De novo assembly of 20 chickens reveals the undetectable phenomenon for thousands of core genes on sub-telomeric regions

Authors:

Ming Li [1†], Congjiao Sun [2†], Naiyi Xu [1†], Peipei Bian [1†], Xiaomeng Tian [1†], Xihong Wang [1†], Yuzhe Wang [3,4†], Xinzheng Jia [5,6], Rasmus Heller [7], Mingshan Wang [8,9], Fei Wang [1], Xuelei Dai [1], Rongsong Luo [1], Yingwei Guo [1], Xiangnan Wang [1], Peng Yang [1], Shunjin Zhang [1], Xiaochang Li [2], Chaoliang Wen [2], Fangren Lan [2], AMAM Zonaed Siddiki [10], Chatmongkon Suwannapoom [11], Xin Zhao [12], Qinghua Nie [13], Xiaoxiang Hu [3*], Yu Jiang [1*], Ning Yang [2*]

Affiliations:

[1] Key Laboratory of Animal Genetics, Breeding and Reproduction of Shaanxi Province, College of Animal Science and Technology, Northwest A&F University, Yangling 712100, China.

[2] National Engineering Laboratory for Animal Breeding and Key Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture and Rural Affairs, China Agricultural University, Beijing 100193, China

[3] State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China

[4] National Research Facility for Phenotypic and Genotypic Analysis of Model Animals (Beijing), China Agricultural University, Beijing 100193, China

[5] Department of Animal Science, Iowa State University, Ames, IA 50011, USA

[6] School of Life Science and Engineering, Foshan University, Foshan 528225, China

[7] Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, Copenhagen N 2200, Denmark.

[8] Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA

[9] Department of Ecology and Evolutionary Biology, University of California Santa Cruz, CA 95064, USA

[10] Department of Pathology and Parasitology, Faculty of Veterinary Medicine, Chittagong Veterinary and Animal Sciences University, Chittagong-4202, Bangladesh

[11] School of Agriculture and Natural Resources, University of Phayao, Phayao, Thailand

[12] Department of Animal Science, McGill University, Montreal, Quebec, Canada

[13] Department of Animal Genetics, Breeding and Reproduction, College of Animal Science, South China Agricultural University, Guangzhou, 510642, Guangdong, China.

†These authors contributed equally

*Corresponding authors. E-mail: nyang@cau.edu.cn (N.Y.), yu.jiang@nwafu.edu.cn (Y.J.), huxx@cau.edu.cn (X.H.).

Abstract:

The gene numbers and evolutionary rates of birds were assumed to be much lower than that of mammals, which in sharp contrast to the huge species number and morphological diversity of birds. It is very necessary to construct a complete avian genome and analyze its evolution.We constructed a chicken pan-genome from 20 *de novo* genome assemblies with high sequencing depth, newly identified 1,335 protein-coding genes and 3,011 long noncoding RNAs. The majority of these novel genes were detected across most individuals of the examined transcriptomes but were accidentally measured in each of the DNA sequencing data regardless of Illumina or PacBio technology. Furthermore, different from previous pan-genome models, most of these novel genes were overrepresented on chromosomal sub-telomeric regions, surrounded with extremely high proportions of tandem repeats, and strongly blocked DNA sequencing. These hidden genes were proved to be shared by all chicken genomes, included many housekeeping genes, and enriched in immune pathways. Comparative genomics revealed the novel genes had three-fold elevated substitution rates than known ones, updating the evolutionary rates of birds. Our study provides a framework for constructing a better chicken genome, which will contribute towards the understanding of avian evolution and improvement of poultry breeding.

Keywords: Chicken, Pan-genome, Missing genes, Nonconical DNA secondary structure, Avian evolution.

Introduction:

The ~10,770 species of birds (Gill et al. 2020) show complex and diverse morphology and behavior, however the currently available avian genomes present a reduced rate of evolution and much lower gene numbers than those of all other tetrapods. The apparent discordance remained a major evolutionary conundrum. Some studies have shown that birds tend to have fewer genes than other tetrapods due to the large segmental deletions in birds (Lovell et al. 2014; Zhang et al. 2014), while other researchers suggested that these missing genes may not have been sequenced (Bornelov et al. 2017; Botero-Castro et al. 2017; Yin et al. 2019; Zhu et al. 2021). Using the advanced sequencing technique and methodology, the Vertebrate Genomes Project (VGP) found lots of genes were missing in previous genomes, even the previous and VGP assemblies were from the same individual animals (Kim et al. 2021; Rhie et al. 2021). It remains unclear how many genes within individual bird species and why some genes are missing in the currently available genomes.

Comprehensive analyses indicate multiple high-quality de novo genome assemblies possess more power to capture the complete set of genes, which leads to the appearance and prevalence of "pan-genome" in various species (Wong et al. 2018; Duan et al. 2019; Tian et al. 2020; Wong et al. 2020). The pan-genome of mammals are typically of the "closed" pattern with a limited number of variable genes (Duan et al. 2019; Li et al. 2019; Tian et al. 2020), which means the number of genes in mammalian species is relatively conserved. While bacteria, fungi, and plants exhibit the characteristic of "open" pattern, with the proportion of core genes size is less than 80% in many species (Golicz et al. 2019). Recent research using population resequencing data found that the core genome of chickens is only 76% (Wang et al. 2021) , which puzzles us because it seems to be inconsistent with the status of chickens in evolution. As the most abundant class of tetrapod vertebrates, the *de novo* based pan-genome of avian has not yet been established, which is essential to solve many biological issues.

Chicken (*Gallus gallus*) as one of the most important farm animals plays a major role in human food production and has been widely used as a model organism in studies of developmental biology, virology, oncogenesis, and immunology (Cooper et al. 1966; Stehelin et al. 1976; Brown et al. 2003; Vogt 2011). In this study, we utilized 20 new high-quality assemblies of diverse chicken breeds to generate the first *de novo* based chicken pan-genome. A total of 159 Mb novel sequences containing 1,335 coding genes that completely absent from the chicken genome were identified, verified, and localized. Importantly, most of the novel genes actually exist in all of the chicken genomes but were prone to be missing in the DNA sequencing leaded by high proportions of tandem repeats and secondary structures. Hence, unwinding complex DNA structures should be one of the most important advances to improve the sequencing quality for the assembly of the complete avian genomes. Our study revealed that the numbers of chicken genes are comparable to those of other tetrapod vertebrates and a new pan-genome pattern of birds.

Results:

Identification and validation of non-redundant novel sequences

Twenty chickens from four continents representing widespread indigenous chicken breeds, commercial broilers and layer lines were sampled for *de novo* genome assembly (**Fig. 1a**, **Supplementary Table 1**). Ten assemblies were constructed by integrating both PacBio (53 to 95×) and Illumina data (45 to 70×), resulting in a contig N50 size ranging from 5.89 to 16.72 Mb (**Supplementary Table 2**). Six of them were further clustered at the chromosome level by using Hi-C (112 to 125×) data (see **Methods**, **Supplementary Fig. 1–4**, **Supplementary Tables 2** and **3**). The remaining ten samples were assembled based on Illumina reads from a combination of libraries with multiple insert sizes, ranging from 500 bp to 5 Kb (with a depth of ~134 × per genome, **Supplementary Table 2**). These ten samples showed a contig N50 size ranging from 80.30 Kb to 137.59 Kb (**Supplementary Table 2**), representing high-quality Illumina genomes (Schatz et al. 2010). The completeness of the 20 assemblies was evaluated through the Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis. Most (from 92.4% to 95.3%) of the 4,915 core genes in the Aves dataset were identified in the 20 assemblies, which is comparable to the percentage in the reference chicken genome (GRCg6a: 95.4%) and thus supports a high-quality genome assembly (**Fig. 1b**, **Supplementary Fig. 5**, and **Supplementary Table 4**).

To identify novel sequences, all 20 *de novo* assemblies were aligned against GRCg6a (see **Methods**, **Supplementary Fig. 6**). For stability, we used GRCg6a from a same red junglefowl as the reference genome in the past two decades, not the newly unpublished GRCg7b from a broiler. The genome length of GRCg6a (1.06 Gb) and GRCg7b (1.05 Gb) are almost the same. Unaligned sequences or sequences with < 90% identity and > 500 bp in length to GRCg6a were retained and potentially contaminating non-Chordata sequences were removed. After these screening, each assembly left 6.10 to 15.40 Mb of novel sequences (**Fig.**

1c). We merged the novel sequences from all 20 assemblies and built a pan-genome of chicken. The pan-genome contains 158.98 Mb of non-redundant novel sequence， which obtained from 45,715 sequences with an average length of 3,478 bp. (**Table 1**, **Supplementary Fig. 7**, and **Supplementary Table 5**). The chicken pan-genome expanded the size of GRCg6a by 14.92% which is the highest percentage among the published vertebrate pan-genomes.

We next validate the reproduced the 158.98 novel sequences. 71.58% novel sequence can be detected in the genomes other individual, including our assemblies and 922 resequenced chicken genomes from previous studies. 44.40% can find orthologs from fourteen other publicly available *Galliformes* genomes. 54.36% can be detected from transcriptomes 263 transcriptomes from multiple tissues from 54 chickens, including 46 transcriptomes obtained from 11 tissues/organs from six individuals in our study and 217 publicly available chicken RNA sequencing (RNA-Seq) datasets from 48 individuals (Cardoso-Moreira et al. 2019) (see **Supplementary Note**, **Fig. 1d**, **Supplementary Fig. 8–10**, **Supplementary Table 5–8**, and **Supplementary Dataset**). In total, 90.97% of the novel sequences were verified in at least one of the above data sources.

Distribution of cryptic novel sequences across chicken individuals

We found that the distribution of the novel sequences is obviously inconsistent in different verified sources. The detection rate of novel sequences in one genome is extremely low, the median is only 0.43% among the 922 resequencing data and 5% among the 20 assemblies (**Fig. 2a**). Among all 159Mb novel sequences, the 10 Illumina assemblies independently detected about 60Mb, containing only 3.44 Mb intersection with PacBio assemblies. Due to the higher detection rate of RNA-Seq, we picked up the transcribed novel sequences according to the 263 transcriptomes for further validation. RNA-Seq confirmed that 60.51% of the transcribed regions of the cryptic novel sequences were shared among more than half of the chicken

genomes (**Supplementary Fig. 11**). In the six individuals with both PacBio genome assembly and transcriptome data, the transcriptomes supported a total of 16,169 novel sequences, 9,200 (56.90%) of which were detected in the transcriptomes of all six individuals. However, 5,711 (62.08%) of the 9,200 sequences were absent in all six PacBio genome assemblies (**Fig. 2b**). By mapping the PacBio reads to the novel sequences, 76.35% and 52.81% of the novel sequences were covered by at least one read across more than half or all PacBio-sequenced individuals, respectively. Moreover, although the GRCg6a assembly did not contain our novel sequences, 6.30% (2,879) of the sequences were covered by the Illumina sequencing reads of the GRCg6a individual with at least 7× coverage (corresponding to 25% of the genome-wide depth) (**Supplementary Fig. 12** and **Supplementary Dataset**). To explain the prevalence of ubiquitously transcribed yet missing novel sequences in the assemblies, we compared the median sequencing depth of the novel sequences with the whole-genome depth in the individuals. We found that the median sequencing depth of the novel sequences was only one-third of the whole-genome depth in the individuals in which the novel sequences were successfully assembled. Furthermore, in the individuals in which a given novel sequence was missing from the assembly, the median sequencing depth of the novel sequences was only one-twentieth of the whole-genome depth, which is insufficient for successful assembly (**Fig. 2c**). Collectively, the results indicated that the novel sequences were most likely present in most or all the chicken genomes but were prone to be missing in the assemblies due to their extremely low DNA sequencing depth.

Cryptic novel sequences have a high content of tandem repeats

We observed a higher GC content in the novel sequences than in the reference genome (57.2% vs. 42.30%). However, the GC content around 60% could not significantly reduce the depth of sequencing. Another influence of sequencing is repeat. The content of tandem repeats (TRs) in

the novel sequences was 79.13%, which is extremely high and significantly higher than in GRCg6a (2.2%; chi-squared test, *P*-value = 0) (**Fig. 2d** and **Table 1**). Other interspersed repeats such as LTR and LINE were low (0.09% in novel sequence vs 9.6% in GRCg6a, **Supplementary Fig. 13**). We predicted the relative importance of TR and GC content in detection rate in assembly using random forest classifier and found the TR content had a greater influence than GC (**Fig. 2d**, **Supplementary Fig. 14**). The TR can form noncanonical DNA structures, such as G-quadruplexes (four-stranded noncanonical DNA/RNA topologies, hereafter referred to as G4 motifs), Z-DNA, A-phased repeats and inverted repeats, which can form cruciforms, triplexes and slipped structures, leading to genomic instability (Zhao et al. 2010) and incapable DNA sequencing (Guiblet et al. 2018). We found these noncanonical structure are highly intersected with TR regions (**Supplementary Fig. 15**). Among these structures, the content of direct repeats (DR) (37.96%) and G4 motifs (37.08%), are the highest in novel sequences, while their occurrence in GRCg6a is only 1.47% and 0.77%. DR and G4 also showed the largest negative correlation with read depth, the novel sequence with DR and G4 motif had only 1/3 and 1/2 read depth of all novel sequences. (**Fig. 2e**). It is worth noting that as particularly stable noncanonical DNA structures, G4 motifs typically form in guanine-rich regions of genomes, which may be one of the reasons why GC-rich sequences are difficult to sequence. We also found that the transcribed regions of novel sequences showed a lower TR content (**Supplementary Fig. 16**), which might be the reason why RNA-Seq resulted in a higher observed frequency than DNA sequencing.

Abundant genes are embedded in novel sequences

Within the novel sequences, the expressible regions are of the most interest for potentially discovering novel candidate genes. To identify novel chicken genes, we performed gene annotation for all 20 assemblies by *de novo* and reference-guided methods using the multi-

tissue transcriptomes (see **Methods**, **Supplementary Table 8**). The median expression level of these putative novel genes was significantly higher than the median expression of GRCg6a-annotated genes (*P*-value = 2.84×10^{-7}) (**Fig. 3a** and **Supplementary Fig. 17**). The chromosome conformation analysis showed that the regions containing novel genes were significantly enriched in the A-compartment (*P*-value = 2.2×10^{-16}) (**Fig. 3b** and **Supplementary Fig. 18**), which is associated with open, expression-active chromatin. Furthermore, the orthologues of the novel genes showed expression levels that were higher than the median levels observed in other species, such as human and mouse (**Supplementary Fig. 19**), suggesting plausible functions and active expression of these genes.

We identified 1,335 novel coding genes with FPKM > 1, and completely missing from GRCg6a (see **Methods**, **Supplementary Fig. 20**, and **Supplementary Table 9**). The novel coding genes were distributed across 1,100 novel sequences, with an average length of 1,047 bp. By searching against the non-redundant (nr) protein database of NCBI (E-value $\leq 1\times10^{-5}$), 969 of the novel coding genes were found to show Chordata protein orthologues, 738 of which belonged to Aves (**Supplementary Table 9**). In addition to novel coding genes, we also identified 3,874 confident transcripts which complemented 1,336 partially missing coding genes in GRCg6a (**Supplementary Fig. 21** and **Supplementary Table 10)**.

To validate the novel coding genes, proteomic analysis of multiple tissues (hypothalamus, spleen and cecal tonsil) was performed via an LC-MS/MS strategy (**Supplementary Table 8**). A total of 255 (19.10%) novel coding genes were confirmed by the existence of corresponding proteins (**Supplementary Table 9**), compared to 6,201 (35.48%) of all the coding genes present in the reference genome. The lower detection rate of novel genes in proteomics may be affected by the differences in protein length and the quality of the protein database used for searching. Notably, after removing novel coding genes less than 1 Kb in

length, the proteomic verification ratio of the remaining novel coding genes increased to 29.11%.

We found that most of the novel coding genes were present and expressed in most chicken breeds. According to the DNA data, 92.47% of the novel sequences containing novel coding genes were supported by at least one PacBio read in each sample (**Supplementary Fig. 22**). According to the comparison of multi-tissue transcriptomes of six individuals, 55.13% and 80.97% of the novel coding genes were detected in all six or at least three individuals, respectively (**Supplementary Table 9**). Based on our sequencing platform, assembly strategy and annotation pipeline, the modelling of the saturation curve by iteratively randomly sampling individuals suggested that the number of novel genes detected by genome assembly did not significantly increase beyond a sample size of ten (**Fig. 3c**). A previous study (Yin et al. 2019) based on the *de novo* assembly of massive chicken transcriptomes increased the number of known chicken coding genes from 17,477 to 17,967 (**Fig. 3d**, **Supplementary Fig. 23**). According to our chicken pan-genome, we found that the total number of chicken coding genes reached at least 19,223 (**Fig. 3d**, **Supplementary Fig. 23**, and **Supplementary Table 11**).

In addition to coding genes, we identified 3,011 long noncoding RNAs (lncRNAs) (see **Methods**, **Supplementary Table 12**). Among these novel lncRNAs, 87.85% were supported by at least one PacBio read in each sample (**Supplementary Fig. 22**). In our multi-tissue transcriptomes of six individuals, 47.72% and 75.09% of novel lncRNA genes were detected in all six or at least three individuals, respectively (**Supplementary Table 12**). The increasing saturation curve of the observed novel lncRNA genes was similar to that of novel coding genes (**Fig. 3c**). Using the same pipeline as in a previous study (Sarropoulos et al. 2019), we showed that the total number of chicken lncRNAs was at least 19,795 (**Fig. 3e**). Therefore, our study revealed that the numbers of both the protein-coding and lncRNA genes of chicken are comparable to those of other tetrapod vertebrates (**Fig. 3d** and **3e**).

Novel sequences and genes are concentrated in sub-telomeric regions with elevated substitution rates

We anchored the novel sequences to GRCg6a based on flanking sequence alignment and chromosome interaction mapping (see **Methods**). A total of 27,966 (61.17%) novel sequences containing 1,043 novel coding genes and 1,567 novel lncRNAs were anchored to GRCg6a by at least one end (**Supplementary Table 5, 9 and 12**). Among these sequences, 6,735 novel sequences containing 388 coding genes were fully anchored by both ends. The fully anchored novel sequences were further classified as insertions, alternate alleles, or multiple alternative alleles (**Supplementary Fig. 24b**, **c**, and **d**) and were dispersed on every chromosome of GRCg6a, filling 72 of 946 gaps in GRCg6a (**Supplementary Fig. 24a**, **e**).

The fully anchored novel sequences and genes were overrepresented on micro-chromosomes (GGA11-38) (< 10 Mb) or the terminal 5 Mb ends of macro-chromosomes (**Fig. 4a**, **Supplementary Fig. 25**), which are termed as sub-telomeric regions. By comparison with the random distribution, we estimated 2.5- (*P*-value < 1×10^{-6}, permutation) and 5-fold (*P*-value < 1×10^{-6}, permutation) increases in fully anchored novel sequences and gene density within sub-telomeric regions, respectively (**Supplementary Fig. 26**). The novel sequences nearly doubled the length of the micro-chromosomes such as chromosomes 16, 25, 30, 31, 32, and 33, adding a total of 421 coding genes (**Supplementary Fig. 24f**, **Supplementary Note**).

It is widely accepted that sub-telomeric regions of the chromosomes and micro-chromosomes of birds exhibit high rates of recombination and mutation (International Chicken Genome Sequencing 2004; Burt 2005; Linardopoulou et al. 2005; Bell et al. 2020). We investigated the evolutionary rates of 160 high-quality orthologues of the novel coding genes by comparing chicken genes with those of human and mouse. The synonymous substitution rate (dS) and nonsynonymous substitution rate (dN) of these novel genes were 3.3- and 2.5-

fold higher than that anchored GRCg6a genes, respectively. And the dN/dS ratio of these novel genes was lower than that of the reference genes. Interestingly, the unlocalized genes of GRCg6a, which may also be located in sub-telomeric regions, showed the similar mutation pattern as the novel genes (**Fig. 4b**). This suggested that the novel coding genes in sub-telomeric regions showed a higher mutation rate.

We next identified novel gene clusters to investigate collinearity. Screening according to the existence of more than 3 novel coding genes within 1 Mb bin across the genome revealed 19 regions containing 201 of 388 fully anchored genes. The 19 gene clusters were all located in sub-telomeric regions (**Fig. 4a**, **Supplementary Fig. 25**). Gene homology analysis indicated that almost all 201 novel coding genes had homologs in mammalian and reptile genomes and showed good collinearity (**Fig. 4d**, **Supplementary Fig. 25**, **Supplementary Table 13**). Some novel gene clusters likely existed in the sub-telomeric regions before the divergence of testudines and avian. However, the significant increase of the TR clusters with high content of noncanonical DNA structures only happened on the bird lineage (**Fig. 4c**). Unlike the previous notion that large segmental deletions occurred in the evolution process (Lovell et al. 2014; Zhang et al. 2014), our results provided a large number of confident new gene clusters in sub-telomeric regions, filling gaps in which genes were often missing due to insufficient sequencing.

Functional assignment of novel regions and genes

Among the novel coding genes that we identified, 176 of them were identified as housekeeping genes in human and mouse (Hounkpe et al. 2021) (**Supplementary Table 9**). Through the annotation and enrichment analyses, we also found that a large number of them were involved in essential biological reactions and pathways, such as metabolism, signal transduction, basic biological functions, the immune system and disease (**Fig. 5a**, **Supplementary Tables 14**, **15 and 16**).

In the novel regions, we dissected chromosome 16 and the sub-telomeric part of chromosome 1 as two examples to reveal their plausible gene arrangement and functions. Chromosome 16 is a micro-chromosome that contains many immune system-related genes (**Fig. 4d**) and spans only 2.84 Mb of GRCg6a. We assembled 3.76 Mb of novel sequences and identified 61 novel coding genes and 80 lncRNA genes on chromosome 16. The novel gene clusters showed good syntenic relationships with other tetrapods (**Fig. 4d**). One of the novel gene clusters showed that birds had experienced regional complications in the cluster and lacked a large number of coding genes (**Fig. 4d**). One novel coding gene, the complement factor B (*CFB*) gene, which is an important immune gene involved in the alternative complement pathway of the immune system (**Supplementary Fig. 27**) and is regulated by the nuclear factor kappa B (NF-κB) pathway, was *de novo* identified on chromosome 16. This gene is highly and uniquely expressed in the liver of chickens and confirmed based on our MS/MS data (**Supplementary Fig. 27**). In addition, we identified two novel ribosomal genes, mitochondrial ribosomal protein S18B (*MRPS18B*) and ribosomal protein S18 (*RPS18*) on chromosome 16 (**Fig. 4d**, **Supplementary Table 13**).

The other novel gene cluster, including the *leptin* gene, located on chromosome 1 (**Supplementary Fig. 25**). Based on RNA-Seq, previous research has shown that the *leptin* gene does exist in the chicken genome, yet it was absent from the chicken reference genome (Seroussi et al. 2017). Interestingly, we found that two divergent haplotypes of the *leptin* gene were assembled from two individuals. The entire gene region and its flanking regions had extremely high TR and G4 motif contents (**Supplementary Fig. 28**). Based on chromosome interaction data, *leptin* was assigned to the distal tip of chromosome 1p, showing collinearity with *SND1* and *LRRC4* (**Supplementary Fig. 25**, **Supplementary Fig. 28**). We found that *leptin* exon 2 was conserved, while exon 1 was variable in chicken. The length of its intron also varied among different chicken individuals (**Supplementary Fig. 28**). Neither of the two exons

showed good homology with other species. In this region, we found another novel gene, ovocleidin-17 (*OC-17*), which plays a key role in avian eggshell biomineralization and is not contained in the reference genome (**Supplementary Table 9**).

Application of the chicken pan-genome in avian influenza

The chicken pan-genome identify more crucial novel genes related to avian diseases resistance that had not been discovered previously. Chickens are susceptible to several diseases that have far-reaching effects on human society, such as avian influenza. Here, we reanalyzed the transcriptome data (Smith et al. 2015) of chicken lung and ileum samples after infection with low pathogenic (H5N2) and highly pathogenic (H5N1) avian influenza virus. Compared with the expression levels observed in the control group, 30 novel coding genes, 65 novel lncRNAs and 79 partially missing genes showed differential expression in these samples (FDR < 0.05) (**Supplementary Tables 9**, **10** and **12**). B cell-related genes (*CD22*, *CD79A*, *PRMT1,* and *SND1*), T cell-related genes (*CD2BP2*), immunoglobulin genes (*IGLL5*), and ribosome genes (*RPS18*) were screened among these differential expression genes (**Supplementary Tables 9** and **17**). Notably, several significantly differentially expressed genes (*AXL*, *HUWE1*, *IKKγ*, *KAT8*, and *KHSRP*) belonged to or were regulated by the NF-κB signaling pathway, which is the master regulator of the immune response to infection due to its role in regulating cytokine and antimicrobial peptide expression (**Fig. 5b**). *RELB*, a subunit of NF-κB, associated with the immune responses to influenza A (Ruckle et al. 2012) and severe acute respiratory syndrome-associated coronavirus (SARS-CoV) (Chen et al. 2006), was identified, anchored and validated in our study (**Supplementary Fig. 29**). Another novel gene, *IKKγ* (**Supplementary Table 9**), a subunit of the IκB kinase complex (IKK), was essential for the activation of NF-κB transcriptional activity. Besides, the newly identified genes *AXL* (Schmid et al. 2016), *CSNK2B* (Marjuki et al. 2008), *DDX39B* (Wisskirchen et al. 2011), *KHSRP* (Liu et al. 2015) and *TP53*

(Wang et al. 2018) have also been reported to play a role in the immune response to influenza A. In total, there were 21 novel coding genes and 7 partially missing coding genes that belonged to or were regulated by the NF-κB signaling pathway (**Fig. 5b**). The NF-κB pathway is essential in defense against viral infections, such as those caused by influenza viruses.

Discussion:

The chicken is the modern descendant of the dinosaurs with its genome firstly sequenced in non-mammalian amniotes (International Chicken Genome Sequencing 2004). Despite several major updates, the completeness of the chicken genome still needs to be improved and the gene number of chicken is still argued. Here, our study suggests that the chicken pan-genome exhibits a more complex mammalian-like closed pattern. More specifically, we identified 1,335 and 3,011 novel coding genes and novel long noncoding genes, respectively, containing mostly core genes, which appear different from previous mammalian pan-genome studies that reported fewer novel genes (Sherman et al. 2018; Golicz et al. 2019; Sherman and Salzberg 2020; Tian et al. 2020). The highly complex noncanonical DNA structure across the novel genes might be the main reason to prevent the efficient genome assembly of identified novel genes in lots of individuals. The accidentally detection of DNA sequencing in the regions of novel sequences due to the secondary DNA structure might be the reason why there are still so many genes missing in the recent high-quality VGP avian assemblies, that may suggest there are still more challenges to complete the avian reference genome. Nevertheless, we increased the number of protein coding genes in chicken to 19,223 and denied the gene loss hypothesis during avian evolution. Although some genes may still hide in some more complex genomic regions and waiting to be discovered. Our study not only revealed the gene number in birds is comparable to that in other tetrapods but also presented a novel closed pattern of avian pan-genome. The

complete avian genomes will greatly improve the studies on comparative genomics and functional genomics research in birds.

It has been believed that both the evolutionary substitution rate and the rate of chromosomal rearrangement in the avian lineage are lower than that of mammals (Burt et al. 1999; Zhang et al. 2014). However, we found that a large number of novel genes that have three times the substitution rate than the known ones, which can greatly increase the average substitution rate of the chicken genome. We find that the novel sequences and genes were concentrated in the sub-telomeric regions of chromosomes, in which the recombination rate tend to be higher (Linardopoulou et al. 2005; Bell et al. 2020). This may drive the base composition evolution via biased gene conversion (Marais 2003) and cause repeat expansions or contractions (Richard and Paques 2000; Polleys et al. 2017), and maybe the critical factor driving the development of the special characteristics in sub-telomeric regions. These genes may have a pivotal role on the formation and development of some unique phenotypes of the dinosaurs-avian branch. For instance, some differentially expressed novel genes were associated with immune response, which may be an ingenious design of the bird immune system to resist viruses with high mutation rates. With the high recombination rate, the novel sequences may represent a large unexplored party of the chicken genetic map, which will contribute to the comprehensive understanding of genetic variation and pinpoint the causal variations of important traits and thus promote the development of chicken breeding.

In conclusion, our chicken pan-genome provides a comprehensive resource and a great platform for the research of avian evolution, functional genomics, and chicken breeding. These results highlight the complexity of species genomes and suggest that many functionally important regions may be cryptic in reference genomes across the tree of life.

Materials and Methods

Sample collection

A total of 20 chicken individuals were collected from all around the world for genomic sequencing. Transcriptome sequencing was also performed in 11 tissues of six individuals, including breast muscle, bursa of Fabriclus, cecal tonsil, Harderian gland, hypophysis, hypothalamus, liver, ovary, spleen, testis, and thymus tissues. Moreover, tandem mass spectrometry (MS/MS) data were generated from three tissues (hypothalamus, spleen, and cecal tonsil) from four of the six individuals by RNA-Seq. The tissue sources and the institutes in charge of the collection are listed in **Supplementary Table 1.** All animal specimens were collected legally in accordance with the policies for the Animal Care and Use Ethics of each institution, making all efforts to minimize invasiveness.

Library construction and genome sequencing

For PacBio continuous long reads (CLR) sequencing, genomic DNA was extracted from chicken liver using a QIAamp DNA Mini Kit (QIAGEN). The integrity of the DNA was determined with an Agilent 4200 Bioanalyzer (Agilent Technologies, Palo Alto, California). Eight micrograms of genomic DNA were sheared using g-Tubes (Covaris), and concentrated with AMPure PB magnetic beads. Each SMRT bell library was constructed using the Pacific Biosciences SMRTbell template prep kit 2.1. The constructed libraries were size-selected on a BluePippin™ system for molecules ≥ 20 kb, followed by primer annealing and the binding of SMRT bell templates to polymerases with the DNA/Polymerase Binding Kit. Sequencing was carried out using P6-C4 chemistry on the Pacific Bioscience Sequel II platform by Annoroad Gene Technology Company.

For short-read DNA sequencing, the genomic DNA of ten samples used for next-generation sequencing (NGS) assembly was extracted from ethylenediaminetetraacetic acid

(EDTA)-anticoagulated blood randomly fragmented. Two paired-end libraries and two mate-pair libraries with insert sizes of 500 bp, 800 bp, 3 Kb and 5 Kb were constructed. All libraries were sequenced on the Illumina HiSeq 2000 platform according to the manufacturer's protocol. After filtering out adapter sequences and low-quality reads, a total of 1.61 Tb (average 134 × coverage of chicken genome) of data were retained for assembly. In addition, the libraries of ten samples used for PacBio sequencing were also constructed using an amplification-free method with an insert size of 350 bp and sequenced on the Illumina XTen platform with paired-end 150 bp sequence reads.

Whole-transcriptome sequencing

For transcriptome analysis, total RNA was extracted using TRIzol extraction reagent (Thermo Fisher). The RNA quality analysis method was the same as DNA quality analysis method described above. Libraries with 250-350 bp insert sizes were prepared using the TruSeq RNA Sample Prep Kit v2 (Illumina, San Diego, CA, USA). To obtain transcriptome profiles, all libraries were sequenced on Illumina XTen system platform using the manufacturer's protocol.

High-throughput chromatin conformation capture (Hi-C) sequencing

Hi-C experiments were performed according to a previously published protocol (Lieberman-Aiden et al. 2009). Hi-C libraries were created from the breast muscle samples of six of the above individuals. All libraries were sequenced on an Illumina HiSeq X Ten sequencer (paired-end sequencing with a 150 bp read length). On average, 127 Gb of data with approximately120-fold genomic coverage and 271,268,477 read pairs could be uniquely aligned to the chicken genome reference sequence (**Supplementary Table 2, Supplementary Table 3**).

441 **Tandem mass spectrometry analysis**

442 The samples were ground into a cell powder in liquid nitrogen and then sonicated in lysis buffer
443 (8 M urea, 1% protease inhibitor cocktail) three times on ice using a high-intensity ultrasonic
444 processor (Scientz). The remaining debris was removed by centrifugation at 12,000 g at 4 °C
445 for 10 min. Thereafter, the supernatant was collected, and the protein concentration was
446 determined with a BCA kit according to the manufacturer's instructions. Then, the protein
447 solution was subjected to trypsin digestion. Next, the tryptic peptides were fractionated by
448 high-pH reverse-phase HPLC using a Thermo Betasil C18 column (5 μm particles, 10 mm ID,
449 and 250 mm length).

450 The tryptic peptides were dissolved in 0.1% formic acid (solvent A) and directly loaded
451 onto a homemade reversed-phase analytical column (15-cm length, 75 μm i.d.). The gradient
452 consisted of an increase from 6% to 23% solvent B (0.1% formic acid in 98% acetonitrile) over
453 26 min, an increase from 23% to 35% over 8 min and then to 80% over 3 min, with holding at
454 80% for the last 3 min, all at a constant flow rate of 400 nL/min in an EASY-nLC 1000 UPLC
455 system. The peptides were introduced to a nanospray ionisation (NSI) source, followed by
456 MS/MS in a Q ExactiveTM Plus system (Thermo) coupled online to the UPLC system.

457 The MS/MS data were processed using the MaxQuant search engine (v.1.5.2.8) (Cox
458 and Mann 2008). Tandem mass spectra were searched against the human UniProt database
459 concatenated with the reverse decoy database. Trypsin/P was specified as the cleavage enzyme,
460 allowing up to 4 missing cleavages. The mass tolerance for precursor ions was set as 20 ppm
461 in the first search and 5 ppm in the main search, and the mass tolerance for fragment ions was
462 set as 0.02 Da. Carbamidomethyl on Cys was specified as a fixed modification and acetylation
463 modifications and oxidation on Met were specified as variable modifications. The FDR was
464 adjusted to < 1%, and the minimum score for modified peptides was set as > 40. For protein
465 identification, peptides containing a minimum of seven amino acids and at least one unique

peptide were required. Only proteins with at least two peptides and at least one unique peptide were considered to have been identified and used for further data analysis.

De novo genome assembly, evaluation and repeat annotation

Assembly based on PacBio SMRT sequencing platform

The raw PacBio SMRT reads were corrected by itself with Canu v1.7 (Koren et al. 2017), and assembled with WTDBG v2.2 (Ruan and Li 2019) to generate the contig layout and edge sequences, and WTPOA-CNS v1.2 was used to obtain the initial consensus in FASTA format. Then, we used minimap2 v2.14-r883 (Li 2018) to map the corrected reads to the consensus, and they were subsequently polished by using WTPOA-CNS v1.2. This process was repeated three times. Next, the consensus sequence obtained in the previous step was mapped by using the NGS reads from the same individual with BWA-MEM v0.7.17-r1188 (Li and Durbin 2010) and then polished with Pilon v1.22 (Walker et al. 2014). This process was repeated three times to obtain the final contigs.

We performed further scaffolding based on the results for six individuals with Hi-C data. Using the final contigs as a reference, we mapped the Hi-C data to the final contigs using Juicer v1.5 (Durand et al. 2016) to obtain the interaction matrix. Finally, 3d-dna v180419 (Dudchenko et al. 2017) was used for scaffolding contigs.

Assembly based on NGS platform

The genomes sequenced on the NGS platform were *de novo* assembled into contigs by using a pipeline that combined the Fermi package (Li 2012) and Phusion assembler (Mullikin and Ning 2003) for 500 bp/800 bp paired-end libraries. For the 3 Kb/5 Kb mate-pair libraries, we used SOAPdenovo (Li et al. 2010) with 77 kmers to build contigs. Furthermore, SSPACE (Boetzer et al. 2011) was used to build scaffolds, and the contigs assembled by Fermi and Phusion were

used for the substitution of sequences and bases and for further rectifying to rectify the local assembly error. After the inspection of the initial scaffolds, gaps were closed using Gap5 (Bonfield and Whitwham 2010) software.

Genome evaluation and annotation

BUSCO v3.0.2 (Simao et al. 2015) was used to assess assembly completeness by estimating the percentage of expected single-copy conserved orthologues captured in our assemblies and reference genome , referring to the lineage dataset aves_odb9 (Creation date: 2016-02-13, number of species: 40, number of BUSCOs: 4,915). Repeat sequences were annotated using RepeatMasker v4.0.8 (with the parameters: -engine ncbi -species 'Gallus gallus' -s -no_is -cutoff 255 -frag 20000). Subsequently, tandem repeats were further annotated using Tandem Repeats Finder v4.07b (Benson 1999) (with the settings 2 7 7 80 10 50 2000 -d -h). In addition, Quadron software (Sahakyan et al. 2017) was used to predict G4 motifs, and only nonoverlapping hits with a score greater than 19 were used for subsequent analysis.

Chicken pan-genome construction

The *de novo* assemblies were aligned to the chicken reference genome (GRCg6a; GCF_000002315.6) using minimap2 (Li 2018) (-cx asm10). Based on the pairwise alignment, unaligned or low-identity sequences (showing more than 10% sequence divergence relative to GRCg6a) were extracted. Then, the adjacent sequences within 200 bp were merged. BLASTN 2.6.0+ (Camacho et al. 2009) (with the parameters -word_size 20 -max_hsps 1 -max_target_seqs 1 -dust no -soft_masking false -evalue 0.00001) was further used to align the unaligned sequences from the previous step to GRCg6a, and the sequences showing identity greater than 90% to GRCg6a sequences were removed. The remaining sequences were merged according to the adjacent regions within 200 bp, and sequences of less than 500 bp in length were removed. Subsequently, the unaligned and low-identity sequences obtained from all of

the assemblies were combined, redundancies were removed with CD-HIT v4.7 (Fu et al. 2012) (parameter: -c 0.9 -aS 0.8 -d 0 -sf 1), and the longest sequence in the cluster was selected as the representative sequence. To further exclude potential contaminants in the dataset, we used BLASTN to compare the non-redundant set with the nr database of NCBI (v20181220). The sequences that were aligned to non-Chordata species were removed from the final novel sequence set (**Supplementary Table 5**).

Observed present or absent analysis of novel sequences in resequenced individuals

The whole-genome resequencing data of 922 chickens (Li et al. 2017; Wang et al. 2020) (**Supplementary Table 6**) were downloaded for the present or absent analysis of novel sequences. To explore whether the different sequencing platforms affected the results, the Illumina sequencing reads of the GRCg6a individual (SRR3954707 (Warren et al. 2017), which were previously used for single-base error correction) were also included in this analysis. The presence and absence of each novel sequence was then determined according to the sequence coverage and depth. First, to obtain high-quality reads and minimize false genotyping results due to low-quality reads supplied by Illumina, we implemented the following quality control procedures to filter the reads before read mapping using Trimmomatic v0.36 (Bolger et al. 2014), and leading or trailing stretches of Ns and bases with a quality score below 3 were trimmed. Then, the reads were scanned using a 4-base wide sliding window and clipped when the average quality per base was below 15, and only reads of 40 nucleotides or longer were finally retained. Second, high-quality paired reads were aligned to GRCg6a using BWA-MEM v0.7.17 (Li and Durbin 2010) with the default parameters, except that "-M" was enabled. The BWA-aligned BAM files were then processed using Picard v2.1 (http://broadinstitute.github.io/picard/), including reads sorted and merged read groups belonging to the same sample and marked duplicates at the sample level. Finally, we estimated

541 the coverage distribution at each called site for each sample using QualiMap v2.2
542 (Okonechnikov et al. 2016).

543 Poorly aligned or unaligned reads were extracted as follows: Samblaster v0.1.24 (Faust
544 and Hall 2014) was used to extract clipped reads and unaligned reads, while sambamba v0.6.8
545 (Tarasov et al. 2015) and SAMTools v1.9 (Li et al. 2009) were used to collect other poorly
546 aligned reads. The paired reads with unaligned/poorly aligned read pairs were extracted using
547 seqtk v1.3-r106 (https://github.com/lh3/seqtk) and were then aligned to the novel sequence set
548 using a previously described process. Novel sequences with a coverage above 0.8 and a depth
549 greater than one-quartered of the whole-genome depth were identified as present.

550

551 **Feature importance analysis**

552 To estimate the influence of GC, G4 motif, and TR contents on the observed frequency of novel
553 sequences, 9,200 novel sequences shared by all individuals were used to construct a random
554 forest model. The sklearn package in Python was used to build the final model and perform
555 classification.

556

557 **Transcribed region annotation and coding potential assessment**

558 The raw RNA-Seq reads were processed to remove adapters, low-quality sequences and
559 sequences with poly A/T tails using Trimmomatic v0.36 (Bolger et al. 2014). The cleaned reads
560 were *de novo* assembled using SPAdes v3.14.1 (Bushmanova et al. 2019). The expression
561 levels of the *de novo* assembled transcripts were quantified by using Kallisto v0.46.2 (Bray et
562 al. 2016). Additionally, the cleaned reads were assembled using a reference-guided method by
563 alignment to the *de novo* genome assemblies using HISAT2 v2.0.3-beta (Kim et al. 2019) with
564 the default parameters, except that "--dta" was enabled. Transcripts including novel splice
565 variants were assembled using StringTie v1.2.2 (Pertea et al. 2015) with the default parameters.

Then, StringTie (--merge) was used to merge all the transcript GTFs obtained from the samples mapped to this assembly to obtain a reference annotation. Finally, all samples were reassembled and quantified using StringTie with the reference annotation to obtain the expression level of each transcript. Notably, the transcripts with fragments per kilobase per million mapped reads (FPKM) ≥ 1 were considered robustly expressed.

Redundancy among genes that were annotated based on the *de novo* and reference-guided methods and intersected with novel sequences was removed with CD-HIT (parameter: -c 0.9 -aS 0.8 -d 0 -sf 1). Then, the remaining genes were searched against the nr database and the genes of GRCg6a using BLASTN 2.6.0+. Genes with no hits to either non-Chordata species or GRCg6a were retained as "novel genes" that were completely absent in the chicken reference genome. Genes showing hits to GRCg6a genes with more than 95% identity were classified as partially missing in the chicken reference genome.

Next, the coding potential of these novel genes was assessed by using CPAT v1.2.3 (Wang et al. 2013) with the default parameters. CPAT uses an alignment-independent logistic regression model to detect coding potential based on sequence features. To select a cut-off for classification, we built hexamer tables and logit models for chicken using chicken CDSs and ncRNA sequences downloaded from Ensembl (release 98) as training data. Then, a two-graph receiver operating characteristic curve was used to determine the optimum cut-off value through 10 random sample validations (**Supplementary Figure 20**). A cut-off of 0.69 was selected to classify the novel genes as potential protein-coding or noncoding genes. Then, the ORFs were searched by using TransDecoder v5.5.0 (http://transdecoder.github.io) and ORFfinder v0.4.3 (https://www.ncbi.nlm.nih.gov/orffinder/) with the default parameters. Genes showing values above the cut-off of the CPAT with a minimum ORF of at least 100 amino acids were classified as novel coding genes. For the remaining novel genes, RNAcode (Washietl et al. 2011) was used to further estimate the coding potential. To prevent the

591 divergent homologous haplotypes that can caused false gene duplications (Ko et al. 2021), we
592 merged novel coding genes that have high similarity with each other or can be annotated to the
593 same gene, and then performed manual check. We generated customized whole-genome
594 alignments for each *de novo* assembly against Japanese quail (GCF_001577835.1), turkey
595 (GCF_000146605.3), and helmeted guineafowl (GCF_002078875.1), which we used to
596 estimate coding potential. We used BLASTX 2.6.0+ (with the parameters '-evalue 0.00001')
597 to translate each novel genes form all six possible reading frames, and the results were
598 compared to known proteins in the nr database. Genes with an E value $\leq 10^{-5}$, alignment length
599 of $\geq$ 10 amino acids and identity $\geq$ 95% were removed from the final potential long noncoding
600 gene set. Only multiple exon genes with more than 200 nucleotides and without any detectable
601 protein-coding potential were classified as novel long noncoding genes.
602
603 **Protein-coding gene annotation**
604 Using the human (*Homo sapiens*) dataset as the background, the novel coding genes were
605 annotated with the annotate module of online KOBAS 3.0 (Xie et al. 2011)
606 (http://kobas.cbi.pku.edu.cn/). The Gene Ontology terms, KEGG pathways, and Reactome
607 pathways of these genes were characterized by using the enrichment module of online KOBAS
608 3.0. $P < 0.05$ was set as the cut-off threshold.
609 InterProScan v5.36-75.0 (Jones et al. 2014) (parameter: -f tsv -dp) was used to classify
610 the protein-coding gene fragments within the novel sequences and the protein-coding genes
611 influenced by the location of novel sequences into protein families. The analysis results of
612 Pfam 32.0 (http://pfam.xfam.org/) were selected to determine the families to which the proteins
613 belonged.
614

615 **Differential expression analysis**

616 The expression levels of each gene obtained from the previous step were used for differential
617 expression analysis. The R language was used to identify differentially expressed genes with
618 the edgeR package (v3.28.1) (Robinson et al. 2010). The fold changes between the two groups
619 were calculated as logFC = log2 (experimental/control group). Benjamini-Hochberg correction
620 was used to correct for multiple comparisons (with a false discovery cut-off of < 0.05). Genes
621 in the two groups a with |logFC| > 2 and q-value < 0.05 were defined as differentially expressed
622 genes.

623

624 **Anchoring novel sequences onto the reference genome**

625 *Flanking sequences*

626 The novel sequences were anchored to GRCg6a based on alignment information between all
627 *de novo* assemblies and GRCg6a. First, the scaffolds of the *de novo* assemblies that contained
628 novel sequences were extracted and anchored on the chromosome/scaffold of GRCg6a which
629 showed the most alignment hits with them. Then, the adjacent flanking sequences (more than
630 100 bp) of the novel sequences aligned to the same chromosome/scaffold were retained for
631 further positioning. If the flanking sequences were perfectly aligned to GRCg6a with no gaps,
632 an identity ≥ 90% and a breakpoint shift of ≤5bp, we recorded the sequences as “placed”. The
633 other alignments were recorded as “ambiguously placed”. The novel sequences with two placed
634 flanking sequences were reported as “localized”. The novel sequences with one or two
635 ambiguously placed flanking sequences were reported as “unlocalized”. The final remaining
636 sequences were reported as “unplaced”. Based on the genome placement information, the
637 localized sequences could be further classified as insertions, alternate alleles, or ambiguous
638 sequences. The insertions introduced only one sequence fragment to the reference genome and
639 were no more than 10 bp in length. For alternate alleles, the novel sequences had to share less

than 90% (or 0%) identity with their counterparts in the reference. Furthermore, the novel sequences and their counterparts had to have comparable lengths, with a length ratio between 1/3 and 3. The remaining sequences that did not meet the above criteria for insertions and alternate alleles were classified as ambiguous sequences.

Chromosome interaction mapping

The preprocessing of paired-end sequencing data, mapping of reads and filtering of mapped di-tags were performed using the Juicer pipeline (version 1.5) (Durand et al. 2016). Briefly, short reads were mapped to the chicken pan-genome using BWA-MEM (version 0.7.17-r1188) (Li and Durbin 2010). Reads with low mapping quality were filtered using Juicer with the default parameters, discarding invalid self-ligated and un-ligated fragments as well as PCR artefacts. Filtered di-tags were further processed with Juicer command line tools to bin ditags (10 kb bins) and to normalize matrices with KR normalization (Knight and Ruiz 2012). We normalized all Hi-C matrices on the same scale by KR normalization, ensuring that any differences between Hi-C data were not attributable to variation in sequence length. The maximum 100-kb bin of each novel sequence interaction (interaction intensity ≥5) was collected as a potential location of novel sequences. Novel sequences that were validated in at least two individuals with Hi-C data and anchored to the same location were remained.

Gene orthology and dN/dS analysis

The integrated toolkit TBtools v1.0 (Chen et al. 2020) was used for collinearity analysis between species. First, the protein sequence of each gene was obtained, and pairwise sequence similarities were calculated using BLASTP with a cut-off of E value ≤ 10^{-10}. Then, syntenic blocks were detected using MCScanX v1.0 (Wang et al. 2012) with the default parameters. OrthoFinder v2.4.0 (Emms and Kelly 2019) was used to identify orthologous genes with the default parameters. Among these genes, 1:1 orthologous genes between different species were

665 used for downstream analysis. Using 1:1 orthologous genes as the input, Codeml in PAML
666 version 4.9d (Yang 2007) was used for dN/dS analysis with the default parameters. The genome
667 assemblies and corresponding annotations used in this analysis as below: Gray short-tailed
668 opossum (GCF_000002295.2), Greater horseshoe bat (GCF_004115265.1), Human
669 (GCF_000001405.39), Western terrestrial garter snake (GCF_009769535.1), Common lizard
670 (GCF_011800845.1), Red-eared slider turtle (GCF_013100865.1), Goodes thornscrub tortoise
671 (GCF_007399415.2), Green sea turtle (GCF_015237465.1), American alligator
672 (GCF_000281125.3), Chinese alligator (GCF_000455745.1), Australian saltwater crocodile
673 (GCF_001723895.1), Okarito brown kiwi (GCF_003343035.1), African ostrich
674 (GCF_000698965.1), Emu (GCA_016128335.1), Golden eagle (GCF_900496995.1), Kakapo
675 (GCF_004027225.2), New Caledonian crow (GCF_009650955.1), Swainson's thrush
676 (GCF_009819885.1), Zebra finch (GCF_008822105.2), Mallard (GCF_015476345.1),
677 Helmeted guineafowl (GCF_002078875.1), Turkey (GCF_000146605.3), Japanese quail
678 (GCF_001577835.2), and Chicken (GCF_000002315.6).
679

680 **Supplementary Materials:**

681 Supplementary Note

682 Supplementary Figure 1 to 29

683 Legends for Supplementary Table 1 to 17

684 Supplementary References

685

686 **Reference:**

687 Bell AD, Mello CJ, Nemesh J, Brumbaugh SA, Wysoker A, McCarroll SA. 2020. Insights into
688 variation in meiosis from 31,228 human sperm genomes. Nature 583:259-264.
689 Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic
690 Acids Res 27:573-580.
691 Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled
692 contigs using SSPACE. Bioinformatics 27:578-579.
693 Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina
694 sequence data. Bioinformatics 30:2114-2120.
695 Bonfield JK, Whitwham A. 2010. Gap5--editing the billion fragment sequence assembly.
696 Bioinformatics 26:1699-1703.
697 Bornelov S, Seroussi E, Yosefi S, Pendavis K, Burgess SC, Grabherr M, Friedman-Einat M,
698 Andersson L. 2017. Correspondence on Lovell et al.: identification of chicken genes
699 previously assumed to be evolutionarily lost. Genome Biol 18:112.
700 Botero-Castro F, Figuet E, Tilak MK, Nabholz B, Galtier N. 2017. Avian Genomes Revisited:
701 Hidden Genes Uncovered and the Rates versus Traits Paradox in Birds. Mol Biol Evol
702 34:3123-3131.
703 Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq
704 quantification. Nat Biotechnol 34:525-527.
705 Brown WR, Hubbard SJ, Tickle C, Wilson SA. 2003. The chicken as a model for large-scale
706 analysis of vertebrate gene function. Nat Rev Genet 4:87-98.
707 Burt DW. 2005. Chicken genome: current status and future opportunities. Genome Res
708 15:1692-1698.
709 Burt DW, Bruley C, Dunn IC, Jones CT, Ramage A, Law AS, Morrice DR, Paton IR, Smith J,
710 Windsor D, et al. 1999. The dynamics of chromosome evolution in birds and mammals.
711 Nature 402:411-413.
712 Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. 2019. rnaSPAdes: a de novo
713 transcriptome assembler and its application to RNA-Seq data. Gigascience 8.
714 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009.
715 BLAST+: architecture and applications. BMC Bioinformatics 10:421.
716 Cardoso-Moreira M, Halbert J, Valloton D, Velten B, Chen C, Shao Y, Liechti A, Ascenção K,
717 Rummel C, Ovchinnikova S, et al. 2019. Gene expression across mammalian organ
718 development. Nature.
719 Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, Xia R. 2020. TBtools: An Integrative
720 Toolkit Developed for Interactive Analyses of Big Biological Data. Mol Plant 13:1194-
721 1202.
722 Chen WJ, Yang JY, Lin JH, Fann CS, Osyetrov V, King CC, Chen YM, Chang HL, Kuo HW, Liao
723 F, et al. 2006. Nasopharyngeal shedding of severe acute respiratory syndrome-associated
724 coronavirus is associated with genetic polymorphisms. Clin Infect Dis 42:1561-1569.
725 Cooper MD, Raymond DA, Peterson RD, South MA, Good RA. 1966. The functions of the
726 thymus system and the bursa system in the chicken. J Exp Med 123:75-102.
727 Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized
728 p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol
729 26:1367-1372.
730 Duan Z, Qiao Y, Lu J, Lu H, Zhang W, Yan F, Sun C, Hu Z, Zhang Z, Li G, et al. 2019. HUPAN:
731 a pan-genome analysis pipeline for human genomes. Genome Biol 20.

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, et al. 2017. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science 356:92-95.

Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. 2016. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst 3:95-98.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20:238.

Faust GG, Hall IM. 2014. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics 30:2503-2505.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150-3152.

Gill F, Donsker D, Rasmussen P. 2020. IOC World Bird List (v10.1).

Golicz AA, Bayer PE, Bhalla PL, Batley J, Edwards D. 2019. Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications. Trends Genet.

Guiblet WM, Cremona MA, Cechova M, Harris RS, Kejnovska I, Kejnovsky E, Eckert K, Chiaromonte F, Makova KD. 2018. Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate. Genome Res 28:1767-1778.

Hounkpe BW, Chenou F, de Lima F, De Paula EV. 2021. HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. Nucleic Acids Res 49:D947-D955.

International Chicken Genome Sequencing C. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695-716.

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236-1240.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37:907-915.

Kim J, Lee C, Ko BJ, Yoo D, Won S, Phillippy A, Fedrigo O, Zhang G, Howe K, Wood J, et al. 2021. False gene and chromosome losses affected by assembly and sequence errors. bioRxiv.

Knight PA, Ruiz D. 2012. A fast algorithm for matrix balancing. IMA Journal of Numerical Analysis 33:1029-1047.

Ko BJ, Lee C, Kim J, Rhie A, Yoo D, Howe K, Wood J, Cho S, Brown S, Formenti G, et al. 2021. Widespread false gene gains caused by duplication errors in genome assemblies. bioRxiv.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 27:722-736.

Li D, Che T, Chen B, Tian S, Zhou X, Zhang G, Li M, Gaur U, Li Y, Luo M, et al. 2017. Genomic data for 78 chickens from 14 populations. Gigascience 6:1-5.

Li H. 2012. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. Bioinformatics 28:1838-1844.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094-3100.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589-595.

780 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R,
781 Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and
782 SAMtools. Bioinformatics 25:2078-2079.
783 Li R, Fu W, Su R, Tian X, Du D, Zhao Y, Zheng Z, Chen Q, Gao S, Cai Y, et al. 2019. Towards
784 the Complete Goat Pan-Genome by Recovering Missing Genomic Segments From the
785 Reference Genome. Front Genet 10:1169.
786 Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, et al. 2010. De
787 novo assembly of human genomes with massively parallel short read sequencing.
788 Genome Res 20:265-272.
789 Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I,
790 Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range
791 interactions reveals folding principles of the human genome. Science 326:289-293.
792 Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM, Trask BJ. 2005. Human
793 subtelomeres are hot spots of interchromosomal recombination and segmental
794 duplication. Nature 437:94-100.
795 Liu AL, Li YF, Qi W, Ma XL, Yu KX, Huang B, Liao M, Li F, Pan J, Song MX. 2015. Comparative
796 analysis of selected innate immune-related genes following infection of immortal DF-1
797 cells with highly pathogenic (H5N1) and low pathogenic (H9N2) avian influenza viruses.
798 Virus Genes 50:189-199.
799 Lovell PV, Wirthlin M, Wilhelm L, Minx P, Lazar NH, Carbone L, Warren WC, Mello CV.
800 2014. Conserved syntenic clusters of protein coding genes are missing in birds. Genome
801 Biol 15:565.
802 Marais G. 2003. Biased gene conversion: implications for genome and sex evolution.
803 Trends in Genetics 19:330-338.
804 Marjuki H, Scholtissek C, Yen HL, Webster RG. 2008. CK2beta gene silencing increases cell
805 susceptibility to influenza A virus infection resulting in accelerated virus entry and higher
806 viral protein content. J Mol Signal 3:13.
807 Mullikin JC, Ning Z. 2003. The phusion assembler. Genome Res 13:81-90.
808 Okonechnikov K, Conesa A, Garcia-Alcalde F. 2016. Qualimap 2: advanced multi-sample
809 quality control for high-throughput sequencing data. Bioinformatics 32:292-294.
810 Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie
811 enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol
812 33:290-295.
813 Polleys EJ, House NCM, Freudenreich CH. 2017. Role of recombination and replication
814 fork restart in repeat instability. DNA Repair (Amst) 56:156-165.
815 Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W,
816 Fungtammasan A, Kim J, et al. 2021. Towards complete and error-free genome assemblies
817 of all vertebrate species. Nature 592:737-746.
818 Richard GF, Paques F. 2000. Mini- and microsatellite expansions: the recombination
819 connection. EMBO Rep 1:122-126.
820 Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for
821 differential expression analysis of digital gene expression data. Bioinformatics 26:139-
822 140.
823 Ruan J, Li H. 2019. Fast and accurate long-read assembly with wtdbg2. Nat Methods.
824 Ruckle A, Haasbach E, Julkunen I, Planz O, Ehrhardt C, Ludwig S. 2012. The NS1 protein
825 of influenza A virus blocks RIG-I-mediated activation of the noncanonical NF-kappaB
826 pathway and p52/RelB-dependent gene expression in lung epithelial cells. J Virol
827 86:10211-10217.

828 Sahakyan AB, Chambers VS, Marsico G, Santner T, Di Antonio M, Balasubramanian S. 2017.
829 Machine learning model for sequence-driven DNA G-quadruplex formation. Sci Rep
830 7:14535.
831 Sarropoulos I, Marin R, Cardoso-Moreira M, Kaessmann H. 2019. Developmental
832 dynamics of lncRNAs across mammalian organs and species. Nature.
833 Schatz MC, Delcher AL, Salzberg SL. 2010. Assembly of large genomes using second-
834 generation sequencing. Genome Res 20:1165-1173.
835 Schmid ET, Pang IK, Carrera Silva EA, Bosurgi L, Miner JJ, Diamond MS, Iwasaki A, Rothlin
836 CV. 2016. AXL receptor tyrosine kinase is required for T cell priming and antiviral
837 immunity. Elife 5.
838 Seroussi E, Pitel F, Leroux S, Morisson M, Bornelov S, Miyara S, Yosefi S, Cogburn LA, Burt
839 DW, Anderson L, et al. 2017. Mapping of leptin and its syntenic genes to chicken
840 chromosome 1p. Bmc Genetics 18:77.
841 Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, Boorgula MP, Chavan S,
842 Vergara C, Ortega VE, et al. 2018. Assembly of a pan-genome from deep sequencing of 910
843 humans of African descent. Nat Genet.
844 Sherman RM, Salzberg SL. 2020. Pan-genomics in the human genome era. Nature Reviews
845 Genetics.
846 Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO:
847 assessing genome assembly and annotation completeness with single-copy orthologs.
848 Bioinformatics 31:3210-3212.
849 Smith J, Smith N, Yu L, Paton IR, Gutowska MW, Forrest HL, Danner AF, Seiler JP, Digard
850 P, Webster RG, et al. 2015. A comparative analysis of host responses to avian influenza
851 infection in ducks and chickens highlights a role for the interferon-induced
852 transmembrane proteins in viral resistance. BMC Genomics 16:574.
853 Stehelin D, Varmus HE, Bishop JM, Vogt PK. 1976. DNA related to the transforming gene(s)
854 of avian sarcoma viruses is present in normal avian DNA. Nature 260:170-173.
855 Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. 2015. Sambamba: fast processing of
856 NGS alignment formats. Bioinformatics 31:2032-2034.
857 Tian X, Li R, Fu W, Li Y, Wang X, Li M, Du D, Tang Q, Cai Y, Long Y, et al. 2020. Building a
858 sequence map of the pig pan-genome from multiple de novo assemblies and Hi-C data. Sci
859 China Life Sci 63:750-763.
860 Vogt PK. 2011. Historical introduction to the general properties of retroviruses.
861 Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q,
862 Wortman J, Young SK, et al. 2014. Pilon: an integrated tool for comprehensive microbial
863 variant detection and genome assembly improvement. PLoS ONE 9:e112963.
864 Wang B, Lam TH, Soh MK, Ye Z, Chen J, Ren EC. 2018. Influenza A Virus Facilitates Its
865 Infectivity by Activating p53 to Inhibit the Expression of Interferon-Induced
866 Transmembrane Proteins. Front Immunol 9:1193.
867 Wang K, Hu H, Tian Y, Li J, Scheben A, Zhang C, Li Y, Wu J, Yang L, Fan X, et al. 2021. The
868 chicken pan-genome reveals gene content variation and a promoter region deletion in
869 IGF2BP1 affecting body size. Mol Biol Evol.
870 Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W. 2013. CPAT: Coding-Potential
871 Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res
872 41:e74.
873 Wang MS, Thakur M, Peng MS, Jiang Y, Frantz LAF, Li M, Zhang JJ, Wang S, Peters J, Otecko
874 NO, et al. 2020. 863 genomes reveal the origin and domestication of chicken. Cell Res
875 30:693-701.

Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, et al. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res 40:e49.

Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, Markovic C, Bouk N, Pruitt KD, Thibaud-Nissen F, et al. 2017. A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. G3 (Bethesda) 7:109-117.

Washietl S, Findeiss S, Muller SA, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF, Goldman N. 2011. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. RNA 17:578-594.

Wisskirchen C, Ludersdorfer TH, Muller DA, Moritz E, Pavlovic J. 2011. The cellular RNA helicase UAP56 is required for prevention of double-stranded RNA formation during influenza A virus infection. J Virol 85:8646-8655.

Wong KHY, Levy-Sakin M, Kwok PY. 2018. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. Nat Commun 9:3040.

Wong KHY, Ma W, Wei CY, Yeh EC, Lin WJ, Wang EHF, Su JP, Hsieh FJ, Kao HJ, Chen HH, et al. 2020. Towards a reference genome that captures global genetic diversity. Nat Commun 11:5482.

Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li CY, Wei L. 2011. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res 39:W316-322.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591.

Yin ZT, Zhu F, Lin FB, Jia T, Wang Z, Sun DT, Li GS, Zhang CL, Smith J, Yang N, et al. 2019. Revisiting avian 'missing' genes from de novo assembled transcripts. BMC Genomics 20:4.

Zhang GJ, Li C, Li QY, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW, et al. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. Science 346:1311-1320.

Zhao J, Bacolla A, Wang G, Vasquez KM. 2010. Non-B DNA structure-induced genetic instability and evolution. Cell Mol Life Sci 67:43-62.

Zhu F, Yin ZT, Wang Z, Smith J, Zhang F, Martin F, Ogeh D, Hincke M, Lin FB, Burt DW, et al. 2021. Three chromosome-level duck genome assemblies provide insights into genomic variation during domestication. Nat Commun 12:5932.

Acknowledgements

We thank High-Performance Computing (HPC) of Northwest A&F University (NWAFU) for providing computing resources. We thank Margarida Cardoso-Moreira for kindly providing the individual source information of the 217 chicken transcriptomes. **Funding:** This project was supported by grant from the National Natural Science Foundation of China (31822052, 31572381) and the National Thousand Youth Talents Plan to Y.J., National Natural Science Foundation of China (No. 31930105) and China Agriculture Research System of MOF and MARA(CARS-40) to N.Y., National Natural Science Foundation of China (31761143014) to Q.N. and the Unit of Excellence 2021, University of Phayao, Thailand (UoE64003) to C.S.

Author contributions

N.Y., Y.J. and X.H. conceived the project and designed the research. M.L., X.T., P.B. and N.X. performed the majority of the analysis with contributions from Y.W., X.D., R.L., Y.G., F.W., Xiangnan W., P.Y., S.Z.. C. Sun, C.W., F.L., X.L., A.S. and C.S. prepared the DNA samples. X.J. and Q.N. provided the genome of Fayoumi. M.L., Xihong W. and N.X. drafted the manuscripts with input from all authors, and Y.J., C. Sun, Y.W., R.H. M.W. and X.Z. revised the manuscript.

Competing interests

The authors declare no competing interests.

Data availability

All the data of our study are publicly available at the NCBI Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) under accession code BioProject: PRJNA573584. The

930 data are available at

931 https://dataview.ncbi.nlm.nih.gov/object/PRJNA573584?reviewer=buapp4f6dl06ls0snslqmitq

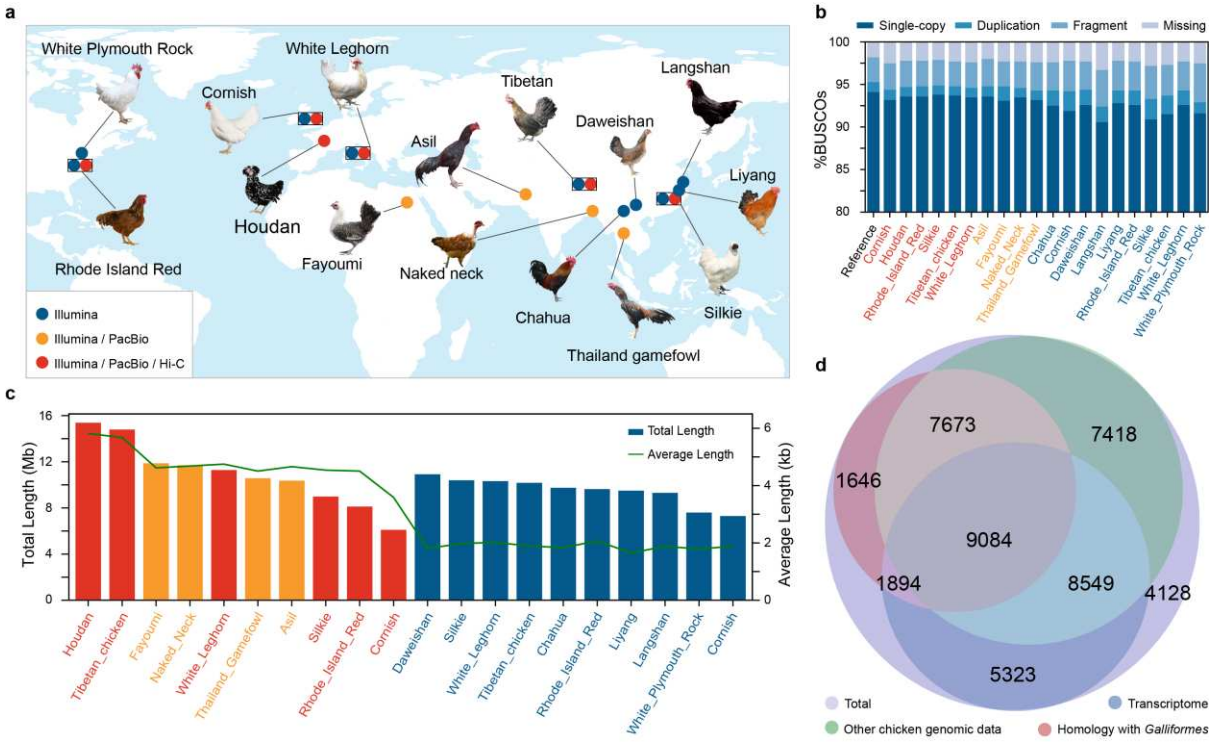
932 h6 and

933 http://animal.nwsuaf.edu.cn/code/index.php/panChicken/loadByGet?address[]=panChicken/

934 Download/Denovo.php.

935

936 **Figure:**



937

938 **Figure 1**

939 **Chicken novel non-redundant sequences identified by twenty *de novo* assemblies. a**,
940 Geographic locations of the original chicken breeds used for *de novo* assembly and their
941 sequencing platforms. The rectangle indicates this breed has two individuals. **b**, Genome
942 assembly completeness assessed by BUSCO. **c**, Length of novel sequences initially obtained
943 from 20 *de novo* assemblies. The polygonal line represents the average length and the column
944 represents the total length. **d**, The number of novel sequences validated by other chicken
945 genomes, homology with Galliformes and transcriptome. The colors of the breed name in **b**
946 and **c** are consistent with **a**.

947

948

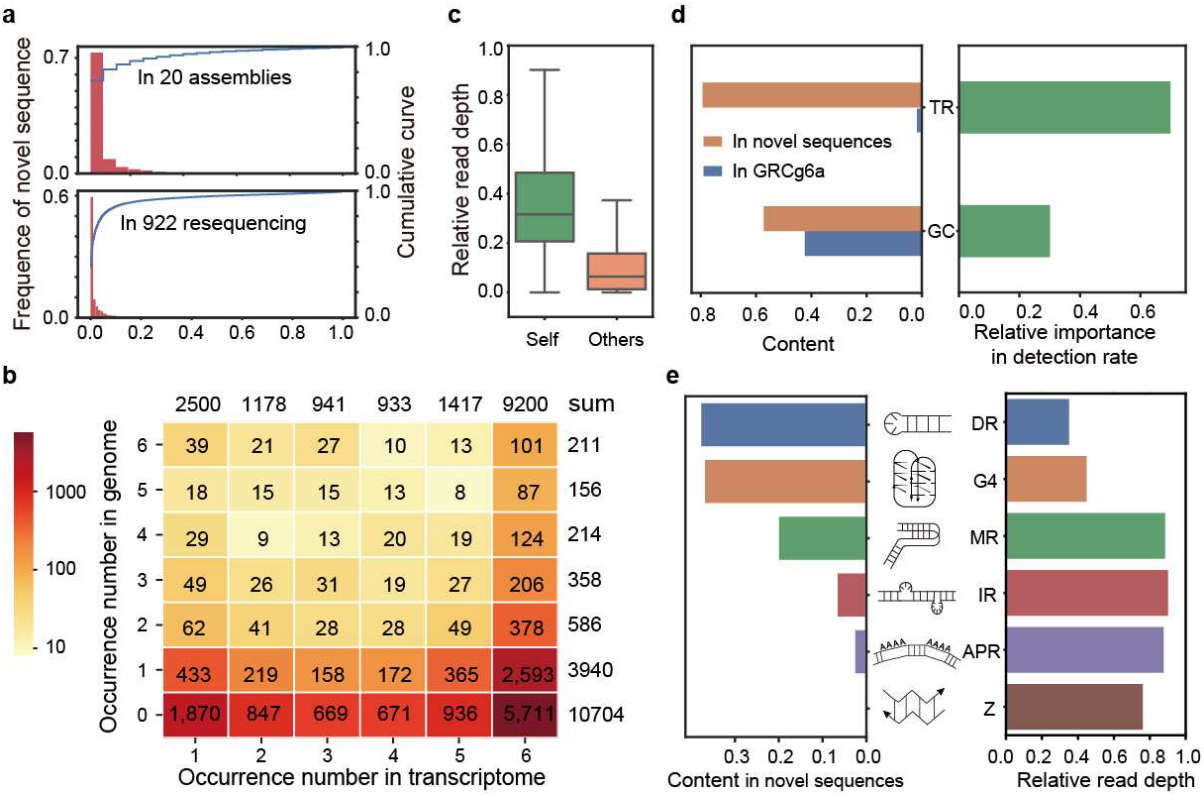


Figure 2

Characterization of novel sequences. a, The distribution and cumulative curve of observed frequencies of novel sequences in 20 assemblies and 922 resequenced individuals. **b**, The observed frequency of the expressed novel sequences in the transcriptomes of six chickens(column) and their corresponding genomes (row). **c**, Relative read depth of novel sequences in the specific assembly which the novel sequence was present (green) and absent (orange) The whole genome read depth was set to one. **d**, Left: the TR and GC content of the GRCg6a and novel sequences, respectively; Right: the feature importance of TR and GC for the detection rate of novel sequences. **e**, Left: the content of noncanonical DNA structures in the novel sequences; Middle: the putative structures of noncanonical DNA; Right: the read depth ratio of novel sequences with or without noncanonical DNA structures.

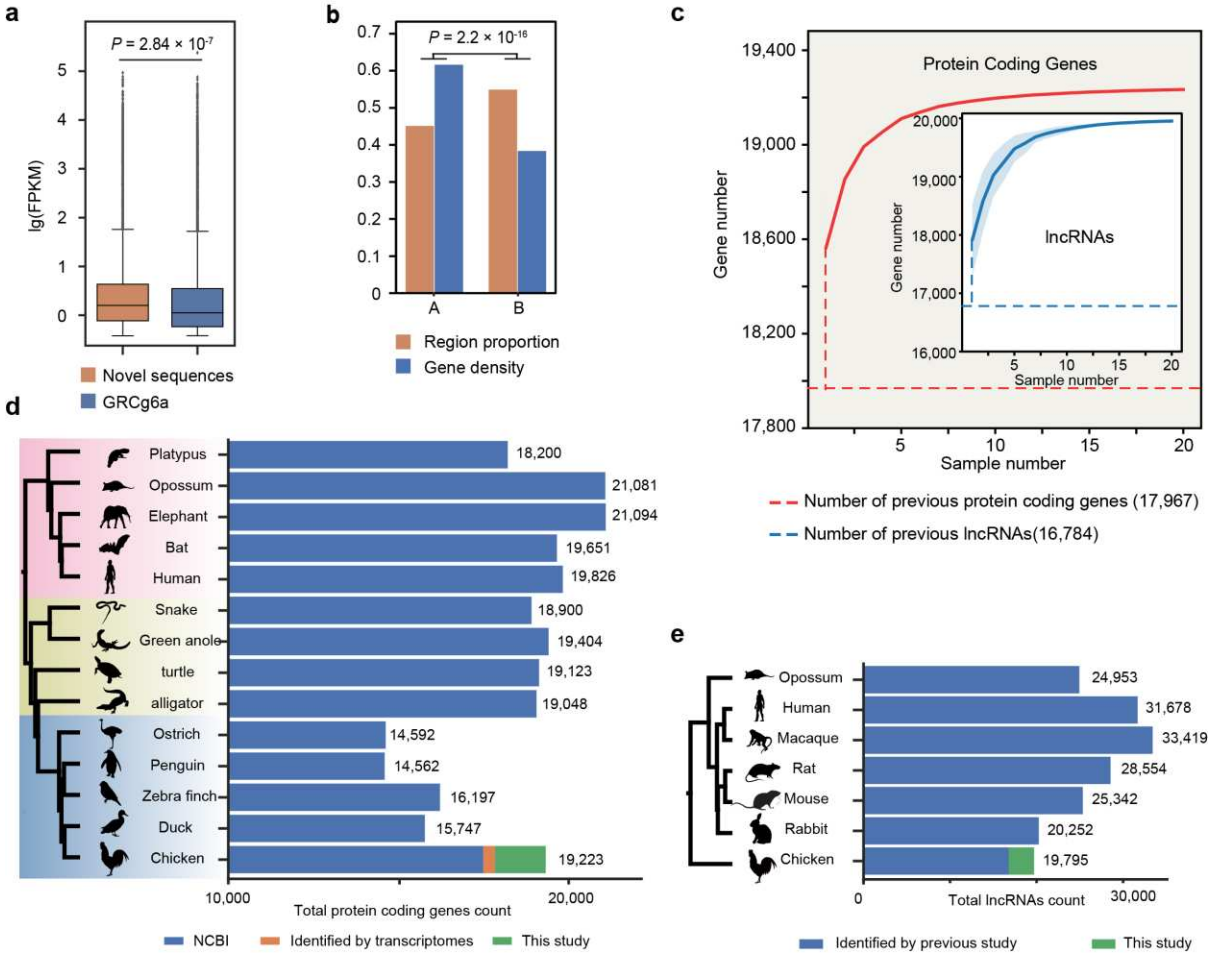


Figure 3

Abundant genes embedded in novel sequences. a, Relative expression of transcripts of reference and novel sequences, respectively. **b**, The region proportion of A/B compartment and the novel gene number proportion in A/B compartment, respectively. **c**, The identification of protein coding genes/lncRNAs increased with sample numbers. The shaded area indicates the 95% confidence interval. **d**, The number of protein coding genes in representative species including mammals, reptiles and birds. Blue, orange and green columns refer to protein coding genes identified by NCBI, Yin et al (2019). and our study, respectively. **e**, Total lncRNAs numbers of mammalian representative species and chicken. Blue and green columns refer to lncRNAs identified by Sarropoulos et al (2019). and this study, respectively.

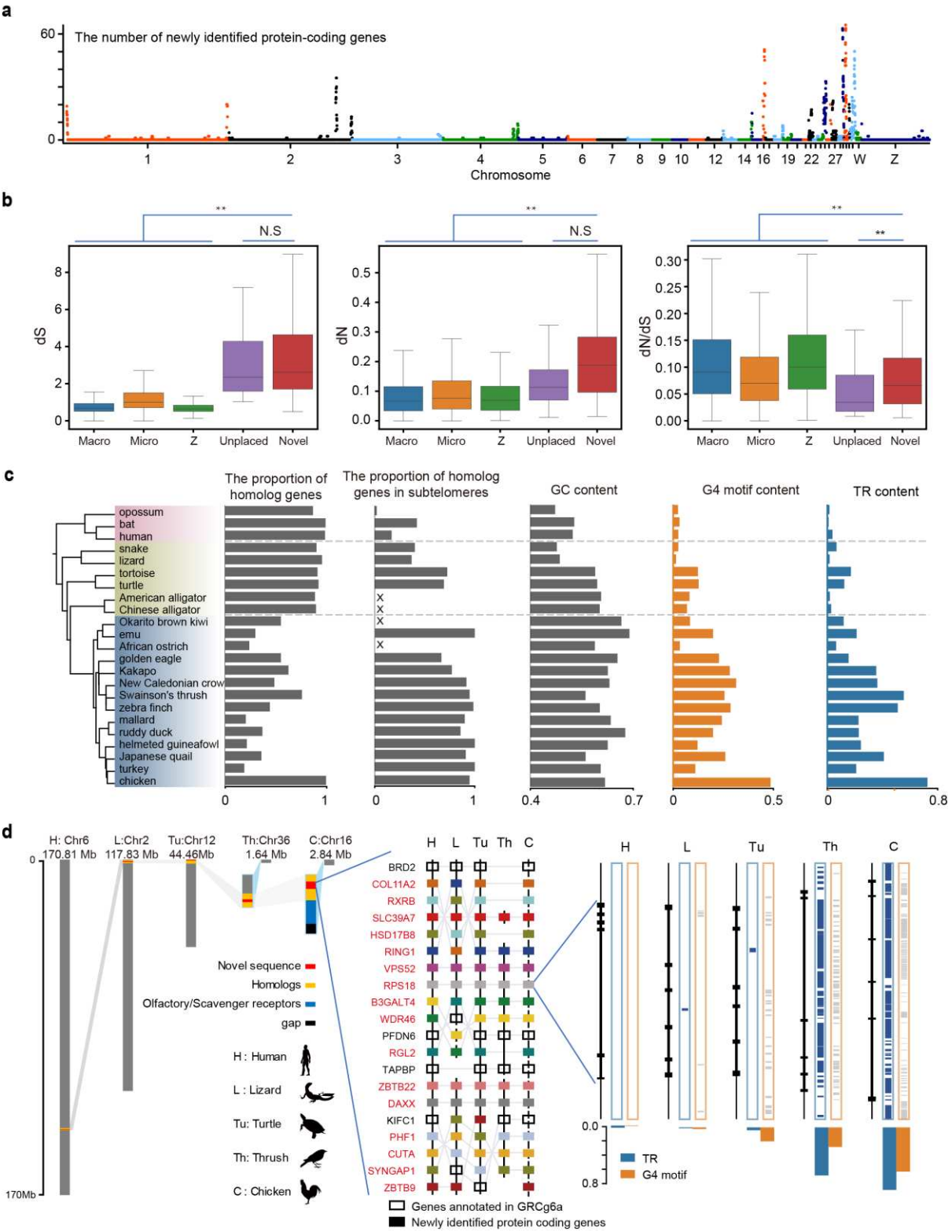


Figure 4

The novel coding genes clustered in the sub-telomere of chromosomes. a, The location of the novel coding gene clusters on chromosomes. **b**, Box plot for dS, dN, dN/dS values of genes on macro-chromosomes, micro-chromosomes, the Z chromosome, unplaced scaffold of

chicken reference genome, and the novel sequences. **c**, The proportion of homologous gene number, homologous gene in sub-telomere, and contents of GC, G4 motif and TR of homologous novel coding genes clusters in chicken and 22 other species. The region containing more than 3 genes are considered as cluster, and genes located within 5-MB of the end of chromosomes are considered as sub-telomeric region. **d**, A detailed synteny conservation of novel coding genes on chromosome 16 of chicken with Mammal (human), Reptilia (lizard, turtle) and Aves (thrush, chicken), respectively. Hollow rectangles represent annotated genes in the genome, and other color rectangles, with gene name in red, represent novel coding genes in chickens.

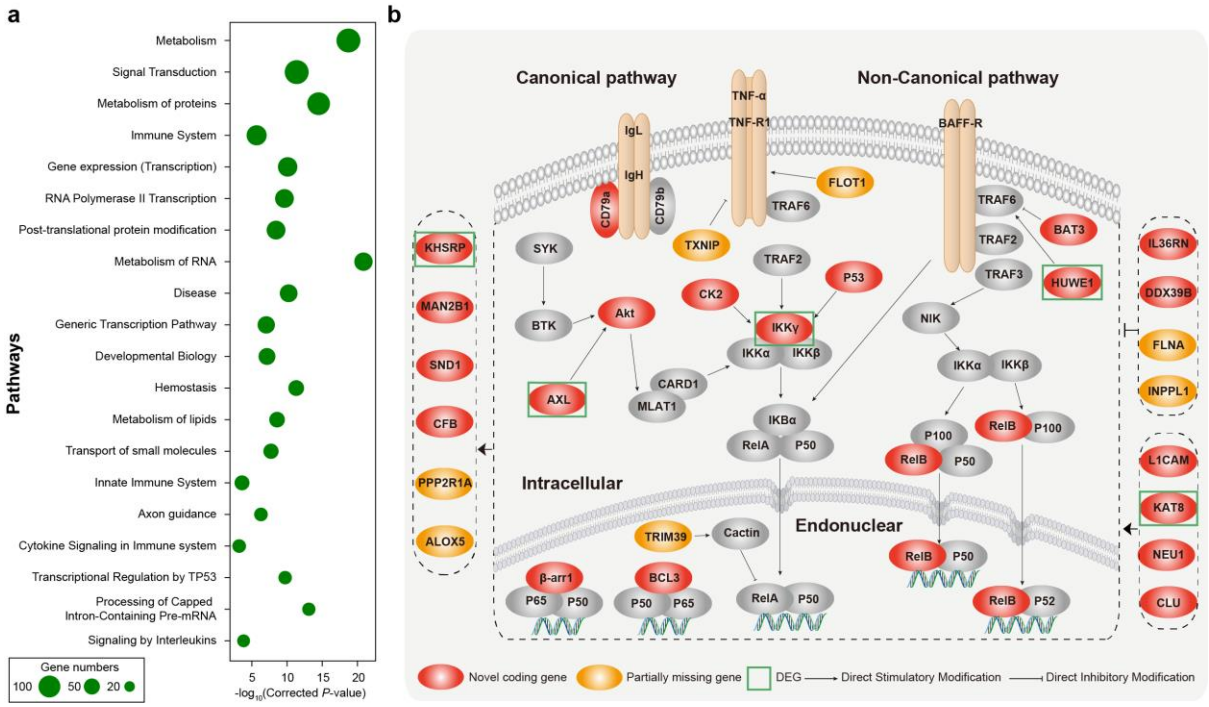


Figure 5

Function enrichment of novel coding genes and the case of NF-κB pathway related novel coding genes. **a**, The top 20 significant Reactome pathway with the largest number of novel coding genes. **b**, Novel coding genes (red) and partially missing gene (yellow) related to NF-κB signaling pathway. The green boxes represent differentially expressed genes (DEGs) in avian influenza virus.

1000 **Table:**

1001 **Table 1. The characteristics of novel sequences in this study.**

1002

Characteristic	
Total novel sequence length (bp)	158,981,245
Total gap length (bp)	1,405,623
Number of novel sequences	45,715
Novel sequence N50 (bp)	6,784
Mean novel sequence mean length	3,478
GC ratio	57.20%
G4 motif content	37.08%
Tandem repeat content	79.13%

1003