



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

## Three chromosome-level duck genome assemblies provide insights into genomic variation during domestication

### Citation for published version:

Zhu, F, Zhong-Tao, Y, Wang, Z, Smith, J, Zhang, F, Martin, FJ, Ogeh, D, Hincke, M, Lin, F-B, Burt, D, Zhou, Z-K, Hou, S-S, Zhao, Q-S, Li, X-Q & Ding, S-R 2021, 'Three chromosome-level duck genome assemblies provide insights into genomic variation during domestication', *Nature Communications*.  
<https://doi.org/10.1038/s41467-021-26272-1>

### Digital Object Identifier (DOI):

[10.1038/s41467-021-26272-1](https://doi.org/10.1038/s41467-021-26272-1)

### Link:

[Link to publication record in Edinburgh Research Explorer](#)

### Document Version:

Publisher's PDF, also known as Version of record

### Published In:

Nature Communications

### General rights









Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

### Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.



# Three chromosome-level duck genome assemblies provide insights into genomic variation during domestication

Feng Zhu<sup>1,10</sup>, Zhong-Tao Yin<sup>1,10</sup>, Zheng Wang<sup>1,10</sup>, Jacqueline Smith <sup>2,10</sup>, Fan Zhang<sup>1,10</sup>, Fergal Martin <sup>3</sup>, Denye Ogeh<sup>3</sup>, Maxwell Hincke <sup>4</sup>, Fang-Bing Lin<sup>1</sup>, David W. Burt <sup>2,5</sup>, Zheng-Kui Zhou <sup>6</sup>, Shui-Sheng Hou<sup>6</sup>, Qiang-Sen Zhao<sup>1</sup>, Xiao-Qin Li<sup>1</sup>, Si-Ran Ding<sup>1</sup>, Guan-Sheng Li<sup>1</sup>, Fang-Xi Yang<sup>7</sup>, Jing-Pin Hao<sup>7</sup>, Ziding Zhang <sup>8</sup>, Li-Zhi Lu<sup>9</sup>, Ning Yang <sup>1</sup> & Zhuo-Cheng Hou <sup>1</sup>✉

Domestic ducks are raised for meat, eggs and feather down, and almost all varieties are descended from the Mallard (*Anas platyrhynchos*). Here, we report chromosome-level high-quality genome assemblies for meat and laying duck breeds, and the Mallard. Our new genomic databases contain annotations for thousands of new protein-coding genes and recover a major percentage of the presumed “missing genes” in birds. We obtain the entire genomic sequences for the C-type lectin (CTL) family members that regulate eggshell biomineralization. Our population and comparative genomics analyses provide more than 36 million sequence variants between duck populations. Furthermore, a mutant cell line allows confirmation of the predicted anti-adipogenic function of NR2F2 in the duck, and uncovered mutations specific to Pekin duck that potentially affect adipose deposition. Our study provides insights into avian evolution and the genetics of oviparity, and will be a rich resource for the future genetic improvement of commercial traits in the duck.

<sup>1</sup>National Engineering Laboratory for Animal Breeding and Key Laboratory of Animal Genetics, Breeding and Reproduction, MARA; College of Animal Science and Technology, China Agricultural University, No. 2 Yuanmingyuan West Rd, Beijing 100193, China. <sup>2</sup>The Roslin Institute & R(D)SVS, University of Edinburgh, Easter Bush, Midlothian EH25 9RG, UK. <sup>3</sup>European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK. <sup>4</sup>Department of Cellular and Molecular Medicine, Department of Innovation in Medical Education, Faculty of Medicine, University of Ottawa, 451 Smyth Road, Ottawa K1H 8M5, Canada. <sup>5</sup>The University of Queensland, St. Lucia, QLD 4072, Australia. <sup>6</sup>Key Laboratory of Animal (Poultry) Genetics Breeding and Reproduction, Ministry of Agriculture and Rural Affairs; State Key Laboratory of Animal Nutrition, Institute of Animal Science, Chinese Academy of Agricultural Sciences, No. 2 Yuanmingyuan West Rd, Beijing 100193, China. <sup>7</sup>Beijing Golden-Star Inc., Beijing 100076, China. <sup>8</sup>State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China. <sup>9</sup>Institute of Animal Husbandry and Veterinary Science, Zhejiang Academy of Agricultural Sciences, Hangzhou 310021, China. <sup>10</sup>These authors contributed equally: Feng Zhu, Zhong-Tao Yin, Zheng Wang, Jacqueline Smith, Fan Zhang. ✉email: [zchou@cau.edu.cn](mailto:zchou@cau.edu.cn)

The duck (*Anas platyrhynchos*) is a major source of meat and eggs for human consumption and is also a significant source of feather down. Recent studies have shown that most domestic duck breeds originated from the Mallard about 2200–2500 years ago<sup>1,2</sup>. The Mallard has been confirmed as a major reservoir for avian influenza A viruses<sup>3–5</sup>, thus making it an important model for zoonotic disease studies. The domestic duck represents an excellent model for dissecting genetic mechanisms underlying domestication, due to its short generation interval, high reproduction ability, and extensive history of artificial selection<sup>1,2</sup>. Intensive artificial selection has resulted in very diverse phenotypes within and between domestic ducks compared with Mallard, due to diversification of body size, reproduction, and plumage color, which has generated two major breed types: laying and meat ducks<sup>2</sup>. The Pekin duck (a meat-type breed) is a world standard breed and is famous for its fast growth rate and superior adipose deposition; on the other hand, the Shaoxing (a laying-type breed) duck is recognized for its reproduction ability and has been widely used in laying duck breeding.

High-quality reference genomes of domesticated breeds and their wild relatives are critically important for understanding the genetic basis of phenotype differences. Alterations in genome size and organization, as well as large structural variations (SV), have been observed between the wild ancestor and domesticated descendants in chicken<sup>6,7</sup> and pig<sup>8</sup> breeds. The current duck reference genomes (BGI 1.0 and CAU1.0) are derived from the Pekin duck<sup>3</sup>, and still require further improvement in assembly quality metrics, for example with regard to fragment sizes and the number of gaps.

All birds, most reptiles, and formerly dinosaurs lay calcareous eggs, which is a successful reproductive adaptation to a desiccating terrestrial environment<sup>9,10</sup>. The avian egg represents the most advanced amniotic egg in oviparous vertebrates<sup>9</sup>. Understanding the genetics of eggshell biomineralization would be a crucial step in our understanding of the evolution of these unique eggshell features. However, there are still few well-assembled avian genomes, which hinders progress in understanding the relationship between genetics and evolution.

Here, we have produced very high-quality genome assemblies of Mallard, Pekin duck, and Shaoxing duck. These assemblies have uncovered genes previously presumed “missing” in birds. For example, we have obtained the entire genomic sequences for the C-type lectin (CTL) family members that regulate eggshell biomineralization. Finally, our work has identified 36.8 million whole-genome level variations (SNP, indels, structural and chromosomal) among wild and domesticated duck populations, and demonstrated the anti-adipogenic function of *NR2F2* and the changes responsible for differences in adipose deposition between meat-type breeds (such as the Pekin Duck) and its wild ancestor, the Mallard.

## Results

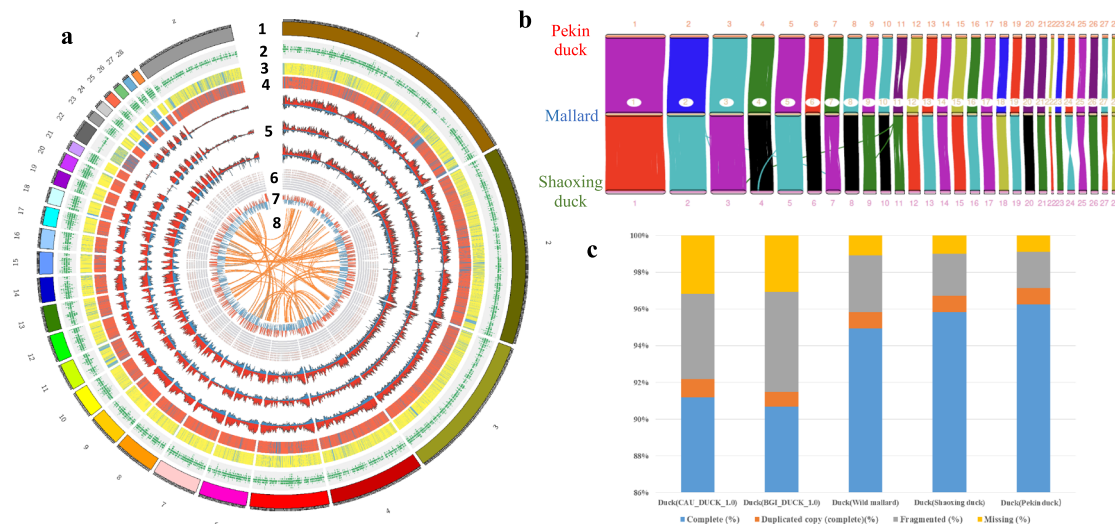
**Genome assembly and annotation.** Three chromosome-level assemblies for Mallard (CAU\_wild\_1.0), laying-type duck (Shaoxing duck, CAU\_Laying\_1.0), and meat-type duck (Pekin duck, CAU\_Pekin\_2.0) were built using a multi-level approach, including four different sequencing and assembly technologies (single-molecule real-time sequencing, PacBio SMRT; BioNano optical mapping; high-throughput chromosome conformation capture techniques, Hi-C; Illumina HiSeq) (Supplementary Fig. 1). A total of 112–114 Gb of PacBio long reads were collected for each duck genome, with approximately 100-fold high-quality sequence coverage for each assembly. A summary of statistics for these genome assemblies and total raw reads are shown in Supplementary Tables 1–6. The contig N50 lengths are 5.46 Mb (Pekin duck), 3.79 Mb (Shaoxing duck), and 4.68 Mb (Mallard). Final scaffold N50 lengths are 76.28 Mb (Pekin duck), 76.92 Mb (Shaoxing duck),

and 77.63 Mb (Mallard). The final assembled genome sizes are 1.19 Gb (Pekin duck), 1.21 Gb (Shaoxing duck), and 1.21 Gb (Mallard). After clustering of high-throughput chromosome conformation capture (Hi-C) data, the largest 40 super-scaffolds (CAU\_wild\_1.0: 41) appear to represent chromosomes consisting of 94.62% (Mallard), 95.52% (Shaoxing duck), and 95.26% (Pekin duck) of all sequences, respectively (Supplementary Figs. 2–6, Supplementary Tables 5 and 6). After genome annotation (see Methods), we obtained 18,490, 18,723, and 18,507 annotated protein-coding genes for Mallard, Shaoxing, and Pekin duck genomes, respectively (Supplementary Tables 7 and 8). We also predicted 1270 (Mallard), 1817 (Shaoxing duck), and 1654 (Pekin duck) noncoding RNA genes (Supplementary Tables 9–11). Moreover, we observed that repetitive element (TE) sequences make up ~17% of the total assembly of each duck genome, with long terminal repeat retrotransposons (LTR-RTs) being the most abundant (~13%) (Supplementary Table 12).

All genomes were evaluated by a number of methods to validate the quality of the assemblies (Supplementary Figs. 2–6). The assembly accuracy and completeness were supported by perfect matches with 221 radiation hybrid-map marker sequences<sup>1,11</sup> (Supplementary Data 1). The genomic collinearity for the three genomes shows that their assembly structure is consistent (Fig. 1a, b). The combination of a variety of sequencing technologies for these genomes significantly improved the assembly quality compared with those previously published. The contiguity of the newly assembled duck genomes (CAU\_wild\_1.0) is 7-fold greater than that of BGI\_duck\_1.0<sup>3</sup> and CAU\_duck\_1.0 (ASM874695v1) assemblies (Supplementary Table 13). The number of gaps in the genomes also decreased to less than 0.26% (391/148961, CAU\_wild\_1.0 vs. BGI\_duck\_1.0) (Supplementary Table 13). Further improvements in genome completeness and accurate assembly of highly complex regions are also seen. The conserved genomic elements in the genome are relatively complete, based on the BUSCO scores (>95%), while the BUSCO score for the BGI\_duck\_1.0 assembly is less than 91% (Fig. 1c). The gene set annotated 12,061 more transcripts and 1131 more protein-coding genes than the BGI\_duck\_1.0 reference gene set (Supplementary Table 14). Among these newly annotated protein-coding genes, we found many functionally important genes as the high assembly quality and full-length transcriptome data were integrated into the annotation pipeline (Supplementary Table 14). Moreover, the distribution of identified TEs in the genome assemblies (average of three assemblies: 16.8%) is much more abundant than those in the previous BGI\_duck\_1.0 assembly (11.5%)<sup>3</sup>. Taken together, these results indicate that the genomes are a marked improvement in contiguity and completeness compared to the previously published duck reference genomes.

**Recovery of “missing genes” from the newly assembled genomes.** The high quality of all three genome assemblies allowed us to address some classic questions in the field of evolution. Previous studies have hypothesized that the avian genome lacks some important functional genes when compared with mammals and amphibians, possibly owing to its adaptation to flight<sup>12,13</sup>. However, recent studies have found evidence of many of these putative “missing genes” through a more detailed genome-wide analysis<sup>12,14,15</sup>.

In this study, we re-checked a group of 571 genes previously thought to be missing from all avian genomes<sup>12,13</sup> (Supplementary Fig. 7). We were able to identify 89 of the 571 “missing” genes, with their complete gene structure in the Mallard genome (Supplementary Table 14). It is worth noting that 5 genes were annotated as pseudogenes and 3 as lncRNA (Supplementary Table 15, Supplementary Fig. 8). In addition, 240 (42.11%) were annotated as paralogous genes (Supplementary Data 2), while 108



**Fig. 1 Overview of the assembly quality and characteristics of the duck genome.** **a** Chromosomal features of three duck genomes with the integration of genetics (from Chr1 to Chr25). 1: Chromosomal length of Mallard genome (Mb); 2: Gene density (100 kb window); 3: Genome collinearity of Shaoxing duck to Mallard, yellow represents the same orientation, blue represents contrary; 4: Genome collinearity of Pekin duck to Mallard, red represents the same orientation, blue represents contrary; 5: the density of SNP and Indels for Mallard, Shaoxing duck and Pekin duck in the reference Mallard genome (100 kb window). Red represents SNP, and blue represents InDels; 6: The distribution of ATAC-seq windows (100 kb window) in fat tissue; 7: The A/B compartments in Mallard genome. 8: The inner lines show syntenic blocks within the Mallard genome. **b** The genome collinearity of the genes among the three assemblies. **c** The 2,586 highly conserved genes in BUSCO dataset were used to search Mallard, Shaoxing duck and Pekin duck genomes. This analysis was carried out with the BUSCO program (version 2) with default settings. BGI\_duck\_1.0 and CAU\_duck\_1.0 are genome assemblies of Pekin duck downloaded from Genbank accessions GCA\_002743455.1 and GCA\_000355885.1, respectively.

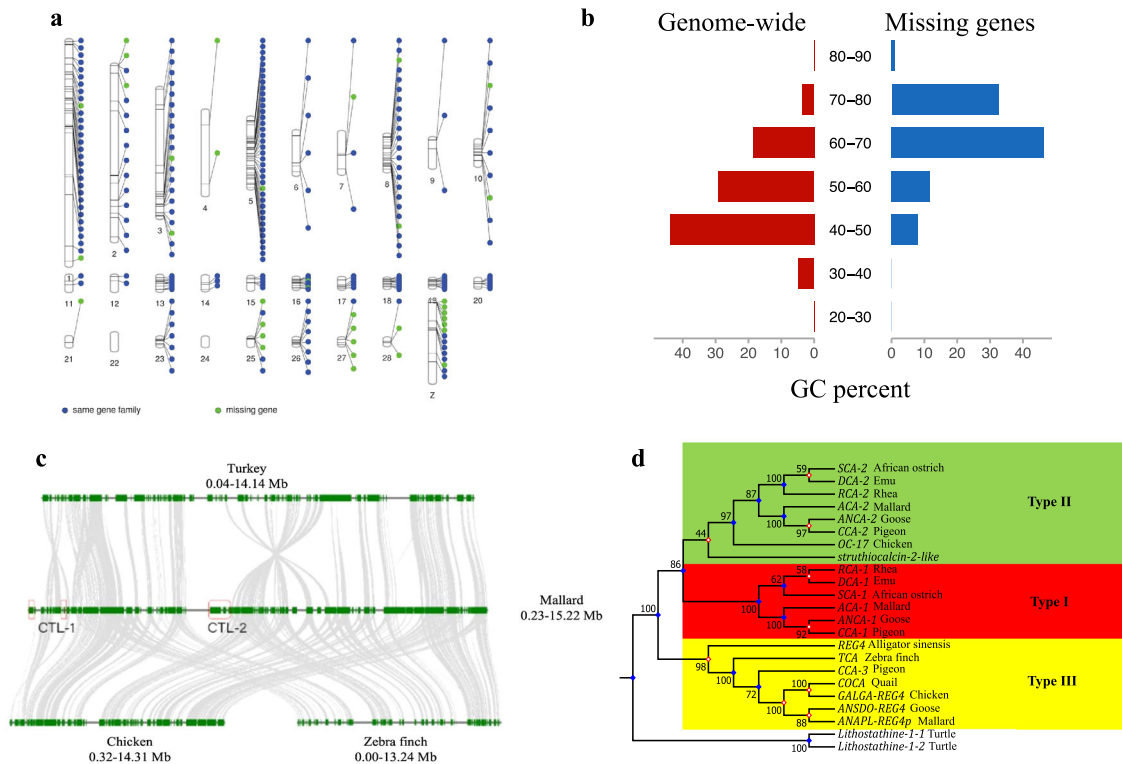
(18.95%) genes could be found in similar sequences in the genome but are not well annotated (Supplementary Data 3). However, 133 (23.33%) genes were still missing in the newly assembled genome with no significant sequence homologies.

According to the distribution of the annotated “missing” genes in the Mallard genome, 66.7% of them were located on the repeat-rich/GC-rich microchromosomes (Fig. 2a, Supplementary Fig. 9). To understand more about the characteristics of these “missing genes”, we compared the GC content of 89 human-Mallard orthologs. The GC content of most (88.76%) genes is higher in Mallard than in humans, especially in the gene encoding *C2orf68*, which has a GC content of 80.59% in Mallard and only 47.29% in humans (Supplementary Fig. 10, Supplementary Table 16). Compared with other protein-coding genes in the Mallard genome, the newly discovered “missing” genes also show the characteristic of high GC content (Fig. 2b). Through quantitative analysis of gene expression, we observed that the “missing” genes have a strong tissue-specific expression in Mallard (Supplementary Fig. 11). We integrated the newly identified missing genes into the Mallard genome and previously recovered missing genes from the transcriptome assembly<sup>15</sup>; only 10 of all missing genes remain to be found in birds, in either the genome or transcriptome assembly (Supplementary Data 2). Therefore, our results and previous studies do not support the missing gene hypothesis in birds. We believe that the main reason that these genes have been classified as “missing” is due to either high GC content or complex genomic structure (e.g., duplications and repeat-rich regions), which has led to misassembly and/or an inability to sequence and assemble<sup>15</sup>.

**Identifying the avian eggshell specific CTL gene family critical for avian eggshell biomineralization.** In addition to identifying “missing genes”, this genome assembly also harbors some critical genes which are unique to birds. Ovocleidin-17 (*OC-17*), a CTL family member, is considered to be a major protein involved in

eggshell calcitic mineralization in the chicken<sup>16</sup>. *OC-17* was the first eggshell-specific protein to be purified and to have its amino acid sequence and protein crystal structure determined<sup>17,18</sup>. However, in spite of the intensive effort, it has been difficult to determine the full-length cDNA and genomic DNA sequences for *OC-17* or any of its orthologs in other avian species. At present, these CTL family members are not annotated in any of the bird reference genomes. Recently, we obtained the full-length cDNA sequence of chicken *OC-17*, using transcriptome assembly and RACE<sup>19</sup>, and determined the expression pattern for *OC-17* in various tissues. However, its genomic sequences and the chromosomal location remains unknown in the chicken or any other avian genome.

In this study, we identified the complete gene structure of two paralogous *OC-17*-like genes (anascalcin-1, *ACA-1*; anascalcin-2, *ACA-2*); these are close together in the same chromosomal region in each of the three duck genome assemblies (Fig. 2c). In order to validate the CTL gene family annotation results in duck and to obtain CTL cDNA sequences in other avian species, we explored de novo assembled transcriptomic data from multiple tissues in five bird species (chicken, duck, pigeon, zebra finch, and goose), and performed BLAST, polymerase chain reaction (PCR), and RACE to verify the newly identified *OC-17*-like cDNAs. We successfully obtained 2-3 similar cDNA sequences from a pigeon (columbacalcin: *CCA-1*, -2, -3), duck (*ACA-1*, -2; ANAPL-REG4p), and goose (Ansercalcic: *ANCA-1*, -2; ANSDO-REG4), with predicted protein sequences that displayed high similarity to the *OC-17* orthologs (Supplementary Fig. 12). According to their sequence similarities, bird *OC-17*-like genes can be classified as Type-I, II, or III (Fig. 2d). The duck proteins (*ACA-1*, *ACA-2*) are orthologs that align with type-I and type-II sequences. There are five exons in each of the *ACA-1* and *ACA-2* genes; the length of their CDS region is 462 and 498 bp, and they encode 154 and 166 amino acids, respectively. The high GC content of the genomic sequences (*ACA-1*, 63.90%; *ACA-2*, 69.24%) and their duplication are probably responsible for the difficulty in obtaining the



**Fig. 2** Characteristics of the CTL gene family and missing genes newly annotated in the Mallard genome. **a** The distribution of the presumed “missing genes” in the Mallard genome; **b** GC content of “missing genes” in Mallard. The figure shows that GC content distribution of missing genes in mallard was significantly higher than the genome background. **c** Multiple alignments of gene annotations in CTL gene family regions of Mallard, chicken, zebra finch, and turkey. In other bird chromosomes, the location and detailed annotations of CTL genes are lacking. **d** Phylogenetic tree of CTL members and its homologs in other birds. There are three types of CTL genes in birds, type I (green), type II (red), and type III (yellow). Ducks contain only type I and type II.

genomic and cDNA sequences associated with these genes. The amino acid sequence similarity between the two types of duck CTL is 56%, and their molecular weights are ~16kD. Compared with eggshell CTL protein sequences of other birds, there are 21 completely conserved amino acid sites in 16 CTL sequences, with 57 semi-conserved amino acid sites and similar secondary structures (Supplementary Figs. 12 and 13). Phylogenetic analysis indicates that the CTL family members have been duplicated multiple times during speciation (Supplementary Fig. 14). Through transcriptomic analysis, we found that expression of CTL protein-coding genes of Type-I and Type-II in Pekin duck and pigeon was highly specific to the uterine segment of the oviduct, the site of eggshell mineralization, whereas expression was negligible in the magnum (responsible for the secretion of egg white), ovary (producing the egg yolk) and other tissues (Supplementary Figs. 15 and 16). In contrast, the Type-III form in pigeons (*CCA-3*) displayed a widespread tissue-specific gene expression pattern (Supplementary Fig. 16). In the zebra finch, where our identified CTL member (Taeniocalcin, TCA) is a Type-III CTL, we observed that it is strongly expressed throughout the GI tract and at very low levels in the uterus (Supplementary Fig. 17). Expression of the Type-I and -II forms is highly specific to eggshell calcification; in the uterus where expression of the Type-I and Type-II genes is high, there is low expression of the Type-III, and vice-versa (Supplementary Fig. 15–17). We investigated the uterine expression patterns of *ACA-1* and *ACA-2* during egg formation in Pekin ducks. Tissues were collected 5–8 h before egg-laying (egg in the uterus during active eggshell calcification) and 5–9 h after egg laying (new forming egg just entering the uterus before the commencement of active calcification) (Supplementary Fig. 18). There was a stage-dependent difference in expression of each duck CTL gene, with

higher expression of both *ACA-1* and *ACA-2* before egg-laying. In the uterus tissues from goose, Pekin duck, and pigeon, the expression levels before ovulation of the Type-I and Type-II CTLs are positively correlated with eggshell quality parameters such as weight (ESW), breaking strength (ESS), and thickness (EST) (Supplementary Fig. 19).

Based on genomic sequences of eggshell-specific CTL genes, we anticipate that researchers can utilize gene markers for genomic selection, conduct genome editing in cell lines or in vivo to determine the functions of CTL gene members, and gain insight into functional consequences of gene duplication during bird evolution. In summary, the duck genome assemblies from this work integrate full annotation for the eggshell CTL family genes, and we have also identified 9 cDNA sequences of CTL family members in five bird species. Moreover, this is the first instance of identifying the type-II CTL gene in the genomes of neognathae birds, since previously the simultaneous presence of Type-I and Type-II proteins have only been observed in eggshells from ratite species<sup>20</sup>.

#### Genomic variations among Mallard and domesticated ducks.

The development of high-quality duck genomes allowed us to identify large structural variants by direct comparative analysis of the three genomes. Genome comparison analysis allowed systematic characterization of presence/absence variations between the two domesticated ducks and Mallard (Supplementary Fig. 20, Supplementary Table 17, Supplementary Data 4–6). We found 350 genes in Shaoxing duck and 551 genes in Pekin duck that are located in or near these presence/absence variations regions. More than 98% of the presence/absence variations are located in the intergenic regions of these genes, with only a few of them

(<2%, 12 genes for Mallard, 20 genes for Pekin duck, and 26 genes for Shaoxing duck) in exonic regions (Supplementary Data 5). Through gene enrichment analysis, we found that most of the presence/absence variations in Mallard are related to morphological development (Supplementary Data 6). In the Pekin duck, 29 of these genes are related to muscle structure development (Supplementary Data 6). By integrating our comprehensive temporal transcriptomic data from multiple-tissues/different time-points in ducks, we observed that the expression level of some genes (*DCN*, *SGCZ*, and *SGCG*) within presence/absence variations in Pekin duck are always higher than their homologs in Mallard at corresponding stages of pectoral muscle development (Supplementary Data 7). This suggests that some presence/absence variations may result in faster muscle development in the Pekin duck compared to the Mallard.

In addition, we identified genome sequence inversions between Mallard and Shaoxing duck (1.87 Mb), and between Mallard and Pekin duck (1.43 Mb) (Supplementary Data 8). We also detected 3820 translocations (1074 intra-chromosome translocations occupying 3.8 Mb, and 2746 inter-chromosome translocations occupying 6.8 Mb) (Supplementary Data 8). In total, 59% of translocations were located in the intergenic region. Our results showed that there are no large translocations and rearrangements between domestic ducks and Mallard, especially between Pekin duck and Mallard (average length 2 kb). To assess the detected SV, we verified each breakpoint by comparing Hi-C data between domestic ducks and Mallard. As a result, one significant translocation was identified at 0.72–0.81 Mb in chromosome 7 between Pekin duck and Mallard (Supplementary Fig. 21), which contains genes ENSAPLG00020017214 and ENSAPLG00020017216. The expression of gene ENSAPLG00020017214, which belongs to the *PWWP2B* protein family, in fat and muscle tissue of Pekin duck was significantly higher than in Mallard (Supplementary Data 7).

Resequencing data for 119 ducks (43 Mallards, 48 Pekin ducks, and 28 Shaoxing ducks) identified many single nucleotide polymorphisms (SNPs) within each variety: 9,232,834 (77 SNPs/kb), 12,845,466 (107 SNPs/kb), and 14,748,593 (122 SNPs/kb) in Pekin duck, Shaoxing duck, and Mallard, respectively. The total SNPs and SNP frequency in the Pekin duck genome are slightly less than that found in Shaoxing duck. We also identified small InDels: 2,711,487 (Pekin duck, 2 InDels/kb) and 4,006,473 (Shaoxing duck, 3.3 InDels/kb). Among these variants, a considerable number of SNPs were found with significantly different frequencies between domesticated ducks and Mallard. There are 311,976 SNPs that are almost fixed in Pekin duck, and 88,961 SNPs in Shaoxing duck (frequency > 0.7 in Pekin duck/Shaoxing duck, frequency < 0.3 in the Mallard; or vice versa, see “Methods”). Among these nearly fixed/lost SNPs in domestic ducks, 51,014 SNPs are located in the promoter/upstream region, while 53,598 SNPs have potential high/moderate-mutation effects as predicted by VEP (Supplementary Data 9).

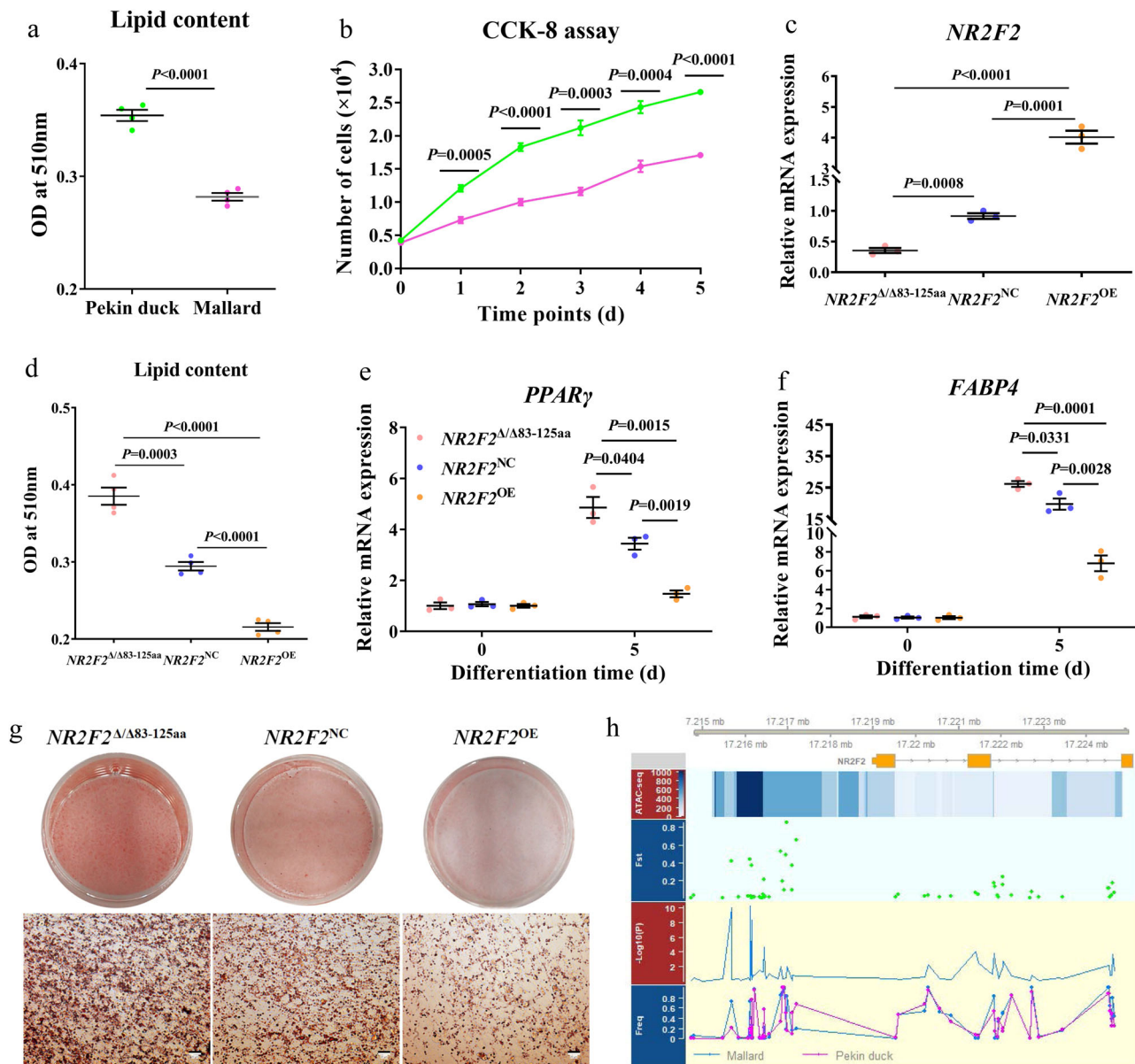
**Regulatory mutations near the *NR2F2* gene affect adipose deposition.** Pekin ducks possess a subcutaneous fat weight/percentage, which is greatly elevated compared to Mallard reared in the same environment, due to intensive artificial selection during the domestication process<sup>21,22</sup>. Fat synthesis and deposition are coordinated processes that take place in the liver and fat tissue, respectively<sup>23</sup>. The growing Pekin duck has a higher liver weight (2-times heavier at 4-weeks of age) and a greater hepatic triglyceride content compared to the Mallard at 2, 4, and 6 weeks of age (Supplementary Fig. 22a, b). Similarly, the triglyceride content of subcutaneous fat tissue in Pekin duck is higher than that in Mallard, especially at 6 weeks (Supplementary Fig. 22c). The expansion of adipose deposits can be driven either by an increase in adipocyte size or by the formation of new adipocytes via precursor

differentiation during adipogenesis<sup>24</sup>. Hence, we tested the difference in proliferation and differentiation potential for subcutaneous preadipocytes between Mallard and Pekin duck. The results showed that higher proliferation and differentiation potential are both key factors that are responsible for excessive subcutaneous fat deposition in Pekin duck (Fig. 3a, b, Supplementary Fig. 22d).

A wide array of studies have revealed the importance of transcription factors in establishing the mature fat-cell phenotype both in vitro and in vivo<sup>24</sup>. In order to systematically identify expression changes in transcription factors of differentiated preadipocytes cultured in vitro in the Pekin duck and Mallard, we performed mRNA-seq analysis on differentiated preadipocytes at multiple time points during adipocyte differentiation (0, 12, 24, 48, and 72 h after induction) (Supplementary Data 7). We found significant differences in the expression of most transcription factors (32/38) known to be associated with adipogenesis (Supplementary Fig. 23). Among these differentially expressed genes (DEGs), most of the transcription factors (19/24) which promote adipocyte differentiation were upregulated in Pekin duck, whereas those (14/14) that inhibit adipocyte differentiation were upregulated in Mallard. Collectively, our results indicate that differences in the expression of transcription factors play a key role in the regulation of adipocyte differentiation. We investigated this further by mRNA-Seq analysis on liver and subcutaneous fat tissue from Mallard and Pekin duck at 2, 4, and 6 weeks of age, and focused on transcription factors that are consistently differentially expressed during fat development. The transcription factors *NR3C4*, *FOXO4*, *GLIS2*, *HOXD10A*, and *RXRG*, were consistently upregulated in Pekin duck, whereas *CDX1*, *FOXO1*, *HNF1B*, *NR2F2*, *TFDP1*, and *ZBTB7A* were consistently upregulated in Mallard. Notably, *NR2F2* expression in Mallard subcutaneous fat tissue, adipocytes and liver is significantly higher than in those of Pekin duck (Supplementary Fig. 24a–c), which is not seen with other transcription factors, indicating that *NR2F2* had a potential effect on subcutaneous fat deposition and metabolism during the domestication process. We investigated the fixation index ( $F_{ST}$ ) in the region around *NR2F2*, using a sliding window approach with a step of 5 kb around *NR2F2* in Pekin duck and Mallard. We found that the region located upstream of the *NR2F2* gene (Chr 12: 17.213–17.218 Mb) showed near fixation of 5 selected loci ( $F_{ST} = 0.49–0.85$ ) from Mallard to Pekin duck (Supplementary Fig. 24d; Supplementary Data 9).

*NR2F2* was initially identified as an activator of the chicken ovalbumin gene<sup>25</sup>, but follow-up studies uncovered functions in adipocyte differentiation in the mouse<sup>26</sup> and in metabolic gene regulation<sup>27</sup>; however, little is known concerning its role in birds. Therefore, we sought to investigate this further by studying the effect of mutations in *NR2F2* in immortalized chicken preadipocytes using CRISPR–Cas9. We obtained a mutant clone with a deletion of 129 bp in the first exon in both *NR2F2* alleles, located between 247 and 375 bp in the CDS region, resulting in a truncated *NR2F2* protein missing 43 amino acids which included the DNA binding domain mutation, and resulted in an mRNA expression level decreased by 60% compared with the control group (Fig. 3c, Supplementary Fig. 24e). Strikingly, cells with the *NR2F2* mutation demonstrated enhanced adipogenic potential (Fig. 3d–g), as shown by increased lipid accumulation and elevated expression of adipocyte markers. On the other hand, overexpression of *NR2F2* prevented adipogenesis, further confirming an anti-adipogenic effect.

To investigate a potential regulatory role for the region defined by the five selected loci detected by  $F_{ST}$  analysis, we performed ATAC-seq analyses to map accessible chromatin regions within this segment. We found that multiple binding sites for regulatory elements were identified in the region encompassing these five selected loci (Chr 12: 17.213–17.218 Mb) (Fig. 3h, Supplementary



**Fig. 3 Comparison of differentiation and proliferation capacity in subcutaneous preadipocytes between Pekin duck and Mallard.** **a** Intracellular lipid content in subcutaneous preadipocytes of Pekin duck and Mallard at day 5 post induction. The oil red O extraction assay was used to measure lipid accumulation. Green: Pekin duck and pink: mallard ( $n = 4$  biological replicates). **b** Cell counting kit-8 assay (CCK8) examines the proliferation of subcutaneous preadipocytes in Pekin duck and Mallard over 5 days. Each cell number is counted by the standard curve established by CCK8 of the respective cells ( $n = 4$  biological replicates). **c** mRNA levels of *NR2F2* were analyzed by Q-PCR in *NR2F2<sup>NC</sup>*, *NR2F2<sup>OE</sup>*, and *NR2F2 $\Delta/\Delta 83-125aa$*  cells. NC negative control, OE overexpression,  $\Delta$  deleted ( $n = 3$  biological replicates). **d** Intracellular lipid content in preadipocytes at day 5 post induction. The oil red O extraction assay was used to measure the lipid accumulation ( $n = 4$  biological replicates). **e, f** mRNA levels of *PPAR $\gamma$*  and *FABP4* were analyzed by Q-PCR at day 0 and 5 post induction ( $n = 3$  biological replicates). **g** Oil Red O staining to assess lipid accumulation at day 5 post induction for *NR2F2<sup>NC</sup>*, *NR2F2<sup>OE</sup>*, and *NR2F2 $\Delta/\Delta 83-125aa$*  cells. The scale bar represents 20  $\mu$ m ( $n = 4$  biological replicates). **h** The distribution of SNPs with a different frequency in *NR2F2* for Pekin duck and Mallard populations. The track below the transcript annotation represents the windows of ATAC-seq. The color depth represents the peak score size. The following tracks are shown separately: Fixation index,  $-\log_{10}$  ( $p$ -value of likelihood ratio test), and allele frequency. Data are presented as mean  $\pm$  SEM. Statistical significance using two-tailed unpaired Students *t*-test for (**a–f**).

Table 18), indicating that this region may play an important role in the transcriptional regulation of the *NR2F2* gene and that *NR2F2* expression changes may be caused by variations in this upstream region.

However, functional verification of this potential causal mechanism is still required. Overall, our results suggest *NR2F2* as an important candidate gene that is responsible for differences in subcutaneous fat deposition between Pekin duck and Mallard.

## Discussion

In this study, we produced three high-quality genomes, for Mallard and two different breeds of domesticated ducks. Our breed-specific reference genomes provide an invaluable resource describing SNPs, indels, and structural variants, which can assist in designing specific genomic selection programs. We identified more than 36 million SNPs, indels, and structural variations that were associated with the domestication process, and these

variations are now resources for illuminating the genetic differences between Mallard and domestic ducks. From a phenotype perspective, domestication has significantly improved the growth rate, reproduction ability, and fat deposition in domestic lines of ducks compared with Mallard. We do find that a large number of genes related to economic traits have undergone significant changes in domestic ducks. Among mutations in gene body regions that are predicted to have potentially high/moderate-functional effects, some mutations are known to cause specific phenotypes, i.e., an insertion in an intron region in the *MITF* gene causes the white feather phenotype in Pekin ducks<sup>1</sup>. The gene showing the most differences is *ZNF469* which has 606 SNPs between Pekin duck and Mallard, while *ZNF469* also expressed significant differences between Pekin duck and Mallard, as well as Shaoxin duck and Mallard<sup>22</sup>. However, we also observed that a high percentage (98.6%, 26,128/26,494) of significantly associated SNPs were located in regulatory or noncoding regions and that the mutation rate in these regulatory regions was much higher than in coding locations (Supplementary Fig. 25). Regulatory changes of this nature can be responsible for phenotypic variation, as demonstrated in a variety of organisms, including humans and chimpanzee<sup>28,29</sup>. However, as relevant duck cell lines are not currently available, we were not able to perform genomic-editing studies to confirm these genetic effects *in vitro*.

Our study demonstrates that great improvements in reference genome assembly can provide new material for the study of bird evolution. Much avian genome data have been released in recent years<sup>30</sup>, which allowed us to discover hundreds of “missing” genes thought to be absent from avian genomes. In our latest genome assembly, we located major missing genes on complex chromosomal regions or on micro-chromosomes, as was recently successful in the chicken assembly<sup>31</sup>. We have resolved the full-length genomic sequences of the eggshell CTL gene family members, which are hypothesized to be critical proteins for calcitic biomineralization in reptilia<sup>16,17</sup>. We also obtained the complete cDNA sequences for CTL protein members in several clades of birds. These newly discovered CTL family members expand our knowledge of eggshell biomineralization. Our preliminary tissue expression data from several species suggested potential diverse functions of different CTL isoforms in avian eggshell biomineralization, which needs to be further explored. These data should facilitate a deeper understanding of the genetic basis of calcitic biomineralization and will be a step forward in illuminating the molecular mechanisms responsible for oviparity.

## Methods

**Ethics.** All experiments with birds were performed under the guidance of ethical regulations from the Animal Care and Use Committee of China Agricultural University, Beijing, China (permit number: SYXK 2007-0023).

### Sample collection, library preparation, and sequencing

**Animals for genome assembly and population genome resequencing.** We randomly selected adult female Pekin ducks, Mallard, and Shaoxing ducks from flocks that were raised under standard feeding regimes for these studies. Genome assembly: fresh blood was used for PacBio sequencing, and breast muscle tissue for Hi-C and Bionano sequencing. Forty-eight (male: 23, female: 25) 42-day-old Pekin ducks were fed with the same diet and maintained under the same lighting conditions as Golden Star Duck Co (Beijing, China), and fresh blood was collected for re-sequencing. The re-sequencing data for 43 Mallard birds were from previously published data<sup>1</sup>. Blood samples from the wing vein of 23 Shaoxing adult female ducks were collected in Shaoxing, China, and stored at  $-20^{\circ}\text{C}$  before DNA extraction. Information relating to sequencing samples is shown in Supplementary Data 10.

**Library preparation and genomic sequencing for genome assembly and population genomics.** Genomic DNA was extracted from the blood samples. At least 5  $\mu\text{g}$  DNA was used for library construction using the Illumina TruSeq DNA Sample Prep Kit (Illumina, CA, USA). DNA was isolated using the DNeasy Blood & Tissue Kit (QIAGEN, ON, Canada). The purified genomic DNA was mechanically disrupted using Bioruptor (Diagenode Inc., NJ, USA) to generate  $\sim 300$  bp inserts. The DNA

fragments were subjected to end repair and A “addition to the 3” end, followed by amplification using the Thermal cycler S1000 (Bio-Rad). The purified library was subjected to quality control using StepOne Plus (Applied Biosystems, MA, USA). Finally, the Nova-seq6000 platform (Illumina, CA, USA) was used to generate paired-end sequencing data with a genome coverage of at least  $10\times$  (Supplementary Data 10).

**Library construction and PacBio sequencing.** To construct sequencing libraries for PacBio sequencing, genomic DNA was fragmented by g-TUBE centrifuged at 2000 r.p.m. for 2 min, treated with end-repair, adapter ligation and exonuclease digestion as recommended by Pacific Biosciences. DNA fragments of about 10–50 kb were selected by BluePippin electrophoresis (Sage Sciences). DNA libraries were sequenced on the PacBio Sequel platform (Pacific Biosciences).

**Data collection for Bionano mapping.** High-molecular-weight DNA was isolated and labeled according to standard BioNano protocols with the single-stranded nicking endonuclease Nt. BssSI<sup>1,32</sup>. The labeled DNA sample was loaded onto the IrysChip nanochannel array. The stretched DNA molecules were imaged with the BioNano Irys system. Raw image data were converted into bnx files; from these, the AutoDetect software (v2.1.4) generated basic labeling and DNA length information.

**Hi-C library construction and data collection.** We created three Hi-C libraries for Mallard, Shaoxing duck, and Pekin duck using the methods described in a similar study<sup>1</sup>. Libraries were subjected to sequencing on the Illumina HiSeq 2000 platform. Information relating to the raw data is given in Supplementary Data 10.

**RNA-Seq samples, RNA-extraction, library construction, and sequencing.** Uterine tissues were collected from three adult Pekin female ducks at two physiological states, before or after egg-laying, in order to analyze the expression changes in CTL genes during the daily egg formation cycle. Birds were sacrificed either 5–8 h before oviposition or 5–9 h after oviposition. Multi-tissue samples from chicken, duck, pigeon, goose, and zebra finch were collected for transcript de novo assembly, as previously described<sup>18</sup>. We collected liver and subcutaneous adipose tissue from Pekin duck and Mallard at 2, 4, and 6 weeks of age. Six biological replicates were collected at each time point for each group. Tissue samples were snap-frozen in liquid nitrogen and then stored at  $-80^{\circ}\text{C}$  until RNA extraction.

RNA from each tissue was extracted individually (10  $\mu\text{g}$  per tissue) using Trizol reagent (Invitrogen, CA, USA) according to the manufacturer’s instructions. The Agilent Bioanalyzer 2100 instrument (Agilent, CA, USA) was used to verify the integrity of the RNA. Approximately, 10  $\mu\text{g}$  of sheared cDNA was prepared for Illumina sequencing according to the manufacturer’s protocols. All samples were sequenced on the Illumina Nova-seq6000 system (Illumina, CA, USA) with 150 bp paired ends. Pekin duck full-length transcriptome data were obtained from our previously published dataset<sup>33</sup> (PRJNA526109).

**Genome assembly.** Falcon<sup>34</sup> was used for constructing initial contigs using the following parameters: `length_cutoff = 13,000` `length_cutoff_pr = 14,000` `pa_DBSplit_option = -x1000 -s250 -a pa_HPCdaligner_option = -v -dal128 -t12 -e.75 -k20 -h320 -l1800 -s1000` `falcon_sense_option = --output_multi --min_idt 0.75 --min_cov 2 --local_match_count_threshold 2 --max_n_read 400 --output_dformat ovlp_DBSplit_option = -x1000 -s200 ovlp_HPCdaligner_option = -v -dal100 -t12 -k18 -h280 -e.96 -l1800 -s1000` `overlap_filtering_setting = --max_diff 50 --max_cov 80 --min_cov 2 --bestn 10`. The initial polishing was performed with Quiver<sup>35</sup> using PacBio-only long reads, and then Pilon<sup>36</sup> (v1.20) was utilized to further correct the PacBio-corrected contigs with accurate Illumina short reads. The BioNano data was first assembled into a consensus map using the IrysView software (with default parameters) with a molecular length threshold of 150 kb and a minimum labels per molecule of 8. Hybrid scaffolding of the PacBio-corrected contigs and the BioNano-based consensus map was performed using the hybrid scaffolding module within IrysView software with the manufacturer’s suggested parameters. After scaffolding, PBJelly from PBSuite<sup>37</sup> (v14.9.9) was performed to close gaps in the hybrid assembly. We re-performed error correction procedures to polish the sequences in the gap regions. Subsequently, the mitochondrial scaffolds or contigs were removed through alignment to mitochondrial references (BGI\_duck\_1.0); any scaffolds or contigs for which at least 80% of the total length was aligned and that showed an identity larger than 90% were discarded as mitochondrial sequences. The Hi-C sequencing data were first aligned to the assembled contigs/scaffolds using the Bowtie2<sup>38</sup> end-to-end algorithm, and then the assembled scaffolds were clustered, ordered, and directed into pseudochromosomes using Lachesis<sup>39</sup>. Finally, the pseudo-chromosomes predicted by Lachesis<sup>39</sup> were cut into bins with an equal length of 100 kb and used to construct a heatmap based on the interaction signals generated by valid mapped read pairs to perform validation and manual correction. The final draft was corrected using GapCloser<sup>40</sup> with default parameters. The 2586 conserved protein models in the BUSCO vertebrata\_odb9 dataset were searched against all assembled genomes by using the BUSCO (version 2) program with default settings<sup>41</sup>.

**Genome annotation.** Annotation of the Mallard assembly (ASM874695v1) was created via the Ensembl gene annotation system. A set of potential transcripts was generated using two major techniques: primarily through alignment of short-read



RNA-seq data and also through gap filling with protein-to-genome alignments of a subset of vertebrate proteins with experimental evidence from UniProt<sup>42</sup> vertebrate proteins. The short-read RNA-seq data was sourced from various samples generated as part of these projects (PRJNA194464 and PRJNA273367)<sup>3,15</sup>. The UniProt vertebrate proteins had experimental evidence for existence at the protein or transcript level. At each locus, low-quality transcript models were removed, and the data were collapsed and consolidated into a final gene model plus its associated non-redundant transcript set. When collapsing the data, priority was given to models derived from transcriptomic data. For each putative transcript, the coverage of the longest open reading frame was assessed in relation to known vertebrate proteins, to help differentiate between true isoforms and fragments. In loci where the RNA-seq data were fragmented or missing, homology data took precedence, with preference given to longer transcripts that had strong intron support from the short-read data. Gene models from the above process were classified into three main types: protein-coding, pseudogene, and long noncoding RNA. Models with hits to known proteins and few structural abnormalities (i.e., canonical splice sites, introns exceeding a minimum size threshold, low level of repeat coverage) were classified as protein-coding. Models with hits to known protein, but having multiple issues in their underlying structure were classified as pseudogenes. Single-exon models, with a corresponding multi-exon copy elsewhere in the genome, were classified as processed pseudogenes. If a model failed to meet the criteria of any of the previously described categories, did not overlap a protein-coding gene, and had been constructed from transcriptomic data, then it was considered as a potential lncRNA. Potential lncRNAs were filtered to remove transcripts that did not have at least two valid splice sites or cover 1000 bp (to remove transcriptional noise). Using LastZ<sup>43</sup>, we generated a whole-genome alignment against the human assembly GRCh38.p12. For each protein-coding gene in the human (Ensembl/Gencode gene set), we projected the coding exons within the canonical transcript to the Mallard. In the case of an exonic overlap on the projected sequence, the longest exon took precedence. If the mapping did not succeed, we selected the next successful projection of the transcript having the longest translation. The annotation of small noncoding genes, particularly miRNAs, were annotated via a BLAST<sup>44</sup> of miRbase<sup>45</sup> against the genome, before passing the results into RNAfold<sup>46</sup>. Poor quality and repeat-ridden alignments were discarded. Other types of small non-coding genes were annotated by scanning Rfam<sup>47</sup> against the genome and passing the results into Infernal<sup>48</sup>.

The genome assemblies of Pekin duck, (JACEUL000000000) and Shaoxing duck (JACEUM000000000) were annotated using the EvidenceModeler (EVM) pipeline<sup>49</sup>. In the homology annotation, Genewise (2.4.1)<sup>49</sup> was used to map the gene sets of Turkey, Chicken, Ostrich, and Zebra finch to the assembled genomes for homology prediction. Augustus (2.5.5)<sup>50</sup>, GlimmerHMM (3.0.4)<sup>51</sup>, SNAP<sup>52</sup>, Geneid(v 1.4.4)<sup>53</sup>, and Genscan<sup>54</sup> were used to de novo predict gene structure. In addition, we used full-length transcriptome data to predict the structure of the transcripts by Pasa (2.3.1)<sup>55</sup> and Cufflink (2.2.1)<sup>56</sup>. Finally, the EvidenceModeler (v1.1.1)<sup>49</sup> was used to integrate the prediction from all sources. The gene sets predicted by the EVM process were searched against public protein databases, using the Blast algorithm, for annotation, including UniProt<sup>42</sup>, Nr<sup>57</sup>, pfam<sup>58</sup>, GO<sup>59,60</sup>, KEGG<sup>61</sup>, and InterPro<sup>62</sup> (Supplementary Table 8).

**Annotation of repeats.** We searched for repetitive sequences in the duck genomes, including tandem repeats and transposable elements (TEs). Tandem Repeats Finder<sup>63</sup> (TRF, v4.09) was employed to annotate the tandem repeats with the following parameters: 2 7 7 80 10 50 2000. Then the TEs were identified at both the DNA and protein levels using a combination of de novo and homology-based approaches. At the DNA level, LTR\_FINDER<sup>64</sup>(v1.0.6) was first used to identify LTR-RTs, and RepeatModeler<sup>65</sup> (v1.0.5) was utilized to construct a de novo repeat library, which comprised a repeat consensus database with classification information. We employed RepeatMasker<sup>65</sup> (v4.0.6) to search for similar TEs in the known Repbase TE library<sup>66</sup> and the de novo repeat library. At the protein level, RepeatProteinMask within the RepeatMasker package was used to search against the TE protein database using a WU-BLASTX engine.

**Genome alignment and gene synteny analysis.** Genome alignment among the three duck genomes was performed using the MUMmer<sup>67</sup> (version 4.00beta2) program with parameter settings `-maxmatch -c 90 -l 40`. The alignments were filtered by running delta-filter with parameter `-1`. SNPs and InDels in the two accessions were extracted by running show-diff in the one-to-one alignment blocks. We also mapped DNA sequencing data (>10×) from the Illumina HiSeq platform for each accession against the other genomes using BWA (version 0.7.10-r789) software. All these variants were annotated using the VEP program<sup>68</sup>. To identify syntenic gene blocks among the duck genomes, we conducted an All-vs.-All blastp (*e*-value < 1e-10, -v 5, -b 5) for each genome pair. The homologous genes were analyzed by the MCScanX package<sup>69</sup> with default settings, except for `gap_penalty -3`. Syntenic blocks were defined as those with at least five syntenic genes.

**Identification of inversions and translocations.** The genomes of Shaoxing and Pekin ducks were aligned with Mallard using MUMmer4<sup>67</sup> (version 4.00beta2) to identify inversions and translocations. The alignment blocks exhibiting inversions were extracted for manual checking. Alignment blocks identified in different

positions were extracted to check their flanking blocks. If alignment blocks had non-colinear flanking sequences, those were retained as putative translocations. The translocations were further divided into inter-chromosomal translocations and intra-chromosomal translocations. Both inversions and translocations were identified if they possessed a length > 100 bp and an identity > 90%.

**Identification of presence/absence variations.** Putative presence/absence variations were identified by extracting unaligned regions between Mallard and domestic ducks from the “show-diff” command in MUMmer4<sup>67</sup> (version 4.00beta2). This approach gave sequences of 8.14 Mb for Mallard, and 7.55 Mb for Shaoxing duck between Mallard and Shaoxing duck, in addition to 4.41 Mb for Mallard and 12.03 Mb for Pekin duck between Mallard and Pekin duck. These sequences were then filtered by discarding those overlapping with gap regions in the respective genome. The remaining sequences above were then filtered by alignment with the contig and scaffold of corresponding genomes using blastn (1e-5) to identify putatively unique regions. The segments with coverage > 50%, and identity > 90% were filtered. Finally, the whole genome re-sequencing data were mapped to the respective genome to confirm the potential presence/absence variations. The one-tailed *t* test was used to determine whether the coverage was significantly different between domestic ducks and Mallard.

**SNP/indel calling.** The variant calling was performed using the Speedseq pipeline<sup>70</sup>. After trimming of low-quality bases using Trimmomatic (version 0.32), the clean data were mapped to the Mallard genome using BWA software<sup>71</sup>. The re-sequencing data for Mallard were downloaded from the NCBI Sequence Read Archive (SRA) database. The clean data were mapped to the wild CAU\_wild\_1.0 genome using BWA software<sup>71</sup>. All the unique mapping data were extracted to identify SNPs and small InDels using freebayes and Samtools<sup>72</sup> (version 0.1.19) programs. Variants were removed with QualByDepth (QD) < 4.0, 300 > depth > 2200, Quality < 30, mapping quality (MQ) < 40.0, MQRankSum < -10, ReadPosRankSum < -7.0, FisherStrand > 60.0, ReadPosRankSum > 7, BaseQRankSum < -6, BaseQRankSum > 6°. Cluster Size and ClusterWindowSize were set to 4 and 10, respectively. For the subsequent analyses, we used only bi-allelic SNPs on autosomes. VEP<sup>73</sup> was used to annotate variants according to their functional categorization, which included the following categories: 5 kb up- and downstream of a gene, intergenic, missense, synonymous, intronic, 3' untranslated regions, 5' untranslated regions, stop gain and stop loss. Variants in the up- and downstream regions and in the 3' UTR/5' UTR regions were merged into single categories.

**Strategy to identify missing genes.** We used the annotated transcripts of the Mallard genome to find sequences homologous to any of the 571 genes previously thought to be missing from the bird genome<sup>13,30</sup>, of which 274 were thought to be missing from all avian genomes. The human protein sequences of the corresponding missing genes were used as query sequences to search for homologies in the newly assembled Mallard genome using the best-reciprocal blast algorithm. We manually checked each matched candidate sequence based on the list of missing genes to distinguish matching paralogous products and alignment errors.

**Detection of selective sweeps.** To detect putative selective sweeps, we first searched the genome for regions with high degrees of differentiation between Mallard and domestic lines. We estimated fixation index ( $F_{ST}$ )<sup>74</sup>, and *p*-value of likelihood ratio test using vcflibs (<https://github.com/vcflibs/vcflibs#vcflib>).

**RNA-seq analysis.** Raw reads were trimmed to remove adapters and low-quality reads, with Trimmomatic (version 0.39)<sup>75</sup>. Trimmed reads were mapped to the Mallard genome using HISAT2<sup>76</sup>. Read counts for each gene were calculated using HTSeq<sup>77</sup> and normalized by library sequencing depth using the TMM method implemented in DESeq2 (v.1.24.0)<sup>78</sup>, after filtering genes with no expression. We used DESeq2 to identify DEGs between Mallard and Pekin duck tissues at different times. Samples that had an average  $R^2$  (the square of the Pearson correlation coefficient) greater than 0.95 when compared with other samples, were accepted as valid biological replicates. Genes with  $|\log_2FC| > 0.584$  and the Benjamini-Hochberg (BH) adjusted *p*-value (adjusted-*p* value) < 0.05 were considered as DEGs.

The de novo assembled transcripts of active uterine tissues during the laying period for five bird species (chicken, duck, goose, pigeon, and zebra finch) were obtained from our previous study<sup>15</sup>. We used BLASTX ( $E = 1e-10$ ) to seek the orthologs of chicken and emu OC-17-like transcripts in the assembled transcript dataset. Once we obtained the best candidate transcripts, we designed primers for each candidate transcript of each species to obtain the cDNA.

**Functional annotation and enrichment.** Gene enrichment analysis for structural variations (within the gene or ~5 kb in the up/downstream of gene) and DEGs were completed with the Metascape<sup>79</sup> (2019-8-14) software using the human reference genome to assign genes to the corresponding terms. Enrichment tests were performed using the hypergeometric test and Benjamini-Hochberg *p*-value correction algorithm as described in Metascape.

**RACE (rapid amplification of cDNA ends) for cloning CTL gene family members.** 5' and 3' RACE were performed using the Smarter<sup>®</sup> RACE 5'/3' Kit (Takara Bio Inc., USA) according to the manufacturer's instructions. RACE-PCR products were obtained with SeqAmp DNA Polymerase (Takara Bio Inc., USA) using the Universal Primer Mix (supplied) and a gene-specific primer (Supplementary Table 19). Products were visualized on a 2% agarose gel and purified by NucleoSpin Gel and PCR Clean-UpKit (Takara Bio Inc., USA). This product was then subcloned into the In-Fusion HD Cloning vector (Takara Bio Inc., USA) and grown in TOP10 *E. coli*. Clones were sequenced with the M13 forward primer.

**Phylogenetic analysis for CTL gene families.** The known bird eggshell CTL proteins were downloaded from Uniprot (<http://www.uniprot.org/>), including chicken Ovocleidin-17 (OC-17, Q9PR88), African ostrich Struthioalcin-1 (SCA-1, P83514) and Struthioalcin-2 (SCA-2, P83515), rhea Rhealcalcin-1 (RCA-1, P84617) and Rhealcalcin-2 (RCA-2, P84618), emu Dromaioalcin-1 (DCA-1, P84615), Dromaioalcin-2 (DCA-2, P84616), Green sea-turtle Lithostathine-1-1 (UY3\_13503, M7B1U1), and Green sea-turtle Lithostathine-1-2 (UY3\_02957, M7BRJ1). The nine cDNA sequences generated in this study were conceptually translated into amino acid sequences and used for the sequence alignments. The CTL proteins translated from the cDNA of this study were blasted against the UniProtKB database. A total of 119 amino acid sequences of representative species in different classes (fishes, amphibians, reptiles, birds, and mammals) were selected from the blast results (*E*-value < 1e−10, Identity > 25%). All sequence information is listed in Supplementary Data 11.

Multiple alignments of the amino acid sequences were done by the MUSCLE algorithm implemented in MEGA<sup>80</sup> (version 10) (MUSCLE, Max Iteration = 8). A phylogenetic tree was constructed using the JTT + G + I model with 100 bootstraps, with the purple sea urchin (echinodermata) used as the outgroup. The crystal structure files of African ostrich Struthioalcin-1 (4UWW.pdb) and chicken Ovocleidin-17 (1GZ2.pdb) were obtained from the RCSB PDB (<https://www.rcsb.org/>) website. ESPript (version 3.0)<sup>81</sup> was used to show the conserved and structural characteristics of the CTL gene family (parameters: MODE, ADV; Strict Global score: 0.5).

**ATAC-seq sequencing and analysis of NR2F2 promoter binding region.** Fifty thousand nuclei from Pekin duck subcutaneous preadipocytes (*N* = 2) before and after oleic acid-induced adipogenic differentiation were transposed using Tn5 transposase as previously described<sup>82</sup>. Briefly, cells were lysed using ice-cold lysis buffer (10 mM Tris-HCl pH = 7.4, 10 mM NaCl, 3 mM MgCl<sub>2</sub> and 0.1% IGEPAL CA-630) and centrifuged at 2400 r.p.m. (500×*g*) for 10 min. The pellet was resuspended in the transposase reaction mix and incubated at 37°C for 30 min. The sample was column-purified and amplified by 15 cycles of PCR before high-throughput sequencing. Each dataset was aligned to the Mallard genome using Bowtie<sup>38</sup>. After alignment, each group of replicates were merged together, sorted, and indexed. Duplicated reads and low mapping quality reads (mapping score < 30) were removed. The merged, filtered, and sorted BAM files were used as the input for HMMRATAC (default parameters)<sup>83</sup>. The motif analysis and annotation were performed using the Homer toolkit<sup>82</sup>.

**Adipogenic transcriptional regulatory network.** The mapping of the adipogenic gene regulatory network is based on the established white fat differentiation cascade regulatory pathway as described in WikiPathways and recent reviews<sup>84–86</sup> (Supplementary Fig. 23a). We then superimposed the expression information as determined from RNA-seq (fold changes compared to day 0 and multiple-testing corrections) on the network (Supplementary Fig. 23a, b).

**Isolation of stromal vascular cells (preadipocytes) from duck subcutaneous fat tissue biopsies.** Preadipocytes were obtained from Pekin duck and Mallard using the same protocol<sup>87</sup>. First, isolated subcutaneous adipose tissues were cut into small pieces of about 1 mm<sup>3</sup> and digested with 1 mg/mL collagenase A (Sigma-Aldrich, MO, USA) in DMEM/F12 (Dulbecco's modified Eagle's medium/Ham's nutrient mixture F-12, Gibco, Gaithersburg, MD, USA) supplemented with 4% BSA, 100 mM Hepes and 150 nM adenosine (Sigma-Aldrich, MO, USA) for 70 min at 37°C in a shaking water bath. The digest was filtered through nylon screens with 70 μm mesh openings, and the mixture was centrifuged at 1500 × *g* for 10 min to remove mature adipocytes and obtain adipose-derived stromal cells. Finally, cells were resuspended in DMEM/F12 supplemented with 10% FBS and 1% antibiotic/antimycotic solution (Gibco, Gaithersburg, MD, USA) for further manipulation.

**Plasmids for constructing NR2F2-editing cell lines.** Using the *chNR2F2* sequence obtained from the NCBI database (Accession: NC\_007129.7), we designed gRNA sequences targeting exon1 of *chNR2F2*, known as sgRNA1: GTT TGTGGGACAAGTCTAG and sgRNA2: GGCAGTACTGACACTGATTG. We synthesized the oligo-DNAs corresponding to these gRNA sequences and annealed them to a T7 promoter-driven Cas9 vector and to a U6 promoter-driven gRNA vector in order to obtain two gRNA-expressing plasmids. In order to construct the *NR2F2*-overexpression vector, the full-length coding sequence of *chNR2F2* (Gene ID: 386585) was amplified from chicken subcutaneous adipose cDNA by PCR, and cloned into the CMV promoter-driven piggyBac and an EFlα promoter-driven GFP plasmid by replacing GFP using *EcoRI* and *SalI* (New England Biolabs,

Ipswich, MA, USA). All plasmids in this study were a gift from Professor S. Wu (State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University). All cloning plasmids were confirmed by sequencing.

**Cell culture and transfection of the immortalized chicken preadipocyte cell line.** A cell line of immortalized chicken preadipocytes (ICP1)<sup>88</sup> was a kind gift of the Poultry Breeding Group of the College of Animal Science and Technology, Northeast Agricultural University, China, and was cultured in DMEM/F12 cell culture medium with 10% FBS, at 37°C with 5% CO<sub>2</sub>. ICP1 preadipocytes were seeded in 6-well plates for further transfection using Lipofectamine 2000 (Invitrogen), and the transfection procedure was performed according to the manufacturer's instructions. After a 48 h recovery period, the cells were supplemented with 3 μg/mL of puromycin (Sigma-Aldrich, MO, USA) to screen out cells which have not been successfully transfected into the plasmids (ie, negative cells). Once the cell clone is formed, cells were harvested using 0.25% trypsin/EDTA (Gibco, Gaithersburg, MD, USA), and the cell density was calculated using a handheld automated cell counter (Millipore, Darmstadt, Germany). Single cells were plated in each well of a 96-well plate by limiting dilution and then cultured for 10 d in the cell culture medium. The medium was replaced every 4 d. Confluent cell colonies were propagated and genotyped by PCR and sequencing. Primer sets used for PCR are listed in Supplementary Table 19.

**Adipogenic differentiation of duck preadipocytes and ICP1 cells.** Adipogenic differentiation of duck subcutaneous preadipocytes and ICP1 cells were induced using the same protocol<sup>89</sup>. The duck preadipocytes and ICP1 cells were expanded in culture using DMEM/F12 cell culture medium with 10% FBS. Cells at passage three to four were induced to differentiate after 2 days of confluence (day 0) with 300 nM oleic acid (Sigma-Aldrich, MO, USA) in DMEM/F12 supplemented with 10% FBS, and 1% antibiotic solution. After 3 days, the medium was changed to a cell culture medium (DMEM/F12 with 10% FBS). The medium was changed every 2 days throughout the differentiation period. Cells were fixed with 10% formalin for 20 min and stained with Oil Red O (Sigma-Aldrich, MO, USA) to examine lipid accumulation. After another wash with PBS, the cell nuclei were counterstained with Hoechst 33342 (Sigma-Aldrich, MO, USA). All experiments were repeated three times, and samples were treated in triplicate. Morphological changes were observed and photographed under an inverted fluorescence microscope (Nikon). Lipid droplet accumulation was measured by the Oil Red O extraction assay, as described by Ramirez et al.<sup>90</sup>

**Q-PCR analysis.** Total RNA was isolated from cells with the EZNA total RNA kit (Omega Bio-Tek, GA, USA) according to the manufacturer's instructions. Quantification of RNA was performed with the Nanodrop 2000 Spectrophotometer (Thermo Fisher Scientific, MA, USA). RNA was reverse transcribed using the PrimeScript RT Master Mix kit (Takara Bio, USA), and used in quantitative PCR reactions containing SYBR-green fluorescent dye (Applied Biosystems, MA, USA). Q-PCR was performed using the ABI-7500 PCR machine. Gene-specific primers were designed using Primer 3 software (version 0.4.0, Howard Hughes Medical Institute). Primer sets are listed in Supplementary Table 40. The relative expression of mRNAs was determined after normalization with GAPDH levels using the 2<sup>−ΔΔCt</sup> method<sup>91</sup>.

**Cell count kit 8 assay for duck preadipocyte proliferation.** Pekin duck and Mallard subcutaneous preadipocytes (4000 per well) were cultivated in 96-well plates, and cell proliferation was detected after 6 days with the cell counting kit-8 (Dojindo, Kumamoto, Japan) at 450 nm using a Model 680 Microplate Reader (Bio-Rad). All the data were acquired by averaging the results from four independent experiments.

**Triglyceride concentration assay for duck liver and fat tissue.** The liver and fat tissue homogenates were digested in RIPA buffer (Thermo Fisher Scientific, MA, USA) containing protease inhibitor cocktail (Sigma-Aldrich, MO, USA), before measurement of protein content using BCA Protein Assay Kit (Sigma-Aldrich, MO, USA). Total triglyceride levels were measured using a Triglyceride Reagent kit (Sigma-Aldrich, MO, USA).

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

genome assembly datasets reported in this study have been deposited in GenBank (NCBI) and BIG with the following accession codes: Mallard genome, PRJNA554956; Pekin duck genome, (GenBank, JACEUL000000000; BIG, GWHANUR000000000); Shaoming duck genome (GenBank, JACEUM000000000; BIG, GWHANUS000000000). All datasets have also been deposited in the Genome Warehouse of the BIG Data Center at the Beijing Institute of Genomics, Chinese Academy of Sciences (<https://ngdc.cnbc.ac.cn/bioproject/>), under the following accession numbers: whole-genome re-sequencing data, CRA002746; whole-genome sequencing data of Shaoming duck, CRA002750 and

CRA002733; dynamic transcriptome sequencing of Mallard liver tissue, CRA002743; dynamic transcriptome sequencing of Mallard skin fat tissue, CRA002755; dynamic transcriptome sequencing of Pekin duck liver tissue, CRA002747; dynamic transcriptome sequencing of Pekin duck skin fat tissue, CRA002754; RNA-Seq of Mallard subcutaneous preadipocytes, CRA002775. The RNA-Seq of Pekin duck subcutaneous preadipocytes<sup>92</sup>, the Iso-Seq of Pekin duck<sup>33</sup>, and whole-genome re-sequencing of Mallard<sup>1</sup> have been reported previously, and the data were deposited into the NCBI database under accession numbers SRX4646736, SRP188279, PRJNA450892, respectively. The raw sequencing data also reported in this paper have been deposited in the Sequence Read Archive (SRA, <https://www.ncbi.nlm.nih.gov/sra>) under NCBI BioProject accession PRJNA465648 and PRJNA554956. The sequence source of the public database were shown below: Uniprot database was downloaded from <https://www.uniprot.org/>; Ensembl/GENCODE gene set of human was downloaded from [http://ftp.ensembl.org/pub/release-103/fasta/homo\\_sapiens/pep/Homo\\_sapiens.GRCh38.pep.all.fa.gz](http://ftp.ensembl.org/pub/release-103/fasta/homo_sapiens/pep/Homo_sapiens.GRCh38.pep.all.fa.gz); Nr database was downloaded from <https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz>; KEGG database was downloaded from <https://www.genome.jp/kegg/>; InterPro was downloaded from <https://www.ebi.ac.uk/interpro/>; Pfam database was downloaded from <http://pfam.xfam.org/>; GO database was downloaded from <http://geneontology.org/>. All data and research materials are available upon reasonable request by contacting the corresponding author.

Received: 1 September 2020; Accepted: 21 September 2021;

Published online: 11 October 2021

## References

- Zhou, Z. et al. An intercross population study reveals genes associated with body size and plumage color in ducks. *Nat. Commun.* **9**, 2648 (2018).
- Zhang, Z. B. et al. Whole-genome resequencing reveals signatures of selection and timing of duck domestication. *Gigascience* **7**, 1–11 (2018).
- Huang, Y. et al. The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. *Nat. Genet.* **45**, 776–783 (2013).
- Olsen, B. et al. Global patterns of influenza A virus in wild birds. *Science* **312**, 384–388 (2006).
- Venkatesh, D. et al. Avian influenza viruses in wild birds: virus evolution in a multihost ecosystem. *J. Virol.* **92**, 599–615 (2018).
- Lawal, R. A. et al. The wild species genome ancestry of domestic chickens. *BMC Biol.* **18**, 13 (2020).
- Piegu, B. et al. Variations in genome size between wild and domesticated lineages of fowls belonging to the *Gallus gallus* species. *Genomics* **112**, 1660–1673 (2020).
- Tian, X. et al. Building a sequence map of the pig pan-genome from multiple de novo assemblies and Hi-C data. *Sci. China Life Sci.* **63**, 750–763 (2020).
- Hincke, M. T. et al. The eggshell: structure, composition and mineralization. *Front. Biosci.* **17**, 1266–1280 (2012).
- Erben, H. K., Hoefs, J. & Wedepohl, K. H. Paleobiological and isotopic studies of eggshells from a declining dinosaur species. *Paleobiology* **5**, 380–414 (1979).
- Rao, M. et al. A duck RH panel and its potential for assisting NGS genome assembly. *BMC Genomics* **13**, 513 (2012).
- Zhang, G. et al. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1320 (2014).
- Lovell, P. V. et al. Conserved syntenic clusters of protein coding genes are missing in birds. *Genome Biol.* **15**, 565 (2014).
- Botero-Castro, F., Figueu, E., Tilak, M. K., Nabholz, B. & Galtier, N. Avian genomes revisited: hidden genes uncovered and the rates versus traits paradox in birds. *Mol. Biol. Evol.* **34**, 3123–3131 (2017).
- Yin, Z. T. et al. Revisiting avian ‘missing’ genes from de novo assembled transcripts. *BMC Genomics* **20**, 4 (2019).
- Reyes-Grajeda, J. P., Moreno, A. & Romero, A. Crystal structure of ovocleidin-17, a major protein of the calcified *Gallus gallus* eggshell: implications in the calcite mineral growth pattern. *J. Biol. Chem.* **279**, 40876–40881 (2004).
- Hincke, M. T., Tsang, C. P., Courtney, M., Hill, V. & Narbaitz, R. Purification and immunochemistry of a soluble matrix protein of the chicken eggshell (ovocleidin 17). *Calcif. Tissue Int.* **56**, 578–583 (1995).
- Mann, K. & Siedler, F. The amino acid sequence of ovocleidin 17, a major protein of the avian eggshell calcified layer. *Biochem. Mol. Biol. Int.* **47**, 997–1007 (1999).
- Zhang, Q. et al. Integrating de novo transcriptome assembly and cloning to obtain chicken Ovocleidin-17 full-length cDNA. *PLoS ONE* **9**, e93452 (2014).
- Mann, K. & Siedler, F. Ostrich (*Struthio camelus*) eggshell matrix contains two different C-type lectin-like proteins. Isolation, amino acid sequence, and posttranslational modifications. *Biochim. Biophys. Acta* **1696**, 41–50 (2004).
- Fan, W. et al. Dynamic accumulation of fatty acids in duck (*Anas platyrhynchos*) breast muscle and its correlations with gene expression. *BMC Genomics* **21**, 58 (2020).
- Chen, L. et al. Transcriptome analysis of adiposity in domestic ducks by transcriptomic comparison with their wild counterparts. *Anim. Genet.* **46**, 299–307 (2015).
- Goodridge, A. G. & Ball, E. G. Lipogenesis in the pigeon: in vivo studies. *Am. J. Physiol.* **213**, 245–249 (1967).
- Ghaben, A. L. & Scherer, P. E. Adipogenesis and metabolic health. *Nat. Rev. Mol. Cell Biol.* **20**, 242–258 (2019).
- Knoll, B. J., Zarucki-Schulz, T., Dean, D. C. & O’Malley, B. W. Definition of the ovalbumin gene promoter by transfer of an ovalglobin fusion gene into cultured cells. *Nucleic Acids Res.* **11**, 6733–6754 (1983).
- Xu, Z., Yu, S., Hsu, C. H., Eguchi, J. & Rosen, E. D. The orphan nuclear receptor chicken ovalbumin upstream promoter-transcription factor II is a critical regulator of adipogenesis. *Proc. Natl Acad. Sci. USA* **105**, 2421–2426 (2008).
- Ashraf, U. M., Sanchez, E. R. & Kumarasamy, S. COUP-TFII revisited: its role in metabolic gene regulation. *Steroids* **141**, 63–69 (2019).
- Franchini, L. F. & Pollard, K. S. Human evolution: the non-coding revolution. *BMC Biol.* **15**, 89 (2017).
- Hubisz, M. J. & Pollard, K. S. Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution. *Curr. Opin. Genet. Dev.* **29**, 15–21 (2014).
- Zhang, G. et al. Genomics: bird sequencing project takes off. *Nature* **522**, 34 (2015).
- Warren, W. C. et al. A new chicken genome assembly provides insight into avian genome structure. *G3 (Bethesda)* **7**, 109–117 (2017).
- Lam, E. T. et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* **30**, 771–776 (2012).
- Yin, Z. T., Zhang, F., Smith, J., Kuo, R. & Hou, Z. C. Full-length transcriptome sequencing from multiple tissues of duck, *Anas platyrhynchos*. *Sci. Data* **6**, 275 (2019).
- Chin, C. S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
- Chin, C. S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
- Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
- English, A. C. et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE* **7**, e47768 (2012).
- Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
- Burton, J. N. et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
- Xu, G. C. et al. LR\_GapCloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *Gigascience* **8**, giy157 (2019).
- Waterhouse, R. M., Seppey, M., Simao, F. A. & Zdobnov, E. M. Using BUSCO to assess insect genomic resources. *Methods Mol. Biol.* **1858**, 59–74 (2019).
- McGinnis, W., Levine, M. S., Hafen, E., Kuroiwa, A. & Gehring, W. J. A conserved DNA sequence in homoeotic genes of the *Drosophila* Antennapedia and bithorax complexes. *Nature* **308**, 428–433 (1984).
- Harris, R.S. Improved pairwise alignment of genomic DNA. *Ph.D. thesis*, The Pennsylvania State University (2007).
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
- Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic Acids Res.* **47**, D155–D162 (2019).
- Gruber, A. R., Lorenz, R., Bernhart, S. H., Neubock, R. & Hofacker, I. L. The Vienna RNA websuite. *Nucleic Acids Res.* **36**, W70–W74 (2008).
- Kalvari, I. et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.* **46**, D335–D342 (2018).
- Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
- Haas, B. J. et al. Automated eukaryotic gene structure annotation using EvidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
- Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
- Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
- Bromberg, Y. & Rost, B. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* **35**, 3823–3835 (2007).
- Blanco, E., Parra, G. & Guigo, R. Using geneid to identify genes. *Curr. Protoc. Bioinform. Chapter 4*, Unit 4.3 (2007).
- Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).

55. Haas, B. J. et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
56. Ghosh, S. & Chan, C. K. Analysis of RNA-Seq data using TopHat and cufflinks. *Methods Mol. Biol.* **1374**, 339–361 (2016).
57. Coordinators, N. R. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **46**, D8–D13 (2018).
58. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
59. The Gene Ontology, C. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).
60. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
61. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
62. Mitchell, A. L. et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **47**, D351–D360 (2019).
63. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
64. Xu, Z. & Wang, H. LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
65. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* Chapter 4 Unit 4 10 (2009).
66. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
67. Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30**, 2478–2483 (2002).
68. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
69. Wang, Y. et al. MCSanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
70. Chiang, C. et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat. Methods* **12**, 966–968 (2015).
71. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
72. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
73. Yourshaw, M., Taylor, S. P., Rao, A. R., Martin, M. G. & Nelson, S. F. Rich annotation of DNA sequencing variants by leveraging the Ensembl Variant Effect Predictor with plugins. *Brief. Bioinform.* **16**, 255–264 (2015).
74. Cockerham, C. C. & Weir, B. S. Covariances of relatives stemming from a population undergoing mixed self and random mating. *Biometrics* **40**, 157–164 (1984).
75. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
76. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
77. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
78. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
79. Zhou, Y. et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10**, 1523 (2019).
80. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
81. Robert, X. & Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42**, W320–W324 (2014).
82. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
83. Tarbell, E. D. & Liu, T. HMMRATAC: a Hidden Markov ModelER for ATAC-seq. *Nucleic Acids Res.* **47**, e91 (2019).
84. Siersbaek, R. & Mandrup, S. Transcriptional networks controlling adipocyte differentiation. *Cold Spring Harb. Symp. Quant. Biol.* **76**, 247–255 (2011).
85. Sarantopoulos, C. N. et al. Elucidating the preadipocyte and its role in adipocyte formation: a comprehensive review. *Stem Cell Rev. Rep.* **14**, 27–42 (2018).
86. Mota de Sa, P., Richard, A. J., Hang, H. & Stephens, J. M. Transcriptional regulation of adipogenesis. *Compr. Physiol.* **7**, 635–674 (2017).
87. Matsubara, Y., Sato, K., Ishii, H. & Akiba, Y. Changes in mRNA expression of regulatory factors involved in adipocyte differentiation during fatty acid induced adipogenesis in chicken. *Comp. Biochem. Physiol. A Mol. Integr. Physiol.* **141**, 108–115 (2005).
88. Wang, W. et al. Immortalization of chicken preadipocytes by retroviral transduction of chicken TERT and TR. *PLoS ONE* **12**, e0177348 (2017).
89. Shang, Z. et al. Oleate promotes differentiation of chicken primary preadipocytes in vitro. *Biosci. Rep.* **34**, e00093 (2014).
90. Ramirez-Zacarias, J. L., Castro-Munozledo, F. & Kuri-Harcuch, W. Quantitation of adipose conversion and triglycerides by staining intracytoplasmic lipids with Oil red O. *Histochemistry* **97**, 493–497 (1992).
91. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2<sup>-</sup>(Delta Delta C(T)) Method. *Methods* **25**, 402–408 (2001).
92. Wang, Z. et al. Dynamics of transcriptome changes during subcutaneous preadipocyte differentiation in ducks. *BMC Genomics* **20**, 688 (2019).

## Acknowledgements

We thank the Poultry Breeding Group of the College of Animal Science and Technology, Northeast Agricultural University, for providing the ICP cell line. The work was supported by the National Waterfowl-Industry Technology Research System (CARS-42), National Nature Science Foundation of China (31972525, 31572388), Beijing Municipal Science & Technology Commission (Z181100002418008), Key-Area Research and Development Program of Guangdong Province (2020B020222003), Wellcome Trust (108749/Z/15/Z). Hincke's participation was additionally supported by the Canadian Natural Sciences and Engineering Research Council (NSERC, Discovery program RGPIN-2016-04410). For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

## Author contributions

Z.C.H., N.Y. and S.S.H. designed the study. F.Z., Z.T.Y., F.Z., Q.S.Z., D.B., Z.Z. and Z.K.Z. performed the analyses of the genome and transcriptome sequence. F.M. and D.O. annotated the Mallard genome. Z.W., F.B.L., X.Q.L., S.R.D., G.S.L., F.X.Y., J.P.H. and L.Z.L. collected the samples and performed the experiments. Z.C.H., F.Z., Z.T.Y., Z.W., and F.Z. wrote the paper. J.S., M.H., D.B. and N.Y. revised the paper. All authors read and approved the final paper.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-26272-1>.

**Correspondence** and requests for materials should be addressed to Zhuo-Cheng Hou.

**Peer review information** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons

Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021