due to the shorter PacBio reads (50 bp~90.8 kbp) and a shallower sequencing depth (36X) than those of Nanopore reads (Supplementary Table 3-3). The contig N50 values of the four assemblies are larger than those of the GRCg6a, GRCg7b and GRCg7w assemblies (17.7~18.8 Mb) (Supplementary Table 3-5), except for Hu chicken. The total length of the contigs (>1 Gb) for each chicken is comparable with those of GRCg6a, GRCg7b and

GRCg7w assemblies (Supplementary Table 3-5). We scaffolded the contigs using Hi-C reads (102~112X) for each chicken (Supplementary Table 3-3), resulting in 308~1,088 scaffolds, with a scaffold N50 of 74.3, 28.9, 62.8 and 71.1 Mb for the Daweishan, Hu, Piao and Wuding chickens, respectively (Supplementary Table 3-5). Using the GRCg7b assembly as the template, we ordered and oriented the scaffolds into 39 autosomal chromosomes, two sex chromosomes W and Z and one mitochondrial genome for each chicken (Supplementary Table 3-4), with a chromosome N50 of 90.5, 90.7, 90.5 and 90.9 Mb for the Daweishan, Hu, Piao and Wuding chickens, respectively (Supplementary Table 3-5).

3.3.2 Evaluation of the quality of the four indigenous chicken genomes

Recently, the VGP consortium proposed six categories of criteria for evaluating the quality of a chromosome-level assembly, including genome (degree of heterozygosity and repeats), continuity, structural accuracy, base accuracy, functional completeness, and chromosomal assignment status [4]. We thus further evaluated the quality of each of our assembled genomes using these criteria (Supplementary Table 3-5). For the genome evaluation, we found that Piao chicken had the highest heterozygosity of 0.9%, Hu chicken the lowest heterozygosity of 0.7%, and both Daweishan and Wuding chicken a middle heterozygosity of 0.8%. Repeats consist of from 19.9 to 20.4% of all the four assembled genomes, which are similar to those of GRCg6a (20.4%), GRCg7b (20.5%) and GRCg7w (20.2%). For the continuity evaluation, both the contig N50 (16.2~25.1 Mb) and chromosome N50 (90.5~90.9 Mb) of our assemblies are comparable with those of GRCg6a (17.7 and 91.3 Mb), GRCg7b (18.8 and 90.9 Mb) and GRCg7w (17.7 and 90.6 Mb). There are 375, 484, 483 and 297 gaps in the Daweishan, Hu, Piao and Wuding

assemblies, respectively, which are substantially fewer than those in GRCg6a (500,945) and are comparable with those in GRCg7b (463) and GRCg7w (409). For the structural accuracy evaluation, we identified reliable blocks and false duplications of the assemblies. As shown in Supplementary Table 3-5, we achieved a reliable block N50 > 13.5 Mb except for Hu chicken (2.9 Mb), and a false duplication rate of 0.3~0.4%. The values of both parameters are comparable to those of the recent VGP assemblies of 16 species of six major vertebrate lineages [4].

For the base accuracy evaluation, we first computed the *k-mer* QVs of our assemblies, which is the log-scaled probability of consensus errors in the assembly [28]. We found that the *k-mer* QVs of the Daweishan, Piao and Wuding chicken assemblies were greater than 41.5 and the value of the Hu chicken assembly was 38.4, suggesting that the consensus base accuracy is greater than 99.99% and 99.90% for the former three assemblies and the Hu chicken assembly [28], respectively, which is comparable to those obtained by the VGP assemblies [4]. We next calculated *k-mer* completeness, which is defined as the fraction of reliable *k-mers* in highly accurate short reads data that are also found in the assembly [28]. As shown in Supplementary Table 3-5, the *k-mer* completeness for all the four assemblies is greater than 92.8%, also comparable to those of the recent VGP assemblies [4]. Since our assemblies are the mosaics of the paternal and maternal homologous chromosomes that differ in heterozygous sites, to further evaluate the completeness of the assemblies, we mapped short reads from each individual chicken to its assembled genome, and found that the mapping rates were greater than 99.2% for all the assemblies. These results indicate that all our four assemblies have achieved high base accuracy.

For the functional completeness evaluation, all of the four assemblies have > 96.5% BUSCO completeness [10], which are comparable to those of GRCg6a (96.6%), GRCg7b (96.6%) and GRCg7w (96.8%). We next mapped the RNA-seq reads from multiple tissues of each chicken to its assembled genome and found that the mapping rates were at least 95.0% for all the four assembled genomes except for Hu chicken (91.3%). These results indicate that the assemblies are of high completeness. For the chromosomal assignment status evaluation, although there are still some unplaced contigs in each of our assemblies, the total length (non-N bp) of the contigs assigned to chromosomes are greater than 98% for all of the four assemblies, which are similar to those of GRCg6a and GRCg7b/w. Thus, most of our contigs are assigned to chromosomes. Additionally, we plotted the Hi-C interaction heatmaps of the autosomes and sex chromosomes (W and Z) of each of the four assemblies. As shown in Figure 3-1, for all the four genomes, most assembled chromosomes including the micro-chromosomes (chr11~39) form a squared box along the main diagonal of the heatmap matrix, with the exception of a few very small chromosomes such as chr31 and chr35. Moreover, the assembled chromosomes in each of the four assembled genomes display high collinearity with those of the GRCg7b (Figure 3-2), indicating that the structures of the assembled genomes are consistent. Our assembled mitochondrial genomes of Daweishan, Hu, Piao and Wuding chickens have a length of 16.0, 16.5, 16.3 and 16.8 kbp, respectively, which are similar to those of GRCg6a (16.8 kbp), GRCg7b (16.8 kbp) and GRCg7w (16.8 kbp). Taken together, these results indicate that we have achieved chromosome-level assembly for all of the four chickens' genomes.

3.3.3 Assembly of missing micro-chromosomes in GRCg6a

The current RJF's reference genome (GRCg6a) lacks assemblies of micro-chromosomes chr29 and chr34~chr39. We found that contigs of these chromosomes were either mistakenly assembled into chr31, chr32 and chr33 or were unplaced. We assembled these missing chromosomes and corrected the mis-assembled chr31, chr32 and chr33 of GRCg6a based on the GRCg7b assembly. Supplementary Table 3-6 summarizes the contig compositions and lengths of these assembled RJF micro-chromosomes in comparison with those of the GRCg7b assembly. Interestingly, except for chr35, chr36 and chr39, these assembled RJF micro-chromosomes are much longer than their corresponding ones in GRCg7b, suggesting that they might be more complete. Chr1 to chr28, chr30, the two sexual chromosomes and the mitochondrial chromosome in GRCg6a were consistent with those of GRCg7b, and thus were kept intact. The assembled GRCg6a chromosomes also display high collinearity with those of the GRCg7b (Figure 3-2), indicating that the structures of the two assembled genomes are consistent.

3.3.4 Varying lengths and high G/C contents of micro-chromosomes

We compared the lengths of the assembled chromosomes of the four indigenous chickens with those of RJF. As shown in Figure 3-3a, the four indigenous chickens and RJF have similar lengths of all the assembled chromosomes, except for micro-chromosomes chr16 and chr29~chr39 that show highly varying lengths. These results indicate that the lengths of all the assembled macro-chromosomes (chr1~10), most micro-chromosomes (chr11~15, chr17~26) and sex chromosome (chrW and chrZ) of the five chickens are consistent, and thus are sufficiently assembled. The lengths of the assembled

chr16 in all the four indigenous chickens are much shorter than that of RJF, suggesting that our assembled chr16 might be incomplete. This might be due to the fact that chr16 contains highly repetitive major histocompatibility complex (MHC) regions, relatively more duplicated genes [44] and higher G/C contents (Figure 3-3b). Our assemblies of chr33, chr35, chr36, chr38 and chr39 are much longer than those of RJF, but our assemblies of chr27, chr29, chr31, chr32, chr34 and chr37 are shorter than those of RJF, suggesting that more efforts are needed in the future to more completely assemble these micro-chromosomes. The difficulty to better assemble these micro-chromosomes might be at least partially due to their higher G/C contents (Figure 3-3b).

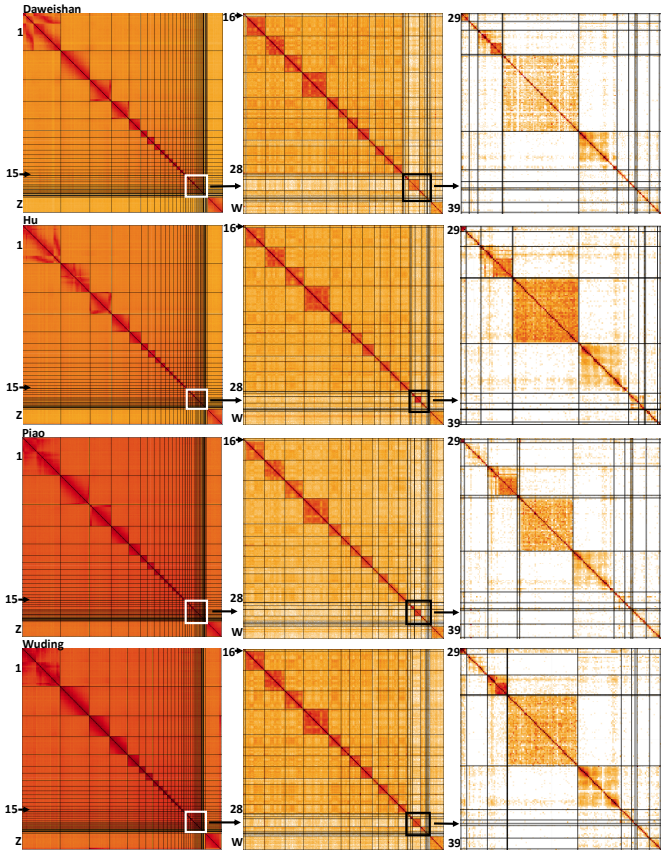

Figure 3-1 Interaction heatmaps of the chromosomes of the four chickens. First column: heatmaps of all the 41 chromosomes of Daweishan, Hu, Piao and Wuding chicken. Second column: a zooming-in view of the heatmaps of chr16~chr28 and sex chromosome W of Daweishan, Hu, Piao and Wuding chicken. Third column: a zooming-in view of the heatmaps of chr29~chr39 of Daweishan, Hu, Piao and Wuding chicken.

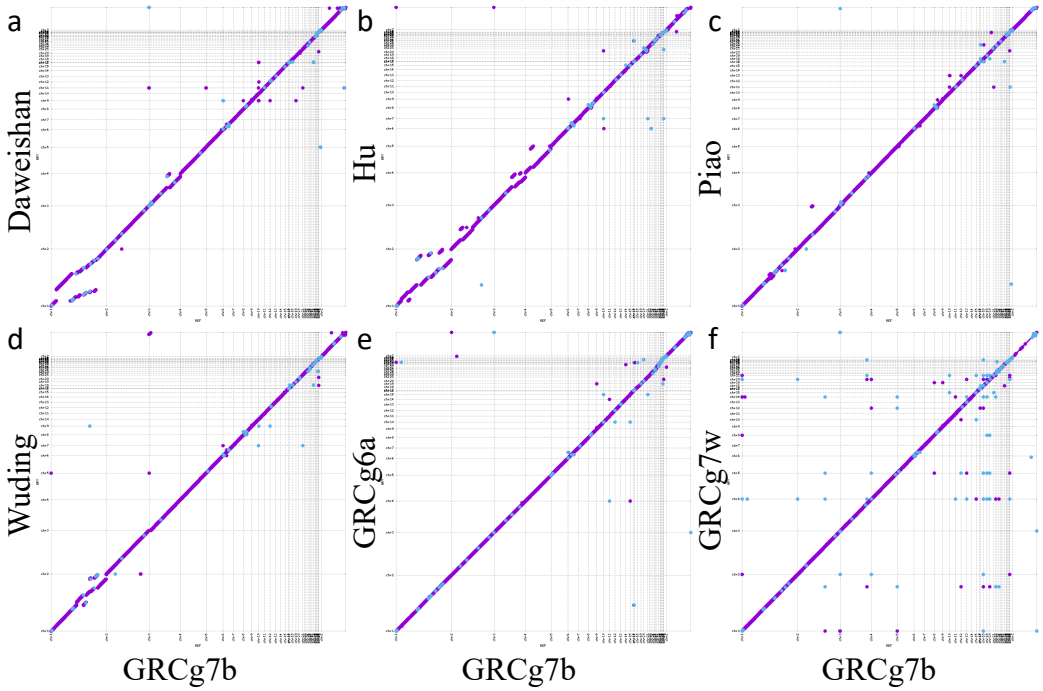

Figure 3-2 Collinearity of chromosomes of the seven chickens

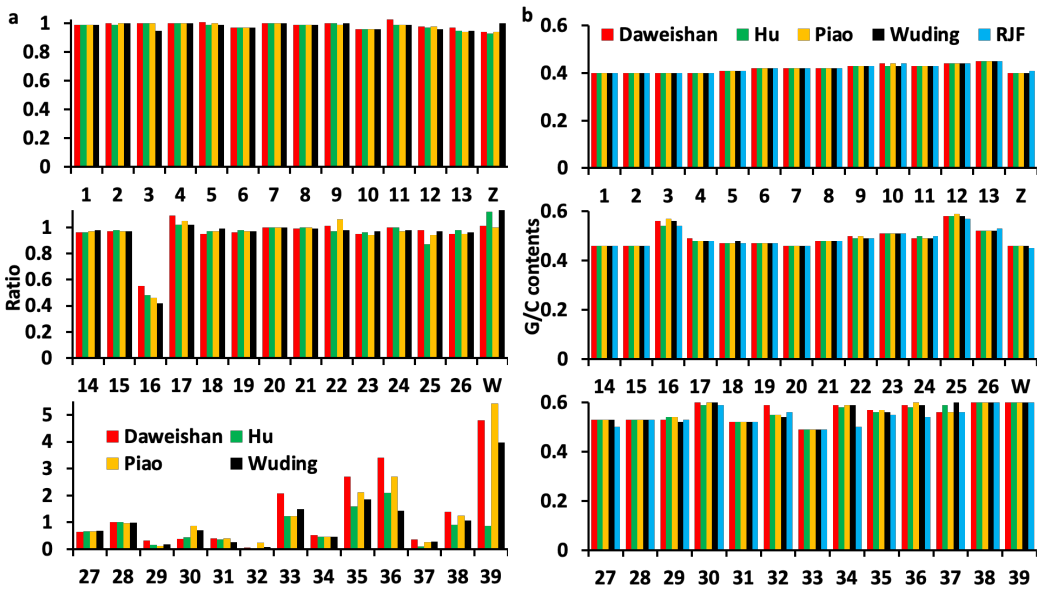

Figure 3-3 Comparison of each chromosome of the five chickens. **a.** Ratio of the length of each chromosome (excluding Ns) of the four indigenous chickens over that of the same chromosome (excluding Ns) of RJF. **b.** G/C contents of each chromosome of the five chickens.

3.3.5 New protein-coding genes are found in the four indigenous chickens

We predicted 17,497~17,718 protein-coding genes in each of the four indigenous chicken genomes (Table 3-1), which is similar to that annotated in GRCg6a (17,485), but fewer than those annotated in GRCg7b (18,024) and GRCg7w (18,016). Specifically, we predicted 16,917~17,141 genes in each genome based on homology to known genes, of which 16,270~16,668 have an intact ORF (intact genes) (Table 3-1), and 473~647 containing either a nonsense or an ORF shift mutation that cannot be fully supported by short DNA reads from the chicken. It is highly likely that such "mutations" might be due to errors in the long reads that cannot be corrected by the short DNA reads, we therefore refer these genes as partially supported genes (Table 3-1). Of the intact and partially supported genes, 16,077~16,455 (98.7%~98.8%) and 463~642 (97.9%~99.2%), respectively, are transcribed in at least one of the tissues examined using RNA-seq (Supplementary Tables 3-7~3-10). Interestingly, we predicted 6~7 genes in each indigenous chicken genome based on homology to pseudogenes in GRCg6a and/or GRCg7b/w (Table 3-1). These genes have an intact ORF that is fully supported by short DNA reads, and 4~7 are transcribed in at least one of the tissues examined (Supplementary Tables 3-7~3-10), thus, they are likely to be functional. For example, the RJF pseudogene *LOC*107049240 at locus chr33:1757489~1758423 with a point deletion is mapped to chr33:2240289~2241224 of Piao chicken, encoding an intact ORF that is supported by 805 short DNA reads as well as large numbers of RNA-seq reads in multiple tissues (Figure 3-4a, Supplementary Table 3-9).

Based on RNA-seq data that can be mapped to the assembled genomes, we identified 12,543~13,930 putative genes in each assembled indigenous chicken genome

(Supplementary Table 3-11). Of these genes, 9,561~10,596 at least partially overlap the homology-supported genes (Supplementary Table 3-11), thus, we do not consider them further. Of the remaining 2,982~3,334 putative genes that do not overlap homology-supported genes, 571~627 contain at least an ORF, and we consider them as RNA-supported new genes for further analysis (Table 3-1). Since some of these RNA-supported new genes in the four assembled genomes are highly similar (identity > 98.5%) to one another, we removed the redundancy and ended up with a total of 1,420 unique new genes in the four genomes, which are not seen in the earlier annotations (GRCg6a, GRCg7b and GRCg7w) (Supplementary Table 3-12). Interestingly, about 24~35 of the new genes found in an indigenous chicken are pseudogenized in the others (Supplementary Table 3-12). Of these 1,420 new genes, 1,277 (511~572 in each assembled genome) are homologous to genes in the NT database, so we refer them to as NT-supported new genes. The remaining 143 (54~63 in each assembled genome) new genes do not have a known homolog, and thus, might be novel genes (Table 3-1 and Supplementary Table 3-12). Most of the NT-supported new genes are mapped to genes in other breeds of *Gallus gallus* or in other avian species (Supplementary Table 3-13).

Table 3-1 Summary of annotated protein-coding genes in chickens

Chicken	Homology-supported genes			Pseudogene-supported		RNA-supported genes			# Originally annotated true genes	# Originally annotated pseudogenes	# True genes added	# Pseudogenes added	# Final true genes	# Transcribed true genes	# Unique true genes (Percentage)	Final pseudogenes		# Tranbribed pseudogenes
	Known gene-supported					RNA-supported new genes		# RNA-supported pseudogenes								# Processed	# Non-processed	
	# Intact genes	# Partially supported genes	# Pseudogenes	# Intact genes	# Pseudogenes	# NT-supported new genes	# Novel genes											
Daweishan	16,668	473	622	6	94	517	54	31	-	-	-	-	17,718	17,494	60 (0.3%)	63	684	713
Hu	16,270	647	486	6	85	511	63	35	-	-	-	-	17,497	17,298	39 (0.2%)	50	556	576
Piao	16,587	513	564	7	83	549	55	35	-	-	-	-	17,711	17,493	55 (0.3%)	55	627	655
Wuding	16,530	483	557	6	86	572	55	24	-	-	-	-	17,646	17,429	51 (0.3%)	48	619	633
RIF(GRCg6a)	-	-	-	-	-	-	-	-	17,485	262	978	280	18,463	-	992 (5.4%)	80	462	-
Broiler(GRCg7b)	-	-	-	-	-	-	-	-	18,024	198	978	276	19,002	-	415 (2.2%)	59	415	-
Layer(GRCg7w)	-	-	-	-	-	-	-	-	18,016	150	962	285	18,978	-	678 (3.6%)	50	385	-

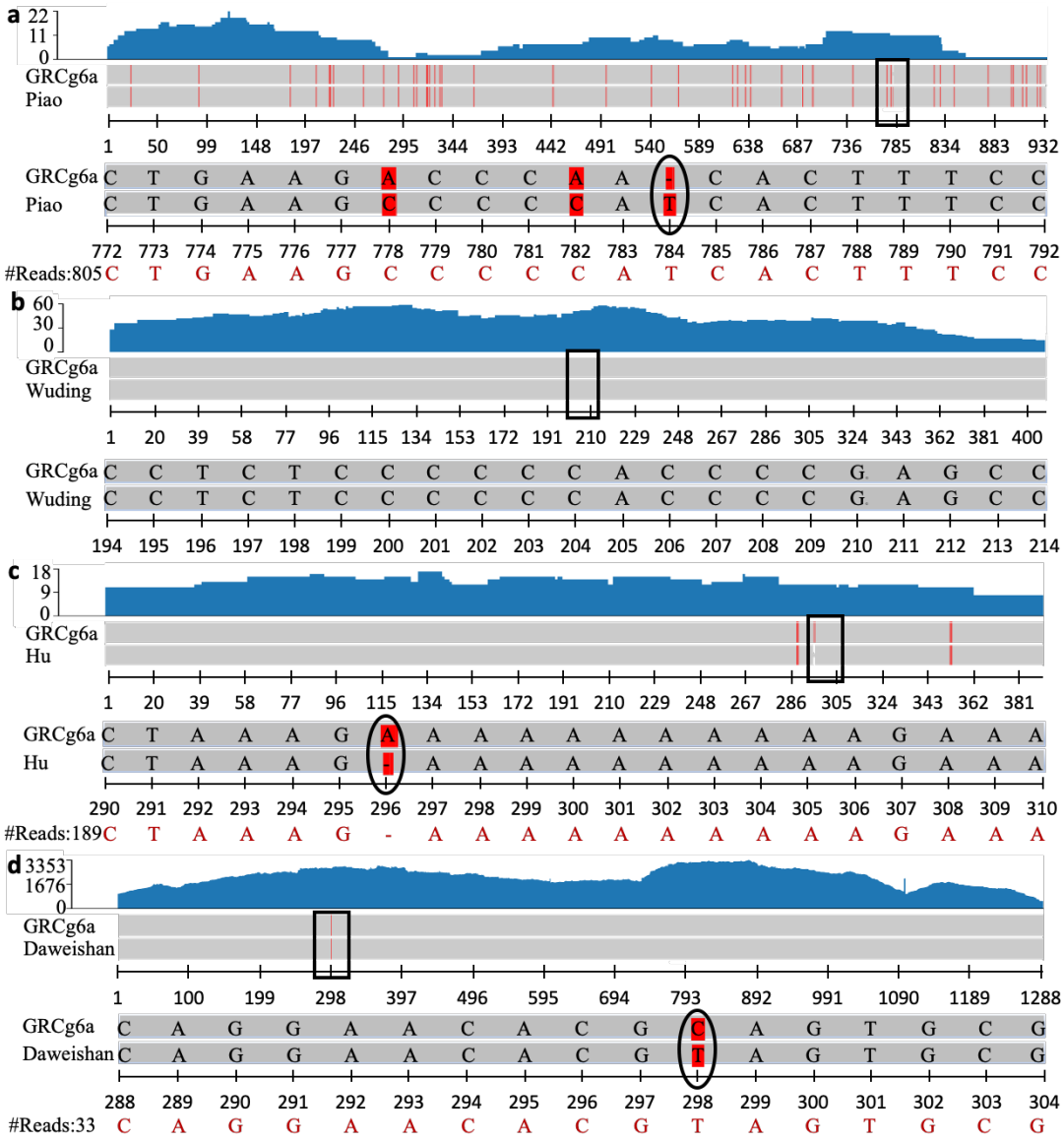

Figure 3-4 Examples of transcribed new genes and pseudogenes. **a.** A protein-coding gene in Piao chicken is predicted based on a pseudogene LOC107049240 in RJF that harbors a point deletion of 'T' at position 784, leading to an ORF shift. **b.** A new gene predicted in Wuding chicken was also encoded in RJF. **c.** A new gene predicted in Hu chicken is pseudogenized in RJF, due to an insertion of 'A' at position 296, leading to an ORF shift. **d.** A pseudogene in Daweishan chicken with a substitution of 'C' in RJF with 'T' at position 298 leading to a nonsense mutation in Daweishan chicken is predicted based on the TNFRSF10B gene in RJF. In all the examples, the first row shows the coverage of RNA-seq reads along the CDSs of the genes in our assemblies; the second row is the comparison of the CDSs of the genes in RJF and their orthologous loci in our indicated assembly; the third row is a zoom-in view of the boxed region in the second row, which shows the mutation at the single nucleotide level; the last row displays the number and a part of DNA short reads that support the mutations in our indicated assembly.

3.3.6 New genes tend to be located at the sub-telomere regions

Although the 1,420 new genes distribute on almost all the assembled chromosomes (Figure 3-5a), they tend to concentrate on micro-chromosomes and unplaced contigs (Figure 3-5b). The new genes also tend to be located at the sub-telomere regions of the chromosomes where G/C contents tend to be high (Figure 3-5c), consistent with a recent report [9]. Moreover, the higher G/C contents of a genome region, the more likely new genes are found in it (Pearson correlation coefficient γ=0.49, Figure 3-5d). Consistent with these facts and the high G/C contents in micro-chromosomes (Figure 3-3b), the new genes tend to have higher G/C contents than the existing genes ($p<4e-4$, K-S test) (Figure 3-5e). This might at least partially explain why the new genes were missed by the earlier annotations, since genome regions, particularly, those in micro-chromosomes with high G/C contents were difficult to be sequenced.

3.3.7 New genes show strong tissue-specific expression patterns

Importantly, both the NT-supported new genes and novel genes show strong tissue-specific expression patterns in all the four chickens (Figure 3-6), suggesting that they are likely authentic genes. To further validate these 1,420 new genes, we randomly selected 42 of them (Figure 3-6) and quantifying their transcription levels in various tissues of the four chicken breeds using RT-qPCR. As shown in Supplementary Tables 3-14~3-17, 21, 17, 10 and 15 of the 42 putative new genes are encoded in Daweishan, Hu, Piao and Wuding chickens, of which 18 (85.71%), 14 (82.35%), 8 (80%) and 15 (100%) were expressed in multiple tissues of the four breeds, respectively (Figure 3-6). Counting on results from all the four breeds, 39 (92.86%) of the 42 selected genes were expressed in multiple tissues of at least one chicken breed. Therefore, most of the predicted new

genes are likely authentic. However, the expression patterns in different tissues of the 39 new genes measured by RT-qPCR are not very similar to those quantified by RNA-seq reads (Figure 3-6).

3.3.8 Our new genes have limited overlaps with these identified previously

Moreover, we compared our 1,420 new genes with the 1,335 new genes recently found in 20 chicken genomes [9]. Our new genes with a mean length of 7,008 bp are much longer than the 1,335 new genes with a mean length of only 1,413 bp (p=2e-171, Wilcoxon rank-sum test) (Figure 3-5f). One of the reason for the discrepancy is that 756 (56.6%) of the 1,335 new genes are mini-ORFs [45] with a CDS length of 100~300 bp, while all of our 1,420 new genes have a CDS length longer than 300 bp, indicating that more than half of the earlier predicted new genes are not bona fide protein-coding genes. A total of 660 (49.4%) of the 1,335 new genes can be mapped to at least one of our indigenous chicken genomes with an identity greater than 98.5%, thus we have assembled their loci in at least one of the four genomes. Of these 660 genes, 246 (37.3%) overlap our predicted genes, while 246 (59.4%) of the remaining 414 genes are mini-ORFs.

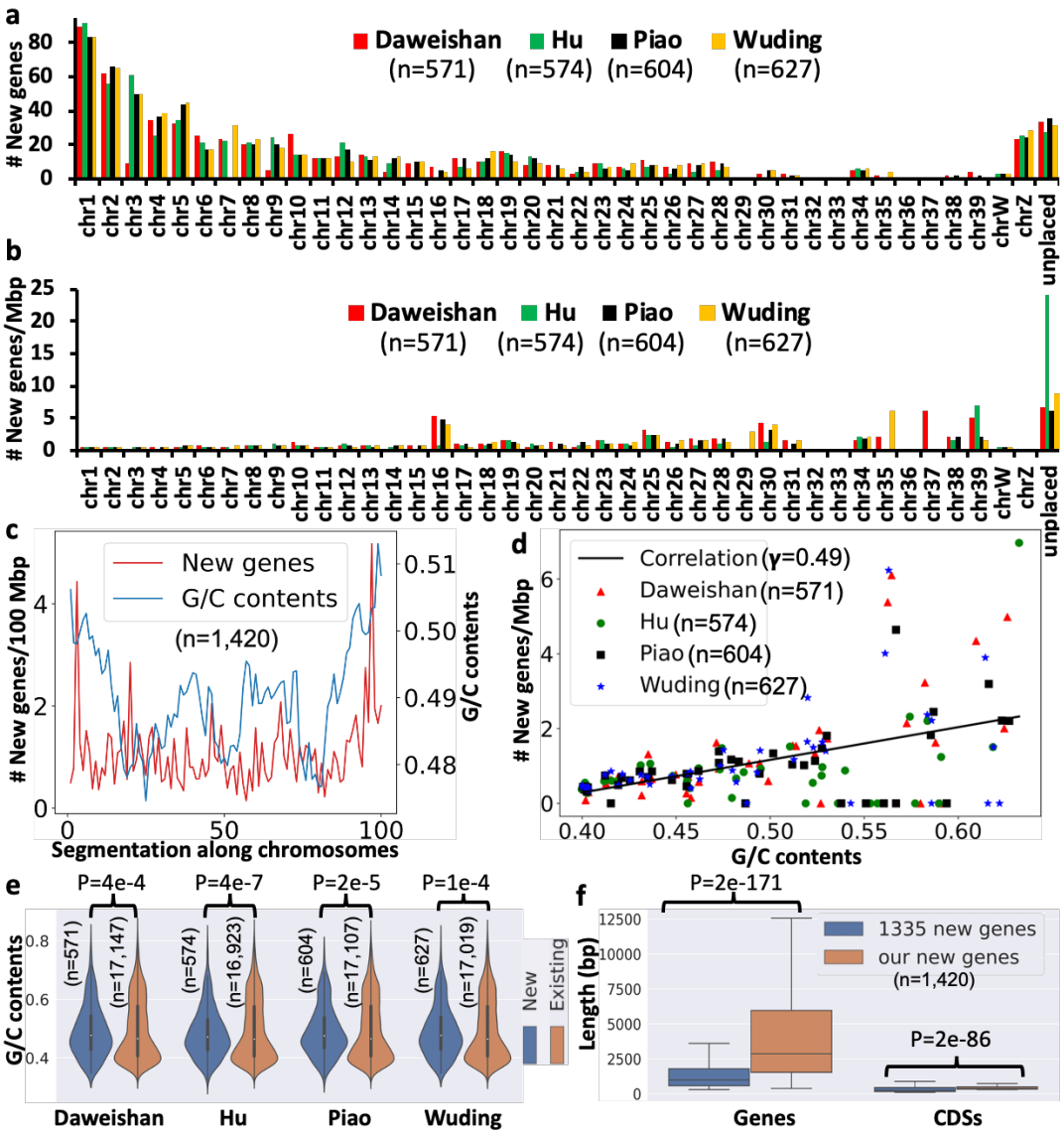

Figure 3-5 Properties of new genes found in the four chickens. **a.** Number of new genes found on each chromosome and unplaced contigs. **b.** Number of new genes per million bp found on each chromosome and unplaced contigs. **c.** Average number of new genes per million bp and average G/C contents along evenly divided 100 segments of the 41 chromosomes in the four chicken genomes. **d.** The relationship between the number of new genes on a chromosome and its G/C contents in the four genomes. The black line is the linear regression of data from the four chickens. **e.** Comparison of G/C contents of the new genes with those of the existing genes. **f.** Comparison of the lengths of our new genes and their CDSs with those of the earlier predicted 1,335 new genes in 20 chickens.

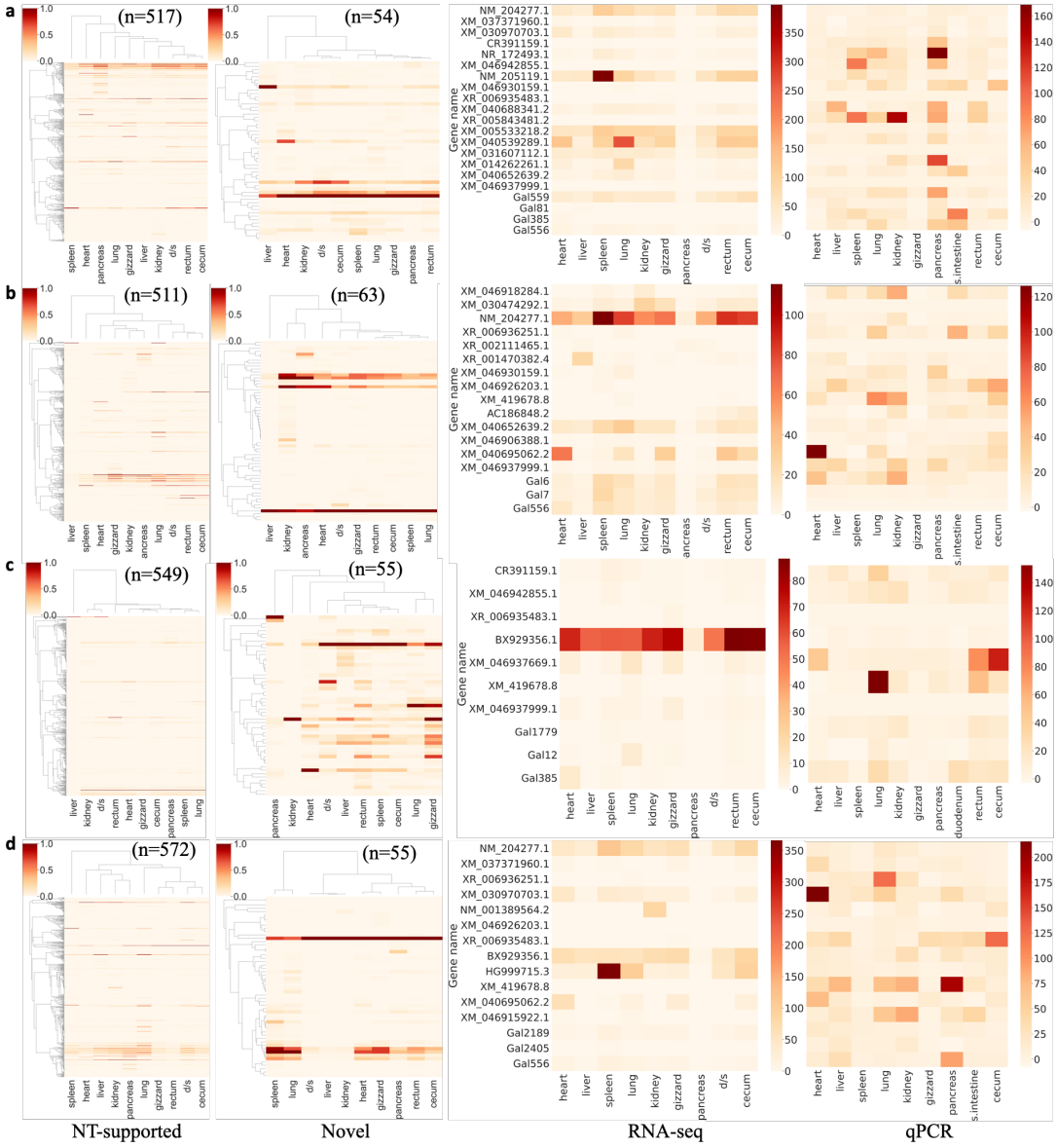

Figure 3-6 Expression levels of RNA-supported new genes in the four chickens. **a.** Expression levels of RNA-supported new genes in different tissues of the Daweishan chicken. **b.** Expression levels of RNA-supported new genes in different tissues of the Hu chicken. **c.** Expression levels of RNA-supported new genes in different tissues of the Piao chicken. **d.** Expression levels of RNA-supported new genes in different tissues of the Wuding chicken. The first column is heatmaps of normalized expression levels of NT-supported new genes in the four chickens. The second column is heatmaps of normalized expression levels of novel genes in the four chickens. The third column is heatmaps of expression levels of the 39 verified new genes measured by RNA-seq reads in the four chickens. The fourth column is heatmaps of expression levels of the 39 verified new genes measured by RT-qPCR in the four chickens. “d/s” represents duodenum/small intestine.

3.3.9 The new protein-coding genes are involved in house-keeping functions

Interestingly, most (1,291 or 90.92%) of our 1,420 new genes found in the indigenous chicken genomes can be mapped to GRCg6a (1,258), GRCg7b (1,254) and GRCg7w (1,247) assemblies (Supplementary Table 3-12). Of these mapped genes, 962~978 have intact ORFs, thus, we consider them as new genes in each of the three earlier assemblies. We thereby increase the number of annotated genes in GRCg6a (18,463), GRCg7b (19,002) and GRCg7w (18,978) by 5.6% (978), 5.4% (978) and 5.3% (962), respectively (Table 3-1). Therefore, GRCg6a and GRCg7b/w encode a comparable number of protein-coding genes to other tetrapods (18,000~25,000) [4, 5] such as humans [46], supporting the previous conclusion inferred for birds in general [47]. Interestingly, the remaining 276~285 new genes mapped to GRCg6a (280), GRCg7b (276) and GRCg7w (285) assemblies are pseudogenes (Supplementary Table 3-12), i.e., they contain at least one nonsense or ORF shift mutations of normal genes. For examples, a new gene of Wuding chicken at chr11:17138466~17139339 is mapped to RJF chr11:17570676~17571083 that encodes an intact ORF (Figure 3-4b); while a new gene of Hu chicken at locus chr2:128256077~128257512 is mapped to RJF locus chr2:129144681~129145071 with an insertion of ‘A’ supported by 189 short reads, leading to an ORF shift (Figure 3-4c). We thereby also substantially increase the number of pseudogenes in GRCg6a (542), GRCg7b (474) and GRCg7w (435) by 106.9% (280), 139.4% (276) and 190.0% (285), respectively (Table 3-1).

The four indigenous chickens and GRCg6a, GRCg7b and GRCg7w display two distinct occurring patterns of the 1,420 new genes in their genomes (Figure 3-7a). Specifically, two thirds (962~978) of the new genes appear in GRCg6a, GRCg7b and

GRCg7w, while the remaining one third (422~458) are either absent or pseudogenized in these three genomes. In contrast, only 40.2~44.2% (571~627) of the new genes occur in each of the four indigenous chickens, while the remaining 55.8~59.8% (793~849) are either absent or pseudogenized in them (Figure 3-7a, Supplementary Table 3-12). Moreover, the new genes are clustered in multiple distinct groups (Figure 3-7a). Although most of the new genes do not have gene ontology (GO) [48] term assignments, they are involved in a total of 40 GO biological pathways (Supplementary Table 3-18), many of which correspond to gene clusters shown in Figure 3-7a. These GO biological pathways are involved in house-keeping functions, including transcription regulation, signal transactions, immunity, cell growth, metabolism and apoptosis, to name a few (Supplementary Table 3-18).

Notably, the commercial chicken (GRCg7b/w) and the RJF (GRCg6a) encode ~1,500 and ~1,000 more protein-coding genes, respectively, than the four indigenous chickens (Table 3-1). Moreover, the four indigenous chickens share more genes with each other and with GRCg7b and GRCg7w than with GRCg6a, while GRCg7b and GRCg7w share more genes with each other than with GRCg6a and with the four indigenous chickens (Figures 3-7b), suggesting that the indigenous chickens are evolutionarily closer to one another than to GRCg6a, GRCg7b and GRCg7w, and that GRCg7b and GRCg7w are more evolutionarily closer to each other than to GRCg6a and the four indigenous chickens. This result is largely consistent with the phylogenetic tree of the seven chickens plus Coturnix japonica as the root (Aves class) (Figure 3-7c), constructed using 6,744 essential protein-coding genes shared by the eight Aves (Supplementary Table 3-19) from the BUSCO aves_odb10 database [10]. In the

phylogenetic tree (Figure 3-7c), the four indigenous chickens form a clade, and GRCg7b and GRCg7w from a clade. Furthermore, there are a varying number of genes unique to Daweishan (60), Hu (39), Piao (55), Wuding (51), GRCg6a (992), GRCg7b (415) and GRCg7w (678) compared among the seven chickens (Table 3-1, Supplementary Table 3-20). The unique genes in each chicken are involved in GO biological pathways that might be good candidates for experimental studies for their roles in forming the chicken's unique traits (Supplementary Table 3-21). For example, FGF (Fibroblast growth factors) signaling , EGF (epidermal growth factor) receptor signaling and PDGF (platelet-derived growth factor) signaling pathways that are involved in stem cell proliferation and growth are found in the genome of Hu chicken with a very large body weight (~6kg); multiple p53-related pathways are found in GRCg7b (broiler) with fast growth rate; Sarcoplasmic/endoplasmic reticulum calcium ATPase 1 pathway is found in GRCg7w (layer) that needs high calcium storage for the formation of egg shells. GRCg6a (RJF) has the largest number (992) of unique genes that are involved in pathways that might favor their wildlife.

3.3.10 "Missing" protein-coding genes are found in chicken genomes

To see whether we have assembled and annotated any of the 274 genes that were widely encoded in reptiles and mammals but were reported missing in avian species in general and another 174 genes that were believed missing in chicken (based on the galGal4 assembly) in particular [11], we mapped their human CDSs to each of the indigenous chicken genomes. We found that 51 and 36 of them, respectively, are among our predicted genes (Supplementary Table 3-22). However, we noted that eight (ITPKC, TEP1, BRSK1, PLXNB3, MAPK3, HIGD1C, KMT5C and PACS1) of the 274 missing

genes in avian species in general and seven (PPP1R12C, CCDC120, ATF6B, ASPDH, DDIT3, ADCK5, GPAA1) of the 174 missing genes in chicken in particular, are annotated in GRCg6a, GRCg7b or GRCg7w, but none of them appear in any of the four indigenous genomes (Supplementary Table 3-22). It has been shown that RNA-seq data from multiple tissues can be used to recover presumed missing genes in birds [49-51]. To see whether more missing genes could be recovered by RNA-seq data collected in various tissues of the chickens, we *de novo* assembled transcripts using RNA-seq reads that could not be mapped to any of the four chicken genomes. Of 34 such predicted CDSs/genes (Supplementary Table 3-23) in the four chickens, three can be mapped to missing genes *MAPK3, SLC25A23* and *HSPB6* (Supplementary Table 3-22A). Notably, MAPK3 is annotated in GRCg7b/w but missing in GRCg6a. Thus, we might be unable to assembly the *MAPK3, SLC25A23* and *HSPB6* loci due probably to their refractory to the sequencing technologies we used to assemble the genomes, and thus these genes are missed by our annotation pipeline.

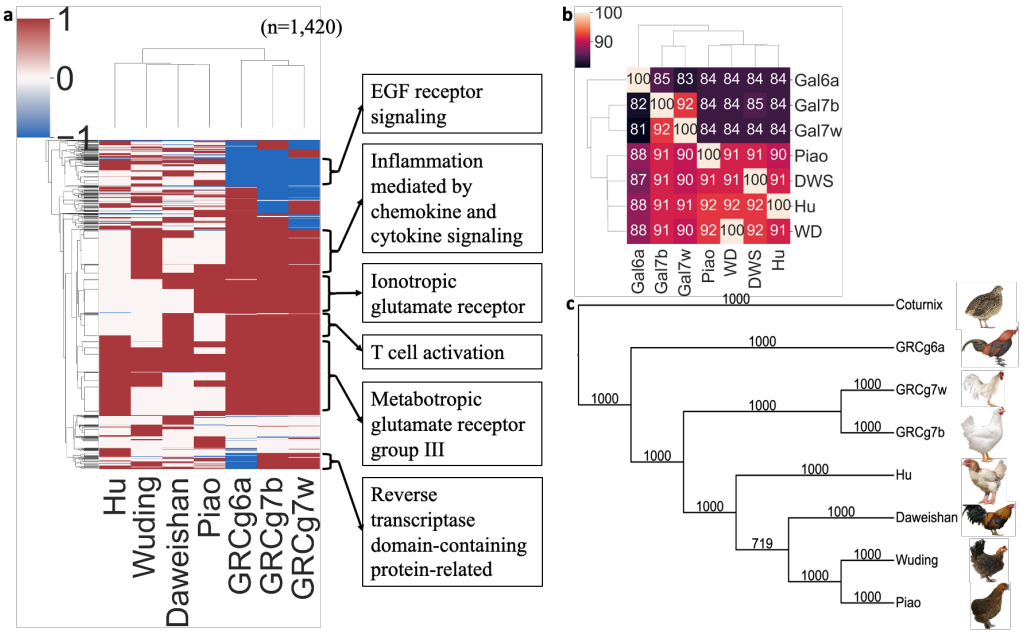

Figure 3-7 Occurring patterns of the 1,420 new genes. **a.** Heatmap of two-way hierarchical clustering of the new genes based on their appearance (1, brown), absence (0, white) and pseudogenization (-1, blue) patterns in the seven chicken genomes. **b.** Heatmap of two-way hierarchical clustering of seven chickens based on the percentage of the genes that each chicken (rows) share with the others (columns). **c.** Consensus neighbor-joining tree, constructed using 6,744 essential protein-coding genes of the seven chickens and coturnix japonica.

3.3.11 Chicken genomes harbor a large number of non-processed pseudogenes

As indicated earlier, we have substantially increased the number of pseudogenes in GRCg6a (542), GRCg7b (474) and GRCg7w (435). To our surprise, we found an even larger number (606~747) of pseudogenes in each of the indigenous chicken genomes (Table 3-1). The pseudogenes in each indigenous chicken are located in almost all the chromosomes. However, the micro-chromosomes (chr11~chr39) and unplaced contigs have higher densities, harboring about half (52.7~54.0%) of the pseudogenes (Figures 3-8a and 3-8b) while they only comprise 27.3~27.9% of the genomes. Likewise, the ratio of the number of pseudogenes/the number of genes tends to be higher on the micro-chromosomes and unplaced contigs (Figure 3-8c). Most (521~653, 86.0%~87.8%) of pseudogenes in each indigenous chicken genome were predicted based on homology to

annotated genes in GRCg6a (322~395), GRCg7b (313~385), GRCg7w (311~398) and our 1,420 new genes (24~35) (Table 3-1, Supplementary Tables 3-24~3-27). The remaining 83~94 (12.2%~14.0%) pseudogenes in each indigenous chicken genome were predicted based on homology to earlier annotated pseudogenes in GRCg6a (49~57), GRCg7b (55~64) and GRCg7w (51~58) (Table 3-1, Supplementary Tables 3-24~3-27). Most pseudogenes (576~713, 94.9~96.0%) in each indigenous chicken genome are transcribed in multiple tissues (Supplementary Tables 3-7~3-10), suggesting that their regulatory systems might be still at least partially functional. For example, the RJF gene *TNFRSF*10B at locus chr22:1415452~1421965 is mapped to locus chr22:1296090~1301696 of Daweishan chicken with a nonsense mutation that is supported by 33 short DNA reads, and the locus is transcribed in multiple tissues (Figure 3-4d, Supplementary Table 3-7). We found a total of 1,577 pseudogenes in the four indigenous chicken genomes (Supplementary Table 3-28), each harboring 606~747 of them (Table 3-1). Most (91.6%~92.8%, 556~684) of the pseudogenes in each indigenous chicken genome are non-processed, i.e., they arose due to direct pseudogenization mutations, while only the remaining 7.2~8.4% (48~63) are processed, i.e., they arose due to retrotransposition followed by pseudogenization mutations. Similar conclusions can be drawn for the pseudogenes found in the GRCg6a and CRCg7b/w assemblies, although the latter two commercial chickens (CRCg7b/w) harbor much fewer (474~435) pseudogenes (Table 3-1).

3.3.12 Pseudogenization mutations are strongly biased to the two ends of CDSs

To see whether the occurrence of pseudogenization mutations in the indigenous chickens is under natural/artificial selection, we plotted the distribution of their first

pseudogenization mutation sites along the CDSs of reference genes. If the pseudogenization mutations are selectively neutral or a result of random genetic drift, they should largely uniformly distribute along the CDSs. As shown in Figure 3-9a, the rate of synonymous mutations in true genes is largely uniformly distributed along the CDSs as expected for neutral mutations, except at the two ends, where the rate decreases, consistent with an earlier report in chickens [52]. The reduced synonymous mutation rate suggests that the two ends of CDSs are under purifying selection, and might harbor functional elements not related to their coding functions, such as transcriptional and post-transcriptional regulatory elements [53].

Interestingly, the first pseudogenization mutations are strongly biased to the two ends of reference CDSs (Figure 3-9a), consistent with earlier reports in chickens [52] and humans [54]. Specifically, 40.2%, 35.5% and 24.3% of first pseudogenization mutations occur at the first 25%, the middle 50% and the last 25% lengths of the CDSs. Almost all the pseudogenes have their first (Figure 3-9b) and last (Figure 3-9c) coding nucleotides aligned with those of reference CDSs, indicating that the biased pseudogenization mutations to the 5'- and 3'-ends are not due to incorrect predictions of the two ends of the pseudogenes. By contrast, the rate of synonymous mutations in the pseudogenes is also largely uniformly distributed along the CDSs including the two ends, indicating that purifying selection on the two ends of pseudogenes (Figure 3-9a) is relaxed.

3.3.13 Biased pseudogenizations might facilitate loss-of-function mutations

To see whether the first pseudogenizations along the CDSs result in loss of function of the genes, we compared evolutionary pressures on true genes with that on pseudogenes using their ratio of the number of non-synonymous mutations over the

number of synonymous mutations (dN/dS). As shown in Figure 3-9d, the pseudogenes have significantly higher dN/dS values than the true genes ($p<6.13e-295$). This is also true when the pseudogenes with the first pseudogenization occurring in the first 25% ($p<8.23e-94$), the middle 50% ($p< 4.67e-148$) or the last 25% ($p<9.58e-67$) of CDSs are compared with the true genes (Wilcoxon rank-sum test). These results suggest that at least most pseudogenes are no longer under purifying selection, and thus might lose gene functions. Moreover, the pseudogenes with the first pseudogenization occurring in the first 25% of CDSs and occurring in the last 25% of CDSs have similar dN/dS values ($p=0.49$), and both have significantly lower dN/dS values ($p<9.19e-5$ and $p<4.51e-4$, respectively) (Wilcoxon rank-sum test) than the pseudogenes with the first pseudogenization occurring in the middle 50% of CDSs. The underlying cause is not clear to us, but might be due to purifying selection on the two ends of CDS before pseudogenization.

Clearly, the closer a pseudogenization mutation to the 5'-end of a CDS, the more likely the loss-of-function to occur. Loss of function of a gene could also occur when critical amino acids or regulatory DNA elements at the 3'-end of the CDS are disrupted by a pseudogenization event. For the former possibility, we noted that the identity of amino acids at the 3'-ends of CDSs elevated in proteins (Figure 3-9a), indicating that the 3'-ends of CDSs may harbor critical amino acids. For the later possibility, as the 3'-UTRs of genes often harbor miRNA binding sites for post-transcriptional regulation [55], we hypothesize that 3'-ends of CDSs may also contain miRNA binding sites, and disruption of such sites in either 3'-ends of CDSs or 3'-UTRs may have functional consequence. To test this, we scanned the pseudogenes' and corresponding reference genes' CDSs and

their 1,000 bp downstream sequences as putative 3'-UTRs for potential miRNA binding sites. As shown in Figure 3-9e, putative 3'-UTRs of both reference genes and pseudogenes have higher density of putative miRNA binding sites than their upstream coding regions, consistent with previously reports [55]; and indeed, the 3'-ends of CDSs of both reference genes and pseudogenes contain putative miRNA binding sites, and their densities decrease toward the 5'-ends of CDSs. Interestingly, pseudogenes have a fewer number of miRNA binding sites in their 3'-ends and 3'-UTRs than do reference genes (Figure 3-9e). We found that this is because pseudogenization mutations could disrupt the miRNA binding sites in the 3'-end of CDSs. However, of the 24.3% (657) first pseudogenization mutations occurring in the last 25% of CDSs, only three (0.5%) disrupt a putative miRNA binding site. Thus, this mechanism only explains a few cases of loss-of-function pseudogenization mutations at the 3'-end of CDSs. Figure 3-10 shows the three cases.

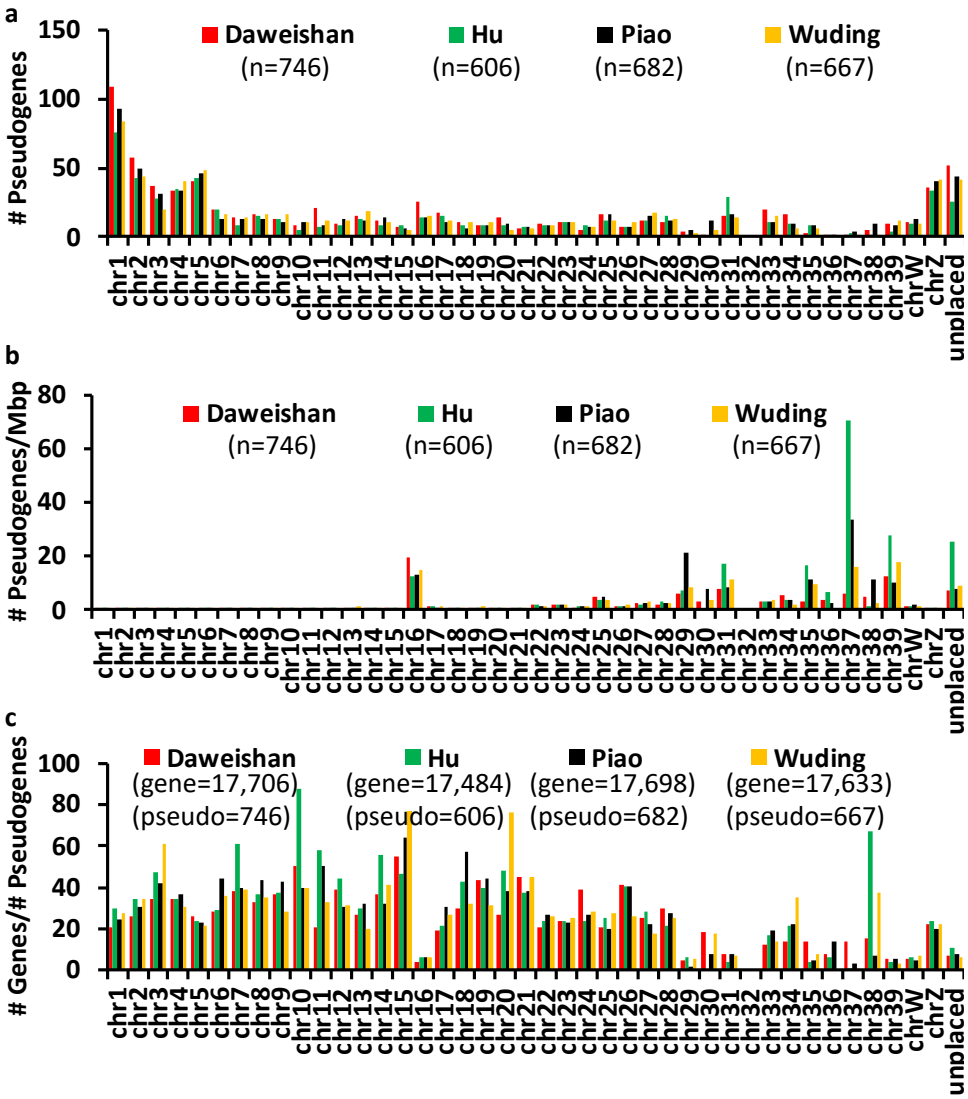

Figure 3-8 Distribution of pseudogenes on each chromosome. **a.** Number of pseudogenes on each chromosome. **b.** Density of pseudogenes on each chromosome. c. Ratio of genes/pseudogenes on each chromosome.

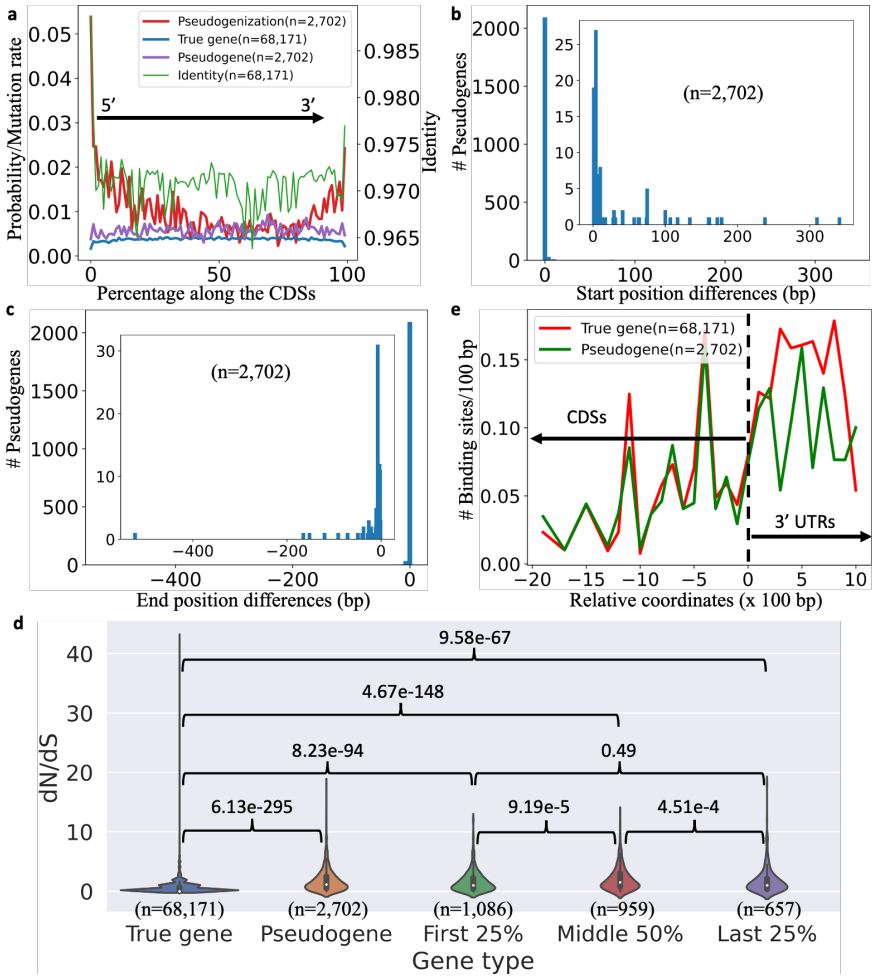

Figure 3-9 Pseudogenization mutations tend to occur at the two ends of CDSs. **a.** Probability of first pseudogenization mutations (red line) in 100 evenly divided CDS segments from the 5'-ends to the 3'-ends of the reference genes of the pseudogenes in the four chickens, mean rates of synonymous mutations in 100 evenly divided CDS segments from the 5'-ends to the 3'-ends of the true genes (blue line) and pseudogenes (purple) in the four chickens, and mean identity of the true genes in 100 evenly divided CDS segments from the 5'-ends to the 3'-ends of the genes (green line). **b.** Start position of the "CDS" of the pseudogenes in the four chickens with respect to the nucleotide positions of their reference genes starting with 0 with the downstream positions being positive integers. **c.** End positions of the "CDS" of the pseudogenes in the four chickens with respect to the nucleotide positions of their reference genes ending with 0 with the upstream positions being negative integers. **d.** Violin plots of the dN/dS values of all true genes, all pseudogenes, pseudogenes with the first pseudogenization occurring in the first 25%, the middle 50% and the last 25% of the CDSs in the four indigenous chickens. **e.** Number of predicted miRNA binding sites per 100 bp along the CDSs and 3' UTRs of the true genes (red line) and pseudogenes (green line). In the figure, '0' represents the end positions of the CDSs, the positive numbers represent the relative positions of 1,000 bp sequences downstream of the end of CDSs, and the negative numbers represent the relative positions of the CDSs with respect to the ends of CDSs.

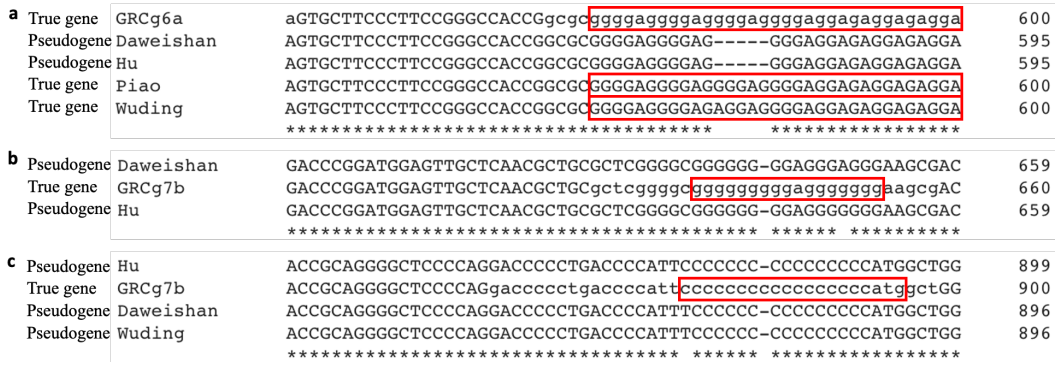

Figure 3-10 Examples of disruptions of miRNA binding sites in pseudogenes. **a.** A putative mir-335 (family: RF00766) binding site at the 3'-end of the CDS (753 bp) of gene LOC112531802 in RJF, Piao and Wuding chickens is disrupted by a 5-bp ORF shift deletion in the pseudogenes of Daweishan and Hu chickens. **b.** A putative mir-1281 (family: RF03493) binding site at the 3'-end of the CDS (762 bp) of gene LOC124417726 in GRCg7b is disrupted by a 1-bp ORF shift deletion in the pseudogenes of Daweishan and Hu chickens. **c.** A putative mir-1249 (family: RF01918) binding site at the 3'-end of the CDS (945 bp) of gene LOC124417207 in GRCg7b is disrupted by a 1-bp ORF shift deletion in the pseudogenes of Daweishan, Hu and Wuding chickens.

3.3.14 Functions of parental genes of most pseudogenes were lost in chickens

To see whether alternative isoforms of the pseudogenes in the four chickens could skip the exons harboring the pseudogenization mutations, we assembled transcripts of all the transcribed pseudogenes. We found that most transcribed pseudogenes had only one type of transcript containing the pseudogenization mutations, while for those that had more than one isoform, very few of them had transcripts that skipped the pseudogenization mutations (Supplementary Tables 3-29~3-32). For example, in Daweishan chicken, there are 684 non-processed pseudogenes (Table 3-1), of which only 139 (20.32%) have alternative splicing transcripts and only one of these 139 genes has transcripts that skip the pseudogenization mutation (Supplementary Table 3-29). In Hu chicken, there are 556 non-processed pseudogenes (Table 3-1), of which only 137 (24.64%) have alternative splicing transcripts and none of them has transcripts that skip the pseudogenization mutations (Supplementary Table 3-30). In Piao chicken, there are 627 non-processed pseudogenes (Table 3-1), of which 131 (20.89%) have alternative

splicing transcripts and only two of them have transcripts that skip the pseudogenization mutations (Supplementary Table 3-31). In Wuding chicken, there are 619 non-processed pseudogenes (Table 3-1), of which only 139 (22.46%) have alternative splicing transcripts and none of them has transcripts that skip the pseudogenization mutations (Supplementary Table 3-32). All these results suggest that almost all the pseudogenes in the four chickens did not skip the exons harboring the pseudogenization mutations. Thus, the functions of parental genes cannot be rescued by alternative isoforms of the pseudogenes. Loss-of-function mutations often occur after a duplication event to eliminate a redundant copy [56]. However, we failed to find an intact paralog for most (88.2%~90.3%) of the non-processed pseudogenes in the same genomes (Supplementary Tables 3-24~3-27), suggesting that the functions of the parental genes were lost in the chickens.

3.3.15 Most pseudogenes arose recently and are fixed in respective populations

The vast majority of the pseudogenes in each of the seven genomes share more than 95% of sequence identity with their functional reference genes (Figure 3-11), indicating that they arose quite recently, and only few with less than 80% of sequence identity with their functional reference genes arose a relatively long time ago. To see whether the first pseudogenization mutation along a pseudogene in the seven chicken genomes are fixed or not in their respective populations, we computed the frequencies of the mutated alleles in the population of the seven breeds based on DNA short reads (sequencing depth: ~20X) from populations of 25, 10, 23, 23, 35, 60 and 56 Daweishan, Hu, Piao, Wuding, RJF, broilers and layers, respectively (Materials and methods). As shown in Figure 3-12, most (~75%) of mutations in the pseudogenes in each chicken

genome are fixed or nearly fixed (allele frequency > 80%) in their respective populations. The high probability of fixation of the pseudogenes suggests that they might be under artificial and/or natural selection. A few examples of fixed or nearly fixed pseudogenes with no functional copies in an indigenous chicken population are shown in Figure 3-13. Specifically, the *OAZ2* gene (CDS length = 573 bp) that is functional in GRCg6a, CRCg7b and CRCg7w, is pseudogenized and fixed in all the four indigenous chicken populations, due to an insertion of a “T” after position 96, resulting an ORF shift (Figure 3-13a). *OAZ*2 coding for an ornithine decarboxylase antizyme plays a role in cell growth and proliferation in animals by regulating intracellular polyamine levels [57, 58]. The *PDCL*3 gene (CDS length = 723 bp) that is functional in GRCg6a, CRCg7b and CRCg7w, is pseudogenized and fixed in the four indigegnous chickens, due to the deletion of an “A” at the first position of the CDS, resulting an ORF shift (Figure 3-13b). *PDCL*3 encoding a phosducin-like protein is involved in G protein signaling and chaperone-assisted protein folding [59]. The *PRDM*16 gene (CDS = 3,828 bp), which is functional in the six other chickens, is pseudogenized and fixed in the Hu chicken population, due to a “C to T" substitution at position 10 of the CDS, resulting a stop-gain (Figure 3-13c). *PRDM*16 coding for a transcription factor controls the differentiation of brown adipocytes. Knockout of *PRDM*16 in mice results in a loss of brown fat and promotes muscle differentiation [60]. The *ZNF*408 gene (CDS=603 bp) is pseudogenized and almost fixed (allele frequency = 0.88) in the Daweishan chicken population, due to a “G to A" substitution at position 56 of the CDS, resulting a stop-gain (Figure 3-13d), but it is functional in the three other indigenous chickens and GRCg6a but missing in GRCg7b and GRCg7w. *ZNF*408 encoding a ten tandem zinc finger transcriptiona factor

plays a role in the development of vasculature [61]. Morpholino-induced knockdown of *ZNF*408 in zebrafish causes defects in developing retinal and trunk vasculature [62]. However, the effects of loss of function of these genes on chicken phenotypes during the course of domestication and evolution remain to be elucidated.

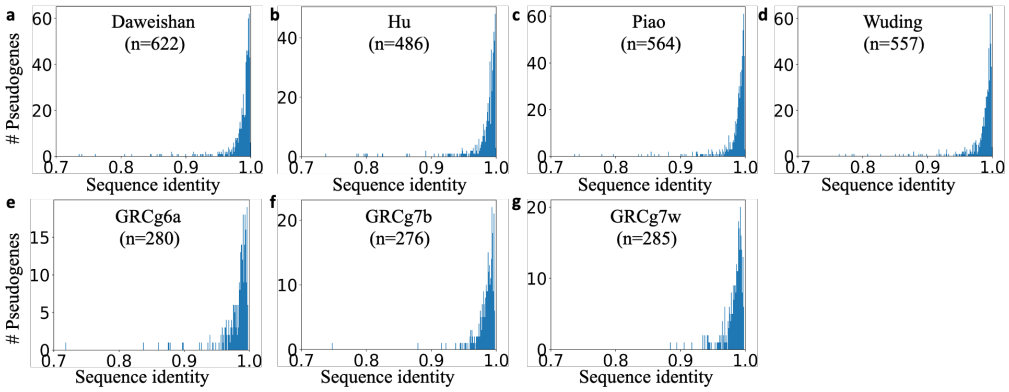

Figure 3-11 Most pseudogenes arise recently. **a~g.** Number of pseudogenes in each of the seven genomes with the indicated identity with their reference genes.

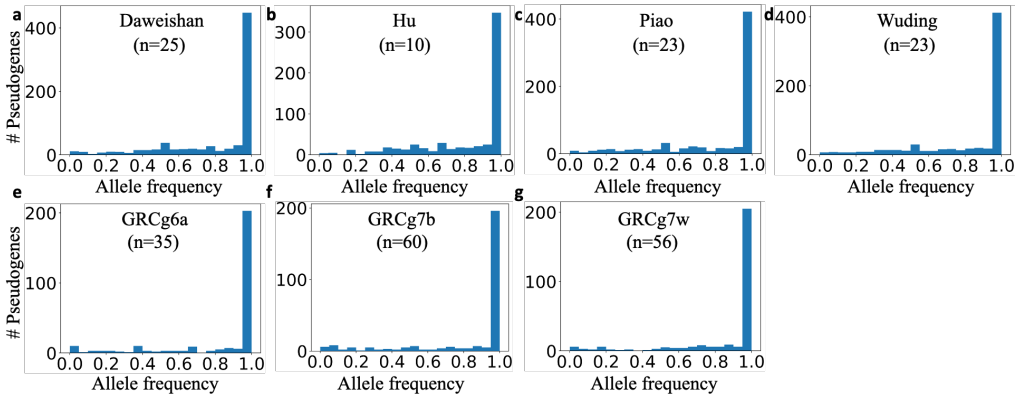

Figure 3-12 Most pseudogenes are fixed in the populations. **a~g.** Number of pseudogenes with the indicated pseudogenization mutation in the seven chicken populations (n= 25, 10, 23, 23, 35, 60 and 56 for the Daweishan, Hu, Piao, Wuding, RJF, broiler and layer populations, respectively).

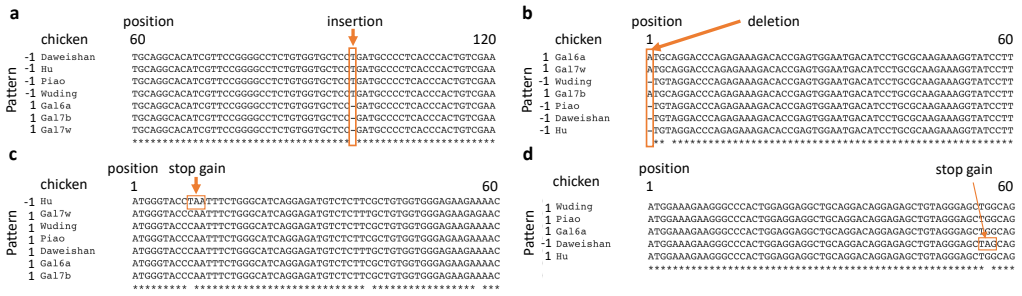

Figure 3-13 Examples of fixed or nearly fixed pseudogenes in chicken populations. **a.** The OAZ2 gene (CDS length=573bp), which is functional in GRCg6a, CRCg7b and CRCg7w, is pseudogenized and fixed in the four indigenous chickens, due to an insertsion of a "T" after position 96 of the CDS, resulting an ORF shift. **b.** The PDCL3 gene (CDS length=723bp), which is functional in GRCg6a, CRCg7b and CRCg7w, is pseudogenized and fixed in the four indigenous chickens, due to deletion of an "A" at the first position of the CDS, resulting an ORF shift. **c.** The PRDM16 gene (CDS=3,828bp), which is functional in the six other chickens, is pseudogenized and fixed in the Hu chicken population, due to a "C to T" substitution at position 10 of the CDS, resulting a stop-gain. **d.** The ZNF408 gene (CDS=603bp), which is functional in the three other indigenous chickens and GRCg6a but missing in GRCg7b and GRCg7w, is pseudogenized and almost fixed (allele frequency=0.88) in the Daweishan chicken population, due to a "G to A" substitution at position 56 of the CDS, resulting a stop-gain.

3.3.16 Loss-of-function mutations affect many important biological pathways

In total, we find 2,293 pseudogenes that appear in at least one of the seven chicken genomes, and 1,365 of which only appear in the four indigenous chickens (Supplementary Table 3-28). These pseudogenes form multiple distinct clusters according to their occurring patterns in the seven chicken genomes. Pseudogenes in the clusters affect critical biological pathways based on GO term assignments to their functional reference genes in RJF or human (Figure 3-14). For instance, some of the 156 pseudogenes that are only shared by the four indigenous chickens affect the hypoxia response via HIF-1 (hypoxia-inducible factor-1) activation pathway; some of the 123 pseudogenes that are only shared by GRCg6a, GRCg7b or GRCg7w affect inotropic glutamate receptor pathway; some of the 247 pseudogenes that are largely unique to Wuding chicken affect the p53 pathway by glucose deprivation; some of the 232

pseudogenes that are largely unique to Piao chicken affect the gonadotropin releasing hormone receptor pathway; some of the 175 pseudogenes that are unique to Daweishan chicken affect the Wnt signaling pathway; and some of the 168 pseudogenes that are unique to Hu chicken affect the insulin/IGF (insulin-like growth factor) pathway-protein kinase B signaling cascade pathway. The pathways affected by the pseudogenes in each chicken are summarized in Supplementary Table 3-33. Among the affected pathways of the indigenous chickens, four are shared by the four chickens, and 3~11 are unique to each of them (Figure 3-14, Supplementary Table 3-33). Additionally, as shown in Figure 3-14c, all the affected pathways in RJF (GRCg6a), the broiler (GRCg7b) and the layer (GRCg7w) are a subset of the union of the affected pathways in the four indigenous chickens, except for the "5-Hydroxytryptamine degradation" pathway in RJF. These pathways might be good candidates for experimental investigation for their roles in the domestication process of the chickens through loss-of-function mutations.

3.3.17 Pseudogenization occurring patterns reflect chicken evolutionary history

Interestingly, most pseudogenes in GRCg6a, GRCg7b and GRCg7w have neither an intact nor pseudogenized copies in the four indigenous chickens (Figure 3-14a), i.e., they are complete lost in the indigenous chickens. In contrast, most pseudogenes in the four indigenous chickens have either an intact or pseudogenized copies in GRCg6a, GRCg7b and GRCg7w (Figure 3-14a). Moreover, the four indigenous chickens share more pseudogenes with each other than with GRCg6a, GRCg7b or GRCg7w, and vice versa (Figures 3-14d and Supplementary Table 3-28). These results suggest again that the indigenous chickens are evolutionarily closer to one another than to GRCg6a, GRCg7b and GRCg7w. To further infer the evolutionary patterns of pseudogenes in the seven

chicken genomes, we constructed a neighbor-joining phylogenetic tree, using the occurring patterns (yes or no) of the pseudogenes. As shown in Figure 3-14e, GRCg7b and GRCg7w form a clade, and the four indigenous chickens form another clade, consistent with the tree constructed using 6,744 essential protein-coding genes in the seven chicken genomes (Figure 3-7c). Therefore, the occurring patterns of pseudogenes in the chickens largely reflect their evolutionary relationships. This result is in contrast to earlier reports that null mutations failed to segregate between wild and domestic chickens or between broilers and layers, and thus selection for loss-of-function mutations had a little role in chicken domestication [12, 13].

Of the 2,293 pseudogenes, 1,919 (83.7%) have a functional copy in at least one of the seven chicken genomes, suggesting that they might lose their functions in a chicken after the divergence from their common ancestor (Supplementary Table 3-28). Of the remaining 374 (16.3%) pseudogene lacking a functional copy in any of the seven chickens, 21 are pseudogenized in all the seven chickens (Supplementary Table 3-28). Interestingly, six of these 21 pseudogenes are caused by the same mutation at the same position of the parent genes in the seven chickens (examples are shown in Figure 3-15), suggesting that they might lose their functions before the divergence from their common ancestor. The remaining 15 are caused by the same or different types of mutations at different positions of the CDSs in the seven chickens (examples are shown in Figure 3-15), suggesting that they might be pseudogenized by convergent evolution. Moreover, we failed to detect either intact or pseudogenized forms of 1,791 (78.1%) of the 2,293 pseudogenes in at least one of the seven chickens (Supplementary Table 3-28), suggesting that they might be completely lost in the relevant chickens. Taken together,

these results suggest that loss-of-function mutations reflect the evolutionary history of the chickens. Thus, pseudogenization might play a critical role in chicken domestication and evolution.

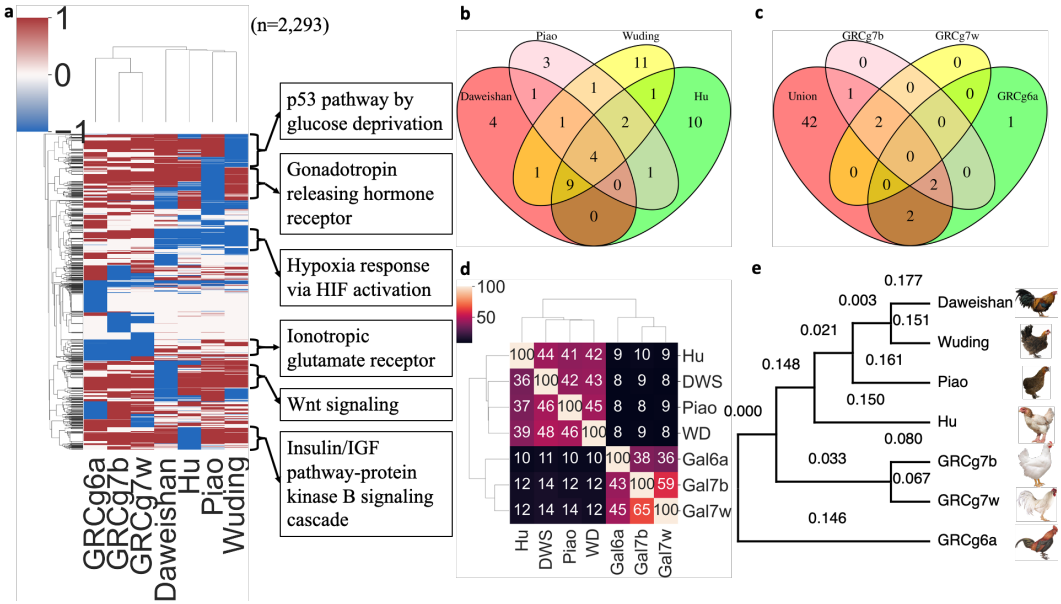

Figure 3-14 Loss-of-functions accompany chicken domestication and evolution. **a.** Heatmap of two-way hierarchical clustering of the pseudogenes based on their appearance as an intact form (1, brown), absence (0, white) and as a pseudogenized form (-1, blue) in the seven chicken genomes. **b.** Venn diagram of the pathways affected by pseudogenes of the four indigenous chickens. **c.** Venn diagram of pathways affected by pseudogenes of the GRCg6a, GRCg7b and GRCg7w genomes compared with the union of the pathways affected by pseudogenes of the four indigenous chickens. **d.** Heatmap of two-way hierarchical clustering of the seven chickens based on the percentage of the pseudogenes that each chicken (rows) share with the others (column). **e.** Neighbor-joining phylogenetic tree, constructed using the occurring patterns of the pseudogenes in the seven chicken genomes.

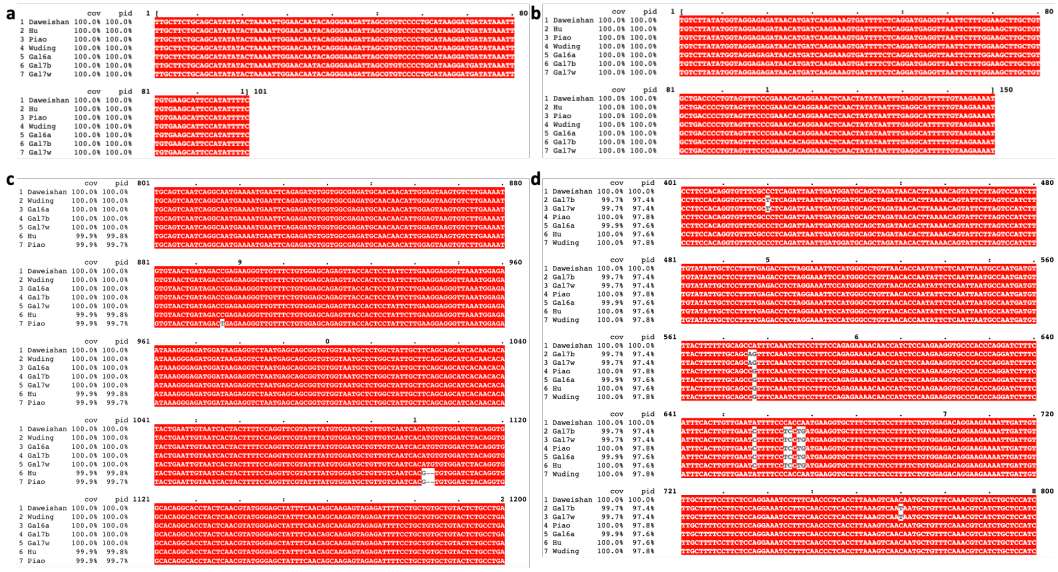

Figure 3-15 Examples of pseudogenes that appear in all the seven chickens. Pseudogenes LOC112531613 (a) and LOC112531630 (b) have the same mutation in the seven chicken genomes, while pseudogenes AMY1AP (c) and CYP2J24P (d) show at least two different mutations in the seven chicken genomes.

3.4 Discussion

Although multiple versions of the RJF genome (GRCg1~6a) [1-3] and commercial chicken genomes (GRCg7b/w) [5] have been released, no high-quality indigenous chicken genome has been assembled. Using a combination of short, long and Hi-C reads, we assembled all the chromosomes of four indigenous chickens with high-quality comparable to the GRCg7b/w assemblies. These assemblies allow us to uncover a large number of new genes (1,420) using RNA-seq reads collected from various tissues of the four indigenous chicken breeds. Most (39 or 92.86%) of the 42 randomly selected putative new genes can be verified by RT-qPCR experiments in multiple tissues of at least one breed, suggesting that they are likely authentic. Interestingly, 1,142 (80.4%) of the 1,420 new genes also are encoded in GRCg6a, GRCg7b or GRCg7w assemblies (Supplementary Table 3-12). The ubiquitous presences of these new genes and their affected GO pathways (Supplementary Table 3-18) indicate that they might play crucial

roles in house-keeping functions of chickens. Only 246 (18.4%) of the 1,335 new genes recently reported by Li et al [9] overlap our predicted new genes. The discrepancy might be due to different numbers of chickens/breeds used and different definitions of a gene adopted. Moreover, we identified 48 of the 274 missing genes in birds in general and 36 of another 174 missing genes in chicken in particular in at least one of the four assembled indigenous genomes (Supplementary Table 3-22). We also uncovered three of the 274 missing genes in birds (Supplementary Table 3-22A) by mining unmapped RNA-seq reads. Our ability to uncover these new genes and missing genes indicates that the approach that we used to sequence, assemble and annotate the indigenous chicken genomes have largely overcome the limitations of the earlier methods. However, our assembled chr16 and some other micro-chromosomes are still not complete enough, and none of our 1,420 new genes recall any of the 274 missing genes in birds in general and additional 174 missing genes in chickens in particular. Therefore, if the remaining presumed missing genes are encoded in the chicken genomes, accurate-enough ultra-long-reads (>100 kbp) from new sequencing technologies that allows the recent telomere to telomere assembly of a hypoploid human genome [63] might be needed to completely assemble the chicken genomes, thereby recovering the still missing genes.

By counting the 1,420 new genes and recovered missing genes, we annotate 17,497~17,718 protein-coding genes in the four indigenous chicken genomes, and increase the numbers of protein-coding genes in GRCg6a, GRCg7b and GRCg7w to 18,463, 19,002, 18,978, respectively. Considering many gene-rich micro-chromosomes are still not fully assembled, the number of annotated genes may increase further once all the micro-chromosomes are fully assembled. Thus, the chicken genomes might encode

similar number of protein-coding genes as other tetrapods (18,000~25,000) [4, 5] such as humans [46], as previously suggested for birds in general [47]. The much larger number of protein-coding genes in the two commercial chickens than in RJF and indigenous chickens might be the results of the breeding programs used to produce the fast-growth and high egg-laying terminal commercial lines for production by taking advantage of hybrid vigor through a series of gene introgression from multiple purebred populations generated by intensive artificial selection [64-67].

Unexpectedly, we found 606~747 pseudogenes in each of the four indigenous chicken genomes. Meanwhile, based on our predicted 1,420 new genes, we also substantially increased the number (435~542) of pseudogenes in the three (GRCg6a and GRCg7b/w) earlier assembled chicken genomes. Only 50~80 of the pseudogenes found in each genome are processed, while most (385~684) of them are not. The small numbers of processed pseudogenes found in the chicken genomes are consistent with the previous results [68, 69], presumably because the chicken's LINE1 like CR1 (L1) elements lack retro transposase activity [68, 69]. However, the large number of non-processed pseudogenes that we found in each chicken genome is in stark contrast to the earlier findings that only a few dozen non-processed pseudogenes were found in human population [70, 71], not to mentioning in a human individual. In a more recent study [72], 165 and 303 non-processed pseudogenes were found in mouse and human populations, respectively. However, these numbers are still much smaller than those we found in the chicken genomes. Thus, we observed a large scale of non-processed pseudogenization in chickens, particularly in the four indigenous chickens, which has not been reported in any tetrapods, to the best of our knowledge.

We provide two pieces of evidence that the pseudogenes are no longer functional as a protein coding gene. Firstly, pseudogenes have elevated dN/dS ratios compared with true genes, no matter where the first pseudogenizations occur along the CDSs, indicating that purifying selection is relaxed on the pseudogenes. Secondly, while synonymous mutation rates are largely uniformly distributed along the CDSs of true genes, they decrease at the two-ends of the CDSs, suggesting that both ends are under stronger purifying selection than the middle parts of the CDSs. However, this is not seen in the two ends of pseudogenes, indicating again that the purifying selection is relaxed on the pseudogenes.

We provide three pieces of evidence that loss of function of genes through pseudogenization is a result of natural/artificial selection. Firstly, unlike synonymous mutations along the true genes and pseudogenes, both of which are largely uniformly distributed along the CDSs as expected for neutral mutations, the first pseudogenization mutations in pseudogenes are strongly biased to the two ends of CDSs. Such biases facilitate eliminating the functions of genes. It is well-known that a promoter can extend into the 5'-end of the CDS, thus mutations in the region may disrupt the promoter of the gene [73]. Moreover, the closer a pseudogenization mutation is toward the 5'-end of the CDS, the greater impact of the mutation could have on the gene function and the more likely the gene would lose its function. Although pseudogenizations at the 3'-ends of CDSs can potentially produce at least partially functional proteins, this is unlikely for at least most of the pseudogenes that we found in the indigenous chicken genomes, since dN/dS ratios for pseudogenes with the first pseudogenization sites occurring in the last 25% and in the first 25% of the CDSs are not significantly different, and both are

significantly higher than those for true genes. In other words, pseudogenizations in the 3'-ends of CDSs can be as efficient as pseudogenizations in the other parts of the CDSs to eliminate the functions of genes. We found that 3'-ends of CDSs might harbor miRNA binding sites, and pseudogenization could disrupt such binding sites, which might change post-transcriptional regulation, and thus, the functions of genes. However, this mechanism can explain a few (0.5%) cases of possible loss-of-function due to pseudogenization at the 3'-ends of CDSs. Thus, missing or disruption of critical amino acids at the 3'-end of CDSs might be the major mechanism of loss-of-function due to pseudogenization at the 3'-ends of CDSs. Secondly, most pseudogenization mutations are fixed in the indigenous chicken, RJF, broiler and layer populations (Figure 3-12), likely due to natural/artificial selection. Most of the pseudogenes might arise after the divergence of the chickens, while only few might lose their functions before the divergence of the chickens from their common ancestor (Figure 3-11). Moreover, most of the pseudogenes are completely lost in at least one of the seven chickens, i.e., fixation of null alleles (Supplementary Table 3-28), suggesting that these genes are not essential for chickens. Thirdly, the patterns of pseudogenization segregate the four indigenous chickens from the two commercial ones and RJF (Figure 3-14e), which is in line with the phylogenetic tree of more 6,000 essential protein-coding genes of the chickens (Figure 3-7c). Taken together, these results suggest that loss-of-function mutations accompany the formation of the chicken breeds, and thus might play critical roles in chicken domestication and evolution. This conclusion is in contrast to an earlier report that loss-of-function mutations played a little role in chicken domestication [12].

Although it has been shown that deleterious mutations might play a role in plants [74, 75] and animals [76, 77] domestication, lost-of-function mutations are not necessarily deleterious. In fact, it has been well documented that loss of certain genes might be the results of adaptation of birds for flight [78-81], of beef cattle for artificial selection [82] and of humans for new abilities [70]. It has been proposed that loss-of-function mutations may be an important factor in rapid evolution as occurred during domestication—the "less is more" hypothesis [83]. The earlier conclusion that fixation of null alleles is not a common mechanism for phenotypic evolution in chicken domestication[12, 13] might be incorrectly drawn because of the low quality of earlier chicken genome assemblies, leading to the failure to detect inactivating mutations such as pseudogenization [84].

Interestingly, there are more pseudogenes in the four indigenous chickens (556~684) than in the two commercial chickens GRCg7b (415) and GRCg7w (385), mainly because many genes that are functional in the commercial chickens are either pseudogenized or completely lost in the indigenous chickens (Supplementary Tables 3-24~3-27). The discrepancy might be due to the domestication processes of the indigenous chickens in their unique ecological niches and artificial environments as well as to the special ways that the terminal commercial chicken lines for production are produced [64-67]. Specifically, large scale pseudogenization might favor the domestication of the indigenous chickens in their unique ecological niches and artificial environments, while the purebred parental lines of commercial chickens, which are produced by intensive artificial selection for fast growth and high egg-laying traits of their offspring through hybrid vapor, might harbor fewer pseudogenes. Although the pseudogenes in the four

indigenous chickens are not essential in their unique ecological niches and artificial environments, it is highly likely that many of their functional versions in the commercial breeds might be crucial for their high productive capabilities in industrialized artificial environments.

3.5 Conclusion

By combining short, long and Hi-C reads, we assembled the genomes of four indigenous chickens with the highest continuity for chicken genome assembly so far. We also assembled the missing chromosomes (chr29 and chr34~chr39) and corrected mis-assembled chromosomes (chr31~chr33) for the RJF genome (GRCg6a). We identified a total of 1,420 new genes in the four indigenous chicken genomes. Most of these new genes also are encoded in the earlier assembled RJF (GRCg6a) and commercial chicken genomes (GRCg7b/w), and might be involved in important biological processes. Counting these new genes, chicken genomes encode similar numbers of protein-coding genes as other tetrapods. Moreover, we identified a total of 2,293 pseudogenes in the seven chicken genomes, affecting many important biological processes. The occurring patterns of these pseudogenes in the genomes suggest that loss-of-function mutations might play critical roles in chicken evolution and domestication.

CHAPTER 4 Artificial selection footprints in domestic chicken genomes

4.1 Background

Chicken (*Gallus gallus*) has been domesticated by human for about 8,000 years [7], and multiple lines of evidence show that red jungle fowl (RJF) is the major ancestor of domestic chicken all over the world [6, 7, 85]. Artificial selection has resulted in numerous chicken breeds with distinct traits in different parts of the world for various purposes, including meat and egg production as well as recreation and ornament. Particularly, intensive systematic artificial selections carried out in a few companies in the last decays have led to highly production-efficient commercial broiler and layer lines used all over the world. Understanding the genetic basis of distinct traits of traditionally bred indigenous chicken as well as of commercialized broilers and layers is crucial to guide breeding programs to further improve the commercial lines and chicken welfare [86]. Besides commercial lines, indigenous chicken breeds are also excellent model systems to study the relationships between genotypes and phenotypes [87]. Indeed, many studies have been done to reveal artificial selection signatures on commercial broilers and layers [8, 12, 65, 88-91] as well as on indigenous chickens [92-98]. These studies have identified genes or quantitative trait loci (QTLs) related to specific traits such as body size [93, 94, 99-104], meat quality [105-107], egg production [108], feathering [109-116], plumage color [117-120], skin color [121], behaviors [122, 123], immunity [124], crest shapes [125], bone traits [126-128], rumpless trait [129-132], and polydactyly [133, 134]. However, genetic bases of many artificially selected traits, in particular, of indigenous chickens, are far from being fully understood. Yunnan, a southwest province of China, is one of the major centers where domestic chickens arise [6], and numerous

chicken breeds have been raised in mountainous areas there. Among these indigenous chicken breeds are Daweishan, Hu, Piao, Wuding and Nine-claw chicken, each with distinct traits. Specifically, Daweishan chickens have a miniature body size (0.5~0.8kg for female and 0.8~1.2kg for male adults); Hu chickens have a large body size (3kg for female and 6kg for male adults) with extraordinarily stout legs; Piao chickens have a short tail (a rumpless phenotype); Wuding chickens have a relatively large body size with colorful feathers and thick fat; and Nine-claw chickens have nine claws with a middle-sized body.

To understand the domestication and genetic basis of the distinct traits of these indigenous chickens, we have re-sequenced 25 Daweishan chickens, 10 Hu chickens, 23 Piao chickens, 23 Wuding chickens and four Nine-claw chickens. By comparing the single nucleotide polymorphisms (SNPs) of these indigenous chicken populations with those of 35 RJF individuals as well as of 60 broiler individuals and 56 layer individuals, we were able to find more artificial selection signatures on the indigenous chickens, broilers and layers using a rigorous statistic model [135] than previously reported [8, 12, 65]. By comparing the selection signatures between the indigenous chicken breeds, RJF, broilers and layers, we have found numerous genomic regions and genes related to the breed-specific traits.

4.2 Methods and materials

4.2.1 Re-sequencing short reads from NCBI SRA

We downloaded genomic short reads for two broiler lines from NCBI Sequence Read Archive (SRA): “Broiler A” (n=40, access number PRJEB15276) and “Broiler B” (n=20, access number PRJEB30270). Broiler A was originally from France, and Broiler

B was from the company Indian River International. We downloaded DNA short reads for three layer lines from NCBI SRA: “Layer A” (n=25, access number PRJEB15189) were white egg layers, “Layer B” (n=25, access number PRJEB30270) were brown egg layers, and “Layer C” (n=6, access number PRJEB30270) were crossbred layers. We downloaded genomic short reads for two RJF populations from NCBI SRA: “RJF A” (n=25, access number PRJEB3027) were from northern Thailand, and “RJF B” (n=10, access number PRJEB3027) were from India.

4.2.2 Re-sequencing of indigenous chicken samples

We re-sequenced 85 indigenous chicken individuals from the Experimental Breeding Chicken Farm of the Yunnan Agricultural University (Yunnan, China), including 25 Daweishan chickens aged 10 months (nine males, 16 females), 10 Hu chickens aged seven months (five males, five females), 23 Piao chickens aged 10 months (11 males, 12 females), 23 Wuding chickens aged 10 months (11 males, 12 females) and four Nine-claw chickens aged 10 months (two males, two females).

4.2.3 Short-reads DNA sequencing

Two milliliters of blood were drawn from the wing vein of each chicken in a centrifuge tube containing anticoagulant (EDTA-2K) and stored at -80℃ until use. Genomic DNA (10μg) in each blood sample was extracted using a DNA extraction kit (DP326, TIANGEN Biotech, Beijing, China) and fragmented using a Bioruptor Pico System (Diagenode, Belgium). DNA fragments around 350 bp were selected using SPRI beads (Beckman Coulter, IN, USA). DNA-sequencing libraries were prepared using Illumina TruSeq® DNA Library Prep Kits (Illumina, CA, USA) following the vendor’s

instructions. The libraries were subject to 150 cycles paired-end sequencing on an Illumina Novaseq 6000 platform (Illumina, CA, USA) at ~30 coverage.

4.2.4 Real-time quantitative PCR (RT-qPCR) analysis

RT-qPCR was performed using the Bio-Rad CFX96 real-time PCR platform (Bio-Rad Laboratories. lnc, America) and SYBR Green master mix (iQTM SYBRGreen ® Supermix, Dalian TaKaRa Biotechnology Co. Ltd. Add). The primers of the eight genes are listed in Supplementary Note 4-1. The β-actin gene was used as a reference. Primers were commercially synthesized (Shanghai Shenggong Biochemistry Company P.R.C). Each PCR reaction was performed in 25 μl volumes containing 12.5 μl of iQ™ SYBR Green Supermix, 0.5 μl (10 mM) of each primer, and 1 μl of cDNA. Amplification and detection of products was performed with the following cycle profile: one cycle of 95 °C for 2 min, and 40 cycles of 95 °C for 15 s, annealing temperature for 30 s, and 72 °C for 30 s, followed by a final cycle of 72 °C for 10 min. The specificity of the amplification product was verified by electrophoresis on a 0.8% agarose gel and DNA sequencing. The $2^{-\Delta Ct}$ method was used to analyze mRNA abundance. All samples were analyzed with at least three replicates, and the mean of these measurements was used to calculate mRNA expression.

4.2.5 Variant calling

We mapped the short reads of each individual chicken to the reference genome (GRCg7b) using BWA (0.7.17) [29] and SAMtools (1.9) [30] with the default settings and called variants for each individual using GATK-HaplotypeCaller (4.1.6) [43] with the default settings. After generating the GVCF files for each individual, we computed allele frequencies in the same chicken breed using the GATK-CombineGVCFs (4.1.6)

tool [43]. For each chicken breed, we removed variants with Quality by depth (QD) < 2, Fisher strand (FS) > 60, Root mean square mapping quality (MQ) < 40, Strand odd ratio (SOR) > 3, Rank Sum Test for mapping qualities (MQRankSum) < -12.5 and Rank Sum Test for site position within reads (ReadPosRankSum) < -8 for SNPs and QD < 2, FS > 200, SOR > 10, Likelihood-based test for the consanguinity among samples (InbreedingCoeff) < -0.8 and ReadPosRankSum < -20 for indels.

4.2.6 Functional annotation of variants

We used the package ANNOVAR [136] to annotate the variants according to their locations on the reference genome into seven categories including 1) intergenic regions, 2) intronic regions, 3) coding regions (synonymous, nonsynonymous, stop gain and stop loss), 4) up/downstream of a gene, 5) splicing sites, 6) 5' untranslated regions (5'UTRs) and 3' UTRs, and 7) non-coding RNA regions. We used the tool Ensembl Variant Effect Predictor (VEP) [137] to predict the impact of amino acid-altering SNPs.

4.2.7 Detection of selective sweeps

The selective sweeps were detected using two methods including genetic differentiation (F_{ST}) [138] and nucleotide diversity (π). We estimated F_{ST} for each comparison of two chicken populations using VCFtools (0.1.16) [139] with a sliding window size 40 kb and a step size 20 kb. We estimated π for each group using VCFtools (0.1.16) [139] with a sliding window size 40 kb and a step size 20 kb, and calculated the absolute value of the difference in nucleotide diversity ($|\Delta\pi|$) in each window for each comparison of two populations. For both F_{ST} and $|\Delta\pi|$, we only used the bi-allelic SNPs on autosomes and sex chromosomes for the analysis. To evaluate the statistic significance of the F_{ST} and π values for a comparison, we generate a Null model by shuffling the

allele frequency data for 100 times while keeping the SNP positions fixed [135]. We then computed F_{ST} and $|\Delta\pi|$ for the permuted windows as well as their means ($\mu(F_{ST\,Null})$ and $\mu(|\Delta\pi|_{Null})$) and standard deviations ($\sigma(F_{ST\,Null})$ and $\sigma(|\Delta\pi|_{Null})$). We computed the Z value for each F_{ST} and $|\Delta\pi|$ values for a comparison by using the following formulas:

$ZF_{ST}(i) = (F_{ST}(i) - \mu(F_{ST\,Null})/\sigma(F_{ST\,Null})$ and

$Z|\Delta\pi(i)| = (|\Delta\pi(i)| - \mu(|\Delta\pi|_{Null})/\sigma(|\Delta\pi|_{Null})$.

We consider a window with either $ZF_{ST}(i) > 6$ or $Z|\Delta\pi| > 3.09$ (P-value < 0.001) to be a putative selective sweep. Since adjacent putative selective sweep windows may overlap with each other, we merged adjacent windows if they overlapped by at least one nucleotide to obtain the discrete selective sweeps (DSSs) for each comparison.

4.2.8 Selective sweeps analysis

To reveal selective sweeps in the domestic chicken populations, we conducted a total of 19 comparisons, including 1) Daweishan chickens VS. RJF, 2) Hu chickens VS. RJF, 3) Piao chickens VS. RJF, 4) Wuding chickens VS. RJF, 5) Nine-claw chickens VS. RJF, 6) Broilers VS. RJF, 7) Layers VS. RJF, 8) Indigenous chickens (Daweishan chickens + Hu chickens + Piao chickens + Wuding chickens + Nine-claw chickens) VS. RJF, 9) Commercial chickens (layers + broilers) VS. RJF, 10) Indigenous chickens VS. Commercial chickens, 11) Daweishan chickens VS. the other six domestic chicken breeds, 12) Hu chickens VS. the other six domestic chicken breeds, 13) Piao chickens VS. the other six domestic chicken breeds, 14) Wuding chickens VS. the other six domestic chicken breeds, 15) Nine-claw chickens VS. the other six domestic chicken breeds, 16) Broilers VS. the other six domestic chicken breeds, 17) Layers VS. the other

six domestic chicken breeds, 18) Daweishan chickens VS. chickens with relatively large body size (Hu chickens, Wuding chickens and Broilers), 19) Broilers VS. Layers.

4.3 Results

4.3.1 Indigenous chicken breeds have higher nucleotide diversity

By using the re-sequencing short reads of 25 Daweishan chickens, 10 Hu chickens, 23 Piao chickens, 23 Wuding chickens, four Nine-claw chickens, 60 broilers, 56 layers and 35 RJFs, we detected 20, 16, 22, 19, 13, 17, 14 and 22 million single nucleotide variants (SNVs) and small indels in the eight chicken groups (Table 4-1), respectively, using the GRCg7b assembly as the template. After the redundant variants in the different groups were removed, there are 33 million variants totally. There are 17, 13, 19, 16, 11, 14, 12 and 19 million bi-allelic SNPs on autosomes and sex chromosomes in the groups of Daweishan, Hu, Piao, Wuding, Nine-claw chickens, broilers, layers and RJFs, respectively (Table 4-1). After removing the redundant ones in the different groups, we ended up with 26 million bi-allelic SNPs on autosomes and sex chromosomes, and our subsequent analyses will be focused on these bi-allelic SNPs in each chicken group (Table 4-1).

When analyzing the mean nucleotide diversity (π) of each chicken group, we found the values of the broilers and the layers are smaller than those of the five indigenous chickens and RJFs (Table 4-1), indicating that the commercial lines are genetically more uniform than the indigenous lines and RJFs, as expected. The relatively low π values of commercial lines might be due to their close mating and linked selection [65]. At the same time, we detected 30 and 19 million SNPs on autosomes and sex chromosomes within the indigenous chicken group (85 individuals totally) and the

commercial chicken group (116 individuals totally), respectively, of which, 23 and 16 million were bi-allelic SNPs for the two chicken groups, respectively (Table 4-1). Consistent with the aforementioned results, the mean π values of the indigenous group and the RJFs group are higher than that of the commercial group (Table 4-1), suggesting that artificial selection in commercial lines is more intensive than in indigenous chickens. Unexpectedly, the mean π value of the indigenous group is higher than that of the RJFs group with two different origins (India and Thailand, Materials and methods, Table 4-1), which might be due to some level of inbreeding on farms of the RJFs since their ancestors' captures.

Table 4-1 Summary of genetic variants in the chicken groups

	Daweishan	Hu	Piao	Wuding	Nine-claw	Broiler	Layer	RJF	Indigenous	Commercial
# Individuals	25	10	23	23	4	60	56	35	85	116
# Variants	20,387,469	15,708,834	21,993,907	19,468,846	12,946,552	16,747,786	14,390,952	22,295,197	29,749,839	18,738,449
# Bi-allelic SNPs on autosomes and sex chromosomes	17,267,774	13,446,705	18,646,530	16,484,447	11,225,485	14,162,751	12,096,556	19,053,911	23,405,295	15,692,214
Mean Pi	4.20E-03	3.90E-03	4.40E-03	3.80E-03	4.20E-03	3.40E-03	2.30E-03	3.80E-03	4.30E-03	3.10E-03

4.3.2 Variants are enriched in non-coding regions

Based on the location of the bi-allelic SNPs, we classified them into seven categories, including intergenic (variants in intergenic regions), intronic (variants in introns), up/downstream (variants within a 1kb region up/downstream of transcription start/end sites), splicing (variants within 2 bp of a splicing junction), UTR 3'/UTR 5' (variants in 3'/5' untranslated regions), ncRNA (variants in non-coding RNA genes) and coding (variants in coding sequences). The relative abundances of the bi-allelic SNPs in each chicken group are shown in Table 4-2. Specifically, of the bi-allelic SNPs in each chicken group, 29.71%~30.46% fall within intergenic regions, 50.44%~51.57% are located in intronic regions, 2.86%~2.97% fall within up/downstream regions, 4.08%~4.24% are located in 3' UTR/5' UTR regions, 0.01% fall within splicing regions,

10.09~10.18% are located in ncRNA regions, and 1.53%~1.80% fall within coding regions (Table 4-2). Therefore, only a small portion (1.53%~1.80%) of the bi-allelic SNPs fall within coding regions, while the vast majority (98.20%~98.47%) are located in non-coding regions. As non-coding regions comprise 96.92% of the reference chicken genome (GRCg7b assembly), the SNPs are enriched in non-coding regions relative to in coding regions.

Among the bi-allelic SNPs in coding regions of each chicken group, 33.33%~38.33% (Daweishan = 35.71%, Hu = 34.55%, Piao = 36.47%, Wuding = 35.71%, Nine-claw = 33.33%, broiler = 34.97%, layer = 35.22%, RJF = 35.80%, indigenous = 38.33%, commercial = 36.36%) are amino acid-altering (AA-altering, i.e., nonsynonymous and stop-gain/loss) SNPs (Table 4-2). Among the nonsynonymous SNPs in each chicken group, most (Daweishan = 86.44%, Hu = 87.50%, Piao = 85.24%, Wuding = 86.44%, Nine-claw = 88.00%, broiler = 87.50%, layer = 87.27%, RJF = 85.96%, indigenous = 63.24%, commercial = 69.49%) are tolerant SNPs and only a small proportion (Daweishan = 13.56%, Hu = 12.50%, Piao = 14.76%, Wuding = 13.56%, Nine-claw = 12.00%, broiler = 12.50%, layer = 12.73%, RJF = 14.04%, indigenous = 36.76%, commercial = 30.51%) are intolerant, which might be harmful variants and thus under purifying selection in the chicken group (Table 4-2).

Table 4-2 Functional annotation of genetic variants in each chicken group

	Daweishan	Hu	Piao	Wuding	Nine-claw	Broiler	Layer	RJF	Indigenous	Commercial
# SNPs in intergenic region (%)	5,192,013 (30.07)	3,995,687 (29.71)	5,636,381 (30.23)	4,963,186 (30.11)	3,353,149 (29.87)	4,237,766 (29.92)	3,628,162 (29.99)	5,803,793 (30.46)	7,107,445 (30.37)	4713,113 (30.03)
# SNPs in Intronic region (%)	8,802,652 (50.98)	6,901,461 (51.32)	9,465,663 (50.76)	8,398,955 (50.95)	5,788,439 (51.57)	7,258,352 (51.25)	6,201,733 (51.27)	9,670,335 (50.75)	11,805,669 (50.44)	8,014,790 (51.07)
# SNPs in up/downstream (%)	513,016 (2.97)	398,721 (2.97)	552,450 (2.96)	489,473 (2.97)	320,904 (2.86)	419,495 (2.96)	355,747 (2.94)	548,631 (2.88)	694,514 (2.97)	464,953 (2.96)
# SNPs in UTR 3'/UTR 5' (%)	719,457 (4.17)	559,799 (4.16)	778,556 (4.18)	683,119 (4.14)	458,168 (4.08)	587,352 (4.15)	495,415 (4.10)	786,835 (4.13)	992,584 (4.24)	652,061 (4.16)
# SNPs in splicing region (%)	1,166 (0.01)	859 (0.01)	1,258 (0.01)	1,150 (0.01)	662 (0.01)	1,055 (0.01)	866 (0.01)	1,276 (0.01)	1,791 (0.01)	1,265 (0.01)
# SNPs in ncRNA (%)	1,748,603 (10.13)	1,368,611 (10.18)	1,894,392 (10.16)	1,672,022 (10.14)	1,132,747 (10.09)	1,428,434 (10.09)	1,222,251 (10.10)	1,935,153 (10.16)	2,382,196 (10.18)	1,587,018 (10.11)
# SNPs in coding region (%)	290,867 (1.68)	221,567 (1.65)	317,830 (1.70)	276,542 (1.68)	171,416 (1.53)	230,297 (1.63)	192,382 (1.59)	307,888 (1.62)	421,096 (1.80)	259,014 (1.65)
# Nonsynonymous-INTOL SNPs (%)	14,069 (0.08)	9,337 (0.07)	16,330 (0.09)	13,415 (0.08)	6,452 (0.06)	10,527 (0.07)	8,594 (0.07)	14,803 (0.08)	57,713 (0.25)	28,147 (0.18)
# Nonsynonymous-TOL SNPs (%)	88,831 (0.51)	66,095 (0.49)	97,721 (0.52)	84,157 (0.51)	49,459 (0.44)	69,943 (0.49)	58,099 (0.48)	93,118 (0.49)	100,729 (0.43)	64,301 (0.41)
# Synonymous SNPs (%)	185,466 (1.07)	144,341 (1.07)	201,060 (1.08)	176,605 (1.07)	114,265 (1.02)	147,808 (1.04)	123,989 (1.02)	197,449 (1.04)	258,938 (1.11)	164,231 (1.05)
# Stop-gain/loss (%)	1,220 (0.01)	862 (0.01)	1,334 (0.01)	1,145 (0.01)	581 (0.01)	962 (0.01)	831 (0.01)	1,290 (0.01)	1,927 (0.01)	1,169 (0.01)
# No prediction (%)	1,281 (0.01)	932 (0.01)	1,385 (0.01)	1,220 (0.01)	659 (0.01)	1,057 (0.01)	869 (0.01)	1,228 (0.01)	1,789 (0.01)	1,166 (0.01)

4.3.3 Indigenous chickens have a high portion of rare nonsynonymous SNPs

We compared the allele frequencies of the SNPs in coding regions in the groups of RJFs, indigenous and commercial populations. As shown in Figure 4-1a, all the three groups have higher portion of rare nonsynonymous SNPs than rare synonymous SNPs, indicating that rare nonsynonymous SNPs tend to be deleterious and thus under purifying selection. The same conclusion has been drawn in an earlier study for commercial chickens [65]. Interestingly, the indigenous chickens have the highest rare allele frequency densities for both nonsynonymous and synonymous SNPs, followed by RJFs and the commercial chickens. The earlier study also noted that RJFs had higher rare allele frequency densities than commercial chickens [65]. Intensive industrial selective breeding of commercial chickens can lead to the loss of rare alleles which might be slightly deleterious, thus the density of rare allele frequency of commercial chickens is the lowest among the three chicken groups. Consistent with our earlier results (Table 4-1), the indigenous chickens had higher rare allele frequency densities than RJFs. On the other hand, the indigenous chickens consist of five chicken breeds with distinct traits from different regions in Yunnan Province, China, thus they have higher nucleotide diversity (Table 4-1) compared with commercial chickens and RJFs, due to harboring more rare alleles (Figure 4-1a). Among the five different indigenous chicken breeds and two commercial chicken breeds, the density of the rare allele frequency of Piao chickens is the highest, and the value of layers is the lowest (Figure 4-2), consistent with their highest and lowest π values, respectively (Table 4-1).

4.3.4 Only a small portion of breed-specific SNPs are fixed

We analyzed the group-specific SNPs in each chicken group, and found that RJFs have the highest number of unique SNPs (2.9 million) among the eight chicken groups (Table 4-3), which is consistent with the finding in the previous study [65], suggesting a loss of ancestral alleles during chicken domestication. Except for Hu chicken and Nine-claw chicken with a small population size (Table 4-1), layers have the lowest number of unique SNPs (455k) and broilers have the second lowest number of unique SNPs (520k) among the eight chicken breeds (Table 4-3), while Daweishan, Piao and Wuding chickens have 1.1, 1.3 and 0.7 million unique SNPs, respectively, indicating once again that genetic diversity of indigenous chickens is higher than those of the layers and the broilers. From 0.83% (RJFs) to 1.39% (Wuding chicken) of the group-specific SNPs are missense mutations (Table 4-3). Most of the group-specific SNPs have allele frequencies lower than 0.5, and only a very small portion (0.05%~0.59%) are close to being fixed (allele frequency > 0.9) in all the eight groups of chickens (Table 4-3). The same is true for the missense SNPs (0.04%~0.48%) (Table 4-3). More details of the group-specific missense SNPs and affected genes are listed in Supplementary Tables 4-1~4-8.

We also compared allele frequency spectrums of the group-specific SNPs in the eight chicken groups (Figure 4-1b). Group-specific alleles of the layers tend to have higher frequencies than those in other groups (except for Hu chicken and Nine-claw chicken with a small population size), consistent with the finding in the earlier study [65]. This might be a result of artificial selection which increase the frequency of favorable alleles in layers.

Table 4-3 Frequencies of group-specific SNPs in the eight groups of chickens

	Daweishan		Hu		Piao		Wuding		Nine-claw		Broiler		Layer		RJF	
Allele frequency	# SNPs	# Missense	# SNPs	# Missense	# SNPs	# Missense	# SNPs	# Missense	# SNPs	# Missense	# SNPs	# Missense	# SNPs	# Missense	# SNPs	# Missense
0~0.1	960,081	11,415	333,183	4,415	1,224,071	14,808	591,131	8,364	0	0	437,138	6,113	329,256	4,896	2,361,887	20,261
0.1~0.2	114,622	1,294	84,008	1,063	78,505	989	83,703	1,051	106,514	1,270	47,251	554	59,783	666	417,060	3,172
0.2~0.3	33,290	396	32,380	399	6,863	133	16,014	201	49,960	487	16,192	175	20,795	237	119,324	790
0.3~0.4	10,522	128	12,865	171	2,534	45	5,418	68	14,077	163	10,124	78	13,923	101	24,587	159
0.4~0.5	4,485	46	7,046	68	2,045	33	2,781	22	6,900	76	5,522	44	11,792	96	5,641	44
0.5~0.6	1,478	17	2,212	19	405	12	1,706	12	0	0	1,339	13	11,469	69	1,184	12
0.6~0.7	926	9	1,796	13	435	6	537	7	2,881	10	725	5	2,549	22	748	11
0.7~0.8	732	12	1,344	12	430	6	410	2	2,136	11	518	4	1,783	9	466	3
0.8~0.9	363	6	991	9	223	5	327	3	1,631	5	336	2	1,165	1	293	2
0.9~1.0	1,173	13	2,827	30	1,001	10	1,199	12	3,783	31	891	4	2,659	16	1,362	10
Sum	1,127,672	13,336	478,652	6,199	1,316,512	16,047	703,226	9,742	187,882	2,053	520,036	6,992	455,174	6,113	2,932,552	24,464

4.3.5 Indigenous chickens are more closely related to one another

To reveal the genetic relationships of the individual chickens, we performed a principal component analysis (PCA) based on occurring patterns of the ~26 million bi-allelic SNPs. As shown in Figure 4-1c, the five indigenous breeds from Yunnan Province, China, are clustered together with RJFs from northern Thailand (RJF A) that is geographically close to Yunnan Province, China, while RJFs from India (RJF B) form a separate cluster nearby. This result suggests that the five indigenous breeds are closed related to one another, and they are also more closely related to RJFs from northern Thailand (RJF A) than to those from India (RJF B). On the other hand, brown egg layers (Layer B) and broilers from Indian River International (Broiler B) form two tight clusters close to the cluster of indigenous chickens and RJF A, while white egg layers (Layer A), crossbred layers (Layer C) and broilers from France (Broiler A) form clusters far away from the cluster of indigenous chickens and RJF A. These results suggest that broilers and layers with different origins have quite different genetic structures although they might have similar productivities.

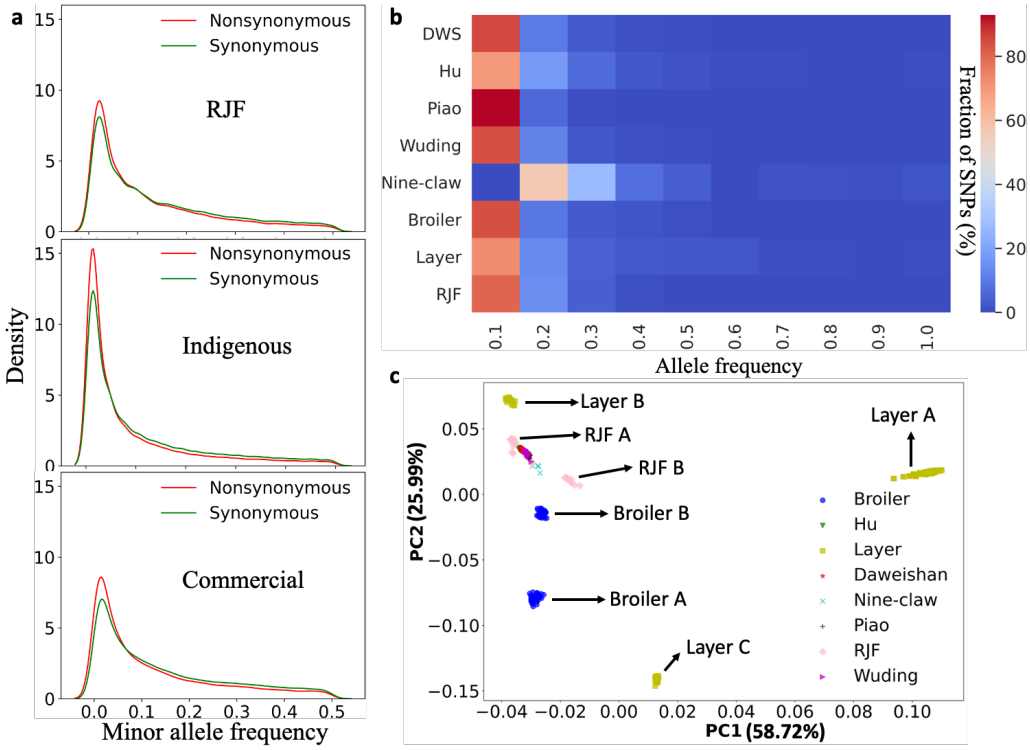

Figure 4-1 Analysis of frequency spectrums of SNPs. **a.** Distribution of minor allele frequency of synonymous and nonsynonymous SNPs in different chicken groups. **b.** Heatmap of allele frequency of group-specific SNPs. **c.** Principal component analysis of chicken population based on the detected 26 million SNPs.

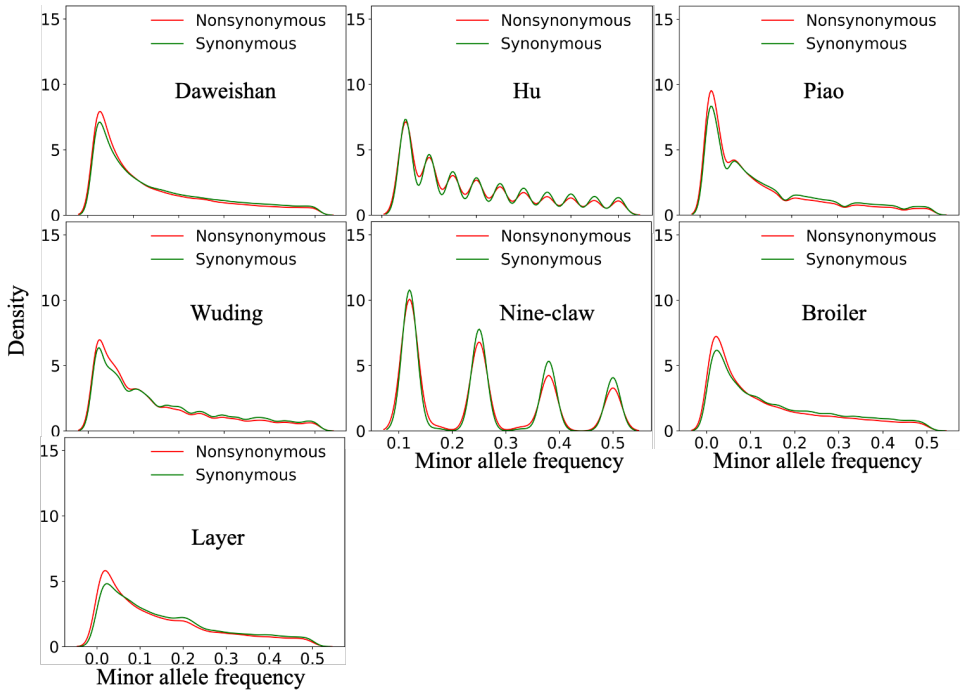

Figure 4-2 Distribution of the minor allele frequency among each chicken breed

4.3.6 A rigorous Null model facilitates sensitive detection of selective sweeps

To detect selection signatures of each chicken breed, we identified selective sweeps [140] along the chromosomes based on the frequencies of the bi-allelic SNPs. We used both genetic differentiation (F_{ST}) and nucleotide diversity (π) to determine the selective sweeps of each chicken breeds. A sliding window of 40 kb with 20 kb step size was used to compute both F_{ST} and π. To identify selective sweeps more sensitively, unlike previous studies [65] that used the sample mean and standard deviation to compute Z-values, we computed ZF_{ST} and $Z|\Delta\pi|$ values for each comparison based on the mean and standard deviation of Null models generated by permutating allele frequencies of the samples [135] (Materials and methods). We consider a window with $ZF_{ST} > 6$ or $Z|\Delta\pi| > 3.09$ (P-value < 0.001) as a selective sweep. Since adjacent windows can overlap with each other, we merged the overlapping selective sweep windows to form a discrete selective sweep (DSS) (Materials and methods). To find selection signatures of the chicken groups, we conducted a total of 19 different comparisons (Table 4-4 and Supplementary Table 4-9). The selective sweep windows identified by either of the two methods are distributed along all the chromosomes with varying densities (Figures 4-3, 4-4, 4-5 and 4-6). We generally detected more DSSs using ZF_{ST} (806~2,125 DSSs) than using $Z|\Delta\pi|$ (110~818 DSSs) for all the 19 comparisons (Table 4-4 and Supplementary Table 4-9) even using a higher ZF_{ST} cutoff, suggesting that ZF_{ST} is more sensitive than $Z|\Delta\pi|$ to identify selective sweeps. However, less than 60% (9.16%~58.23%) of the DSSs identified by $Z|\Delta\pi|$ overlap with those identified by ZF_{ST} (Supplementary Tables 4-10~4-28) in the 19 comparisons, indicating that the results of the two methods are largely complementary with each other. We thus take their union as the final prediction

of DSSs (Supplementary Tables 4-10~4-28). We finally identified 1,073~2,745 DSSs consisting of 1,998~7,284 windows containing 528~2,147 genes in each of the 19 comparisons (Table 4-4 and Supplementary Table 4-9). Therefore, we find much more selective sweeps than the previous study using a mixture model (~60 selective sweep windows of 40 kb size) [65]. The DSSs in the 19 comparisons have a varying length ranging from 40 kbp to 2,240 kbp with a median length of 60 kbp, and 91.71% of them are shorter than 140 kbp (Figure 4-7a). The total length of the DSSs in each comparison consist of 5.85% (Nine Claw VS RJF) ~ 18.88% (Broiler VS Layer) of the reference genome (GRCg7b assembly) (Figure 4-7b). In general, comparisons with broilers alone as one of the two compared groups tend to have a high portion of genome under selection (Figure 4-7b), suggesting that broilers have gone through most extensive selection.

Table 4-4 Summary of putative selective sweeps and DSSs

Breeds	# Analyzed windows		# Windows passed critical cut-off		# Discrete selective sweeps (# genes)		# Final putative selective sweep windows (# DSSs, # genes in DSSs)
	ZF_{ST}	$Z\|\Delta Pi\|$	$ZF_{ST} > 6$	$Z\|\Delta Pi\| > 3.09$	$ZF_{ST} > 6$	$Z\|\Delta Pi\| > 3.09$	
Daweishan VS. RJF	51,857	51,789	4,157	595	1,268 (978)	263 (790)	4,619 (1,550, 1706)
Hu VS. RJF	51,822	51,719	3,407	1,601	1,054 (783)	645 (802)	4,415 (1,811, 1,425)
Piao VS. RJF	51,860	51,789	4,057	383	1,313 (845)	152 (647)	4,381 (1,483, 1,455)
Wuding VS. RJF	51,853	51,782	3,752	1,554	1,255 (811)	764 (892)	4,962 (2,077, 1,617)
Nine-claw VS. RJF	51,769	51,615	1,640	497	806 (398)	262 (176)	1,998 (1,073, 528)
Broiler VS. RJF	51,643	51,756	6,366	2,150	1,784 (1,477)	818 (646)	7,012 (2,745, 1,661)
Layer VS. RJF	51,808	51,725	3,808	791	1,521 (896)	407 (404)	4,192 (1,926, 1,139)
Indigenous VS. RJF	51,886	51,793	4,294	301	1,183 (817)	130 (528)	4,572 (1,319, 1,333)
Commercial VS. RJF	51,658	51,767	5,063	1,146	1,787 (1,191)	544 (584)	5,540 (2,382, 1,544)

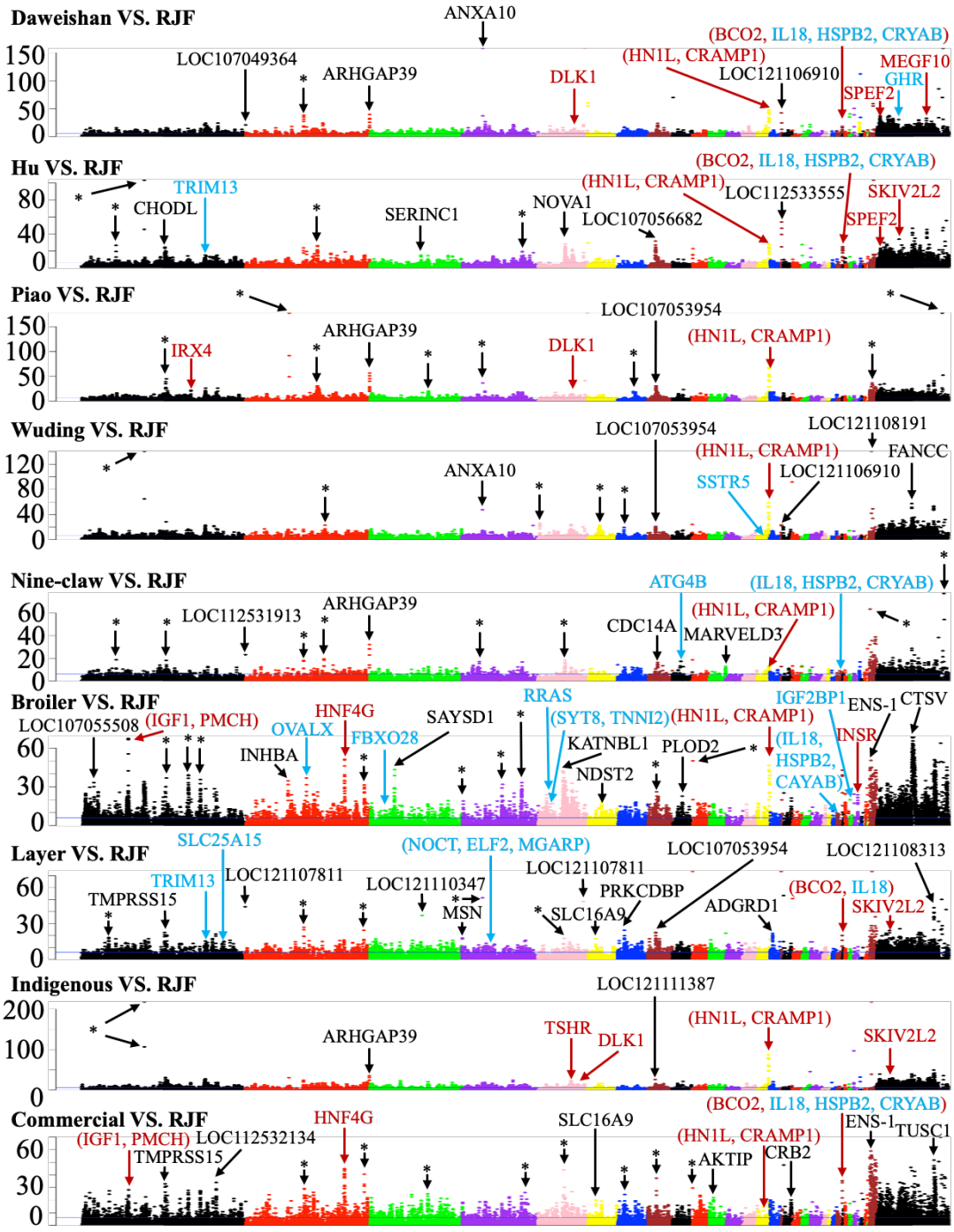

Figure 4-3 Manhattan plots of ZF_{ST} values for the indicated comparisons. The blue horizontal line indicates the ZF_{ST} cutoff = 6. Examples of genes in significant selective sweep windows are shown in different color. Genes that have been previously reported in selective sweep windows are shown in red, genes in our predicted selective sweep windows potentially related to the specific traits of each chicken breed are shown in blue, and genes in novel selective sweep windows with extremely high ZF_{ST} values are shown in black. Asterisk represents selective sweep windows lacking annotated genes.

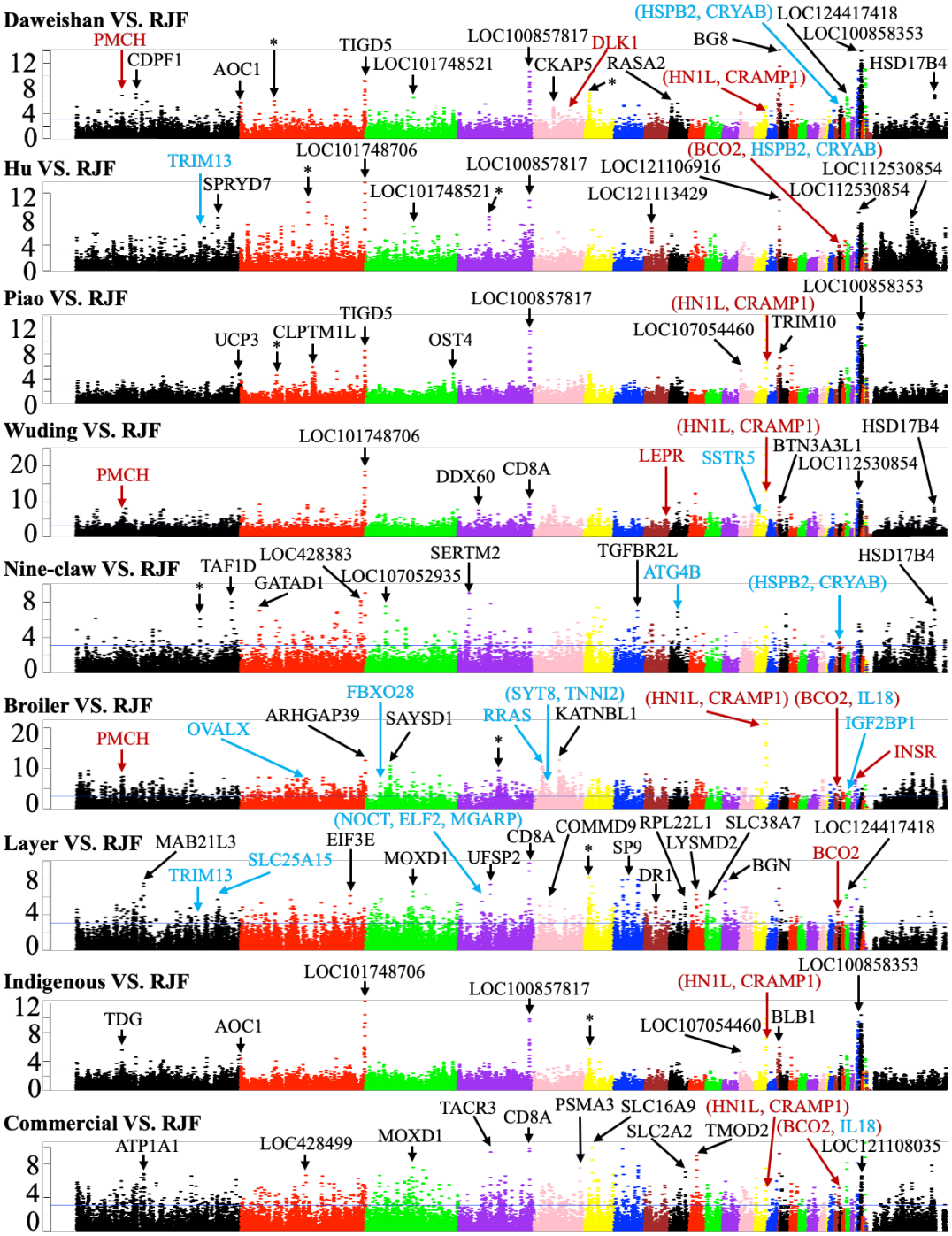

Figure 4-4 Manhattan plots of $Z|\Delta\pi|$ values for the indicated comparisons. The blue horizontal line indicates the $Z|\Delta\pi|$ cutoff = 3.09. Examples of genes in significant selective sweep windows are shown in different color. Genes that have been previously reported in selective sweep windows are shown in red, genes in our predicted selective sweep windows potentially related to the specific traits of each chicken breed are shown in blue, and genes in novel selective sweep windows with extremely high $Z|\Delta\pi|$ values are shown in black. Asterisk represents selective sweep windows lacking annotated genes.

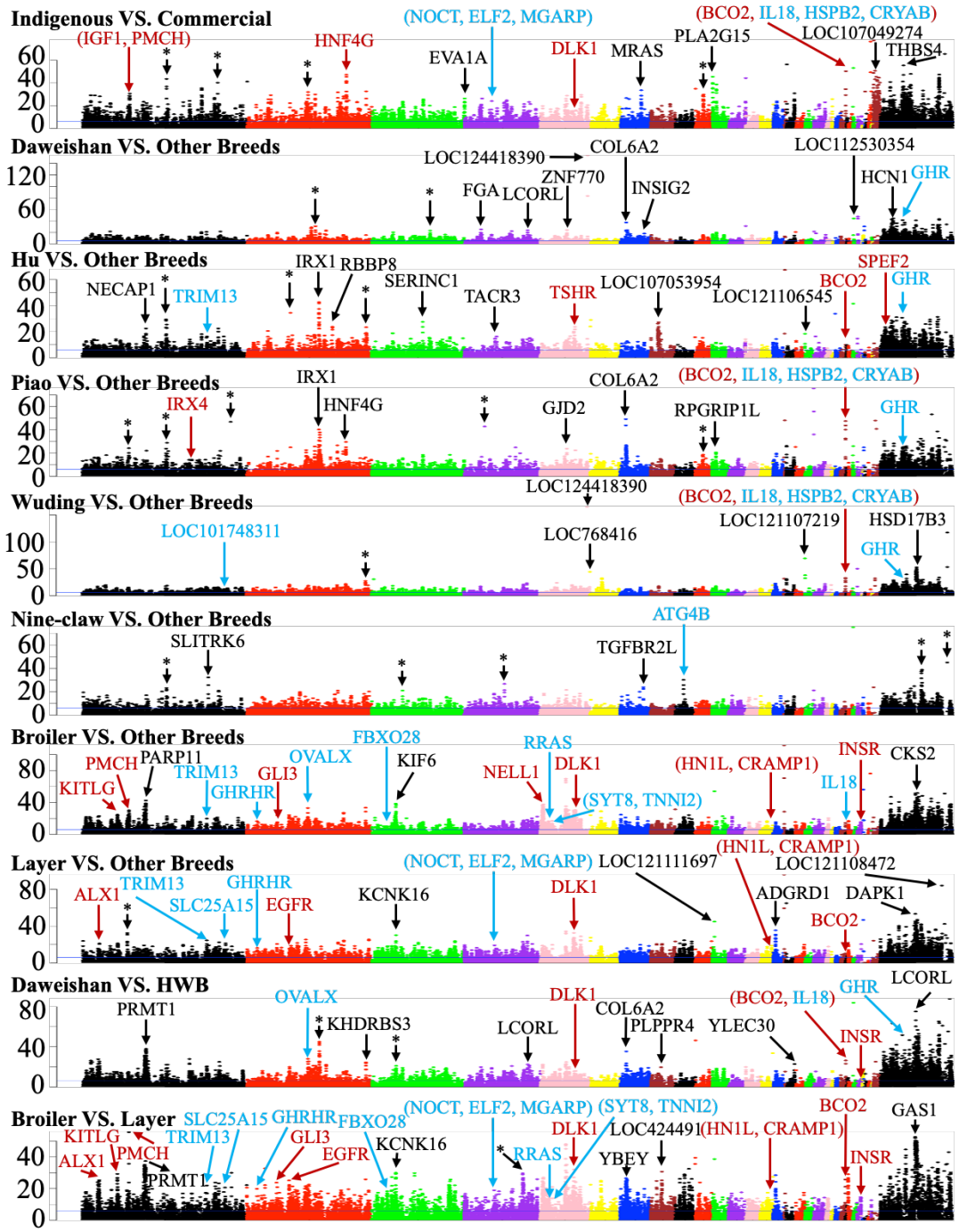

Figure 4-5 Manhattan plots of ZF_{ST} values for the indicated comparisons. The blue horizontal line indicates the ZF_{ST} cutoff = 6. Examples of genes in significant selective sweep windows are shown in different color. Genes that have been previously reported in selective sweep windows are shown in red, genes in our predicted selective sweep windows potentially related to the specific traits of each chicken breed are shown in blue, and genes in novel selective sweep windows with extremely high ZF_{ST} values are shown in black. Asterisk represents selective sweep windows lacking annotated genes.

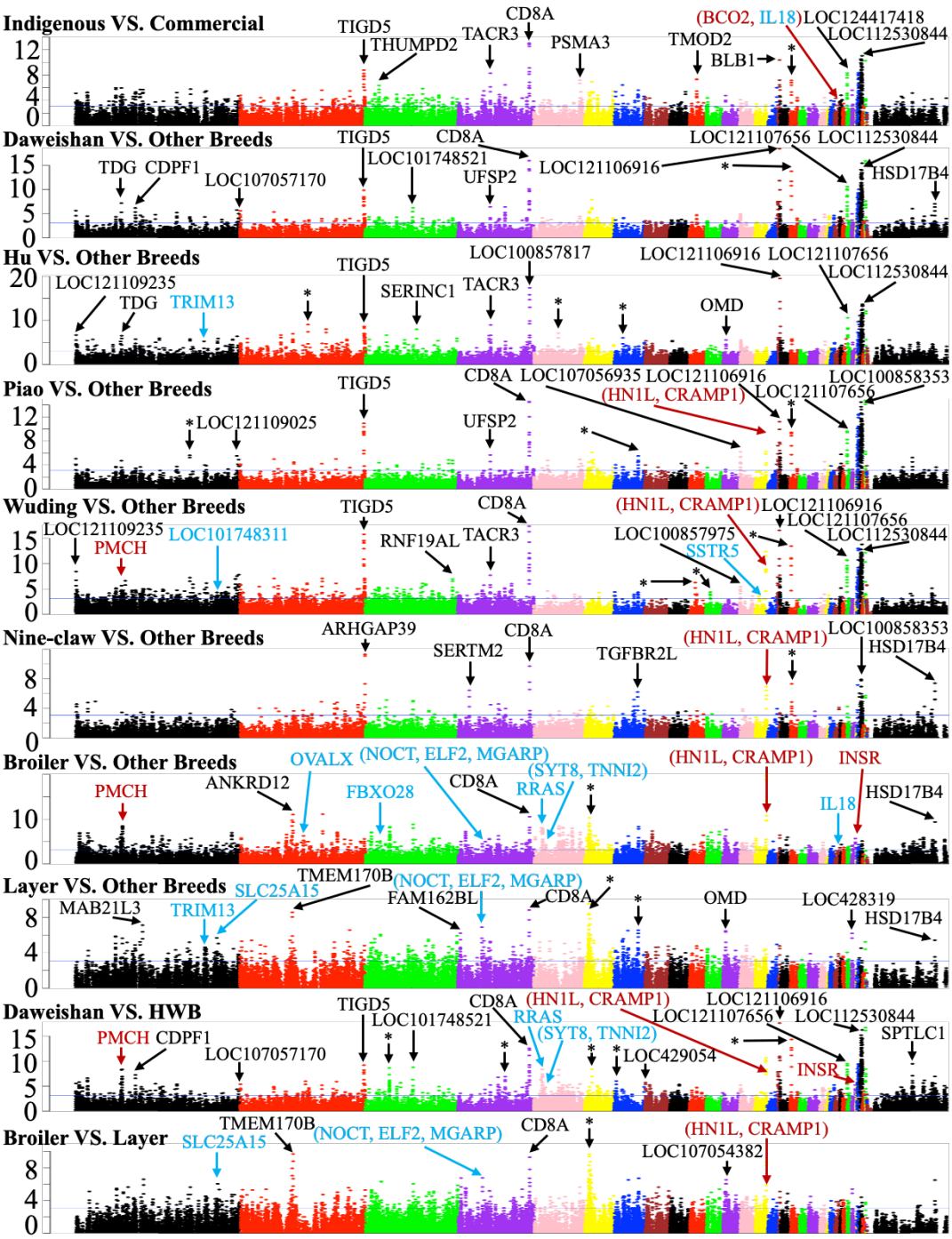

Figure 4-6 Manhattan plots of $Z|\Delta\pi|$ values for the indicated comparisons. The blue horizontal line indicates the $Z|\Delta\pi|$ cutoff = 3.09. Examples of genes in significant selective sweep windows are shown in different color. Genes that have been previously reported in selective sweep windows are shown in red, genes in our predicted selective sweep windows potentially related to the specific traits of each chicken breed are shown in blue, and genes in novel selective sweep windows with extremely high $Z|\Delta\pi|$ values are shown in black. Asterisk represents selective sweep windows lacking annotated genes.

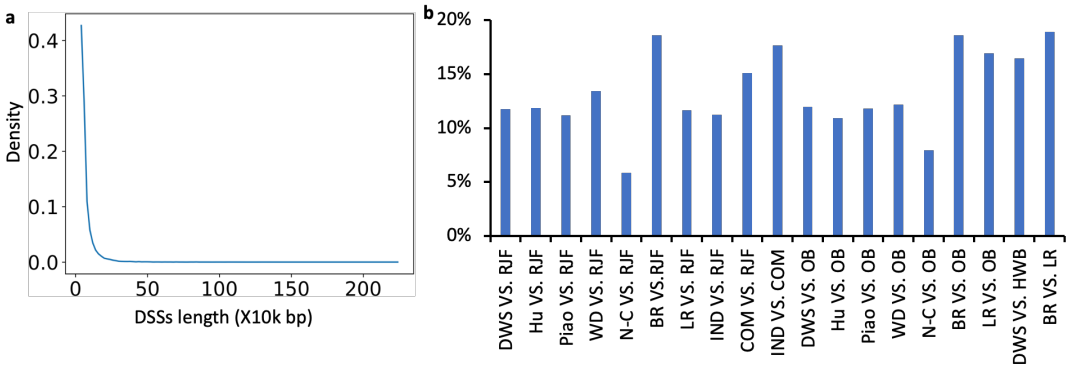

Figure 4-7 Summary of the DSSs lengths. **a**. Distribution of the lengths of the DSSs pooled from the 19 comparisons. **b.** Ratio of the DSSs lengths in each comparison with respect to the length of the reference genome (GRCg7b assembly). Abbreviations: DWS for Daweishan, WD for Wuding, N-C for Nine-claw, BR for Broiler, LR for Layer, IND for indigenous, COM for commercial, OB for Other Breeds and HWB for Hu+Wuding+Broiler.

4.3.7 The 19 comparisons reveal selection signatures of the chicken groups

Firstly, to find the genetic differences between artificial selection and natural selection, we compared each chicken breed (including indigenous chicken group and commercial chicken group) with the RJFs. As summarized in Table 4-4 and Supplementary Tables 4-10~4-18, we identified varying numbers of selective sweeps (1,998~7,012) and DSSs (1,073~2,745) involving 528~1,706 genes for the eight comparisons, suggesting that these chicken breeds might have gone through different levels of artificial selection. For example, the Broiler VS. RJF comparison yields the highest numbers of selective sweeps (7,012) and DSSs (2,745), suggesting again that broilers might have gone through the most intensive artificial selection. Among the five indigenous breeds, Wuding chickens might have gone through the most intensive artificial selection with the highest numbers of selective sweeps (4,962) and DSSs (2,077). In addition, we identified more selective sweeps (5,540) and DSSs (2,382) for the Commercial VS. RJF comparison than those (4,572 selective sweeps and 1,319 DSSs) in Indigenous VS. RJF comparison (Table 4-4), suggesting that the commercial

chickens have gone through more intensive artificial selections than indigenous chickens as generally understood.

Secondly, to reveal the genetic differences between traditional selection and industrial selection, we compared the indigenous chicken group with the commercial chicken group and identified a large number of selective sweeps (6,735) and DSSs (2,532) involving 2,147 genes (Supplementary Tables 4-9 and 4-19). This result suggests that indigenous chickens and the commercial chickens have gone through quite different artificial selection routes as commonly understood.

Thirdly, to reveal uniquely selective sweeps of each chicken breed, we compared each domestic chicken breed with the rest domestic chicken breeds and found that broilers and layers have much higher number of uniquely selective sweeps (7,239 and 6,401, respectively) and DSSs (2,514 and 2,479, respectively) than the indigenous breeds (2,801~4,555 and 1,363~1,890, respectively) (Supplementary Tables 4-9 and 4-20~4-26).

Fourthly, to reveal possible selective sweeps underlying the miniature body size of Daweishan chicken, we compared Daweishan chicken with the group of Hu chicken, Wuding chicken and Broiler (HWB), with a relatively large body size, and identified 6,359 selective sweeps and 2,263 DSSs including 1,911 genes (Supplementary Tables 4-9 and 4-27).

Finally, to find the selection difference between broilers and layers, we compared the two groups and identified 7,284 selective sweeps and 2,630 DSSs including 1,802 genes (Supplementary Tables 4-9 and 4-28). For the similar comparison in a previous study [65], only 41 selective sweeps (40 kb) were identified. Therefore, we identified substantially more selective sweeps. This might be because we use a more rigorous Null

model for normalizing F_{ST} and $\Delta\pi$, while the previous study employed a mixture model [65].

4.3.8 Amino-acid altering SNPs are enriched in the selective sweeps

To identify selective sweeps that might be responsible for the formation of a chicken breed, we took the union of DSSs found in comparisons with breed alone as one of the compared group, e.g, for Daweishan chicken, we took the union of DSSs in comparisons Daweishan VS. RJF, Daweishan VS. Other Breeds and Daweishan VS. HWB; and for Hu chicken, we took the union of DSSs in comparisons Hu VS. RJF and Hu VS. Other Breeds; and so on. We identified from 1.22 million (Nine-claw chicken) to 4.06 million (broilers) SNPs in the union of DSSs in each domestic chicken breed (Table 4-5). Among these SNPs, only 1.28% (Nine-claw chicken) ~2.02% (Hu chicken) are located in coding regions, while the remaining vast majority (97.98%~98.72%) fall in non-coding regions (Table 4-5). As non-coding regions comprise 96.92% of the reference chicken genome (GRCg7b assembly), as in the case of all the bi-allelic SNPs (Table 4-2), the SNPs in the DSSs are also enriched in non-coding regions relative to in coding regions. Among the SNPs in coding regions, 36.30%~45.23% are amino-acid altering, which are higher than the corresponding values of all the bi-allelic SNPs (33.33%~38.33%) (Table 4-2), suggesting that amino-acid altering SNPs are enriched in the selective sweeps relative to all the bi-allelic SNPs ($p = 0.005$, K-S test).

Table 4-5 Functional annotation of SNPs in DSSs of each domestic chicken breed

	Daweishan	Hu	Piao	Wuding	Nine-claw	Broiler	Layer
# Total bi-allelic SNPs	3,657,983	1,919,925	2,997,593	3,020,337	1,219,065	4,058,599	2,862,137
# SNPs in intergenic region (%)	1,237,910 (33.84)	662,275 (34.49)	1,026,302 (34.24)	982,969 (32.55)	424,878 (34.85)	1,279,806 (31.53)	890,623 (31.12)
# SNPs in Intronic region (%)	1,753,832 (47.95)	914,031 (47.61)	1,420,985 (47.40)	1,491,637 (49.39)	587,664 (48.21)	2,030,548 (50.03)	1,455,882 (50.87)
# SNPs in up/downstream (%)	115,633 (3.16)	66,121 (3.44)	98,454 (3.28)	99,913 (3.31)	32,073 (2.63)	108,623 (2.68)	75,384 (2.63)
# SNPs in UTR 3'/UTR 5' (%)	137,075 (3.75)	68,060 (3.54)	106,584 (3.56)	112,934 (3.74)	40,400 (3.31)	151,470 (3.73)	101,979 (3.56)
# SNPs in splicing region (%)	323 (0.01)	163 (0.01)	248 (0.01)	270 (0.01)	86 (0.01)	312 (0.01)	196 (0.01)
# SNPs in ncRNA (%)	347,114 (9.49)	170,583 (8.88)	285,264 (9.52)	276,744 (9.16)	118,388 (9.71)	431,263 (10.63)	299,387 (10.46)
# SNPs in coding region (%)	66,096 (1.81)	38,692 (2.02)	59,756 (1.99)	55,870 (1.85)	15,576 (1.28)	56,577 (1.39)	38,686 (1.35)
# Nonsynonymous-INTOL SNPs (%)	2,966 (0.08)	1,620 (0.08)	2,898 (0.10)	2,445 (0.08)	586 (0.05)	2,247 (0.06)	1,530 (0.05)
# Nonsynonymous-TOL SNPs (%)	25,045 (0.68)	15,125 (0.79)	23,725 (0.79)	21,523 (0.71)	5,561 (0.46)	18,136 (0.45)	12,407 (0.43)
# Synonymous SNPs (%)	37,058 (1.01)	21,352 (1.11)	32,281 (1.08)	31,094 (1.03)	9,332 (0.77)	35,823 (0.88)	24,429 (0.85)
# Stop-gain/loss (%)	427 (0.01)	252 (0.01)	409 (0.01)	371 (0.01)	73 (0.01)	252 (0.01)	204 (0.01)
# No prediction (%)	600 (0.02)	343 (0.02)	443 (0.01)	437 (0.01)	24 (0.00)	119 (0.00)	116 (0.00)

4.3.9 Our predicted selective sweeps are supported by experimental data

To evaluate our detected selective sweeps, we first compared them (Supplementary Tables 4-10~4-28) with the 15,439 QTLs in the chicken QTL database [141]. We find that 90.5%~98.3% putative DSSs in each of our comparisons overlap one or more QTLs in the chicken QTLdb (Table 4-6 and Supplementary Table 4-29), suggesting that our approach of finding selective sweeps achieves high precision (or positive prediction values). On the other hand, we find that 23.9%~41.6% QTLs in chicken QTLdb overlap one or more our predicted DSSs in each of our comparisons (Table 4-6 and Supplementary Table 4-29), and 11,449 (74.2%) QTLs in the chicken QTLdb overlap one or more of our predicted DSSs in different comparisons, suggesting that our approach of finding selective sweeps is quite sensitive.

As an additional validation of our detected selective sweeps, we next compared the genes in our predicted DSSs with those that have been reported to be under selection during chicken domestication process, and we describe a few examples here. It has been shown that the *BCO2* locus is involved in the yellow skin trait in domestic chickens [142], and we confirmed this results in several of our comparisons, including Daweishan VS. RJF (F_{ST} support, ZF_{ST} = 11.27), Daweishan VS. HWB (F_{ST} support, ZF_{ST} = 14.79), Hu VS. RJF (Two methods support, ZF_{ST} = 18.26, $Z|\Delta\pi|$ = 4.86), Hu VS. Other Breeds (F_{ST} support, ZF_{ST} = 6.27), Piao VS. Other Breeds (F_{ST} support, ZF_{ST} = 29.36), Wuding VS. Other Breeds (F_{ST} support, ZF_{ST} = 22.18), Broiler VS. RJF ($\Delta\pi$ support, $Z|\Delta\pi|$ = 5.08), Broiler VS. Layer (F_{ST} support, ZF_{ST} = 9.90), Layer VS. RJF (Two methods support, ZF_{ST} = 15.66, $Z|\Delta\pi|$ = 3.17), Layer VS. Other Breeds (F_{ST} support, ZF_{ST} = 8.70), Indigenous VS. Commercial (Two methods support, ZF_{ST} = 26.75, $Z|\Delta\pi|$ = 3.67)

and Commercial VS. RJF (Two methods support, ZF_{ST} = 27.16, $Z|\Delta\pi|$ = 4.44) (Figures 4-3, 4-4, 4-5 and 4-6). The *TSHR* locus is known to be involved in regulation of reproduction and metabolic functions in commercial chickens [12], and we found that the locus was in selective sweeps in comparisons Hu VS. Other Breeds (F_{ST} support, ZF_{ST} = 11.75) and Indigenous VS. RJF (F_{ST} support, ZF_{ST} = 13.43) (Figures 4-3 and 4-5). It has been reported that the *HNF4G* and *IGF1* loci are involved in growth regulation in chicken [12, 88], we found that the two loci were in selective sweeps in comparisons Broiler VS. RJF (For *HNF4G*, F_{ST} support, ZF_{ST} = 24.27; For *IGF1*, F_{ST} support, ZF_{ST} = 32.54), Commercial VS. RJF (For *HNF4G*, F_{ST} support, ZF_{ST} = 26.25; For *IGF1*, F_{ST} support, ZF_{ST} = 15.51) and Indigenous VS. Commercial (For *HNF4G*, F_{ST} support, ZF_{ST} = 24.51; For *IGF1*, F_{ST} support, ZF_{ST} = 17.86) (Figures 4-3 and 4-5). It has been reported that the *PMCH* locus is related to regulation of appetite and metabolic functions [12, 143], and we found that the locus was in the selective sweeps in several of our comparisons including Daweishan VS. RJF ($\Delta\pi$ support, $Z|\Delta\pi|$ = 3.25), Daweishan VS. HWB ($\Delta\pi$ support, $Z|\Delta\pi|$ = 6.04), Wuding VS. RJF ($\Delta\pi$ support, $Z|\Delta\pi|$ = 4.82), Wuding VS. Other Breeds ($\Delta\pi$ support, $Z|\Delta\pi|$ = 4.07), Broiler VS. RJF (Two methods support, ZF_{ST} = 32.54, $Z|\Delta\pi|$ = 6.05), Broiler VS. Other Breeds (Two methods support, ZF_{ST} = 19.59, $Z|\Delta\pi|$ = 6.34), Broiler VS. Layer (F_{ST} support, ZF_{ST} = 27.21), Commercial VS. RJF (F_{ST} support, ZF_{ST} = 15.51) and Indigenous VS. Commercial (F_{ST} support, ZF_{ST} = 17.86) (Figures 4-3, 4-4, 4-5 and 4-6). It has been shown that the *INSR* locus is related to the growth of chicken by encoding a critical component in insulin signaling [12], and we found that the locus was in the selective sweeps in comparisons Daweishan VS. HWB (Two methods support, ZF_{ST} = 7.37, $Z|\Delta\pi|$ = 4.46), Broiler VS. RJF (Two methods

support, ZF_{ST} = 15.30, $Z|\Delta\pi|$ = 5.23), Broiler VS. Other Breeds (Two methods support, ZF_{ST} = 11.61, $Z|\Delta\pi|$ = 4.86) and Broiler VS. Layer (F_{ST} support, ZF_{ST} = 9.78) (Figures 4-3, 4-4, 4-5 and 4-6). It has been shown that the *NELL1* locus is related to the skeletal integrity in broiler [88], and we found that the locus was in the selective sweeps of the Broiler VS. Other Breeds comparison (F_{ST} support, ZF_{ST} = 22.06) (Figure 4-5). It has been reported that the *IRX4* locus is related to the rumpless trait of Piao chicken [129], and we found that the locus was in the selective sweeps in comparisons Piao VS. RJF (F_{ST} support, ZF_{ST} = 13.79) and Piao VS. Other Breeds (F_{ST} support, ZF_{ST} = 12.12) (Figures 4-3 and 4-5). The other selective sweep loci found in the previous studies are also confirmed by our results, such as *ALX1*, *KITLG*, *EGFR*, *DLK1*, *JPT2* (annotated as *HN1L* in GRCg7b), *CRAMP1* and *GLI3* loci, which are related to the general domestication process of chicken [65], the *SKIV2L2* locus that is related to pigmentation [65], and the *LEPR*, *MEGF10* and *SPEF2* loci, which are related to production-oriented selection [65] (Figures 4-3, 4-4, 4-5 and 4-6). Taken together, all these results suggest that our approach is highly reliable to find selective sweeps in domestic chickens.

Table 4-6 Summary of putative DSSs overlapped with chicken QTLs

Breeds	# Putative DSSs	# Putative DSSs overlap QTLs in chicken QTLdb (%)	# QTLs in chicken QTLdb overlap our putative DSSs (%)
Daweishan VS. RJF	1,550	1,472 (95.0%)	5,526 (35.8%)
Hu VS. RJF	1,811	1,702 (94.0%)	5,247 (34.0%)
Piao VS. RJF	1,483	1,342 (90.5%)	4,801 (31.1%)
Wuding VS. RJF	2,077	1,979 (95.3%)	5,381 (34.9%)
Nine-claw VS. RJF	1,073	990 (92.3%)	3,686 (23.9%)
Broiler VS. RJF	2,745	2,601 (94.8%)	5,463 (35.4%)
Layer VS. RJF	1,926	1,837 (95.4%)	5,334 (34.5%)
Indigenous VS. RJF	1,319	1,266 (96.0%)	5,010 (32.5%)
Commercial VS. RJF	2,382	2,237 (93.9%)	5,086 (32.9%)

4.3.10 Novel selective sweeps are found in the chicken breeds

In addition to confirming many previously identified selective sweeps containing genes related to chicken domestication as described above, we also find numerous novel selective sweeps containing genes (Supplementary Tables 4-10~4-28) or in gene deserts.

We now highlight a few of them with extremely high ZF_{ST} and/or $Z|\Delta\pi|$ values in each comparison (Figures 4-3, 4-4, 4-5 and 4-6). Gene *ARHGAP39* on chromosome 2 is in the selective sweeps with extremely high ZF_{ST} and/or $Z|\Delta\pi|$ value in comparisons Daweishan VS. RJF, Piao VS. RJF, Nine-claw VS. RJF, Indigenous VS. RJF, Nine-claw VS. Other Breeds and Broiler VS. RJF. *ARHGAP39* plays important roles in cell cytoskeletal organization, growth, differentiation, neuronal development and synaptic functions [144]. Gene *TIGD5* on chromosome 2 is in the selective sweeps with extremely high $Z|\Delta\pi|$ value in comparisons Daweishan VS. RJF, Piao VS. RJF, Daweishan VS. Other Breeds, Hu VS. Other Breeds, Piao VS. Other Breeds, Wuding VS. Other Breeds, Daweishan VS. HWB and Indigenous VS. Commercial. *TIGD5* encodes the tigger transposable element-derived protein 5 and is important for nucleic acid binding [145]. Gene *KCNK16* on chromosome 3 is in the selective sweeps with extremely high ZF_{ST} value in comparisons Layer VS. Other Breeds and Broiler VS. Layer. *KCNK16* encodes a rapidly activating and non-inactivating outward rectifier K^+ channel [146]. Gene *CD8A* on chromosome 4 is in the selective sweeps with extremely high $Z|\Delta\pi|$ value in comparisons Wuding VS. RJF, Layer VS. RJF, Commercial VS. RJF, Daweishan VS. Other Breeds, Piao VS. Other Breeds, Wuding VS. Other Breeds, Nine-claw VS. Other Breeds, Broiler VS. Other Breeds, Layer VS. Other Breeds, Indigenous VS. Commercial and Daweishan VS. HWB. *CD8A* encodes the T-cell surface glycoprotein CD8 alpha chain precursor and plays essential roles in immune response [147]. Gene *COL6A2* on chromosome 7 is in the selective sweeps with extremely high ZF_{ST} value in comparisons Daweishan VS. Other Breeds, Daweishan VS. HWB and Piao VS. Other Breeds. The gene encodes the collagen alpha-2(VI) chain precursor which act as a cell-binding protein

[148]. Besides the genes mentioned above, we also indicate in Figures 4-3, 4-4, 4-5 and 4-6 many other genes located in novel selective sweeps with extremely high ZF_{ST} and/or $Z|\Delta\pi|$ values in multiple comparisons such as: *ANXA10* on chromosome 3, gene *LOC*107053954 on chromosome 8, gene *SLC16A9* on chromosome 6, gene *ENS-1* on chromosome W and gene *HSD17B4* on chromosome Z. It is interesting to experimentally investigate the roles of these genes in the domestication and breeding of each chicken breed.

In Figures 4-3, 4-4, 4-5 and 4-6, we also label a few examples of selective sweeps in gene desserts, with extremely high ZF_{ST} and/or $Z|\Delta\pi|$ values in multiple comparisons. It is highly likely that these selective sweeps might harbor non-coding functional sequences such as *cis*-regulatory modules of distal genes.

4.3.11 Selective sweeps related to each chicken breed

In addition to finding numerous novel selective sweeps containing genes in each comparison (Supplementary Tables 4-10~4-28), we also identify numerous unique selective sweeps/DSSs that are only seen in comparisons with a breed alone as one of the two compared groups or selective sweeps/DSSs containing genes with interesting functions. These selective sweeps/DSSs might contain genes (Supplementary Tables 4-30~4-36) whose functions are related to the specific traits of the chicken breed. Specifically, for Daweishan chicken, we identified 44 putative genes in the selective sweeps that might be related to its unique traits including the small body size (Supplementary Table 4-30). For example, the *GHR* (growth hormone receptor) gene is located in a selective sweep window on chromosome Z, which overlaps body weight QTLs and shank length QTLs. The gene is in the selective sweep windows identified in

comparisons with Daweishan chicken alone as one of the two compared groups (Daweishan VS. RJF, ZF_{ST} = 17.71; Daweishan VS. Other Breeds, ZF_{ST} = 17.69; Daweishan VS. HWB, ZF_{ST} = 19.51) (Figures 4-3 and 4-5). It has been reported that loss-of-function mutations in *GHR* was related to sex-linked dwarfism in chicken [149]. We analyzed the SNPs in the *GHR* gene body for each chicken breeds using the GRCg7b assembly as the template and found 79 unique SNPs in the gene of Daweishan chicken, which were not present in the other chicken breeds (Hu, Piao, Wuding, Nine-claw, Broiler, Layer and RJF). Among these 79 unique SNPs, 68 are in intronic regions, 10 are in UTRs and one is nonsynonymous SNP that leads to a CGG to TGG (R to W) mutation, which is tolerant. The substitution allele has a frequency of 0.76, thus it is only nearly fixed. As no fixed potential amino acid-altering mutation could be found in the *GHR* coding regions, we hypothesize that *GHR* gene might be related to the small body size of Daweishan chicken through changes in its regulatory sequences in the window, resulting in a decrease in its expression. To test this hypothesis, we measured the *GHR* expression levels in the liver, kidney, leg muscle and breast muscle of Daweishan chickens and broilers with large body size. To our surprise, Daweishan chickens had significantly higher *GHR* expression levels in all the four tissues than the broilers (Figure 4-8). It remains to be elucidated how the higher *GHR* expression level is related to the small body size of Daweishan chickens.

For Hu chicken, we identified 14 putative genes in selective sweeps (Supplementary Table 4-31) that might be related to its unique traits including the very stout legs. Specifically, gene *TRIM13* in a selective sweep on chromosome 1 (Hu VS. RJF, ZF_{ST} = 7.49 and $Z|\Delta\pi|$ = 7.19; Hu VS. Other Breeds, ZF_{ST} = 10.01 and $Z|\Delta\pi|$ =

4.58) (Figures 4-3, 4-4, 4-5 and 4-6) overlaps shank circumference QTLs, and there are two nonsynonymous SNPs in the gene body which are fixed (allele frequency = 1). Thus, it is interesting to experimentally investigate the role of *TRIM13* in the development of the very stout legs of Hu chicken.

For Piao chicken, we identified six putative genes in selective sweeps (Supplementary Table 4-32) that might be related its unique traits including the rumpless trait. Of these six genes, *IRX4* in a selective sweep on chromosome 2 was reported to be related to the rumpless trait of Piao chicken in a previous study [129], and we also found that the selective sweeps were only identified by the comparisons Piao VS. RJF (ZF_{ST} = 13.79) and Piao VS. Other Breeds (ZF_{ST} = 12.12) (Figures 4-3 and 4-5, Supplementary Tables 4-12 and 4-23). Thus, it is highly likely that *IRX4* is related to the rumpless trait of Piao chicken. At the same time, the previous study also identified genes *IL18*, *HSPB2*, and *CRYAB* to be related to the rumpless trait of Piao chicken. Although we also found these three genes in the selective sweeps for the comparison Piao VS. Other Breeds (Supplementary Table 4-23 and Figure 4-5), these three genes were also present in the selective sweeps for comparisons with a breed having normal tails alone as one of the two compared groups, such as Daweishan VS. RJF (Supplementary Table 4-10, Figures 4-3 and 4-5), Hu VS. RJF (Supplementary Table 4-11, Figures 4-3 and 4-5), Nine-claw VS. RJF (Supplementary Table 4-14, Figures 4-3 and 4-5), Broiler VS. RJF (Supplementary Table 4-15, Figures 4-3 and 4-5), Layer VS. RJF (Supplementary Table 4-16, Figure 4-3) and Daweishan VS. HWB (Supplementary Table 4-21, Figure 4-3). Therefore, these three genes might not be related to the rumpless trait of Piao chicken.

For Wuding chicken, we identified 18 putative genes in selective sweeps (Supplementary Table 4-33) that might be related to its unique traits including colorful feathers and thick fat. Specifically, gene *SSTR5* in a selective sweep on chromosome 14 (Wuding VS. RJF, ZF_{ST} = 7.38 and $Z|\Delta\pi|$ = 5.16; Wuding VS. Other Breeds, $Z|\Delta\pi|$ = 3.33) (Figures 4-3, 4-4 and 4-6) overlaps body weight QTLs, however, there are no nonsynonymous SNPs in its gene body. Gene *LOC*101748311 in a selective sweep on chromosome 1 (Wuding VS. Other Breeds comparison, ZF_{ST} = 9.70 and $Z|\Delta\pi|$ = 3.59) (Figures 4-5 and 4-6) overlaps the feather density QTLs and comb length QTLs and there are two nonsynonymous SNPs in its gene body, but their allele frequencies are very low (< 0.2). It is likely that both genes might be related to Wuding chicken's traits by changes in regulatory regions, which warrants further experimental studies.

For Nine-claw chicken, we identified seven putative genes in selective sweeps (Supplementary Table 4-34) that might be related to its unique traits. Specifically, gene *ATG4B* on chromosome 9 (Nine-claw VS. RJF, ZF_{ST} = 7.00 and $Z|\Delta\pi|$ = 3.60; Nine-claw VS. Other Breeds, ZF_{ST} = 7.56) (Figures 4-3, 4-4 and 4-5) overlaps egg production rate QTLs, but there are no nonsynonymous SNPs in its gene body.

For Broilers, we have identified 151 putative genes in selective sweeps (Supplementary Table 4-35) that might be related to its unique traits including the fast growth rate. Of these genes, *GHRHR* on chromosome 2 (Growth hormone-releasing hormone receptor) (Broiler VS. Other Breeds, ZF_{ST} = 10.36; Broiler VS. Layer, ZF_{ST} = 7.09) (Figure 4-5) is well-known for its role in determining growth rate and body size via regulating the growth hormone (GH) level in blood [150], however, there are no nonsynonymous SNPs in its gene body; *IGF2BP1* on chromosome 27 (insulin-like

growth factor 2 mRNA-binding protein 1) (Broiler VS. RJF, ZF_{ST} = 7.69 and $Z|\Delta\pi|$ = 3.98) (Figures 4-3 and 4-4) may affect growth rate via regulating insulin-like growth factor 2 level [151], but there are no nonsynonymous SNPs in its gene body. The *IGF2BP1* locus also overlaps the claw percentage QTLs, shank length QTLs, claw weight QTLs, drumstick and thigh weight QTLs, breastbone crest length QTLs, body weight QTLs, body slope length QTLs and femur bending strength QTLs. The result is consistent with a recent report that mutations in the promoter region of the *IGF2BP1* gene can affect chicken body size [99]. In addition, the following genes are also interesting as they overlap white striping QTLs, abdominal fat percentage QTLs, wooden breast QTLs and body weight QTLs, and thus might be related to the large body and fast growth rate of broilers, including *OVALX* on chromosome 2 (Broiler VS. RJF, ZF_{ST} = 9.41 and $Z|\Delta\pi|$ = 5.78; Broiler VS. Other Breeds, ZF_{ST} = 8.45 and $Z|\Delta\pi|$ = 4.18) (Figures 4-3, 4-4, 4-5 and 4-6), *RRAS* on chromosome 5 (Broiler VS. RJF, ZF_{ST} = 14.87 and $Z|\Delta\pi|$ = 7.53; Broiler VS. Other Breeds, ZF_{ST} = 11.73 and $Z|\Delta\pi|$ = 6.00) (Figures 4-3, 4-4, 4-5 and 4-6), *SYT8* on chromosome 5 (Broiler VS. RJF, ZF_{ST} = 6.90 and $Z|\Delta\pi|$ = 5.87; Broiler VS. Other Breeds, ZF_{ST} = 6.89 and $Z|\Delta\pi|$ = 3.85) (Figures 4-3, 4-4, 4-5 and 4-6), *TNNI2* on chromosome 5 (Broiler VS. RJF, ZF_{ST} = 6.90 and $Z|\Delta\pi|$ = 5.87; Broiler VS. Other Breeds, ZF_{ST} = 6.89 and $Z|\Delta\pi|$ = 3.85) (Figures 4-3, 4-4, 4-5 and 4-6) and *FBXO28* on chromosome 3 (Broiler VS. RJF, ZF_{ST} = 10.25 and $Z|\Delta\pi|$ = 5.30; Broiler VS. Other Breeds, ZF_{ST} = 8.38 and $Z|\Delta\pi|$ = 4.90) (Figures 4-3, 4-4, 4-5 and 4-6). All these genes either has no nonsynonymous SNPs or the allele frequencies of the nonsynonymous SNPs are very low. Thus, it is highly likely that they might be related to the broilers' traits through changes in their regulatory regions. To test this hypothesis, we

measured the expression levels of *GHRHR, IGF2BP1* and *OVALX* in the liver, kidney, leg muscle and breast muscle of Daweishan chickens with small body size and broilers. As shown in Figure 4-8, *GHRHR* had higher expression levels in liver of broilers than in the same tissue of Daweishan chickens; *IGF2BP1* had higher expression levels in liver, kidney and breast muscle of broilers than in the respective same tissues of Daweishan chickens; and *OVALX* had higher expression levels in liver and kidney of broilers than in the same tissues of Daweishan chickens. On the other hand, Daweishan chickens had a higher expression level of *IGF2BP1* in leg muscles and a higher expression level of *OVALX* in breast muscles (Figure 4-8). The high expression levels of the three genes in the liver and kidney might be related to the high metabolic and growth rates of broilers.

For layers, we identified 36 genes in selective windows (Supplementary Table 4-36) that might be related to its unique traits including larger number of egg-production. Specifically, gene *NOCT* (Nocturnin), *ELF2* (ETS-related transcription factor Elf-2) and *MGARP* (mitochondria localized glutamic acid rich protein) are all located in the same selective sweep on chromosome 4 (Layer VS. RJF, ZF_{ST} = 9.84 and $Z|\Delta\pi|$ = 4.61; Layer VS. Other Breeds, ZF_{ST} = 10.80 and $Z|\Delta\pi|$ = 5.24; Broiler VS. Layer, ZF_{ST} = 9.23 and $Z|\Delta\pi|$ = 5.20) (Figures 4-3, 4-4, 4-5 and 4-6). *NOCT* is expressed in retina and many other tissues, and its expression shows circadian rhythm [152, 153]. *NOCT* is known to be involved in adipogenesis, osteogenesis, and obesity in mice [154]. It has been shown that *MGARP* is involved in the synthesis of estrogen in ovary, and its expression is under the control of the hypothalamic-pituitary-gonadal (HPG) axis [155]. It has been reported that *ELF2* plays a role in cell proliferation [156]. Moreover, gene *SLC25A15* on chromosome 1 (mitochondrial ornithine transporter 1) (Layer VS. RJF, ZF_{ST} = 7.23 and

$Z|\Delta\pi|$ = 3.96; Layer VS. Other Breeds, ZF_{ST} = 9.68 and $Z|\Delta\pi|$ = 3.82; Broiler VS. Layer, ZF_{ST} = 10.52 and $Z|\Delta\pi|$ = 3.21) (Figures 4-3, 4-4, 4-5 and 4-6) overlaps the oviduct length QTLs, thus might be related to the egg-production rate of layers. Thus, these four genes might be related to the layers' unique traits. However, the four genes either have no nonsynonymous SNPs or the allele frequencies of nonsynonymous SNPs are very low. Thus, it is highly likely that they might be related to the high egg-production of layer through changes in their regulatory regions. To test this hypothesis, we measured the expression levels of *ELF2, MGARP*, *NOCT* and *SLC25A15* in 10 tissues of layers and Wuding chickens with low-egg production rate (Figure 4-9). Interestingly, the expression levels of the four genes in layers is higher than those in Wuding chickens in all the 10 tissues, except *MGARP* in liver and *NOCT* in pituitary (Figure 4-9). Moreover, as the expression of *NOCT* shows circadian rhythm [69, 70], we measured its expression level in 10 tissues of layers and Wuding chickens in 24 hours with 4 hours interval. As shown in Figure 4-10, in almost all the 10 tissues, the expression level of *NOCT* was lower after 16:00, and it started to increase at 4:00, was peaked at 8:00, and then decreased. In most tissues, layers had significantly higher expression levels than Wuding chickens. The higher expression level of these four genes in relevant tissues in layers might be related to their high egg-production trait. It is interesting to experimentally investigate roles of variations in the regulatory regions of these genes in the high egg-production related traits of layers, such as the lack of brooding behaviors, egg-laying circadian rhythm and high demand for light.

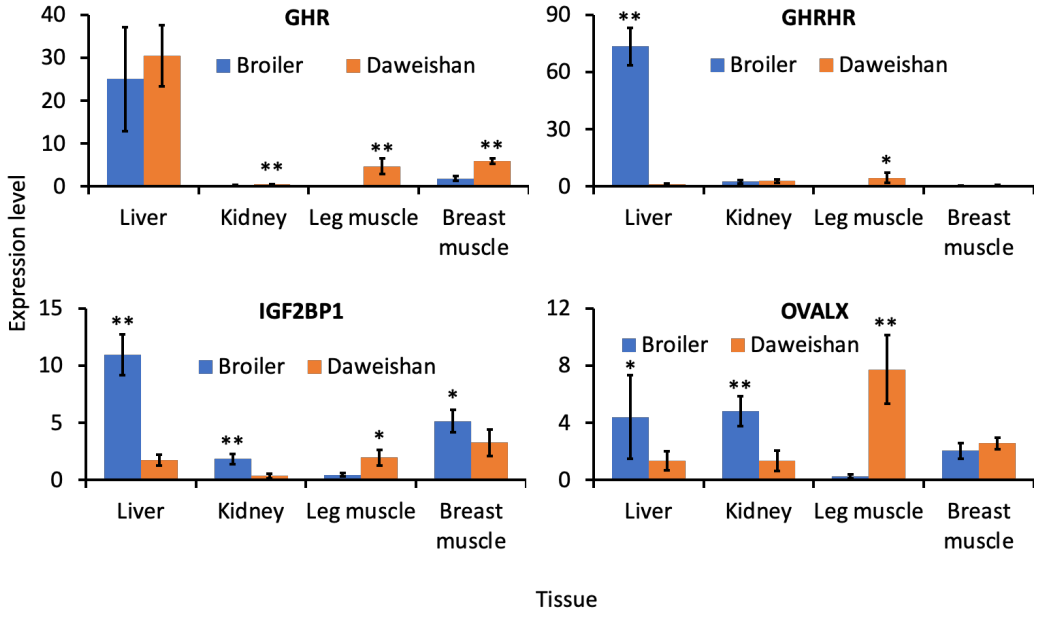

Figure 4-8 Expression levels of the indicated four genes. *p<0.05, **p<0.01, two-tailed t-test.

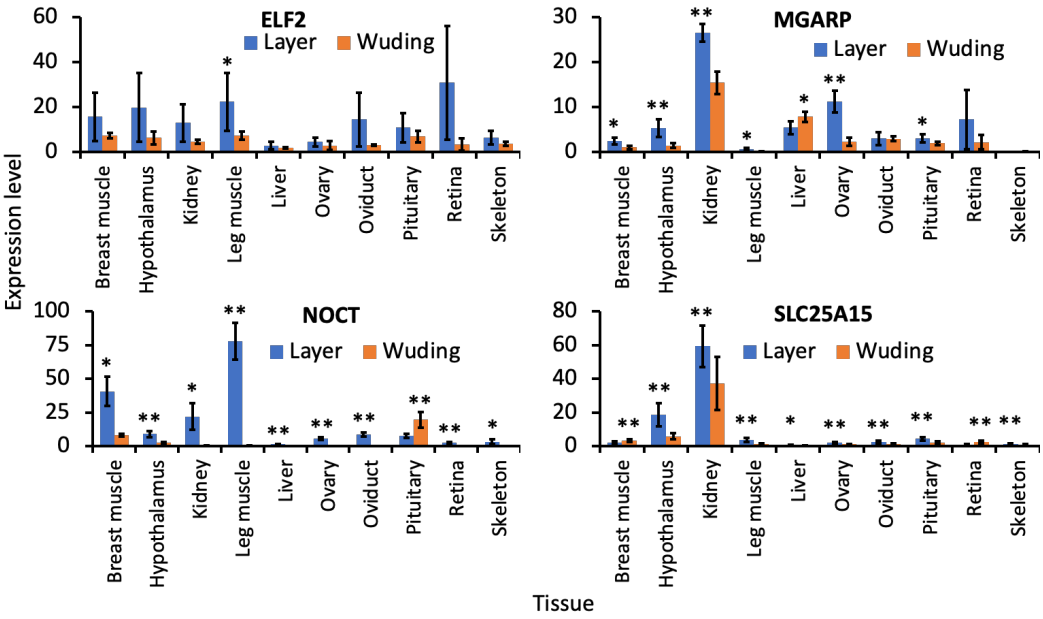

Figure 4-9 Expression levels of the indicated four genes. *p<0.05, **p<0.01, two-tailed t-test.

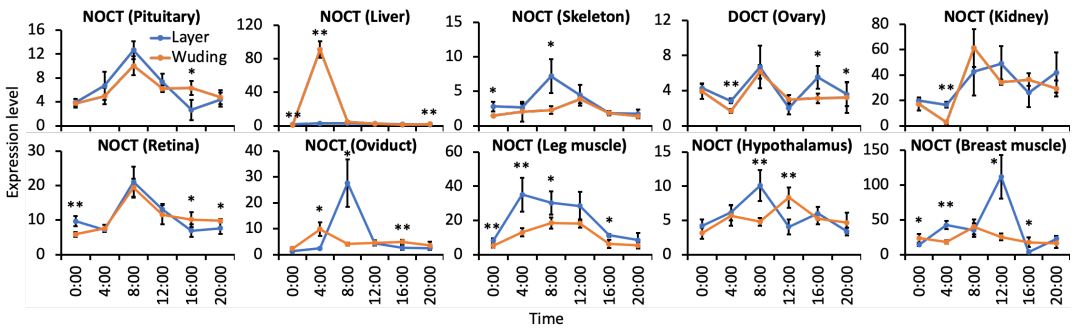

Figure 4-10 Expression level of gene NOCT in different time. *p<0.05, **p<0.01, two-tailed t-test.

4.4 Discussion

Next generation sequencing technology makes it possible to re-sequence a large number of individuals for a species for genome-wide studies. In 2021, NCBI released more complete domestic chicken (*Gallus gallus*) genome assemblies (GRCg7b and GRCg7w), providing good reference genomes for this economically, medically and evolutionally important bird. Using the GRCg7b assembly as the template, we have called the variants in populations of eight chicken breeds including 25 Daweishan chickens, 10 Hu chickens, 23 Piao chickens, 23 Wuding chickens, four Nine-claw chickens, 60 broilers, 56 layers and 35 RJFs. By comparing the putative selective sweeps of Daweishan, Hu, Piao, Wuding, Nine-claw chicken, broilers and layers with respect to others and RJFs (19 comparisons, Table 4-4 and Supplementary Table 4-9), we identified putative selective sweeps and genes that might be related to the specific-traits of each chicken breed or groups (Supplementary Tables 4-30~4-36). Remarkably, the vast majority (90.5%~98.3%) of our identified DSSs in each of our 19 comparisons overlap QTLs in the chicken QTLdb (Table 4-6 and Supplementary Table 4-29), while 74.2% QTLs in the chicken QTLdb overlap our identified DSSs in different comparisons. Moreover, we also confirm many previously identified genes under artificial selection.

Thus, we have achieved very high prediction precision (or positive prediction values) and sensitivity.

More importantly, our analyses also result in many new findings. Firstly, we identify a much larger number of selective sweeps/DSSs and genes related to the specific traits of broilers and layers than the previous study [65]. We attribute the difference to the different statistic models used in the two studies. More specifically, we use a more rigorous Null model by generating 100-sets of windows with the allele frequencies randomly permutated [135]. Using the mean and standard deviation of the Null model, we compute ZF_{ST} and $Z|\Delta\pi|$ for each window in each comparison. In contrast, the previous study [65] used the mean and standard deviation of the F_{ST} and $|\Delta\pi|$ values of the windows to compute the ZF_{ST} and $Z|\Delta\pi|$, which is not a rigorous Null model. Thus, the previous study might underestimate the number of selective sweeps. Consequently, we identify ~2,500 putative DSSs containing ~1,800 genes for the broilers and ~2,000 putative DSSs containing ~1,000 genes for the layers (Table 4-4 and Supplementary Table 4-9), which included almost all the only 90 and 66 putative selective sweep windows (40 kb) found in broilers and layers, respectively, in the previous study [65].

Secondly, we negate several genes found in a previous study [129] to be related to the rumpless trait of Piao chicken based on our results from multiple comparisons with or without Piao chicken. More specifically, in addition to *IRX4*, the previous study also claimed that *IL18*, *HSPB2*, and *CRYAB* [129] might be related to the rumpless trait of Piao chicken. Although we also find that gene *IRX4* presents in putative selective sweeps only in the Piao VS. RJF and Piao VS. Other Breeds comparisons (Supplementary Tables 4-12 and 4-23), thus it might be related to the rumpless trait of Piao chicken. However,

genes *IL18*, *HSPB2*, and *CRYAB* present in selective sweeps in not only the comparison related to Piao chicken (Piao VS. Other Breeds, Supplementary Table 4-23), but also in comparisons with chickens having a normal tail alone as a group, such as Daweishan, Wuding, Nine-claw chicken, broilers and layers (Supplementary Tables 4-10, 4-11, 4-14, 4-15, 4-16 and 4-21). Thus, these three genes might not be related to the rumpless trait of Piao chicken.

Thirdly, our analyses provide many novel selective sweeps containing genes that might be related to artificial selection of unique traits of each chicken breed (Supplementary Tables 4-30~4-36), and some are quite interesting, thus warranting further experimental investigations. For example, it is interesting to test the roles of the genes *NOCT* and *MGARP* in the high egg-production of layers.

Finally, we find that although SNPs in selective sweeps are more likely to alter amino acids than expected, many genes in selective sweeps often lack fixed amino acid-altering mutations. These genes might affect the traits of chicken breeds by changing their expression levels through changes in their cis-regulatory regions. Consistently, we found that all the eight genes that we examined were significantly differentially expressed in tissues of chicken breeds where these genes were in putative selective sweeps from those of chicken breeds. Due to the lack of a more complete map of *cis*-regulatory modules and their constituent transcription factor binding sites in the chicken reference genome, it is difficult to further pin down the sites that affect the expression of these genes and related organism traits. Therefore, it is pressing to map out the *cis*-regulatory elements in the chicken reference genome as has been done for *C. elegans*[157, 158], *D. melanogaster* [157, 159], mice [160, 161] and humans[161, 162].

4.5 Conclusion

In this study, we analyzed re-sequencing data from 85 individuals of five indigenous chicken breeds and 116 individuals of commercial chickens (broilers and layers) and 35 individuals of red jungle fowl. We find 33 million genetic variants in the populations, and identify more selective sweeps and affected genes for each chicken breed using a rigorous statistic model than previously reported. The identified variants are more likely located in non-coding regions while those in coding regions are largely tolerant, suggesting that most of the variants might affect *cis*-regulatory sequences. In agreement, 98.3% of our identified selective sweeps overlap known chicken QTLs, suggesting that our results are highly reliable. Meanwhile, 74.2% known QTLs overlap our predicted selective sweeps, indicating that our approach is very powerful for predicting selective sweeps. Our predicted selective sweeps and affected genes include most of previously identified ones. We also identify candidate selective sweeps and genes that might be related to the unique traits of each breed.

CHAPTER 5 High quality genome assembly of a white eared pheasant individual

5.1 Background

Crossoptilon, belonging to the Phasianidae family in the Galliformes order, is a rare but important genus endemic in China [186]. There are four species in the *Crossoptilon* genus, including Tibetan eared pheasant (TB) (*C. harmani*), white eared pheasant (WT) (*C. crossoptilon*), blue eared pheasant (BL) (*C. auritum*) and brown eared pheasant (BR) (*C. mantchuricum*) [187, 188]. These species are found in coniferous forests, mixed broadleaf-conifer forests and alpine scrubs in various parts of China with very different altitudes [189]. TB and WT are believed to be conspecifics and sympatrically inhabit in muntane forest with a high altitude of 3,000~5,000 m. TB is distributed in southeastern Tibet and adjacent northern India (3,000~5,000 m) [190], while WT is found in Qinghai, Sichuan, Yunnan and Tibet (3,000~4,300 m) [187]. BL and BR are closely related, but allopatrically inhabit. BL is only encountered in the mountains of Qinghai, Gansu, Sichuan and Ningxia, with an intermediate altitude of 1,500~3,000 m [187], while BR is mainly distributed in mountains in Shanxi and Hebei and near Beijing, with a low altitude of 20~1,000 m [187]. Since these species inhibit from low to intermediate and high attitudes, they are excellent models to study the genetic bases for similar species to adapt to different altitude levels.

Moreover, all the four species are in varying degree of danger of extinction due to human hunting and erosion of their habitants as a result of deforestation and urbanization. They all are registered as national key protection animals in China. Particularly, BR is listed by the International Union for Conservation of Nature (IUCN) as vulnerable. Thus, these species also provide good models to reveal footprints on the genome of a species

due to rapid decline of its population size. Indeed, Wang, et al. [191] recently assembled a BR genome and used it as the reference to compare nucleotide variations in the genomes of three fragmental BR populations that are in danger of extinction with those in the genomes of a BL population that are of less concern of extinction. However, this assembly cannot provide a high-quality reference genome for the *crossoptilon* species due to its small contig N50 (0.11 Mb).

To further understand the biology of these eared pheasants, particular, genetic basis of their adaptation to different levels of altitude, we re-sequenced 10 WT individuals and assembled the genome for one of them at chromosome-level with much higher quality than the earlier assembled BR genome [191] (contig N50 19.6 Mb VS. 0.11 Mb). We annotated and compared the genes and pseudogenes in the WT and BR genomes, and found that former contained more genes (16,315 VS. 15,003), while the latter harbored more pseudogenes (1,519 VS. 1,976). The unique genes in WT often become unique pseudogenes in BR and vice versa. Moreover, we compared the selective sweeps in the genomes of WT, BR and BL using re-sequencing data of 10 WT, 41 BR and 12 BL individuals, respectively. We found that the unique genes and pseudogenes in WT and BR are mainly involved in and affect, respectively, the same pathways that include genes in selective sweeps in each of the three species. These pathways are mainly related to cardiovascular, energy metabolic, neuronal, and immune functions. Therefore, it appears that the three species might adapt to different altitudes by altering the functions of the same pathways related to the four functional categories. Our assembled WT genome and re-sequencing data can also be valuable resources for studying the biology, evolution and developing conservation strategies of these endangered species.

5.2 Methods and materials

5.2.1 Bird populations

For the WT populations, blood samples of a total of 11 WT individuals (five males and six females) aged 10 months were collected from Diqing Tibet Autonomous prefecture, Yannan, China. A female individual was collected from the same area for whole genome assembly and its relevant tissues were subject to Illumina paired-end DNA short reads sequencing, PacBio long reads sequencing, Hi-C paired-end short reads sequencing, and paired-end RNA-seq of 20 tissues (Heart, Liver, Spleen, Lung, Kidney, Pancreas, Gizzard, Glandular, Crops, Ovary, Abdominal fat, Rectum, Duodenum, Cecum, Skin, Small intestine, Brain, Cerebellum, Chest muscle, Leg muscle) and mixed tissues for gene annotation. All other 10 individuals were subject to Illumina paired-end DNA short reads sequencing.

For the BR populations, re-sequencing data of 41 BR individuals were downloaded from the China National Genomics Data Center (CNGDC) (accession number: PRJCA003284) [191]. The assembled genome of a BR individual and RNA-seq data from the individual's blood, developing primaries and developing tail feathers tissues were downloaded from CNGDC (accession number: PRJCA003284) [191]. For the BL populations, re-sequencing data of 11 BL individuals were downloaded from CNGDC (accession number: PRJCA003284) [191]. A blood sample was collected from a captured BL individual in a zoom and was subject to Illumina paired-end DNA short reads sequencing.

5.2.2 Short reads DNA sequencing

Two milliliters of blood were drawn from the wing vein of each bird in a centrifuge tube containing anticoagulant (EDTA-2K) and stored at -80℃ until use. Genomic DNA (10μg) in each blood sample was extracted using a DNA extraction kit (DP326, TIANGEN Biotech, Beijing, China) and fragmented using a Bioruptor Pico System (Diagenode, Belgium). DNA fragments around 350 bp were selected using SPRI beads (Beckman Coulter, IN, USA). DNA-sequencing libraries were prepared using Illumina TruSeq® DNA Library Prep Kits (Illumina, CA, USA) following the vendor's instructions. The libraries were subject to 150 cycles paired-end sequencing on an Illumina Novaseq 6000 platform (Illumina, CA, USA) at 102X coverage.

5.2.3 PacBio long reads DNA sequencing

High molecular weight DNA was extracted from each blood sample using NANOBIND® DNA Extraction Kits (PacBio, CA, USA) following the vendor's instructions. DNA fragments of about 25 kb were size-selected using a BluePippin system (Sage Science, MA, USA). Sequencing libraries were prepared for the DNA fragments using SMRTbell® prep kits (PacBio, CA, USA) following the vendor's instructions, and subsequently sequenced on a PacBio Sequel II platform (PacBio, CA, USA) at 91X coverage.

5.2.4 RNA-seq reads sequencing

One to two grams of various tissues and mixed tissues were collected from the selected female WT individual in a centrifuge tube and immediately frozen in liquid nitrogen, then stored at -80℃ until use. Total RNA from each tissue sample were extracted from each tissue or mixed tissues using TRlzol reagents (TIANGEN Biotech,

Beijing China) according to the manufacturer's instructions. RNA-sequencing libraries for each tissue collected from the individual were prepared using Illumina TruSeq® RNA Library Prep Kits (Illumina, San Diego) following the vendor's instructions. The libraries were subject to 150 cycles paired-end sequencing on an Illumina Novaseq 6000 platform at a sequencing depth of 281X.

5.2.5 Hi-C reads sequencing

Five milliliters of blood were drawn from the wing vein of the selected WT individual in a Streck Cell-free DNA BCT collecting vessel (Streck Corporate, USA), and stored at 4℃ and used in 24 hours. Hi-C libraries were constructed using Phase Genomics' Animal Hi-C kit following the vendor's instructions and subsequently sequenced on an Illumina's Novaseq 6000 platform at a sequencing depth of 81X.

5.2.6 Real-time quantitative PCR (RT-qPCR) analysis

RT-qPCR was performed using the Bio-Rad CFX96 real-time PCR platform (Bio-Rad Laboratories. lnc, America) and SYBR Green master mix (iQTM SYBRGreen ® Supermix, Dalian TaKaRa Biotechnology Co. Ltd. Add). The primers of the 17 randomly selected putative new genes are listed in Supplementary Note 5-1. The β-actin gene was used as a reference. Primers were commercially synthesized (Shanghai Shenggong Biochemistry Company P.R.C). Each PCR reaction was performed in 25 μl volumes containing 12.5 μl of iQ™ SYBR Green Supermix, 0.5 μl (10 mM) of each primer, and 1 μl of cDNA. Amplification and detection of products was performed with the following cycle profile: one cycle of 95 °C for 2 min, and 40 cycles of 95 °C for 15 s, annealing temperature for 30 s, and 72 °C for 30 s, followed by a final cycle of 72 °C for 10 min. The specificity of the amplification product was verified by electrophoresis on a 0.8%

agarose gel and DNA sequencing. The 2−ΔCt method was used to analyze mRNA abundance. All samples were analyzed with at least three replicates, and the mean of these measurements was used to calculate mRNA expression.

5.2.7 Contig assembling and scaffolding

We used the PacBio long reads longer than 5,000 bp to assemble the contigs using Wtdbg (2.5) [57], and polished the contigs using Wtdbg (2.5) [57] with Illumina DNA short reads for the white eared pheasant. Then we used SALSA [58, 59] to bridge the contigs and obtain the scaffolds with Hi-C short reads. We filled the gaps in the scaffolds using PBJelly [60] with the PacBio long reads, and then made two rounds of polish by firstly using Racon (1.4.21) [61] with PacBio long reads and secondly using NextPolish (1.4.0) [62] with Illumina DNA short reads from the selected white eared pheasant individual.

5.2.8 Quality evaluation of assemblies

We masked the repeats for the assembly of the white eared pheasant using WindowMasker (2.11.0) [63] to get the repeat rate, and estimated the heterozygosity of the assembly using Jellyfish (2.3.0) [64] and GenomeScope [65]. To estimate the continuity of the assembly, we used QUAST (5.0.2) [66] to calculate the contig N50 and scaffold N50. To estimate the structural accuracy, we used Asset [8] to calculate the reliable block N50 and used BUSCO (5.1.3) [9] to calculate the false duplication rate for the assembly. To estimate the base accuracy, we used Merqury (1.3) [10] to calculate the k-mer QV and k-mer completeness for the assembly, used BWA (0.7.17) [67] to map the short reads of the selected white eared pheasant individual to the assembly, and used SAMtools (1.10) [68] to analyze the mapping results. To estimate the functional

completeness, we used BUSCO (5.1.3) [9] to assess the completeness of the assembly against the avian gene set. To plot the heatmap of the scaffolds of the assembly, we mapped the Hi-C paired-end short reads to the assembly using BWA (0.7.17) [67], used SAMtools (1.10) [68] and Pairtools (0.3.0) [69] to analyze the mapping results, and used Higlass [70] to plot the heatmap for the assembly.

5.2.9 Protein-coding gene annotation

To annotate the protein-coding genes and pseudogenes in the assembled genomes, we used a combination of reference-based and RNA-based method. For the reference-based method, we used the CDSs of 53 well-represented Aves in NCBI (Supplementary Table 5-1) as the templates. We mapped all the CDSs isoforms of genes in the 53 Aves to the WT and BR assemblies using Splign (2.0.0) [71]. For each template gene whose CDSs can be mapped to an assembly, we concatenated all the mapped parts on the assembly and checked whether the resulting sequence forms an intact ORF (the length is an integer time of three and contain no stop codon in the middle). If yes, we called it an intact gene. If the CDSs of the template gene can be mapped to multiple loci on the assemblies, we choose the locus with the highest mapping identity. If the target sequence did not form an intact ORF (the length is not an integer time of three or/and contain one or more stop codons in the middle), we mapped the Illumina DNA short reads from the same individual to the assembly allowing no mismatch and gaps using bowtie (2.4.1) [72]. If the sequence can be supported by at least 10 short reads at each nucleotide position, we consider the sequence as a pseudogene; otherwise, we called it a partially supported gene since the pseudogenization might be caused by the sequencing errors in long reads that cannot be corrected by DNA short reads.

For the RNA-based method, we first mapped all the RNA-seq reads from multiple tissues of WT and BR into the rRNA database SILVA_138 [73] and filtered out the mapped reads. We then mapped the unaligned reads of each pheasant to their corresponding assemblies using STAR (2.7.0c) [74]. Based on the mapping results, we assembled the mapped reads for each pheasant into transcripts using Trinity (2.8.5) genome-guided method [75]. Then we mapped the transcripts of each pheasant to their corresponding assemblies using Splign (2.0.0) [71] and removed those that partially overlapped the genes, pseudogenes predicted by the reference-based method or the non-coding RNA genes (see below). For the remaining transcripts, if we could find a longest ORF with at least 300 bp, we called it a protein-coding gene. If multiple ORFs are found in a transcript, we select the longest one.

For the non-coding RNA genes, we annotated the tRNA, miRNA, rRNA, snoRNA, telomerase RNA and SRP RNA using infernal (1.1.2) [76] with Rfam (v.14) database [77] as the reference for each of the two pheasants.

5.2.10 Single nucleotide variants calling

To calculate the fixation rate of the pseudogenes in the population of each of the two pheasants, we called the SNPs in the pseudogenes of each of the two pheasants using the DNA short reads from the species population (n=10 for WT and n=40 for BR) using GATK (4.1.6) [78].

To call the variants for the populations of BL, BR and WT, we first mapped the short reads of each individual to the WT genome using BWA (0.7.17) [67] and SAMtools (1.9) [68] with the default settings, and then we called the variants for each individual using GATK-HaplotypeCaller (4.1.6) [78] and merged the variants of each individual

from the same species using GATK-CombineGVCFs (4.1.6) [78] with the default settings. We removed the SNPs with Quality by depth (QD) < 2, Fisher strand (FS) > 60, Root mean square mapping quality (MQ) < 40, Strand odd ratio (SOR) > 3, Rank Sum Test for mapping qualities (MQRankSum) < -12.5 and Rank Sum Test for site position within reads (ReadPosRankSum) < -8 and indels with QD < 2, FS > 200, SOR > 10, Likelihood-based test for the consanguinity among samples (InbreedingCoeff) < -0.8 and ReadPosRankSum < -20.

5.2.11 Population structure and PCA

Principal component analysis was performed using plink (1.90) [79] using the 16 million SNPs in the three species. The same SNPs were used to infer the population structure using ADMIXTURE (1.3.0) [24].

5.2.12 Population diversity estimation

The genetic diversity was measured by the number of heterozygous SNPs per base pair in each species. The pairwise nucleotide diversity was computed on synonymous and nonsynonymous sites in each species. The Tajima's D was estimated using ANGSD (0.921) [80].

5.2.13 Selective sweeps detection

Selective sweeps were detected using two methods including genetic differentiation (F_{ST}) [138] and difference in nucleotide diversity ($\Delta\pi$). We estimated F_{ST} for each comparison using VCFtools (0.1.16) [139] with a sliding window of 40 kb and the step size of 20 kb. We estimated π for each species using VCFtools (0.1.16) [139] with a sliding window of 40 kb and the step size of 20 kb and calculated the difference in nucleotide diversity ($\Delta\pi$) in each window for each comparison. For F_{ST} method, we

defined windows with top 10% F_{ST} score as the putative selective sweeps, and for $\Delta\pi$ method, we defined windows with bottom 5% or top 5% $\Delta\pi$ score as the putative selective sweeps. We consider windows supported by both of the two methods as the final selective sweeps. For the selective sweeps in each comparison, we used the value of $\Delta\pi$ to help identify the selective sweeps of each species. Windows with $\Delta\pi$ score in the bottom 5% represent selective sweeps of the minuend species, and windows with $\Delta\pi$ score in the top 5% represent selective sweeps of the subtrahend species.

5.3 Results

5.3.1 High-quality genome assembly of a WT individual

We generated Illumina paired-end short reads (102X), PacBio long reads (91X) and Hi-C paired-end short reads (81X) for a female WT individual (Supplementary Table 5-2). Using the short and long sequencing reads, we assembled the genome into 805 contigs with a contig N50 of 19.63 Mb. The total length of the contigs is 1.02 Gb, comparable to those of the chicken (*Gallus gallus*) genome assemblies GRCg6a and GRCg7b/w as well as of the previously assembled BR genome [191] (1.01Gb) (Table 5-1). Using the Hi-C paired-end short reads (Supplementary Table 5-2), we further assembled the contigs into 643 scaffolds with a scaffold N50 of 29.59 Mb (Table 5-1). We assessed the quality of the assembly using the criteria proposed by the VGP consortium [4], and compared it with chicken assemblies GRCg6a and GRCg7b/w, the best-studied bird genomes. These criteria include genome features (heterozygosity and repeat rates), continuity (assembly size, N50 and gaps), structure accuracy (reliable block N50 and false duplication rate), base accuracy (k-mer QV, k-mer completeness and short reads mapping rate) and functional completeness (BUSCO completeness) (Table 5-1).

The heterozygosity rate of the WT is 0.54%, and its repeat rate is 20.6%, both are comparable to those of the GRCg6a and GRCg7b/w assemblies (Table 5-1). For the continuity, the contig N50 (19.6 Mb) of the assembly is slightly larger than those of the GRCg6a and GRCg7b/w assemblies. The scaffold N50 (29.6 Mb) of the assembly is slightly larger than that of the GRCg6a assembly, but smaller than those of the GRCg7b/w assemblies (Table 5-1). For the number of gaps, there are only six gaps in our assembly, which is much fewer than those of the GRCg6a and GRCg7b/w assemblies (Table 5-1), indicating that our assembly is almost gapless. For the structural accuracy, the reliable block N50 [27] of our assembly (14.6 Mb) is comparable to those of Avian genomes assembled by the recent VGP consortium [4]. The false duplication rate [10] of our assembly (0.3%) is slightly smaller than those of the GRCg6a and GRCg7b/w assemblies (Table 5-1), indicating that the structural accuracy of our assembly is very high. For the base accuracy, the k-mer QV of our assembly is 42.0, suggesting that the consensus base accuracy is greater than 99.99% [28] (Table 5-1). The k-mer completeness (defined as the fraction of reliable k-mers in highly accurate short reads data that are also found in the assembly [28]) of our assembly is 95.3% (Table 5-1), which is comparable to those of the recent VGP assemblies [4]. To further evaluate the base accuracy, we mapped the Illumina short reads of the WT individual to the assembly and found that 99.3% short reads can be mapped to the assembly (Table 5-1), suggesting that our assembly is of high base accuracy. For the functional completeness, we achieved a larger BUSCO completeness [10] value (97.2%) than those of the GRCg6a and GRCg7b/w assemblies (Table 5-1), suggesting that our assembly is of high functional completeness. To further check whether our assembly is at chromosome-level, we plotted

the Hi-C interaction heatmap of the scaffolds. As shown in Figure 5-1a, almost all the scaffolds form a square at the diagonal of the heatmap, indicating that our assembly is at chromosome-level, although we lack genetic marks to sort them into specific chromosomes.

Table 5-1 Evaluation of the genome assemblies

	Genome		Continuity				Structural Acc.		Base Acc.			Func. Comp
Breed	Het (%)	Rep (%)	Size (Gb)	# Contigs (Contig N50) (Mb)	# Scaffolds (Scaffold N50) (Mb)	# Gaps	Reliable block N50 (Mb)	False duplications (%)	k-mer QV	k-mer Comp. (%)	Short reads Comp. (%)	BUSCO Comp. (%)
WT	0.54	20.6	1.023	805 (19.6)	643 (29.6)	6	14.6	0.3	42.0	95.3	99.3	97.2
GRCg6a	-	20.6	1.056	1,402 (17.7)	464 (20.8)	500,945	-	0.4	-	-	-	96.6
GRCg7b	-	20.6	1.050	677 (18.8)	214 (90.9)	463	-	0.4	-	-	-	96.6
GRCg7w	-	20.2	1.046	685 (17.7)	276 (90.6)	409	-	0.4	-	-	-	96.8

5.3.2 Annotation of genes in the WT and BR genomes

Using the coding DNA sequences (CDSs) of 53 well-represented Aves (Supplementary Table 5-1) as the templates and the RNA-seq data from various tissues of the WT and BR individuals, we annotated 16,315 and 15,003 protein-coding genes and 1,519 and 1,976 pseudogenes in the WT and BR genomes, respectively (Table 5-2). The vast majority (15,565 and 14,727) of the genes and all the pseudogenes in the WT and BR genomes were predicted based on the CDSs in the reference genomes, while a small portion (750 and 276) were predicted based on their respective 20 and three available RNA-seq datasets (RNA-based genes) (Table 5-2). The smaller number of RNA-based genes (276) predicted in BR might be due to the small number of RNA-seq datasets used. Of the 750 RNA-based genes predicted in WT, 743 can be mapped to the BR genome. Most of the reference-based putative genes (14,815 and 14,518) in both species have an intact open reading frame (ORF) (intact genes) while the remaining small numbers (750 and 209) contain stop-gain or ORF-shift mutations that are not supported by the corresponding short DNA reads, thus, we called them partially supported genes, as the observed stop-gains and ORF-shifts might be caused by sequencing errors, particularly,

in long reads. The vast majority of the intact genes (99.01%), partially supported genes (97.73%) and pseudogenes (95.13%) in WT were transcribed in at least one of the 20 tissues we examined (Supplementary Table 5-3). These percentages (96.54%, 93.78% and 90.84%) are smaller in BR (Supplementary Table 5-4) as RNA-seq data from only three tissues are available [191]. The RNA-based putative genes were not seen in the 53 reference avian genomes, thus, they are likely new genes. Most (694 and 259) of these putative new genes in both species can be mapped to the NT database (NT-supported new genes), and the remaining small numbers (56 and 17) cannot be mapped to the NT database, thus they are likely novel genes (Table 5-2).

Table 5-2 Gene annotation of the two species

Breeds	Reference-based			RNA-based		# Total genes	# Total pseudogenes
	# Intact	# Partial	# Pseudogenes	# NT-supported	# Novel		
WT	14,815	750	1,519	694	56	16,315	1,519
BR	14,518	209	1,976	259	17	15,003	1,976
Common	14,057		1,040	121	0	14,178	1,040

5.3.3 New genes found in both WT and BR are likely functional

Interestingly, 121 genes are shared by the 750 and 276 new genes we found in the WT and BR assemblies, respectively (Table 5-2). Thus, these shared genes are likely authentic as the same transcriptional noise are unlikely to occur in two different species. Moreover, as all the new genes do not have GO term assignment, to further validate these new genes in both species, we analyzed their expression levels in the 20 and three different tissues from WT (Supplementary Table 5-3) and BR (Supplementary Table 5-4), respectively. We found that the expression levels of the new genes in both species were highly tissue-specific (Figures 5-1b, 5-1c, 5-1d and 5-1e), indicating that they might be authentic and functional. In addition, we randomly selected 17 new genes out of the 750 new genes in WT and measured their expression levels in 16 tissues using RT-qPCR.

We found that 15 (88.24%) of them were transcribed in at least one of the tissues examined (Supplementary Table 5-5, Figure 5-1f), further suggesting that most of the new genes at least in WT are likely to be authentic, although the expression patterns of some genes are different from those seen in the RNA-seq data due probably to the higher sensitivity of RT-qPCR than that of RNA-seq (Figure 5-1g).

To see whether the genes that we annotated in the two species contain any of the 274 genes that were widely encoded in reptiles and mammals but were reported missing in avian species [11], we compared the missing genes with the genes annotated in the two species and found that we annotated 12 and eight of them in WT and BR, respectively (Supplementary Table 5-6).

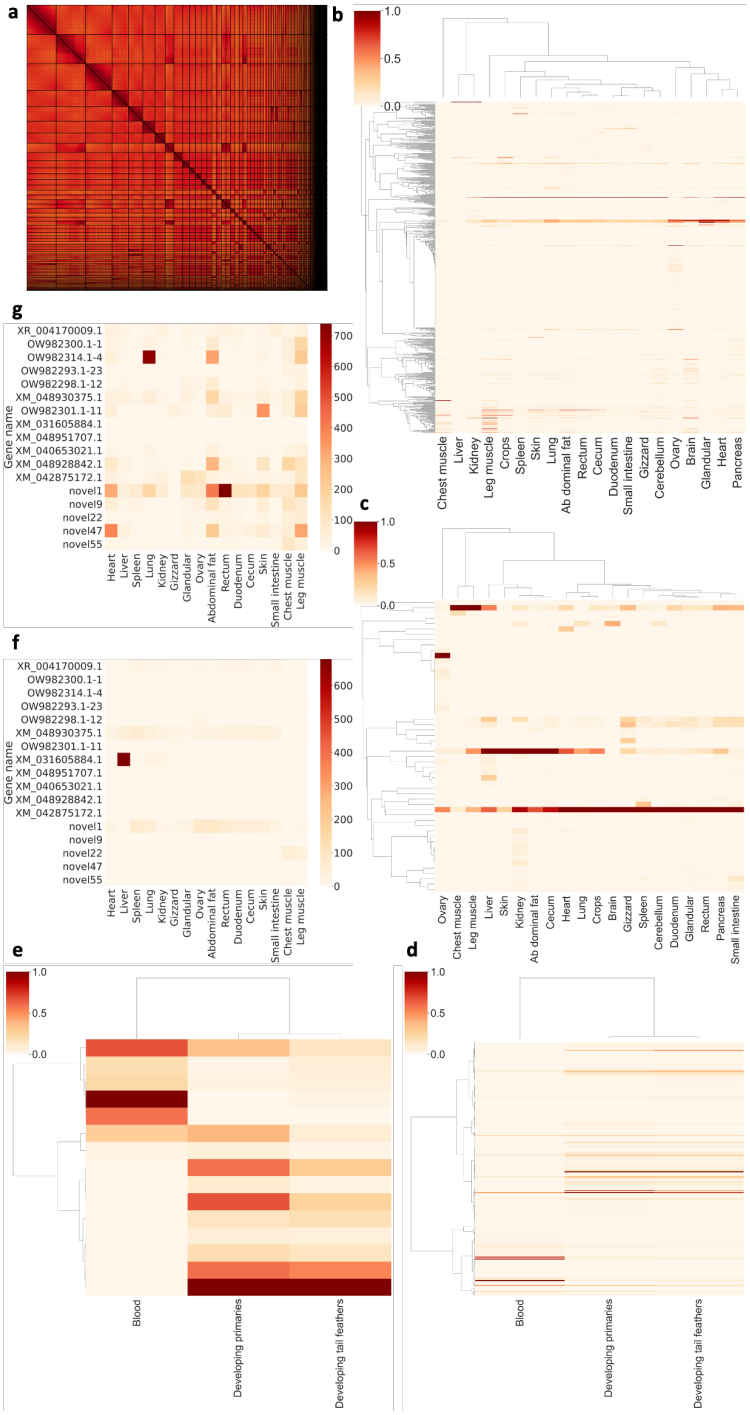

Figure 5-1 Heatmaps of the assembly and new genes of the two species. **a.** Hi-C interaction heatmap of the scaffolds of the WT assembly. **b.** Expression pattern of the NT-supported new genes in 20 tissues of WT. **c.** Expression pattern of the novel genes in 20 tissues of WT. **d.** Expression pattern of the NT-supported new genes in three tissues of BR. **e.** Expression pattern of the novel genes in three tissues in BR. **f.** Expression pattern of the 17 selected new genes in WT measured by RT-qPCR. **g.** Expression pattern of the 17 selected new genes from RNA-seq calculation in WT.

5.3.4 Numerous pseudogenes are found in WT and BR

As indicted earlier, we annotated 1,519 and 1,976 pseudogenes in the WT and BR genomes, respectively. To see whether or not the pseudogenization alleles are fixed in the respective population, we mapped the short reads from 10 WT individuals and 41 BR individuals to the corresponding genomes and found that most of the first pseudogenization alleles along the genes were fixed in the population of the two species (Figures 5-2a and 5-2b), suggesting that these pseudogenes might contribute to the phenotypic traits of the two species. We then analyzed the positions of the first pseudogenization alleles of the pseudogenes along the CDSs. We reason that if these pseudogenization mutations are selectively neutral, then they should uniformly distribute along the CDSs. Indeed, the synonymous mutations in true genes in the two species, which are generally assumed to be selectively neutral, are largely uniformly distributed long the CDSs, except at the two ends, where the numbers of synonymous mutations decrease, consistent with an earlier report in chickens [52] (Figures 5-2c and 5-2d). The reduced synonymous mutations suggest that the two ends of CDSs might harbor functional elements not related to their coding functions, such as transcriptional and post-transcriptional regulatory elements [53]. In contrast, pseudogenizations are strongly biased to the 5'-ends (~40%) and 3'-ends (~20%) in both species (Figures 5-2c and 5-2d). The same phenomenon was also found in the other species such as human [52, 54] and chickens [52]. Therefore, the strongly biased pseudogenizations to the 5'-end and 3'-end in the two species strongly suggest that they are under positive selection. Obviously, the closer a pseudogenization mutation to the 5'-end of a CDS, the larger the affected part of the translated peptide, and thus, the more likely the mutation would eliminate the

protein's function. Although pseudogenization mutations close to the 3'-ends of CDSs might still produce at least partially functional proteins, the mutations might also disrupt regulatory elements at the 3'-ends, and thus may have functional consequence. Therefore, the biased pseudogenization mutations to the two ends suggest that natural selection tends to eliminate the functions of the parent genes of pseudogenes.

Next, we checked possible functional alternative transcripts of the pseudogenes found in both species. In WT, 727 of the 1,519 pseudogenes have alternative transcripts, but only 10 (1.4%) of them have functional alternative transcripts (Supplementary Table 5-7). In BR, 1,712 have alternative transcripts, but only 17 (1.0%) of them have functional alternative transcripts (Supplementary Table 5-8). These results suggest that most pseudogenes in both species do not have functional alternative transcripts and thus most of them have lost the functions of their parent genes.

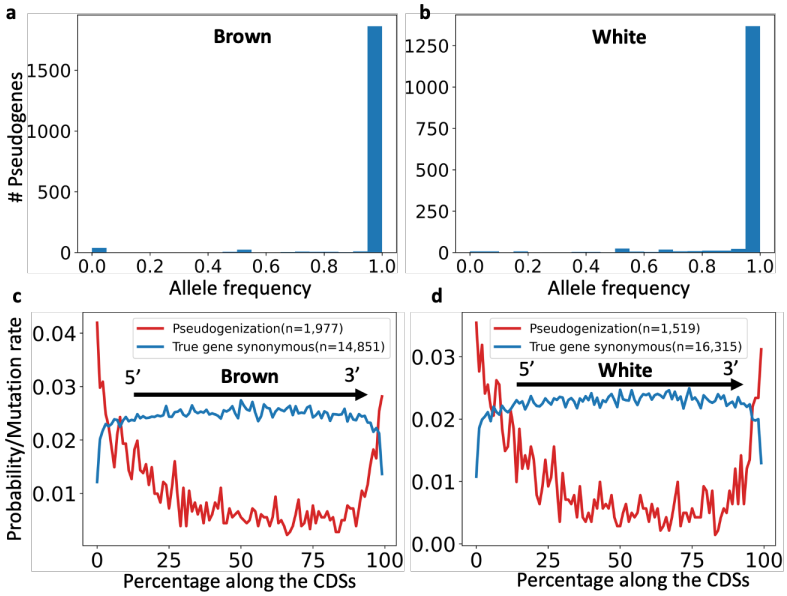

Figure 5-2 Distribution and fixation rate of pseudogenes in WT and BR. **a.** Number of pseudogenization alleles with indicated frequencies in the BR population. **b.** Number of the pseudogenization alleles with indicated frequencies in the WT populations. **c.** Distribution of the first pseudogenization alleles along the CDSs of BR. **d.** Distribution of the first pseudogenization alleles along the CDSs of WT.

5.3.5 Unique genes and pseudogenes are identified in WT and BR

Of the annotated genes and pseudogenes in the WT and BR genomes, 14,178 genes and 1,040 pseudogenes are shared by the two species, respectively (Table 5-2), while 2,137 genes and 479 pseudogenes are unique to WT, and 825 genes and 936 pseudogenes are unique to BR (Supplementary Tables 5-9~5-12). Of the 2,137 unique genes in WT, 658 (30.8%) become unique pseudogenes in BR, while 70.3% (658/936) of the unique pseudogenes in BR are unique genes in WT. Of the 479 unique pseudogenes in WT, 173 (36.1%) are unique genes in BR, while 21.0% (173/825) unique genes in BR become unique pseudogenes in WT. To see whether the unique genes and pseudogenes in each species are related to their unique traits and evolutionary pressure from their niches, we assigned the gene ontology (GO) [48] terms to the unique genes and pseudogenes in each species based on their homologs in humans. Although most of the unique genes and pseudogenes in both WT and BR did not have GO term assignments, those in WT were involved in 69 and 25 biological pathways (Supplementary Table 5-13), respectively, while those in BR were involved in 27 and 59 biological pathways, respectively (Supplementary Table 5-13). Not surprisingly, pathways involving unique genes in WT are often shared with those (59) affected by unique pseudogenes in BR (Figure 5-3), and vice versa, since unique genes with GO pathway assignments in WT or BR are often pseudogenized in BR or WT (Figure 5-3). Interestingly, these unique genes and unique pseudogenes are mainly involved in and affect, respectively, pathways related to four functional categories (Supplementary Table 5-13).

1) Cardiovascular functions: For example, the unique gene *PIK3C2G* in BR, which is involved in HIF activation pathway [192-194], is a unique pseudogene in WT.

Unique genes *SH2D2A*, *PRKCD*, *PRKD3* and *NOS3* in WT, which are involved in the VEGF signaling pathway that play critical roles in proliferation, survival and migration of blood endothelial cells required for angiogenesis pathway [195-197], are missing in BR. Unique genes *SH2D2A*, *MAP3K1*, *PRKCD*, *WNT1*, *PRKD3*, *NOS3*, *EPHA3*, *NOTCH1* and *GRB14* in WT, which are involved in Angiogenesis pathway, are missing (*SH2D2A*, *PRKCD*, *PRKD3*, *NOS3* and *EPHA3*) or pseudogenized (*MAP3K1*, *WNT1*, *NOTCH1* and *GRB14*). Unique genes *DOK1* and *PIK3C2G* in BR, which are involved in the Angiogenesis pathway, are pseudogenized in WT. Unique genes *PRKCD*, *NOS3* and *NPR2* (Endothelin signaling pathway) in WT are missing (*PRKCD* and *NOS3*) or pseudogenized (*NPR2*) in BR. Unique gene *ARRB1* (Angiotensin II-stimulated signaling through G proteins and beta-arrestin pathway) in BR is missing in WT. And unique gene *VWF* (Blood coagulation pathway) in WT is pseudogenized in BR (Figure 5-3).

2) Energy metabolism: For example, unique gene *PFKM* in BR, which is involved in glycolysis, is pseudogenized in WT. Unique genes *PRKCD*, *CACNB1* and *CACNB3* in WT, which are involved in Thyrotropin-releasing hormone receptor signaling pathway, are missing (*PRKCD* and *CACNB3*) or pseudogenized (*CACNB1*) in BR. Unique genes *IRS2* and *TSC2* (Insulin/IGF pathway-protein kinase B signaling cascade pathway) in WT are missing (*TSC2*) or pseudogenized (*IRS2*) in BR. And unique gene *TSC2* (p53 pathway by glucose deprivation) in WT is missing in BR (Figure 5-3).

3) Neuronal functions: For example, unique genes *CACNA1D*, *CACNB1* and *CACNB3* (Beta1/2 adrenergic receptor signaling pathway) in WT are missing (*CACNA1D* and *CACNB3*) or pseudogenized (*CACNB1*) in BR. Unique gene *ADRB1* (Beta1/2 adrenergic receptor signaling pathway) in BR is pseudogenized in WT. Unique genes

GRIK1 and *GRIN2C* (Metabotropic glutamate receptor group I pathway) in WT are pseudogenized in BR. Unique gene *CACNB1* (Metabotropic glutamate receptor group III pathway) in WT is pseudogenized in BR. Unique gene *GABBR1* (GABA-B receptor II signaling pathway) in WT is missing in BR. Unique genes *CAMK2G*, *GRIK1*, *SHANK2*, *GRIN2C* and *CAMK2B* (Ionotropic glutamate receptor pathway) in WT are missing (*SHANK2* and *CAMK2B*) or pseudogenized (*CAMK2G*, *GRIK1* and *GRIN2C*) in BR. And unique genes *CACNA1D*, *MYO15A* and *CACNB1*(Nicotinic acetylcholine receptor signaling pathway) in WT are missing (*CACNA1D*) or pseudogenized (*MYO15A*, *CACNB1*) in BR (Figure 5-3).

4) Immunity: For example, unique gene *MAP3K1* (T cell activation pathway) in WT is pseudogenized in BR. Unique gene *PIK3C2G* (T cell activation pathway) in BR is pseudogenized in WT. Unique genes *CD79A* and *PRKCD* (B cell activation pathway) in WT are missing in BR. Unique genes *PREX1*, *CAMK2G*, *ITGB7*, *VWF*, *TYK2*, *IL6* and *CAMK2B* (Inflammation mediated by chemokine and cytokine signaling pathway) in WT are missing (*ITGB7*, *VWF*, *TYK2* and *CAMK2B*) or pseudogenized (*PREX1*, *CAMK2G*, *VWF* and *IL6*) in BR. And unique gene *ARRB1* (Inflammation mediated by chemokine and cytokine signaling pathway) in BR is missing in WT (Figure 5-3).

The enrichment of the unique genes and pseudogene in the four functional categories might be due to the strikingly different altitudes of the niches of the two species, posing strikingly different pressures on these related functional systems that have been shown to be involved in high altitude-related adaptation [198]. Moreover, unique genes or pseudogenes in the two species are also involved in pathways that might be related to the adaptation of the two species to other aspects of their distinct ecological

niches, and in the case of BR, also as a consequence of population decline. For example, unique genes *CACNA1D*, *PRKCD*, *CACNB1* and *CACNB3* in WT, which are involved in Oxytocin receptor mediated signaling pathway that is involved in social behaviors [199], are missing (*CACNA1D*, *PRKCD* and *CACNB3*) or pseudogenized (*CACNB1*) in BR; unique genes *IRS2*, *SMAD4*, *FOSB*, *CACNA1D*, *MAP3K1*, *PRKCD*, *MAP2K7*, *EP300*, *NR5A1*, *NPR2*, *MAP3K5* and *CAMK2B* in WT are missing (*SMAD4*, *FOSB*, *CACNA1D*, *PRKCD*, *MAP2K7*, *NR5A* and *CAMK2B*) or pseudogenized (*IRS2*, *MAP3K1*, *EP300*, *NPR2* and *MAP3K5*) in BR, these genes are in involved in Gonadotropin releasing hormone receptor pathway that regulate reproduction.

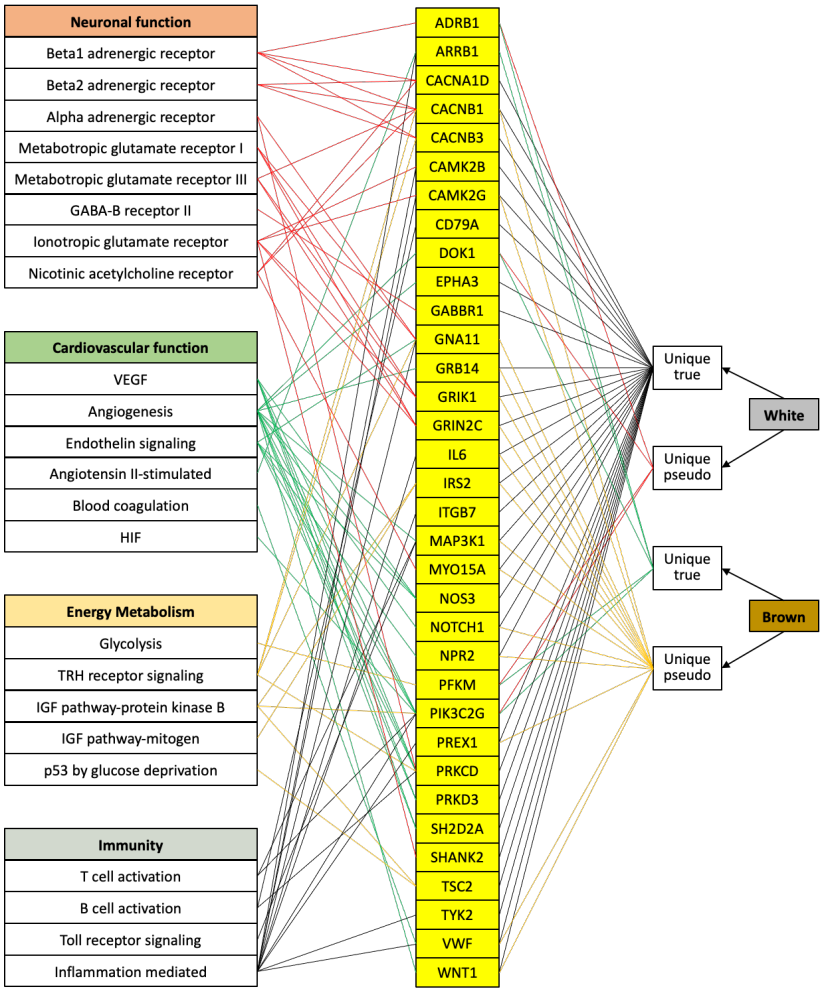

Figure 5-3 Pathways involved in unique genes/pseudogenes in WT and BR

5.3.6 Fragmental populations are clustered distinctly

To further understand the genetic basis of altitude adaptation, in addition to the comparative analysis of the assembled genomes of WT (high altitude) and BR (low altitude), we also compared SNP spectrums of WT (10 individuals), BR (41 individuals) and BL (12 individuals) populations. We included the BL population in the analysis as they inhabit at intermediate altitude (1,000~1,500 m). Figure 5-4a shows the geographical locations where the populations were sampled. Specifically, three BR subpopulations were from Shannxi, Shanxi and Hebei & Beijing provinces, respectively [191]; the BL populations were from Gansu province [191]; and the WT populations from Yunnan province. Using our assembled genome of the WT individual as the template, we identified 10, six and nine million SNPs in the BL, BR and WT populations, respectively (Table 5-3). After removing the redundancy, we ended up with 16 million SNPs in the populations of the three species (Table 5-3). Based on the 16 million SNPs, we performed principal component analysis (PCA) of the population of the three species. As shown in Figure 5-4b, individuals of the same species form distinct clusters. Notably, each of the three BR subpopulations forms a distinct compact subcluster, which is consistent with the previous result [191]. We also run the admixture algorithm [200] on the called SNPs. As shown in Figure 5-4c, when two clusters were set, individuals of BL and BR were grouped into one cluster, which is consistent with previous phylogenetic analysis [190], while WT individuals formed the other cluster. When three clusters were set, individuals of each species formed a distinct cluster, consistent with the result from PCA. These results suggest that the populations of the three species are strongly structured.

Table 5-3 Summary of the SNPs in each species

Breeds	# individuals	# Total SNPs
BL	12	10,147,889
BR	41	6,849,438
WT	10	9,004,245
All	63	16,459,558

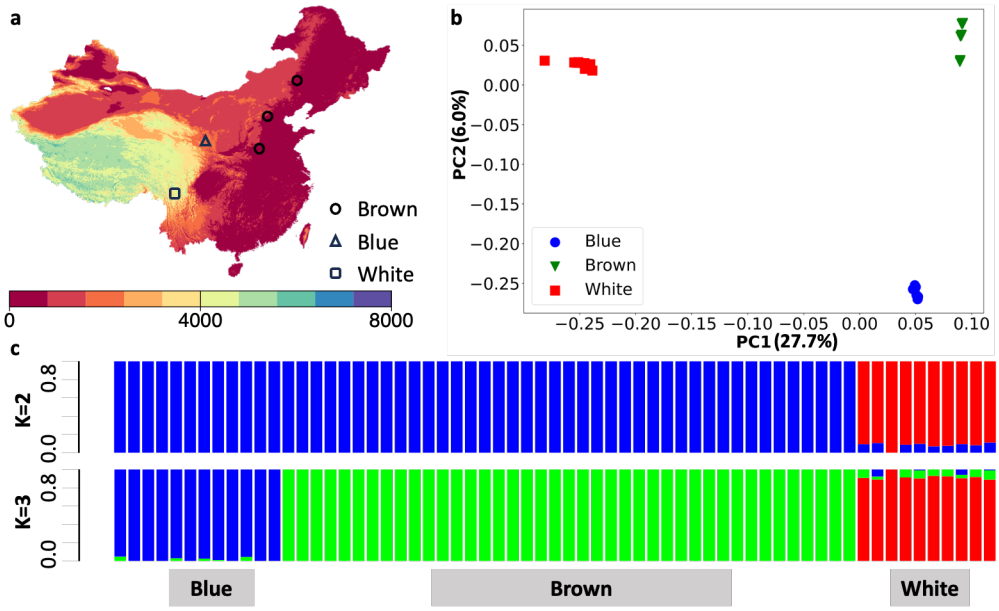

Figure 5-4 Distribution and population structure of WT, BL and BR. **a.** Distribution of the populations of three species used in this study. The color of the map indicates altitude level (m). **b.** Principal component analysis of the populations of the three species based on their SNP spectrum. **c.** Genetic structure of the individuals of the three species estimated using ADMIXTURE.

5.3.7 Low genetic diversity might explain the vulnerability of BR

To evaluate genetic diversity of individuals in each population, we first computed pairwise nucleotide diversity (π) in a 40 kb window with 20 kb step size for each population. As shown in Figure 5-5a, BR has the lowest mean π value (3.61e-4), followed by BL (1.60e-3), and WT (2.63e-3). We next computed the heterozygous SNP rate in each individual of each species. As shown in Figure 5-5b, WT has the highest

mean heterozygous SNP rate (1.36e-3), followed by BL (6.73e-4), and BR (1.84e-4). We then calculated the pairwise nucleotide diversity for 0-fold and 4-fold degeneration SNPs in each species population. As shown in Figure 5-5c, BR has the lowest nucleotide diversity (2.67e-5) for 0-fold degeneration sites, followed by BL (3.34e-5) and WT (4.28e-5). The same pattern is seen in the nucleotide diversity for 4-fold degeneration sites of the three species (Figure 5-5d). To see whether BR accumulates more missense mutations, we computed the ratio of nucleotide diversity of 0-fold degeneration sites over that of 4-fold degeneration sites ratio in each of the three species (Figure 5-5e). We found that BR has the highest ratio value (1.15), followed by BL (0.93) and WT (0.90), suggesting that BR contains more deleterious mutations than the other two species, which is a genomic reflection of the vulnerability of BR. Moreover, the BR population have the highest positive Tajima's D value among the three species (Figure 5-5f), suggesting that they have undergone a more dramatic decrease in population size, and have fewer rare alleles. All these results indicate that BR has the lowest population genetic diversity among the three eared pheasants, and that BR has undergone a more serious population bottleneck than the other two species. Taken together, these results strongly suggest that WB is in higher risk of extinction than the other two species, which is consistent with the previous results [191].

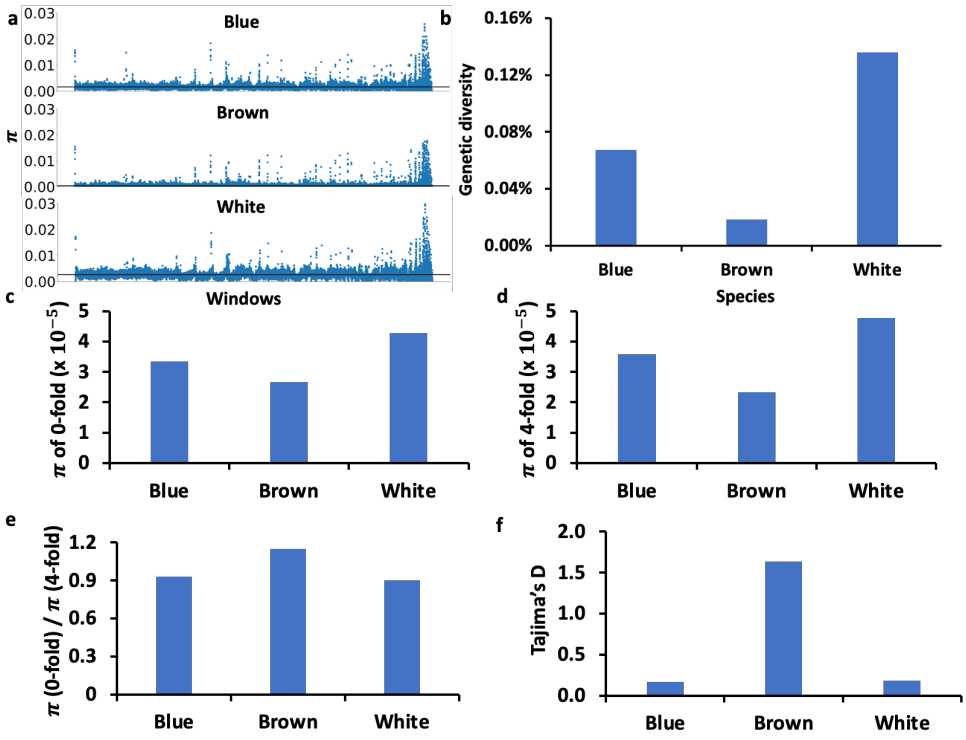

Figure 5-5 Summary of population genetics statistics. **a.** Nucleotide diversity in 40-kb windows with 20-kb step size of the three species using the WT assembly as the reference. **b.** Genetic diversity of each species. **c.** Nucleotide diversity on 0-fold degeneration sites of each species. **d.** Nucleotide diversity on 4-fold degeneration sites of each species. **e.** Nucleotide diversity of 0-fold degeneration sites over nucleotide diversity of 4-fold degeneration sites of each species. **f.** Genome-wide Tajima's D of each species.

5.3.8 Altitude adaptation-related pathways are under selection

To identify genomic regions and genes that might be related to the adaptation of the three species to their different ecological niches, particularly, to different altitudes, we identified selective sweeps in each species by comparing its SNPs spectrum with those of the other two species using both the genetic differentiation (F_{ST}) and difference in nucleotide diversity ($\Delta\pi$) parameters. Figure 5-6 shows the distribution of F_{ST} and $\Delta\pi$ as well as their values for each 40 kb genome windows for the three pairwise comparisons. We consider a window with a top 10% F_{ST} values and a top 5% or a bottom 5% $\Delta\pi$ values as a selection sweep. Since adjacent selective sweeps can overlap with one

another, we merged the overlapping ones as a discrete selection sweep (DSS) in each species. For the comparison of BL and BR (BL VS. BR), we identified 22 and 194 selective sweeps, 14 and 151 discrete selective sweeps (DSSs) containing six and 175 genes in BL and BR, respectively (Table 5-4, Supplementary Table 5-14). For the comparison of BL and WT (BL VS. WT), we identified 103 and 611 selective sweeps, 81 and 350 DSSs containing 76 and 341 genes in BL and WT, respectively (Table 5-4 and Supplementary Table 5-15). For the comparison of BR and WT (BR VS. WT), we identified 120 and 478 selective sweeps, 85 and 252 DSSs containing 78 and 254 genes, respectively (Table 5-4 and Supplementary Table 5-16) (Figure 5-6).

Although most of the genes in the selective sweeps identified in each species do not have GO term assignments, we found that they are mostly involved in the same pathways in the four functional categories as are the unique genes and unique pseudogenes in the WT and BR (Figure 5-3). Specifically, for BL, genes in selective sweeps identified in the 'BL VS. BR' and 'BL VS. WT' comparisons are involved in the same number of 22 GO pathways (Table 5-4). Compared to BR, six genes in the selection sweeps in BL have GO pathway assignments (Supplementary Table 5-14). Among these genes, *MAPK1* is involved in cardiovascular functions such as angiogenesis pathway, VEGF signaling pathway [195-197], Endothelin signaling pathway [201-203], and Angiotensin II-stimulated signaling through G proteins and beta-arrestin pathway (Figure 5-7). Compared to WT, 76 genes in the selection sweeps in BL have GO pathway assignments (Supplementary Table 5-15). Among these genes, *PRKAR2B* is involved in neuronal functions such as Beta1/2 adrenergic receptor signaling pathway, Metabotropic glutamate receptor group III pathway, GABA-B receptor II signaling pathway and

Endothelin signaling (Figure 5-7). As the cardiovascular system plays a crucial role in long term response to hypoxia [204], selection on cardiovascular function-related pathways through *MAPK1* and *PRKAR2B* genes in BL might be related to its higher altitude niches. Selection of neuronal related pathways through *PRKAR2B* might be also beneficial for BL to survive in its ecological niches.

For BR, genes in the selective sweeps identified in the 'BL VS. BR' and 'BR VS. WT' comparisons are involved in 27 and 32 GO pathways, respectively (Table 5-4). Compared to BL, 175 genes in the selection sweeps in BR have GO pathway assignments (Supplementary Table 5-14). Among these genes, some are involved in cardiovascular functions, such as *PLA2G4A* (VEGF pathway and Angiogenesis pathway), *PLA2G4A* and *ITPR2* (Endothelin signaling pathway), *ITPR2* (Angiotensin II-stimulated signaling through G proteins & beta-arrestin pathway), *PREP* (Vasopressin synthesis pathway [205]), *PLA2G4A* (Oxidative stress response pathway); some are involved in immunity such as *PLA2G4A*, *ITPR2* and *COL6A6* (Inflammation mediated by chemokine and cytokine signaling pathway), *ITPR2* (B cell activation pathway); some are involved in neuronal functions, such as *RYR2* (Beta1/2 adrenergic receptor signaling pathway), *SLC17A6* (Ionotropic glutamate receptor pathway) (Figure 5-7). Compared to WT, 78 genes in the selection sweeps in BR have GO pathway assignment (Supplementary Table 5-16). Some of these genes are involved in cardiovascular functions, such as *PRKCD* (VEGF pathway, Angiogenesis pathway, Endothelin signaling pathway) and *CEP* (Vasopressin synthesis pathway). *PRKCD* is also involved in energy metabolism (Thyrotropin-releasing hormone receptor signaling pathway), immunity (B-cell activation pathway) and neuronal functions (Alpha adrenergic receptor signaling pathway) (Figure

5-7). Selection on these cardiovascular, immunity, energy metabolism pathways might be beneficial for BR to adapt to low altitude.

For WT, genes in the selective sweeps identified in the 'BL VS. WT' and 'BR VS. WT' comparisons are involved in 62 and 48 GO pathways, respectively (Table 5-4). Compared to BL, 341 genes in the selection sweeps in WT have GO pathway assignments (Supplementary Table 5-15). Of these genes, some are involved in cardiovascular functions such as *ARNT2* and *PPARG* (Hypoxia induced factor pathway [192-194, 206]), *RASA1*, *FZD1* and *MAPK8* (Angiogenesis pathway), *PLCB1* (Endothelin signaling pathway, Angiotensin II-stimulated signaling through G proteins and beta-arrestin pathway), *MAPK8* (Oxidative stress response pathway); some are involved in energy metabolism pathway, such as *CACNB2* and *PLCB1* (Thyrotropin-releasing hormone receptor signaling pathway), *RASA1* (Insulin/IGF pathway-mitogen activated protein kinase kinase/MAP kinase cascade pathway), *PPP2CB* (p53 pathway by glucose deprivation); some are involved in neuronal functions, such as *CACNB2* (Beta1/2 adrenergic receptor signaling pathway), *SEMA4D* (Axon guidance mediated by semaphorins pathway); *PLCB1* (Alpha adrenergic receptor signaling pathway), *GRIN3A* and *GRIK5* (Metabotropic glutamate receptor group I pathway), *GRIN3A*, *GRIK5*, *SLC1A1* and *GRIK2* (Metabotropic glutamate receptor group III pathway), *GABBR2* (GABA-B receptor II signaling pathway), *GRIN3A*, *GRIK5*, *SLC1A1* and *GRIK2* (Ionotropic glutamate receptor pathway), *CACNB2* and *ACTA2* (Nicotinic acetylcholine receptor signaling pathway); and some are involved in immunity, such as *PPP3CB*, *MAPK8* and *PTPRC* (T/B cell activation pathway), *MAPK8* and *UBE2V1* (Toll receptor signaling pathway and Interferon-gamma signaling pathway), *ACTA2* and *PLCB1*

(Inflammation mediated by chemokine and cytokine signaling pathway) (Figure 5-7). Compared to BR, 254 genes in the selection sweeps in WT have GO pathway assignments (Supplementary Table 5-16). Of the these genes, some are involved in cardiovascular functions, such as *CUL2* (Hypoxia induced factor pathway[192-194, 206]), *MAPK8* (Oxidative stress response pathway [207-211]); *TEK*, *RASA1*, *FZD1* and *MAPK8* (Angiogenesis pathway), and *VWF* (Blood coagulation pathway); some are involved in energy metabolism, such as *PDPK1* (Insulin/IGF pathway-protein kinase B signaling cascade), *RPS6KA1* and *RASA1* (Insulin/IGF pathway-mitogen activated protein kinase kinase/MAP kinase cascade pathway[212-217]); Some are involved in immunity, such as *MAPK8* (Interferon-gamma signaling pathway and Toll receptor signaling pathway); *PPP3CB*, *MAPK8* and *PTPRC* (T/B cell activation pathway and Toll receptor signaling pathway and Interferon-gamma signaling pathway[218-220]), *CAMK2G*, *VWF*, *MYO3B* and *PDPK1* (Inflammation mediated by chemokine and cytokine signaling pathway [221, 222]); some are involved in neuronal functions, such as *GRIN3A*, *GRIK5* (Metabotropic glutamate receptor group I pathway), *GRIN3A*, *GRIK5* and *GRIK2* (Metabotropic glutamate receptor group III pathway); *GABBR2* (GABA-B_receptor_II_signaling pathway), *CAMK2G*, *GRIN3A*, *GRIK5* and *GRIK2* (Ionotropic glutamate receptor pathway), and *MYO3B* (Nicotinic acetylcholine receptor signaling pathway) (Figure 5-7). Since WT's high-altitude niches with low oxygen pressure and scarcity of food pose a great challenge to its cardiovascular system and energy metabolism, selection on the hypoxia response- and energy metabolism-related pathways compared to the other two species living in relatively lower altitude niches might help

WT to adapt to the high-altitude niches. Selection on immunity- and neuronal function-related pathways is also beneficial for WT to adapt to their unique ecological niches.

Table 5-4 Summary of selective sweeps in different comparisons

Comparison	BL			BR			WT		
	# SS (# DSS)	# Genes in DSS	# Pathways involved	# SS (# DSS)	# Genes in DSS	# Pathways involved	# SS (# DSS)	# Genes in DSS	# Pathways involved
BL VS. BR	22 (14)	6	22	194 (151)	175	27	NA	NA	NA
BL VS. WT	103 (81)	76	22	NA	NA	NA	611 (350)	341	62
BR VS. WT	NA	NA	NA	120 (85)	78	32	478 (252)	254	48

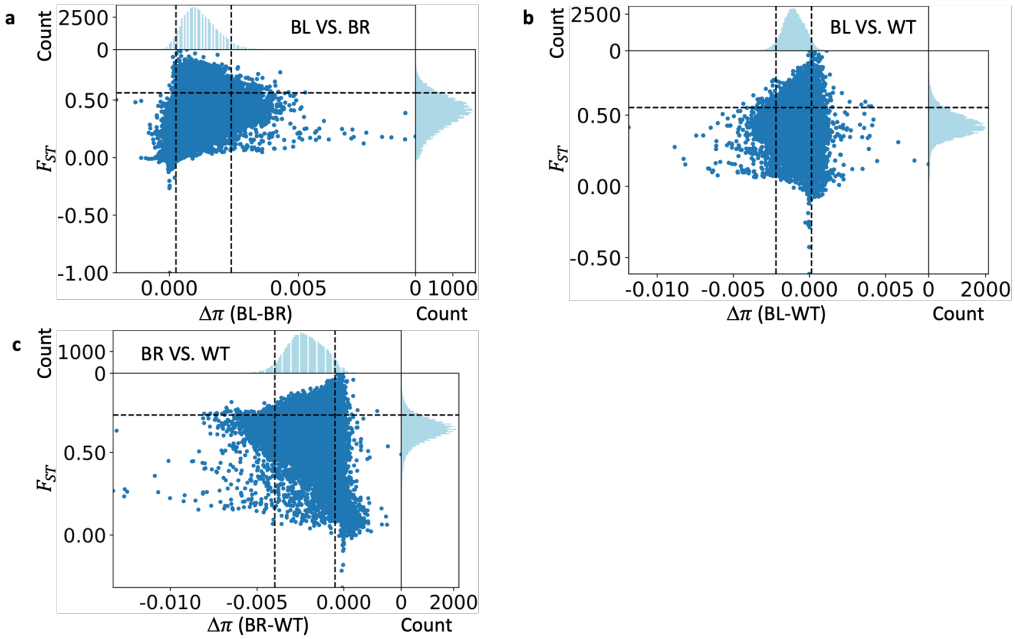

Figure 5-6 Distribution of the F_{ST} and $\Delta\pi$ values in each comparison. **a.** "BL VS. BR" comparison. **b.** "BL VS. WT" comparison. **c.** "BR VS. WT" comparison. In each subfigure, the top panel represents the distribution of $\Delta\pi$ values and the bottom right one represents the distribution of F_{ST}values. Each dot in the bottom left panel represents a sliding window with its $\Delta\pi$ and F_{ST} values being its coordinates in the plot. The dashed lines represent the cut-off for the F_{ST} and $\Delta\pi$.

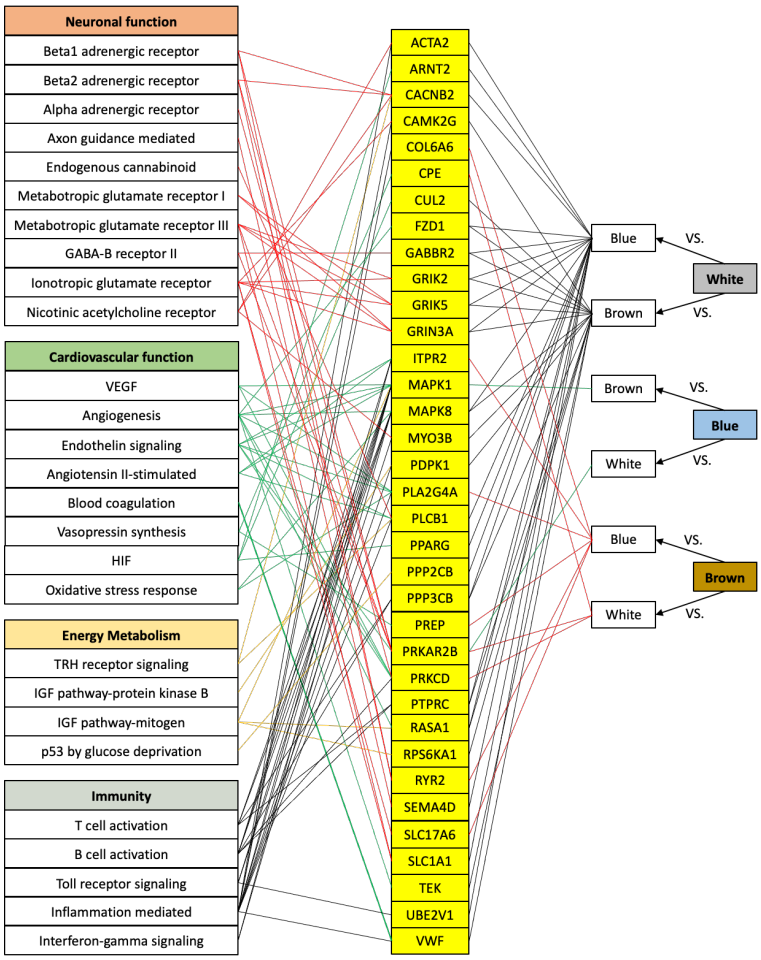

Figure 5-7 Pathways involved in genes in selection sweeps of each species

5.4 Discussion

By using the Illumina short reads, PacBio long reads and Hi-C reads, we assembled the genome for a WT individual. By using the criteria proposed by the VGP project [4] to evaluate quality of the assembly, we found it is of high continuity and accuracy (Table 5-1). The Hi-C interaction heatmap indicates that our assembly is at chromosome-level (Figure 5-1a). Thus, we provide an almost gapless reference genome for the *crossoptilon* species. We annotated and compared the genes and pseudogenes of our assembled WT genome and the previously assembled BR genome [191]. Our annotation covers 12 and eight bird "missing" genes [11], in WT and BR genomes, respectively. At the same time, we annotated 750 and 276 new genes in WT and BR,

respectively, which are not seen in the 53 reference genomes (Table 5-2). RT-qPCR validation suggests that most of the new genes are authentic at least in WT (Supplementary Table 5-5 and Figure 5-1f). These results indicate that our approach of annotating genes is accurate and complete enough.

We found 1,519 and 1,976 pseudogenes in WT and BR, respectively, and most of them are fixed in their corresponding population (Figures 5-2a and 5-2b). This is the first time to find such large numbers of pseudogenes in the *crossoptilon* species, to the best of our knowledge. We also found that the first pseudogenization alleles in the pseudogenes tend to be located in the 5' end (40%) and 3' end (20%) along the CDSs in both species (Figures 5-2c and 5-2d), which is also observed in other species in the previous studies [52, 54]. Therefore, most pseudogenes might lose their functions by disrupting most part of the peptides via pseudogenization occurring near the 5'-ends and affecting the regulatory elements via pseudogenization occurring near the 3'-ends. The larger number of pseudogenes in BR than in WT might be related to the rapid decline of the BR population size, resulting less genetic diversity and more missense mutations.

We found that the WT and BR genomes share most of genes and pseudogene, but they also contain many unique genes and unique pseudogenes. Many unique genes in WT became pseudogenes in BR, and vice versa (Figure 5-3). These unique genes and unique pseudogenes are mainly related to four functional categories: cardiovascular, energy metabolic and immune functions. These alterations of these functions are widely known to be related to adaptation of high altitude [223-231] might be related to the adaptation to different altitude of the two species. For example, we found that gene *PIK3C2G* in BR became a pseudogene in WT. This gene is involved in many pathways, including the HIF

pathway, Endothelin signaling pathway, Angiogenesis pathway, VEGF signaling pathway, T cell activation pathway and Inflammation mediated by chemokine and cytokine signaling pathway (Figure 5-3). Thus, *PIK3C2G* might be related to the adaptation of BR to low altitude and of WT to high altitude for cardiovascular functions and immunity. Gene *CACNB1* in WT became a pseudogene in BR. This gene is involved in the Beta1/2 adrenergic receptor signaling pathway, Metabotropic glutamate receptor group III pathway, Nicotinic acetylcholine receptor signaling pathway and Thyrotropin-releasing hormone receptor signaling pathway (Figure 5-3). Thus, *CACNB1* might be related to the adaptation of WT to high altitude and of BR to lower altitude for neuronal and energy metabolic functions.

By analyzing re-sequencing data of WT, BL and BR populations, we found that the BR population have the lowest genetic diversity and excessive loss-of-function mutations among the three species, consistent with the earlier report [191]. Moreover, we found that the assembled BR genome contain fewer genes but more pseudogenes than those in our assembled WT genome. Although some missing genes and unique pseudogenes in BR might be related to the adaption to their ecological niches, including lower altitude, some might be resulted from the long-term decline of its population.

Interestingly, by pairwise comparisons of the SNP spectrum of WT, BL and BR populations, we found that genes in the selective sweeps in each species converged largely on the same GO pathways of the four functional categories as are the unique genes and unique pseudogenes in the WT and BR genomes. Therefore, the selection on these genes might be related to the adaptation of different altitude. For example, in the "BL VS. BR" comparison, gene *MAPK1* in BL is included in a selective sweep. This

gene is involved in the VEGF signaling pathway, Angiogenesis pathway, Endothelin signaling pathway, Angiotensin II-stimulated signaling through G proteins and beta-arrestin pathway, Insulin/IGF pathway-mitogen activated protein kinase kinase/MAP kinase cascade pathway and many pathways of the immunity (Figure 5-7). Thus, selection on *MAPK1* might be related to the adaptation of BL to intermediate altitude. In the "WT VS. BL" comparison, genes *ARNT2* and *PPARG* are in selective sweeps, and in "WT VS. BR" comparison, gene *CUL2* is in a selection sweep. All these three genes are involved in the HIF pathway [206] (Figure 5-7), thus selection on these three genes might be related to the adaptation of WT to high altitude.

It is not surprising that the unique genes/pseudogenes in WT and BR and genes in the selective sweeps of BL, BR and WT converge on the same pathways involved in cardiovascular, energy metabolic, neuronal and immune functions. On the one hand, high altitude niches with a low oxygen pressure, low temperature and less availability of food would pose a direct pressure on WT's cardiovascular system and energy metabolism functions, while such pressure would be somewhat relaxed for BL and even more relaxed for BR. On the other hand, other altitude related ecological factor might exert different pressures on the pheasants for their neuronal functions for food foray and escaping for predators, and for immune functions for different pathogens. Indeed, it has been shown that high altitude niches can have effects on the neuronal functions such as sensory perception in domestic yaks (Bos grunniens) [232] and olfaction in Tibetan pig and wild boars [224] as well as in ground tit (Parus humilis) [225]. It has also been shown that high altitude niches can have effects on immune functions in ground tit [225]. Thus, it appears that the three eared pheasant species might adapt to highly varying altitudes by

loss-of-function mutation and fine-tuning of genes in the same set of pathways involved in the four functional categories. The unique genes in a species and their missing or pseudogenization in other species as well as selection on genes involved in these pathways in each species might be beneficial for it to adapt to its unique ecological niches.

5.5 Conclusion

By using the Illumina short reads, PacBio long reads and Hi-C data, we assembled the genome for a WT individual at chromosome-level with high quality, providing a reference assembly for the *crossoptilon* species. By annotating genes and pseudogenes of the WT assembly and a BR assembly from a previous study, we identified 2,137 unique genes and 479 unique pseudogenes in WT, and 825 unique genes and 936 unique pseudogenes in BR. These unique genes and pseudogenes in the two species are mainly involved in four categories of GO pathways, including cardiovascular, metabolism, neuronal and immunity functions. By calling the SNPs for the populations of BL, BR and WT, analyzing the selective sweeps in each of the three species and assigning the GO pathways for the genes in the selective sweeps of the three species, we found that these genes are also mainly involved in the four categories of pathways. Pathways in the four categories of functions and involved genes might be related to the adaptation to different altitude of the three species.

CHAPTER 6 Conclusion

In this dissertation, we developed a user-friendly vertebrate genome assembly pipeline using a combination of Nanopore/PacBio long reads, Illumina short reads and Hi-C reads, allowing high quality chromosome-level genome assembling. We also developed a user-friendly accompanying gene annotation pipeline using a combination of homology-based and RNA-based method approaches, allowing more accurate protein-coding gene and pseudogene annotations.

Using these pipelines, we assembled the genomes of four indigenous chickens with high-quality at chromosome-level and annotated the protein-coding genes and pseudogenes in each assembly. We identified a total of 1,420 new protein-coding genes in the four indigenous chickens, most of which are involved in many important biological pathways. Most of the new genes also are encoded or pseudogenized in the GRCg6a and GRCg7b/w assemblies. Thus, chickens encode similar number of protein-coding genes as other tetrapods do. At the same time, we also identified an unexpected large number of pseudogenes in the four chickens, most of which have lost their functions. We found that the occurring patterns of the pseudogenes reflect the phylogenetic relationships of the chickens, suggesting that loss-of-function mutations play important roles in chicken domestication and evolution.

Moreover, using a rigorous model, we found more putative selective sweeps and protein-coding genes that might be related to the specific traits of each chicken breed than previous studies. Most of these genes do not have nonsynonymous mutations, so they might affect chicken phenotypes by altering their expression levels, and we verified a few such cases.

Finally, we assembled the genome of a WT individual with high quality at chromosome-level, providing a good resource to study the *crossoptilon* species. By annotating and comparing the genes and pseudogenes in the WT genome and a previously assembled BR genome, we identified unique genes and pseudogenes in each species. The unique genes and pseudogenes in each species are mainly involved in the same GO pathways in four functional categories, including cardiovascular, metabolism, neuronal and immune functions. Moreover, by analyzing the selective sweeps in BL, BR and WT populations, we found that genes in the selective sweeps of the three species are also involved in the same pathway in the four functional categories. Thus, the three species might adapt to different altitudes by natural selection on different genes in the same pathways.

To summarize, in this dissertation, we have achieved four aims. Firstly, we developed a user-friendly vertebrate genome assembly pipeline and an accompanying gene annotation pipeline. They work back-by-back, allowing high-quality chromosome-level genome assembly and accurate protein-coding genes annotations. Secondly, we assembled the genomes of four indigenous chickens and found that loss-of-function mutations play important roles in chicken domestication and evolution. Thirdly, we called the SNPs in the population of five indigenous chickens, broilers, layers and RJFs and found more putative selective sweeps and genes related to the specific traits of each chicken breed than previous studies. Finally, we assembled the genome of a WT individual, compared genes and pseudogenes in the WT and a BR individual as well as selection sweeps in populations of BL, BR and WT. We found that the three species

might adapt to different altitudes by natural selection on different genes in the same pathways related to cardiovascular, metabolism, neuronal and immune functions.

REFERENCES

1. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature, 2004. **432**(7018): p. 695-716.
2. Warren, W.C., et al., A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. G3 (Bethesda), 2017. **7**(1): p. 109-117.
3. Schmid, M., et al., Third Report on Chicken Genes and Chromosomes 2015. Cytogenet Genome Res, 2015. **145**(2): p. 78-179.
4. Rhie, A., et al., Towards complete and error-free genome assemblies of all vertebrate species. Nature, 2021. **592**(7856): p. 737-746.
5. VGP NCBI Gallus gallus Annotation Release 106. https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Gallus_gallus/106/, 2021.
6. Wang, M.S., et al., 863 genomes reveal the origin and domestication of chicken. Cell Res, 2020. **30**(8): p. 693-701.
7. Lawal, R.A., et al., The wild species genome ancestry of domestic chickens. BMC Biol, 2020. **18**(1): p. 13.
8. Gheyas, A.A., et al., Functional classification of 15 million SNPs detected from diverse chicken populations. DNA Res, 2015. **22**(3): p. 205-17.
9. Li, M., et al., De Novo Assembly of 20 Chicken Genomes Reveals the Undetectable Phenomenon for Thousands of Core Genes on Microchromosomes and Subtelomeric Regions. Mol Biol Evol, 2022. **39**(4).
10. Simão, F.A., et al., BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 2015. **31**(19): p. 3210-2.
11. Lovell, P.V., et al., Conserved syntenic clusters of protein coding genes are missing in birds. Genome Biol, 2014. **15**(12): p. 565.
12. Rubin, C.J., et al., Whole-genome resequencing reveals loci under selection during chicken domestication. Nature., 2010. **464**(7288): p. 587-91. doi: 10.1038/nature08832. Epub 2010 Mar 10.
13. Wong, G.K., et al., A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. Nature., 2004. **432**(7018): p. 717-22.
14. Miao, Y.W., et al., Chicken domestication: an updated perspective based on mitochondrial genomes. Heredity (Edinb), 2013. **110**(3): p. 277-82.
15. Ruan, J. and H. Li, Fast and accurate long-read assembly with wtdbg2. Nat Methods, 2020. **17**(2): p. 155-158.
16. Ghurye, J., et al., Scaffolding of long read assemblies using long range contact information. BMC Genomics, 2017. **18**(1): p. 527.
17. Ghurye, J., et al., Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput Biol, 2019. **15**(8): p. e1007273.
18. English, A.C., et al., Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One, 2012. **7**(11): p. e47768.
19. Vaser, R., et al., Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res, 2017. **27**(5): p. 737-746.
20. Hu, J., et al., NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics, 2020. **36**(7): p. 2253-2255.

21. Altschul, S.F., et al., Basic local alignment search tool. J Mol Biol, 1990. **215**(3): p. 403-10.
22. Jackman, S.D., et al., ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Res, 2017. **27**(5): p. 768-777.
23. Morgulis, A., et al., WindowMasker: window-based masker for sequenced genomes. Bioinformatics, 2006. **22**(2): p. 134-41.
24. Marçais, G. and C. Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics, 2011. **27**(6): p. 764-70.
25. Vurture, G.W., et al., GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics, 2017. **33**(14): p. 2202-2204.
26. Gurevich, A., et al., QUAST: quality assessment tool for genome assemblies. Bioinformatics, 2013. **29**(8): p. 1072-5.
27. Guan, D. https://github.com/dfguan/asset.
28. Rhie, A., et al., Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol, 2020. **21**(1): p. 245.
29. Li, H. and R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 2009. **25**(14): p. 1754-60.
30. Li, H., et al., The Sequence Alignment/Map format and SAMtools. Bioinformatics, 2009. **25**(16): p. 2078-9.
31. Dobin, A., et al., STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 2013. **29**(1): p. 15-21.
32. Goloborodko, A., https://github.com/open2c/pairtools. 2019.
33. Kerpedjiev, P., et al., HiGlass: web-based visual exploration and analysis of genome interaction maps. Genome Biol, 2018. **19**(1): p. 125.
34. Kapustin, Y., et al., Splign: algorithms for computing spliced alignments with identification of paralogs. Biol Direct, 2008. **3**: p. 20.
35. Langmead, B. and S.L. Salzberg, Fast gapped-read alignment with Bowtie 2. Nat Methods, 2012. **9**(4): p. 357-9.
36. Quast, C., et al., The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res, 2013. **41**(Database issue): p. D590-6.
37. Grabherr, M.G., et al., Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol, 2011. **29**(7): p. 644-52.
38. Nawrocki, E.P. and S.R. Eddy, Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics, 2013. **29**(22): p. 2933-5.
39. Kalvari, I., et al., Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res, 2021. **49**(D1): p. D192-d200.
40. Sievers, F. and D.G. Higgins, Clustal Omega for making accurate alignments of many protein sequences. Protein Sci, 2018. **27**(1): p. 135-145.
41. Felsenstein, J., Comparative methods with sampling error and within-species variation: contrasts revisited and revised. Am Nat, 2008. **171**(6): p. 713-25.
42. Krüger, J. and M. Rehmsmeier, RNAhybrid: microRNA target prediction easy, fast and flexible. Nucleic Acids Res, 2006. **34**(Web Server issue): p. W451-4.
43. McKenna, A., et al., The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res, 2010. **20**(9): p. 1297-303.

44. Cheng, Y. and D.W. Burt, Chicken genomics. Int J Dev Biol, 2018. **62**(1-2-3): p. 265-271.
45. Guerra-Almeida, D., D.A. Tschoeke, and R. Nunes-da-Fonseca, Understanding small ORF diversity through a comprehensive transcription feature classification. DNA Res, 2021. **28**(5).
46. Salzberg, S.L., Open questions: How many genes do we have? BMC Biol, 2018. **16**(1): p. 94.
47. Botero-Castro, F., et al., Avian Genomes Revisited: Hidden Genes Uncovered and the Rates versus Traits Paradox in Birds. Mol Biol Evol, 2017. **34**(12): p. 3123-3131.
48. Ashburner, M., et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000. **25**(1): p. 25-9.
49. Yin, Z.T., et al., Revisiting avian 'missing' genes from de novo assembled transcripts. BMC Genomics, 2019. **20**(1): p. 4.
50. Hron, T., et al., Hidden genes in birds. Genome Biol, 2015. **16**(1): p. 164.
51. Bornelöv, S., et al., Correspondence on Lovell et al.: identification of chicken genes previously assumed to be evolutionarily lost. Genome Biol, 2017. **18**(1): p. 112.
52. Derks, M.F.L., et al., A survey of functional genomic variation in domesticated chickens. Genet Sel Evol, 2018. **50**(1): p. 17.
53. Parmley, J.L. and L.D. Hurst, How do synonymous mutations affect fitness? Bioessays, 2007. **29**(6): p. 515-9.
54. Ng, P.C., et al., Genetic variation in an individual human exome. PLoS Genet, 2008. **4**(8): p. e1000160.
55. Jonas, S. and E. Izaurralde, Towards a molecular understanding of microRNA-mediated gene silencing. Nat Rev Genet, 2015. **16**(7): p. 421-33.
56. Hurles, M., Gene duplication: the genomic trade in spare parts. PLoS Biol, 2004. **2**(7): p. E206.
57. Hayashi, S., Y. Murakami, and S. Matsufuji, Ornithine decarboxylase antizyme: a novel type of regulatory protein. Trends Biochem Sci, 1996. **21**(1): p. 27-30.
58. Kim, S.W., et al., Regulation of cell proliferation by the antizyme inhibitor: evidence for an antizyme-independent mechanism. J Cell Sci, 2006. **119**(Pt 12): p. 2583-91.
59. Willardson, B.M. and A.C. Howlett, Function of phosducin-like proteins in G protein signaling and chaperone-assisted protein folding. Cell Signal, 2007. **19**(12): p. 2417-27.
60. Seale, P., et al., PRDM16 controls a brown fat/skeletal muscle switch. Nature, 2008. **454**(7207): p. 961-7.
61. Karjosukarso, D.W., et al., An FEVR-associated mutation in ZNF408 alters the expression of genes involved in the development of vasculature. Hum Mol Genet, 2018. **27**(20): p. 3519-3527.
62. Collin, R.W., et al., ZNF408 is mutated in familial exudative vitreoretinopathy and is crucial for the development of zebrafish retinal vasculature. Proc Natl Acad Sci U S A, 2013. **110**(24): p. 9856-61.
63. Nurk, S., et al., The complete sequence of a human genome. Science, 2022. **376**(6588): p. 44-53.

64. Guo, Y., et al., Researching on the fine structure and admixture of the worldwide chicken population reveal connections between populations and important events in breeding history. Evol Appl, 2022. **15**(4): p. 553-564.
65. Qanbari, S., et al., Genetics of adaptation in modern chicken. PLoS Genet, 2019. **15**(4): p. e1007989.
66. Rishell, W.A., Breeding and genetics--historical perspective. Poult Sci, 1997. **76**(8): p. 1057-61.
67. Emmerson, D.A., Commercial approaches to genetic selection for growth and feed conversion in domestic poultry. Poult Sci, 1997. **76**(8): p. 1121-5.
68. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature., 2004. **432**(7018): p. 695-716.
69. Yu, Z., et al., Analysis of the role of retrotransposition in gene evolution in vertebrates. BMC Bioinformatics, 2007. **8**: p. 308.
70. Wang, X., W.E. Grus, and J. Zhang, Gene losses during human origins. PLoS Biol, 2006. **4**(3): p. e52.
71. MacArthur, D.G., et al., A systematic survey of loss-of-function variants in human protein-coding genes. Science, 2012. **335**(6070): p. 823-8.
72. Sisu, C., et al., Transcriptional activity and strain-specific history of mouse pseudogenes. Nat Commun, 2020. **11**(1): p. 3695.
73. Aguezzoul, M., A. Andrieux, and E. Denarier, Overlap of promoter and coding sequences in the mouse STOP gene (Mtap6). Genomics, 2003. **81**(6): p. 623-7.
74. Renaut, S. and L.H. Rieseberg, The Accumulation of Deleterious Mutations as a Consequence of Domestication and Improvement in Sunflowers and Other Compositae Crops. Mol Biol Evol, 2015. **32**(9): p. 2273-83.
75. Lu, J., et al., The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. Trends Genet, 2006. **22**(3): p. 126-31.
76. Cruz, F., C. Vilà, and M.T. Webster, The legacy of domestication: accumulation of deleterious mutations in the dog genome. Mol Biol Evol, 2008. **25**(11): p. 2331-6.
77. Xie, X., et al., Accumulation of deleterious mutations in the domestic yak genome. Anim Genet, 2018. **49**(5): p. 384-392.
78. Daković, N., et al., The loss of adipokine genes in the chicken genome and implications for insulin metabolism. Mol Biol Evol, 2014. **31**(10): p. 2637-46.
79. Haimson, B., et al., Natural loss of function of ephrin-B3 shapes spinal flight circuitry in birds. Sci Adv, 2021. **7**(24).
80. Mello, C.V. and P.V. Lovell, Avian genomics lends insights into endocrine function in birds. Gen Comp Endocrinol, 2018. **256**: p. 123-129.
81. Krchlíková, V., et al., Repeated MDA5 Gene Loss in Birds: An Evolutionary Perspective. Viruses, 2021. **13**(11).
82. Grobet, L., et al., Molecular definition of an allelic series of mutations disrupting the myostatin function and causing double-muscling in cattle. Mamm Genome, 1998. **9**(3): p. 210-3.
83. Olson, M.V., When less is more: gene loss as an engine of evolutionary change. Am J Hum Genet, 1999. **64**(1): p. 18-23.
84. Andersson, L., Molecular consequences of animal breeding. Curr Opin Genet Dev, 2013. **23**(3): p. 295-301.

85. Fumihito, A., et al., One subspecies of the red junglefowl (Gallus gallus gallus) suffices as the matriarchic ancestor of all domestic breeds. Proc Natl Acad Sci U S A, 1994. **91**(26): p. 12505-9.
86. Cheng, H.W., Breeding of tomorrow's chickens to improve well-being. Poult Sci, 2010. **89**(4): p. 805-13.
87. Burt, D.W., Emergence of the chicken as a model organism: implications for agriculture and biology. Poult Sci., 2007. **86**(7): p. 1460-71.
88. Elferink, M.G., et al., Signatures of selection in the genomes of commercial and non-commercial chicken breeds. PLoS One, 2012. **7**(2): p. e32720.
89. Qanbari, S., et al., A high resolution genome-wide scan for significant selective sweeps: an application to pooled sequence data in laying chickens. PLoS One, 2012. **7**(11): p. e49525.
90. Fan, W.L., et al., Genome-wide patterns of genetic variation in two domestic chickens. Genome Biol Evol, 2013. **5**(7): p. 1376-92.
91. Boschiero, C., et al., Genome-wide characterization of genetic variants and putative regions under selection in meat and egg-type chicken lines. BMC Genomics, 2018. **19**(1): p. 83.
92. Guo, X., et al., Whole-genome resequencing of Xishuangbanna fighting chicken to identify signatures of selection. Genet Sel Evol, 2016. **48**(1): p. 62.
93. Wang, M.S., et al., Comparative population genomics reveals genetic basis underlying body size of domestic chickens. J Mol Cell Biol, 2016. **8**(6): p. 542-552.
94. Wang, M.S., et al., An Evolutionary Genomic Perspective on the Breeding of Dwarf Chickens. Mol Biol Evol, 2017. **34**(12): p. 3081-3088.
95. Zhang, M., et al., Genome-wide scan and analysis of positive selective signatures in Dwarf Brown-egg Layers and Silky Fowl chickens. Poult Sci, 2017. **96**(12): p. 4158-4171.
96. Bortoluzzi, C., et al., The effects of recent changes in breeding preferences on maintaining traditional Dutch chicken genomic diversity. Heredity (Edinb), 2018. **121**(6): p. 564-578.
97. Lawal, R.A., et al., Whole-Genome Resequencing of Red Junglefowl and Indigenous Village Chicken Reveal New Insights on the Genome Dynamics of the Species. Front Genet, 2018. **9**: p. 264.
98. Wang, Q., et al., Whole-genome resequencing of Dulong Chicken reveal signatures of selection. Br Poult Sci, 2020. **61**(6): p. 624-631.
99. Wang, K., et al., The Chicken Pan-Genome Reveals Gene Content Variation and a Promoter Region Deletion in IGF2BP1 Affecting Body Size. Mol Biol Evol, 2021. **38**(11): p. 5066-5081.
100. Kerje, S., et al., The twofold difference in adult size between the red junglefowl and White Leghorn chickens is largely explained by a limited number of QTLs. Anim Genet, 2003. **34**(4): p. 264-74.
101. McElroy, J.P., et al., Identification of trait loci affecting white meat percentage and other growth and carcass traits in commercial broiler chickens. Poult Sci, 2006. **85**(4): p. 593-605.
102. Wang, Y., et al., Genetic Dissection of Growth Traits in a Unique Chicken Advanced Intercross Line. Front Genet, 2020. **11**: p. 894.

103. Wang, Y., et al., Multiple ancestral haplotypes harboring regulatory mutations cumulatively contribute to a QTL affecting chicken growth traits. Commun Biol, 2020. **3**(1): p. 472.
104. Zhang, H., et al., Haplotype-based genome-wide association studies for carcass and growth traits in chicken. Poult Sci, 2020. **99**(5): p. 2349-2361.
105. Jennen, D.G., et al., Confirmation of quantitative trait loci affecting fatness in chickens. Genet Sel Evol, 2005. **37**(2): p. 215-28.
106. Bihan-Duval, E.L., et al., Identification of genomic regions and candidate genes for chicken meat ultimate pH by combined detection of selection signatures and QTL. BMC Genomics, 2018. **19**(1): p. 294.
107. Moreira, G.C.M., et al., Integration of genome wide association studies and whole genome sequencing provides novel insights into fat deposition in chicken. Sci Rep, 2018. **8**(1): p. 16222.
108. Zhao, X., et al., Identification of candidate genomic regions for chicken egg number traits based on genome-wide association study. BMC Genomics, 2021. **22**(1): p. 610.
109. Yang, S., et al., Whole-genome resequencing reveals genetic indels of feathered-leg traits in domestic chickens. J Genet, 2019. **98**(2).
110. Bortoluzzi, C., et al., Parallel Genetic Origin of Foot Feathering in Birds. Mol Biol Evol, 2020. **37**(9): p. 2465-2476.
111. Li, J., et al., Mutations Upstream of the TBX5 and PITX1 Transcription Factor Genes Are Associated with Feathered Legs in the Domestic Chicken. Mol Biol Evol, 2020. **37**(9): p. 2477-2486.
112. Qiu, M., et al., Whole-genome resequencing reveals aberrant autosomal SNPs affect chicken feathering rate. Anim Biotechnol, 2022. **33**(5): p. 884-896.
113. Zhang, X., et al., Analysis of a genetic factors contributing to feathering phenotype in chickens. Poult Sci, 2018. **97**(10): p. 3405-3413.
114. Zhao, J., et al., Identification of candidate genes for chicken early- and late-feathering. Poult Sci, 2016. **95**(7): p. 1498-1503.
115. Chen, B., et al., Deletion in KRT75L4 linked to frizzle feather in Xiushui Yellow Chickens. Anim Genet, 2022. **53**(1): p. 101-107.
116. Guo, X., et al., A parallel mechanism underlying frizzle in domestic chickens. J Mol Cell Biol, 2018. **10**(6): p. 589-591.
117. Huang, T., et al., A quantitative trait locus on chromosome 2 was identified that accounts for a substantial proportion of phenotypic variance of the yellow plumage color in chicken. Poult Sci, 2020. **99**(6): p. 2902-2910.
118. Huang, X., et al., Genome-wide genetic structure and selection signatures for color in 10 traditional Chinese yellow-feathered chicken breeds. BMC Genomics, 2020. **21**(1): p. 316.
119. Nie, C., et al., Genomic Regions Related to White/Black Tail Feather Color in Dwarf Chickens Identified Using a Genome-Wide Association Study. Front Genet, 2021. **12**: p. 566047.
120. Zhang, G.W., et al., A new dominant haplotype of MC1R gene in Chinese black plumage chicken. Anim Genet, 2017. **48**(5): p. 624.
121. Li, D., et al., Breeding history and candidate genes responsible for black skin of Xichuan black-bone chicken. BMC Genomics, 2020. **21**(1): p. 511.

122. Mehlhorn, J. and S. Caspers, The Effects of Domestication on the Brain and Behavior of the Chicken in the Light of Evolution. Brain Behav Evol, 2020. **95**(6): p. 287-301.
123. Luo, W., et al., Genome diversity of Chinese indigenous chicken and the selective signatures in Chinese gamecock chicken. Sci Rep, 2020. **10**(1): p. 14532.
124. Zou, A., et al., Accumulation of genetic variants associated with immunity in the selective breeding of broilers. BMC Genet, 2020. **21**(1): p. 5.
125. Li, J., et al., The crest phenotype in domestic chicken is caused by a 197 bp duplication in the intron of HOXC10. G3 (Bethesda), 2021. **11**(2).
126. Li, Y.D., et al., A combination of genome-wide association study and selection signature analysis dissects the genetic architecture underlying bone traits in chickens. Animal, 2021. **15**(8): p. 100322.
127. Kondoh, D., et al., Morphological variations of caudal skeleton between three chicken breeds. J Vet Med Sci, 2022. **84**(9): p. 1225-1229.
128. Wang, Y., et al., Associations between variants of bone morphogenetic protein 7 gene and growth traits in chickens. Br Poult Sci, 2018. **59**(3): p. 264-269.
129. Wang, Y.M., et al., Integrating Genomic and Transcriptomic Data to Reveal Genetic Mechanisms Underlying Piao Chicken Rumpless Trait. Genomics Proteomics Bioinformatics, 2021. **19**(5): p. 787-799.
130. Xu, J., et al., Mapping of Id locus for dermal shank melanin in a Chinese indigenous chicken breed. J Genet, 2017. **96**(6): p. 977-983.
131. Freese, N.H., et al., A novel gain-of-function mutation of the proneural IRX1 and IRX2 genes disrupts axis elongation in the Araucana rumpless chicken. PLoS One, 2014. **9**(11): p. e112364.
132. Noorai, R.E., et al., Genome-wide association mapping and identification of candidate genes for the rumpless and ear-tufted traits of the Araucana chicken. PLoS One, 2012. **7**(7): p. e40974.
133. Chu, Q., et al., Association of SNP rs80659072 in the ZRS with polydactyly in Beijing You chickens. PLoS One, 2017. **12**(10): p. e0185953.
134. Zhang, Z., et al., Parallel Evolution of Polydactyly Traits in Chinese and European Chickens. PLoS One, 2016. **11**(2): p. e0149010.
135. Churchill, G.A. and R.W. Doerge, Empirical threshold values for quantitative trait mapping. Genetics, 1994. **138**(3): p. 963-71.
136. Wang, K., M. Li, and H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res, 2010. **38**(16): p. e164.
137. McLaren, W., et al., The Ensembl Variant Effect Predictor. Genome Biol, 2016. **17**(1): p. 122.
138. Reynolds, J., B.S. Weir, and C.C. Cockerham, Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics, 1983. **105**(3): p. 767-79.
139. Danecek, P., et al., The variant call format and VCFtools. Bioinformatics, 2011. **27**(15): p. 2156-8.
140. Sabeti, P.C., et al., Positive natural selection in the human lineage. Science, 2006. **312**(5780): p. 1614-20.

141. Hu, Z.L., C.A. Park, and J.M. Reecy, Bringing the Animal QTLdb and CorrDB into the future: meeting new challenges and providing updated services. Nucleic Acids Res, 2022. **50**(D1): p. D956-d961.
142. Eriksson, J., et al., Identification of the yellow skin gene reveals a hybrid origin of the domestic chicken. PLoS Genet, 2008. **4**(2): p. e1000010.
143. Shimada, M., et al., Mice lacking melanin-concentrating hormone are hypophagic and lean. Nature, 1998. **396**(6712): p. 670-4.
144. Moon, S.Y. and Y. Zheng, Rho GTPase-activating proteins in cell regulation. Trends Cell Biol, 2003. **13**(1): p. 13-22.
145. Nusbaum, C., et al., DNA sequence and analysis of human chromosome 8. Nature, 2006. **439**(7074): p. 331-5.
146. Girard, C., et al., Genomic and functional characteristics of novel human pancreatic 2P domain K(+) channels. Biochem Biophys Res Commun, 2001. **282**(1): p. 249-56.
147. Littman, D.R., et al., The isolation and sequence of the gene encoding T8: a molecule defining functional classes of T lymphocytes. Cell, 1985. **40**(2): p. 237-46.
148. Gerhard, D.S., et al., The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). Genome Res, 2004. **14**(10b): p. 2121-7.
149. Agarwal, S.K., L.A. Cogburn, and J. Burnside, Dysfunctional growth hormone receptor in a strain of sex-linked dwarf chicken: evidence for a mutation in the intracellular domain. J Endocrinol, 1994. **142**(3): p. 427-34.
150. Jia, J., et al., Selection for growth rate and body size have altered the expression profiles of somatotropic axis genes in chickens. PLoS One, 2018. **13**(4): p. e0195378.
151. Cao, J., Q. Mu, and H. Huang, The Roles of Insulin-Like Growth Factor 2 mRNA-Binding Protein 2 in Cancer and Cancer Stem Cells. Stem Cells Int, 2018. **2018**: p. 4217259.
152. Baggs, J.E. and C.B. Green, Nocturnin, a deadenylase in Xenopus laevis retina: a mechanism for posttranscriptional control of circadian-related mRNA. Curr Biol, 2003. **13**(3): p. 189-98.
153. Stubblefield, J.J., J. Terrien, and C.B. Green, Nocturnin: at the crossroads of clocks and metabolism. Trends Endocrinol Metab, 2012. **23**(7): p. 326-33.
154. Hughes, K.L., E.T. Abshire, and A.C. Goldstrohm, Regulatory roles of vertebrate Nocturnin: insights and remaining mysteries. RNA Biol, 2018. **15**(10): p. 1255-1267.
155. Zhou, M., et al., The expression of a mitochondria-localized glutamic acid-rich protein (MGARP/OSAP) is under the regulation of the HPG axis. Endocrinology, 2011. **152**(6): p. 2311-20.
156. Chung, I.H., et al., ChIP-on-chip analysis of thyroid hormone-regulated genes and their physiological significance. Oncotarget, 2016. **7**(16): p. 22448-59.
157. Kudron, M.M., et al., The ModERN Resource: Genome-Wide Binding Profiles for Hundreds of Drosophila and Caenorhabditis elegans Transcription Factors. Genetics, 2018. **208**(3): p. 937-949.

158. Van Nostrand, E.L. and S.K. Kim, Integrative analysis of C. elegans modENCODE ChIP-seq data sets to infer gene regulatory interactions. Genome Res, 2013.
159. Negre, N., et al., A cis-regulatory map of the Drosophila genome. Nature, 2011. **471**(7339): p. 527-31.
160. Shen, Y., et al., A map of the cis-regulatory sequences in the mouse genome. Nature, 2012. **488**(7409): p. 116-20.
161. Moore, J.E., et al., Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature, 2020. **583**(7818): p. 699-710.
162. Gerstein, M.B., et al., Architecture of the human regulatory network derived from ENCODE data. Nature, 2012. **489**(7414): p. 91-100.
163. Sanger, F., S. Nicklen, and A.R. Coulson, DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
164. Mardis, E.R., Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet, 2008. **9**: p. 387-402.
165. van Dijk, E.L., et al., Ten years of next-generation sequencing technology. Trends Genet, 2014. **30**(9): p. 418-26.
166. Schadt, E.E., S. Turner, and A. Kasarskis, A window into third-generation sequencing. Hum Mol Genet, 2010. **19**(R2): p. R227-40.
167. Clarke, J., et al., Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol, 2009. **4**(4): p. 265-70.
168. Eisenstein, M., Oxford Nanopore announcement sets sequencing sector abuzz. Nat Biotechnol, 2012. **30**(4): p. 295-6.
169. Belton, J.M., et al., Hi-C: a comprehensive technique to capture the conformation of genomes. Methods, 2012. **58**(3): p. 268-76.
170. Koren, S., et al., Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res, 2017. **27**(5): p. 722-736.
171. Chin, C.S., et al., Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods, 2013. **10**(6): p. 563-9.
172. Stanke, M. and S. Waack, Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics, 2003. **19 Suppl 2**: p. ii215-25.
173. Majoros, W.H., M. Pertea, and S.L. Salzberg, TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics, 2004. **20**(16): p. 2878-9.
174. Blanco, E., G. Parra, and R. Guigó, Using geneid to identify genes. Curr Protoc Bioinformatics, 2007. **Chapter 4**: p. Unit 4.3.
175. Burge, C. and S. Karlin, Prediction of complete gene structures in human genomic DNA. J Mol Biol, 1997. **268**(1): p. 78-94.
176. Korf, I., Gene finding in novel genomes. BMC Bioinformatics, 2004. **5**: p. 59.
177. Lomsadze, A., et al., Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res, 2005. **33**(20): p. 6494-506.
178. Keilwagen, J., F. Hartung, and J. Grau, GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. Methods Mol Biol, 2019. **1962**: p. 161-177.

179. Brůna, T., et al., BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom Bioinform, 2021. **3**(1): p. lqaa108.
180. Vaser, R., et al., Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res, 2017. **27**(5): p. 737-746.
181. Li, H., Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 2018. **34**(18): p. 3094-3100.
182. Quinlan, A.R. and I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 2010. **26**(6): p. 841-2.
183. Chen, M.M., et al., The GFF3toolkit: QC and Merge Pipeline for Genome Annotation. Methods Mol Biol, 2019. **1858**: p. 75-87.
184. Langmead, B., et al., Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics, 2019. **35**(3): p. 421-432.
185. Langmead, B., et al., Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol, 2009. **10**(3): p. R25.
186. Li, X., Y. Huang, and F. Lei, Comparative mitochondrial genomics and phylogenetic relationships of the Crossoptilon species (Phasianidae, Galliformes). BMC Genomics, 2015. **16**: p. 42.
187. Xin, L., Z. Guangmei, and G. Binyuan, A preliminary investigation on taxonomy, distribution and evolutionary relationship of the eared pheasants, Crossoptilon. Dong wu xue bao.[Acta zoologica Sinica], 1998. **44**(2): p. 131-137.
188. Zheng, Z., A complete checklist of species and subspecies of the Chinese birds. 1994: Science Press.
189. Lu, T., The rare and endangered wild chicken in Chain. 1991, Fuzhou: Fujian Science and Technology Press.
190. Wang, P., et al., The role of niche divergence and geographic arrangement in the speciation of Eared Pheasants (Crossoptilon, Hodgson 1938). Mol Phylogenet Evol, 2017. **113**: p. 1-8.
191. Wang, P., et al., Genomic Consequences of Long-Term Population Decline in Brown Eared Pheasant. Mol Biol Evol, 2021. **38**(1): p. 263-273.
192. Semenza, G.L., Targeting HIF-1 for cancer therapy. Nat Rev Cancer, 2003. **3**(10): p. 721-32.
193. Giaccia, A., B.G. Siim, and R.S. Johnson, HIF-1 as a target for drug development. Nat Rev Drug Discov, 2003. **2**(10): p. 803-11.
194. Bárdos, J.I. and M. Ashcroft, Hypoxia-inducible factor-1 and oncogenic signalling. Bioessays, 2004. **26**(3): p. 262-9.
195. Matsumoto, T. and L. Claesson-Welsh, VEGF receptor signal transduction. Sci STKE, 2001. **2001**(112): p. re21.
196. Cross, M.J., et al., VEGF-receptor signal transduction. Trends Biochem Sci, 2003. **28**(9): p. 488-94.
197. Ferrara, N., H.P. Gerber, and J. LeCouter, The biology of VEGF and its receptors. Nat Med, 2003. **9**(6): p. 669-76.
198. Pamenter, M.E., et al., Cross-Species Insights Into Genomic Adaptations to Hypoxia. Front Genet, 2020. **11**: p. 743.
199. Chatterjee, O., et al., An overview of the oxytocin-oxytocin receptor signaling network. J Cell Commun Signal, 2016. **10**(4): p. 355-360.

200. Alexander, D.H., J. Novembre, and K. Lange, Fast model-based estimation of ancestry in unrelated individuals. Genome Res, 2009. **19**(9): p. 1655-64.
201. D'Orléans-Juste, P., et al., Synthesis and degradation of endothelin-1. Can J Physiol Pharmacol, 2003. **81**(6): p. 503-10.
202. Neylon, C.B., Vascular biology of endothelin signal transduction. Clin Exp Pharmacol Physiol, 1999. **26**(2): p. 149-53.
203. Masaki, T., et al., Subcellular mechanisms of endothelin action in vascular system. Eur J Pharmacol, 1999. **375**(1-3): p. 133-8.
204. Cowburn, A.S., et al., Cardiovascular adaptation to hypoxia and the role of peripheral resistance. Elife, 2017. **6**.
205. Arima, H., et al., Regulation of vasopressin synthesis and release by area postrema in rats. Endocrinology, 1998. **139**(4): p. 1481-6.
206. Graham, A.M. and K.G. McCracken, Convergent evolution on the hypoxia-inducible factor (HIF) pathway genes EGLN1 and EPAS1 in high-altitude ducks. Heredity (Edinb), 2019. **122**(6): p. 819-832.
207. Martindale, J.L. and N.J. Holbrook, Cellular response to oxidative stress: signaling for suicide and survival. J Cell Physiol, 2002. **192**(1): p. 1-15.
208. Davis, R.J., Signal transduction by the c-Jun N-terminal kinase. Biochem Soc Symp, 1999. **64**: p. 1-12.
209. Kyriakis, J.M. and J. Avruch, Mammalian mitogen-activated protein kinase signal transduction pathways activated by stress and inflammation. Physiol Rev, 2001. **81**(2): p. 807-69.
210. Tibbles, L.A. and J.R. Woodgett, The stress-activated protein kinase pathways. Cell Mol Life Sci, 1999. **55**(10): p. 1230-54.
211. Leppä, S. and D. Bohmann, Diverse functions of JNK signaling and c-Jun in stress response and apoptosis. Oncogene, 1999. **18**(45): p. 6158-62.
212. Claeys, I., et al., Insulin-related peptides and their conserved signal transduction pathway. Peptides, 2002. **23**(4): p. 807-16.
213. Van Obberghen, E., et al., Surfing the insulin signaling web. Eur J Clin Invest, 2001. **31**(11): p. 966-77.
214. De Meyts, P. and J. Whittaker, Structural biology of insulin and IGF1 receptors: implications for drug design. Nat Rev Drug Discov, 2002. **1**(10): p. 769-83.
215. Kim, J.J. and D. Accili, Signalling through IGF-I and insulin receptors: where is the specificity? Growth Horm IGF Res, 2002. **12**(2): p. 84-90.
216. White, M.F., IRS proteins and the common path to diabetes. Am J Physiol Endocrinol Metab, 2002. **283**(3): p. E413-22.
217. Nakae, J., Y. Kido, and D. Accili, Distinct and overlapping functions of insulin and IGF-I receptors. Endocr Rev, 2001. **22**(6): p. 818-35.
218. Myung, P.S., N.J. Boerthe, and G.A. Koretzky, Adapter proteins in lymphocyte antigen-receptor signaling. Curr Opin Immunol, 2000. **12**(3): p. 256-66.
219. Krogsgaard, M., et al., Linking molecular and cellular events in T-cell activation and synapse formation. Semin Immunol, 2003. **15**(6): p. 307-15.
220. Kane, L.P., J. Lin, and A. Weiss, Signal transduction by the TCR for antigen. Curr Opin Immunol, 2000. **12**(3): p. 242-9.
221. Lukacs, N.W., Role of chemokines in the pathogenesis of asthma. Nat Rev Immunol, 2001. **1**(2): p. 108-16.

222. Vicente-Manzanares, M., et al., The leukocyte cytoskeleton in cell migration and immune interactions. Int Rev Cytol, 2002. **216**: p. 233-89.
223. Ji, L.D., et al., Genetic adaptation of the hypoxia-inducible factor pathway to oxygen pressure among eurasian human populations. Mol Biol Evol, 2012. **29**(11): p. 3359-70.
224. Li, M., et al., Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. Nat Genet, 2013. **45**(12): p. 1431-8.
225. Qu, Y., et al., Ground tit genome reveals avian adaptation to living at high altitudes in the Tibetan plateau. Nat Commun, 2013. **4**: p. 2071.
226. Wang, M.S., et al., Genomic Analyses Reveal Potential Independent Adaptation to High Altitude in Tibetan Chickens. Mol Biol Evol, 2015. **32**(7): p. 1880-9.
227. Zhu, X., et al., Divergent and parallel routes of biochemical adaptation in high-altitude passerine birds from the Qinghai-Tibet Plateau. Proc Natl Acad Sci U S A, 2018. **115**(8): p. 1865-1870.
228. Liu, J., et al., Genetic signatures of high-altitude adaptation and geographic distribution in Tibetan sheep. Sci Rep, 2020. **10**(1): p. 18332.
229. Xiong, Y., et al., Physiological and genetic convergence supports hypoxia resistance in high-altitude songbirds. PLoS Genet, 2020. **16**(12): p. e1009270.
230. Zhang, Y., et al., Identification of key HIF-1α target genes that regulate adaptation to hypoxic conditions in Tibetan chicken embryos. Gene, 2020. **729**: p. 144321.
231. Xu, D., et al., A single mutation underlying phenotypic convergence for hypoxia adaptation on the Qinghai-Tibetan Plateau. Cell Res, 2021. **31**(9): p. 1032-1035.
232. Qiu, Q., et al., The yak genome and adaptation to life at high altitude. Nat Genet, 2012. **44**(8): p. 946-9.

APPENDIX A: Link of supplementary materials

Supplementary materials are available at https://github.com/Siwen-W/Dissertation.