Genomics Proteomics Bioinformatics 21 (2023) 483–500

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

The First Crested Duck Genome Reveals Clues to Genetic Compensation and Crest Cushion Formation

Guobin Chang [1,2,#], Xiaoya Yuan [1,#], Qixin Guo [1,#], Hao Bai [2,#], Xiaofang Cao [3,#], Meng Liu [3,#], Zhixiu Wang [1], Bichun Li [1], Shasha Wang [1], Yong Jiang [1], Zhiquan Wang [4], Yang Zhang [1], Qi Xu [1], Qianqian Song [1], Rui Pan [1], Lingling Qiu [1], Tiantian Gu [1], Xinsheng Wu [1], Yulin Bi [1], Zhengfeng Cao [1], Yu Zhang [1], Yang Chen [1], Hong Li [3], Jianfeng Liu [5], Wangcheng Dai [6], Guohong Chen [1,2,*]

[1] *Key Laboratory of Animal Genetics and Breeding and Molecular Design of Jiangsu Province, College of Animal Science and Technology, Yangzhou University, Yangzhou 225009, China*
[2] *Joint International Research Laboratory of Agriculture and Agri-Product Safety, Ministry of Education, Institutes of Agricultural Science and Technology Development, Yangzhou University, Yangzhou 225009, China*
[3] *Novogene Bioinformatics Institute, Beijing 100080, China*
[4] *Department of Agricultural, Food, and Nutritional Sciences, University of Alberta, Edmonton, AB T6G 2R3, Canada*
[5] *College of Animal Science and Technology, China Agricultural University, Beijing 100091, China*
[6] *Zhenjiang Tiancheng Agricultural Science and Technology Co., Ltd, Zhenjiang 210034, China*

Received 26 January 2021; revised 5 July 2023; accepted 15 August 2023
Available online 29 August 2023

Handled by Peng Cui

KEYWORDS

Genetic compensation;
Genome assembly;
Chinese crested duck;
Crest cushion;
Genome adaptive evolution

Abstract The **Chinese crested (CC) duck** is a unique indigenous waterfowl breed, which has a **crest cushion** that affects its survival rate. Therefore, the CC duck is an ideal model to investigate the **genetic compensation** response to maintain genetic stability. In the present study, we first generated a chromosome-level genome of CC ducks. Comparative genomics revealed that genes related to tissue repair, immune function, and tumors were under strong positive selection, indicating that these adaptive changes might enhance cancer resistance and immune response to maintain the genetic stability of CC ducks. We also assembled a Chinese spot-billed (Csp-b) duck genome, and detected the structural variations (SVs) in the **genome assemblies** of three ducks (*i.e.*, CC duck, Csp-b duck, and Peking duck). Functional analysis revealed that several SVs were related to the immune system of

* Corresponding author.
E-mail: ghchen2019@yzu.edu.cn (Chen G).
# Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2023.08.002

CC ducks, further strongly suggesting that genetic compensation in the anti-tumor and immune systems supports the survival of CC ducks. Moreover, we confirmed that the CC duck originated from the mallard ducks. Finally, we revealed the physiological and genetic basis of crest traits and identified a causative mutation in *TAS2R40* that leads to crest formation. Overall, the findings of this study provide new insights into the role of genetic compensation in adaptive evolution.

Introduction

Organisms have developed dynamic buffer systems during evolution to maintain normal development in the presence of certain genetic mutations [1–4]. Organisms adapt to their environments by genomic fine-tuning during their evolution. Recently, the genetic compensation response (GCR), a new mechanism supporting genomic robustness, was found in zebrafish [5,6], mice [7], and rockcress [8] by gene knockout mutations. In a sense, the organism developed a lethal phenotype caused by harmful mutations, or in these instances, resulting in 'similar to gene knockout'. Under the action of long-term natural selection and artificial selection, the GCR causes a series of genetic compensation mutations, thereby promoting genetic stability to maintain the organism. Over time, compensation mutations may lead to a series of phenotypic changes that offset the lethal phenotype to maintain the population.

The Chinese crested (CC) duck is a unique breed with complex feather-protruding traits collectively called the crest. Although feather crests are widely distributed in birds (such as cockatoos, gray-crowned cranes, and great-crested grebes), there are significant differences in shape and physiological mechanisms. Almost all birds with crest traits exhibit a distinct crown formed by prominent feathers. The crest cushion of the CC duck consists of soft tissue protuberances covered by feathers and skin. Although the presence of a crest does not affect survival in most crested birds, crested ducks are an exception. Previous studies of Hochbrutflugenten (HBTcr) ducks, which are crested duck breeds in Germany, have shown that crested ducks have high prenatal and postnatal mortalities, exhibiting motor incoordination in the wild due to incomplete skull closure [9–11]. Although the phenotype composition of the crest cushion and the fertilization rate in HBTcr and CC ducks were similar, the survival rate of CC ducks was significantly higher (more than 95%) after birth, and they showed good motor coordination. Therefore, the formation mechanism of the crest cushion and the genomic compensation for the effect of the crest cushion on the CC duck have gathered considerable interests in CC duck research. However, resolving this issue has proven challenging because the crest trait is phenotypically complex. Nevertheless, the CC duck is an ideal example to help explain the function of the GCR in maintaining genetic stability. Specifically, genome assembly may be the best solution to address these issues. However, the genomic resources for ducks are limited, with published genome sequences limited to Peking (PK) duck, mallard duck, and Shaoxing duck in the National Center for Biotechnology Information database [12–14]. In addition, these genomes cannot reveal the basis of crest cushion formation at the genomic level.

To explore the physiological and genetic bases underlying the formation of crest cushions, we first assembled a high-quality CC duck genome and a Chinese spot-billed duck (Csp-b duck; *Anas zonorhyncha*) genome. These genomes were compared with those of other wild and domesticated ducks to investigate shifts in structural variations (SVs) and genes under adaptive evolution. Our results provide valuable insights for understanding the role of the GCR in adaptive evolution and provide a valuable genomic resource for future genome-wide analyses of economically important traits in poultry.

Results

Genome assembly of the CC duck

A 28-week-old female CC duck was selected for genome sequencing and assembly (Figure S1; Table S1). The genome size was estimated to be 1.26 Gb based on the *k*-mer distribution (Figure S2; Table S2). To generate a high-quality reference genome for CC duck, a total of 85.06 Gb (~ 75.97×) PacBio long reads were assembled using FALCON v0.7 [15], and this assembly was then polished using Quiver (smrtlink v6.0.1) [16]. Thereafter, 10X Genomics (~ 79.15×) was used to connect contigs into super-scaffolds with the software FragScaff [17], which resulted in a 1.13-Gb assembly (CC_duck_v1.0) with an N50 contig size of 3.24 Mb and an N50 scaffold size of 7.61 Mb (Table S3). Approximately 88.65-Gb (~ 79.15×) Illumina paired-end reads were used to polish the assembly with Pilon v1.18. Using the high linkage genetics map, 1216 scaffolds were anchored and oriented onto 37 autosome chromosomes using CHROMONMER [18] (Figure S3). The remaining scaffolds were organized into the CC duck Z chromosomes based on their sequence similarity with the Z chromosomes of the published duck genome (CAU_duck1.0) by MUMmer v3.23. The final assembly yielded an N50 scaffold size of 73.74 Mb and an N50 contig size of 3.24 Mb, and ~ 94.10% of the assembled genome was anchored onto the 38 chromosomes (*i.e.*, 37 autosome chromosomes and one Z chromosome) (**Figure 1**; Table S4).

To assess the quality and integrity of the genome assembly, short paired-end reads were aligned with the assembly. Overall, 96.68% of the paired-end reads could be mapped to the genome, suggesting the high integrity of our assembly genome (Table S5). Benchmarking Universal Single-Copy Orthologs (BUSCO) [19] showed that 97.7% (2527/2586) of vertebrate single-copy orthologous genes were captured in our assembly, which was comparable or even better than that in published duck genomes (Table S6). A total of 17,425 protein-coding genes were predicted in the CC duck genome by combining *de novo* prediction, homology-based prediction, and RNA sequencing (RNA-seq) data (Table S7). In addition, to help explore the origin and adaptive evolution of CC ducks, we assembled another duck genome (Csp-b duck) with ~ 85.81× paired-end reads using SOAPdenovo [20]. Finally, we generated a 1.10-Gb assembly with an N50 scaffold size of 675.96 kb (Table S8) and predicted 15,278 protein-coding genes based on homologous comparison approaches (Table S9).

Figure 1 Genomic landscape of the CC duck genome

The circles illustrate (from outside to inside): chromosome ideograms for CC duck (Mb) (**A**); distribution of contig lengths for CC duck chromosomes [three tracks from outside to inside: contigs with length > 1 Mb (orange rectangles); contigs with 500 kb < length ≤ 1 Mb; contigs with length ≤ 500 kb] (**B**); distribution of SNP density (non-overlapping, 500-bp window size) (**C**); distribution of GC content (non-overlapping, 500-bp window size) (**D**); distribution of gene content (non-overlapping, 500-bp window size) (**E**); distribution of repeat content (non-overlapping, 500-bp window size) (**F**); heatmap showing the expression levels of genes in crested head tissue (represented by FPKM values) (**G**); heatmap showing the expression levels of genes in skin tissue (represented by FPKM values) (**H**); distribution of miRNA numbers (non-overlapping, 500-bp window size) (**I**). CC, Chinese crested; SNP, single nucleotide polymorphism; FPKM, fragments per kilobase of transcript per million mapped reads.

Historical population structure reveals the origin of the CC duck

The CC duck is a unique domesticated duck breed with a crest cushion in China. According to historical records, the first documented origin of the CC duck can be traced back to the early Ming Dynasty in China (1368 AD), and the chances are that they might have been present earlier (Figure S4). To explore the origin of the crest cushion, we obtained data from

three wild duck breeds [mallard duck from Ningxia Hui Autonomous Region (MDN), mallard duck from Zhejiang Province (MDZ), and Csp-b] and two domesticated duck breeds (PK duck and CC duck) (Table S10). After excluding linked single nucleotide polymorphism (SNP) loci that could potentially bias clustering results, we built a neighbor-joining (NJ) tree using 39 samples. The NJ tree assigned all samples to three major groups (the wild duck, PK duck, and CC duck groups) (**Figure 2**A). These clustered results were also supported by principal component analysis (PCA; Figure 2B). Additionally, we used FRAPPE to explore the genetic composition of each group after initially removing potential bias caused by missing loci [21]. The sample clusters were evaluated using an ad hoc statistic (ΔK). The domesticated ducks were separated from the wild ducks when the cluster number K was set to 2. The ΔK value reached its maximum at $K = 3$, indicating the uppermost structural level. At the same time, the MDZ was also separated from the MDN. At $K = 4$, the clusters revealed that the PK duck shared gene flow with the CC duck (Figure 2C). The results of the NJ tree, PCA, and FRAPPE indicate that there is gene flow between the PK and CC ducks. We inferred that the CC duck might have been domesticated independently from the MDZ.

Broad-scale population collection and management of CC ducks are critical for population recovery. Such efforts are challenging because the historical population scale of CC ducks is unclear. To infer the ancient demographic history of the CC ducks, the PopSizeABC method, which is based on approximate Bayesian computation, was used to predict the effective population size (*Ne*) of CC ducks, PK ducks, and MDZ ducks over the past 100,000 years. Over this period, we found that the population size of CC (Figure 2D), MDZ (Figure 2E), and PK (Figure 2F) ducks varied significantly in the degree of fluctuations, followed by a short period of relative population stability before the near extinction of the CC duck in the past 100 years. The recent demographic pattern implies that the CC duck experienced a population increase over the past 70 years through human protection beginning in the 20th century.

Gene evolution related to the GCR and crest trait formation of CC duck

To reveal the genomic signatures of the GCR in the adaptive evolution of the CC duck, its genome and 13 other published species genomes (Table S11) were selected for gene family clustering analysis using OrthoMCL software [22], which identified 19,605 gene families, including 3089 single-copy gene families. Based on these 3089 single-copy genes, we constructed a

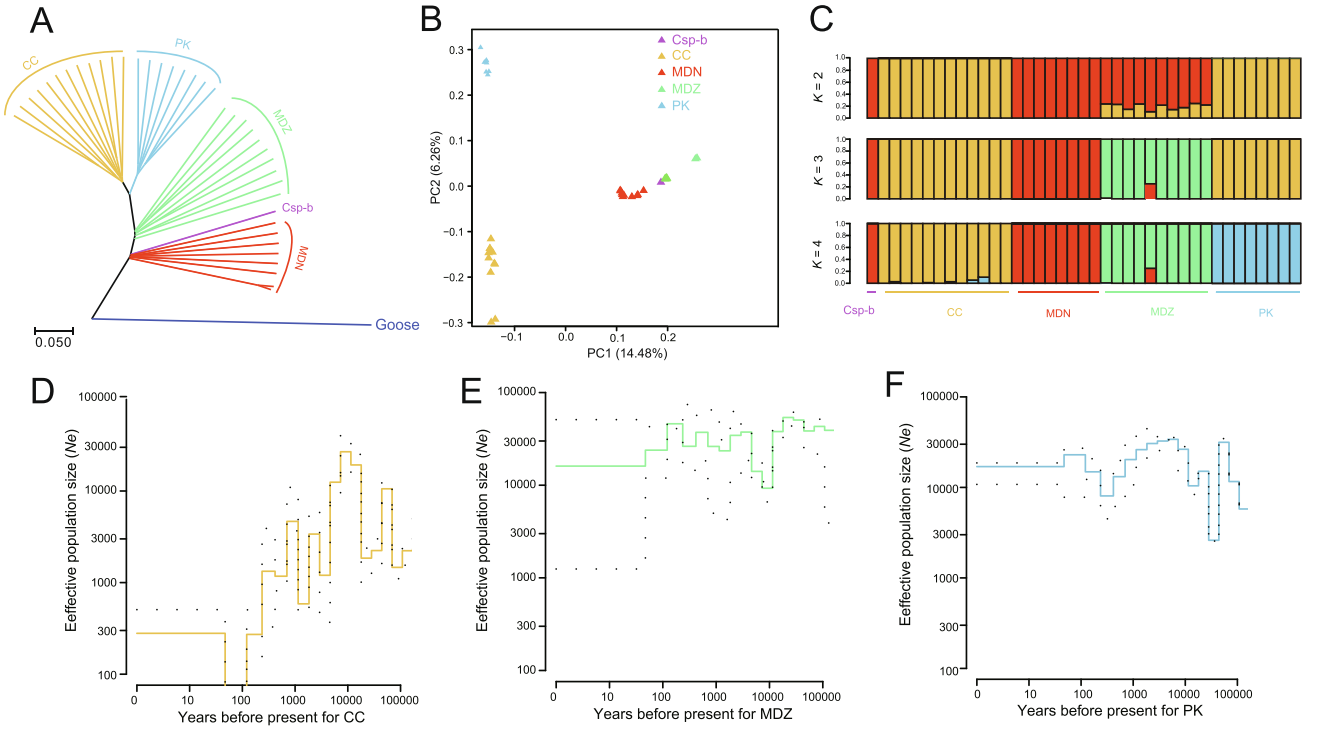

Figure 2 Population structure, genomic landscape of duck divergence, and population size estimate

A. Phylogenetic tree of all duck samples inferred from the whole-genome tag SNPs with goose (*Anser cygnoides domesticus*) as an outgroup. The NJ method was used with 1000 bootstraps. **B.** PCA of the duck samples. **C.** Population structure of all individuals ($K = 2$, 3, and 4). The population origin of each individual is indicated on the x-axis. Each individual is represented by a bar that is segmented into colors based on the ancestry proportions given the assumption of K populations. **D.–F.** The effective population size (*Ne*) for CC (D), MDZ (E), and PK (F) ducks inferred using PopSizeABC v2.1. A 90% confidence interval is indicated by dotted lines. PCA, principal component analysis; PC, principal component; NJ, neighbor-joining; Csp-b, Chinese spot-billed; MDN, mallard duck from Ningxia Hui Autonomous Region; MDZ, mallard duck from Zhejiang Province; PK, Peking.

phylogenetic tree and estimated the divergence time of these 14 species. Phylogenomic analysis showed that the CC duck diverged from the goose ~ 23.3 million years ago (MYA) [23] — slightly earlier than the previous molecular-based estimate of 20.8 MYA for ducks and geese [24] (**Figure 3**A). Interestingly, we found that the crest cushion also exists in every branch of the gray-crowned crane (*Balearica regulorum*), great-crested grebe (*Podiceps cristatus*), hoatzin (*Opisthocomus hoazin*), and little egret (*Egretta garzetta*). Other species in a previous phylogenetic tree of all birds also possess a crest, such as crested cockatoos, crested pigeons, and emperor penguins, but the type and function of their crests may vary [25]. Therefore, we considered that the crest might be widespread in all types of birds, although the crested characters were removed or preserved in some birds under natural selection or human intervention.

Gene family evolution

Gene family expansion and contraction were examined using CAFE software [26]. Compared with the most recent common

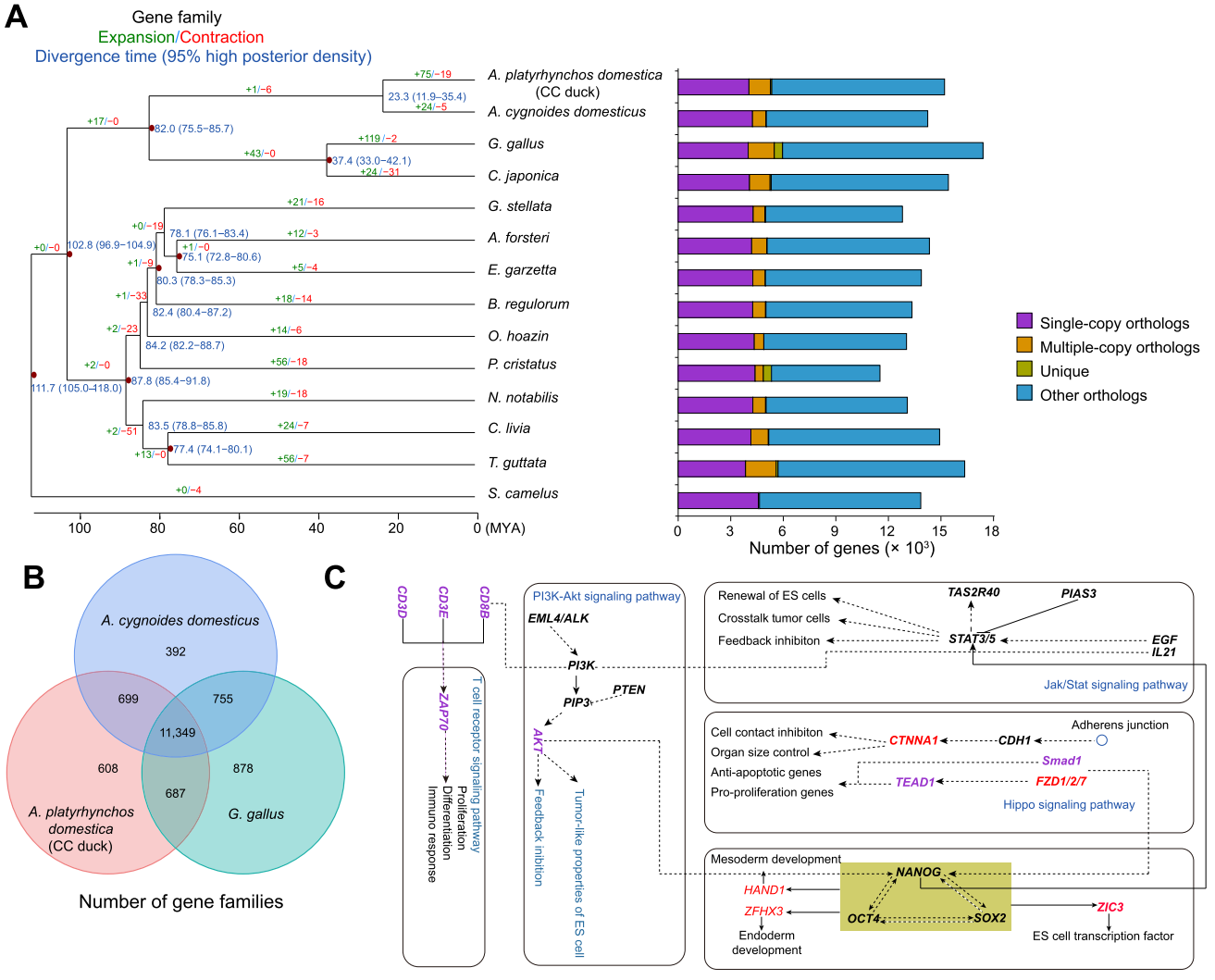

Figure 3 Comparative genomic analysis reveals clues to the genetic compensation of CC duck

A. Phylogenetic tree constructed using single-copy orthologs. The estimated divergence time is shown in the middle of the branch (blue). The numbers of expansion genes (green) and contraction genes (red) are shown above the branch. The gene family cluster is shown to the right of the phylogenetic tree. **B.** Lineage-specific genes of CC duck (*A. platyrhynchos*) identified by comparison with chicken (*G. gallus*) and goose (*A. cygnoides domesticus*). **C.** Proposed signaling pathways for the protective mechanism and crest formation mechanism in CC ducks. The PSGs in CC duck are indicated in purple, and expanded genes are indicated in red. MYA, million years ago; PSG, positively selected gene; *A. platyrhynchos domestica*, *Anas platyrhynchos domestica*; *A. forsteri*, *Aptenodytes forsteri*; *B. regulorun*, *Balearica regulorum*; *C. japonica*, *Coturnix japonica*; *C. livia*, *Columba livia*; *E. garzetta*, *Egretta garzetta*; *G. gallus*, *Gallus gallus*; *G. stellata*, *Gavia stellata*; *N. notalilis*, *Nestor notabilis*; *O. hoazin*, *Opisthocomus hoazin*; *P. cristatus*, *Podiceps cristatus*; *S. camelus*, *Struthio camelus*; *T. guttata*, *Taeniopygia guttata*.

ancestor (MRCA), we identified 75 expanded gene families and 19 contracted gene families in CC ducks (Figure 3A). Furthermore, these expanded gene families were mainly enriched in Gene Ontology (GO) terms, including cell adhesion, intracellular non-membrane-bound organelles, Wnt-activated receptor activity, and interleukin-1 receptor binding. Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses predicted that these genes were involved in the Hippo signaling pathway, cell adhesion molecules, gap junctions, and signaling pathways regulating the pluripotency of stem cells. We speculated that these expanded genes in CC ducks might potentially participate in special phenotypic evolution, such as the crest cushion (Tables S12–S15). In addition, the tripartite motif-containing 39 (*TRIM39*) and tripartite motif-containing 7 (*TRIM7*) genes, which are parts of the contracted genes, have been implicated in the immune system [24,27]. The TRIM gene family has been shown to be involved in some tumor mechanisms owing to their E3-ubiquitin ligase activity [28]. These results might provide the basis for phenotypic plasticity and compensate for the effect of the crest cushion. Compared with the chicken (*Gallus gallus*) and goose (*Anser cygnoides domesticus*) genomes, we identified 608 gene families specific to CC duck (Figure 3B), many of which were involved in cell adhesion molecules, focal adhesion, and calcium ion binding, especially tissue repair and tumor formation pathways (Tables S16 and S17), suggesting that these genes might play an essential role in cancer development, diffusion, and tissue repair. Collectively, we considered that human intervention led to adaptive evolution in the protective mechanisms of certain species.

Positive selection of genes involved in the anti-tumor response

Positive selection has undoubtedly played an important role in the evolution of animals, especially in maintaining some endangered species. We identified 479 positively selected genes (PSGs) in the CC duck lineage. Functional enrichment analyses showed that these PSGs were significantly associated with genomic stability and tumor formation, and were assigned terms including mismatch repair, cellular response to DNA damage stimulus, DNA double-strand break repair, telomere maintenance via telomerase, and cancer, which may be the underlying basis for the crest cushion (Tables S18 and S19). Importantly, we found that several key genes were under positive selection, such as epidermal growth factor (*EGF*), phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit delta (*PIK3CD*), and phosphoinositide-3-kinase regulatory subunit 4 (*PIK3R4*), which are involved in the PD-L1 expression and the PD-1 checkpoint pathway, providing further evidence that the evolution of the crest cushion relies on tumor formation. Furthermore, we found that the Fraser extracellular matrix complex subunit 1 (*FRAS1*) gene was also under positive selection in CC ducks, which provides evidence that *FRAS1* is associated with hair curliness [29]. In addition, certain genes [*e.g.*, golgin, RAB6-interacting (*GORAB*), and Fas cell surface death receptor (*FAS*)] in the p53 pathway were also positively selected. These findings suggest that CC ducks might have enhanced *GORAB* expression and reduced mouse double minute 2 homolog (*MDM2*) expression during evolution, thereby promoting p53 escape and activating the apoptosis pathway [30]. We also found some proto-oncogenes under strong positive selection in CC ducks, such as key genes in the PI3K-Akt signaling pathway [*PIK3CD*, *PIK3R4*, collagen type VI alpha 1 chain (*COL6A1*), *EGF*, laminin subunit alpha 1 (*LAMA1*), and von Willebrand factor (*VWF*)] (Figure 3C). These results suggest that genetic complementation mutations might have occurred at the genomic level. The effect of this compensation variation was amplified by artificial protection, allowing the CC duck to continue to survive or even expand its population.

In addition, to identify the expression levels of PSGs during crest cushion development, we compared the crest region and adjacent frontal skin tissues at each important embryonic development stage to identify differentially expressed genes (DEGs; Figure S5). For the crest cushion, we identified 176, 207, 203, 233, 296, and 401 DEGs in each developmental stage, respectively. Based on the KEGG enrichment analysis, we found that almost all DEGs were enriched in fatty acid biosynthesis and metabolism, tumor formation and anti-tumor response, tissue repair, and neural cell development pathways. Furthermore, the PSGs, such as connective tissue growth factor (*CTGF*), fatty acid synthase (*FASN*), homeobox D10 (*HOXD10*), syndecan 3 (*SDC3*), and four and a half LIM domains 2 (*FHL2*), were DEGs between the crest region and adjacent frontal skin tissues, which were enriched in osteoclast differentiation, cell adhesion molecules, fatty acid biosynthesis and metabolism, and microRNA in cancer pathways. Overall, positive selection analysis and gene expression results suggest that crest cushion formation is largely related to neural cells, skin tissue, bone, and fatty tissue.

Genomic signatures reveal the domestication of CC ducks

To identify genomic regions influenced by the domestication of CC ducks, we compared the genomes of MDZ and CC duck populations representing different geographic regions using cross-population extended haplotype homozygosity (XP-EHH) [31]. We identified 1151 putative selective sweeps in the CC duck genome compared with MDZ genome (XP-EHH score > 4.409; *Z*-test, $P < 0.01$) (**Figure 4**A). To further identify genome-wide signatures of domestication selection, we calculated the fixation index (F_{ST}) values between MDZ and CC ducks. In total, we identified 919 putative selective sweeps in CC duck genome compared with MDZ genome ($F_{ST} > 0.504$, top 1%) (Figure 4B). As genomic regions targeted by artificial selection may be expected to have decreased levels of genetic variation, we also measured and plotted nucleotide diversity (π) along their genomes. Selecting the windows with the top 1% diversity ratios, *i.e.*, low diversity in the two mallard ducks but high diversity in the CC ducks, we found 1023 potential artificial selection windows in CC duck genome compared with MDZ genome (Figure 4C). Combining the results of the three methods (F_{ST}, π, and XP-EHH), we obtained 51 putative selective regions covering 30 genes involved in the CC duck domestication process (Table S20). Among these genes, we found that dynamin 3 (*DNM3*), which encodes an activator of p53, was under selection. *DNM3* is a member of the dynamin gene family, which possesses mechanochemical properties involved in actin-membrane processes, is predominantly expressed in the brain, and is associated with Sézary's syndrome, (a lymphoproliferative disorder). Nanog homeobox (*NANOG*), a gene under positive selection in CC ducks, is a key factor in the specification of

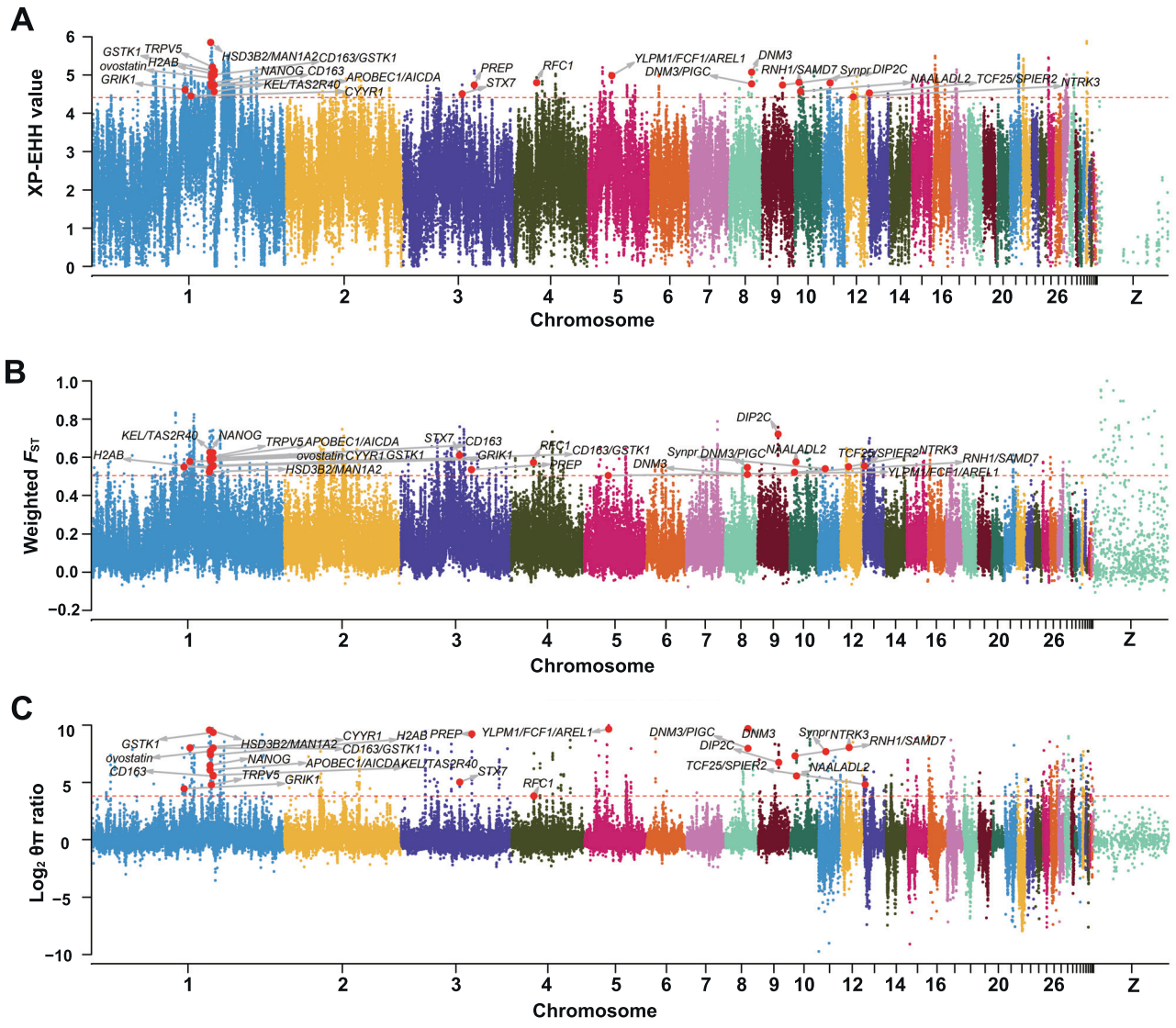

Figure 4 Genomic selection signatures during CC duck domestication
Distribution of XP-EHH values (**A**), population differentiation (F_{ST}) (**B**), and $\log_2$ θπ ratio (**C**) between CC and MDZ ducks using a 20-kb sliding window and 10-kb step. The dotted line represents the significant threshold: F_{ST}, top 1%; XP-EHH value, $P < 0.01$ (Z-test); $\log_2$ θπ ratio, $P < 0.01$ (Z-test). XP-EHH, cross-population extended haplotype homozygosity.

early embryonic pluripotent cells. If *NANOG* is ablated *in vivo*, it will directly affect the fate determination of embryonic stem cells. In addition, a previous study suggests that *NANOG* inhibits apoptosis and promotes cell cycle arrest mainly via p53 regulation [32]. In addition, transient receptor potential cation channel subfamily V member 5 (*TRPV5*), which is the key gene regulating the homeostatic balance of calcium, is also under positive selection in CC ducks. The function of these genes under positive selection during CC duck domestication suggests that regulatory elements may also play a role in the GCR of the crest trait formation.

SV detection reveals the essentials of genome adaptive evolution and genetic compensation

Genome-level evolution and SV accumulation provide an impetus for the adaptive genome evolution of species. Genomic SVs can have a pronounced phenotypic impact, disrupting gene function and modifying gene dosage, whereas some large SVs can lead to large body mutations, including neurodevelopmental disorders and unique trait formation. The CC duck has specific phenotypic traits in the crest cushion and immune levels compared with the PK duck and Csp-b duck. To explore the reasons for these differences at the genomic level, we used the same approach as that of Li and colleagues [33]. We identified 9369 SVs, including 1935 insertions, 4118 deletions, and 3316 inversions in the CC duck assembly. These SVs correspond to 71.91% (6737/9369) of the previous SVs from Illumina short-read genome sequencing, and 28.09% (2632/9369) of the SVs were novel. We found 1541 species-specific genes to be embedded or almost completely contained (> 50% overlap of gene length) in the missing sequences of the PK duck assembly. We explored functional enrichment for the SVs and species-specific genes

from CC ducks using the *ClusterProfiler* package of R v4.1.0 packages [34] and the Gene Ontology (GO) terms revealed by the clustering tool REVIGO (Figures S6–S8) [35]. We identified 35 GO terms that were significantly overrepresented [false discovery rate (FDR) < 0.05] in more than one gene (Table S21). Notably, there were some GO terms related to tissue repair, including cell adhesion, homophilic cell adhesion, and cell communication. These functions may be related to the unique crest traits of the CC ducks. The candidate genes contained several genes related to the immune system and signal transduction, which may have played important roles in the sub-phenotype of the crest traits in CC duck, including ephrin type-A receptor 1 (*EPHA1*; a key factor required for angiogenesis and regulating cell proliferation), RUNX family transcription factor 2 (*RUNX2*; mutations in this gene have been found to be associated with the bone development disorder cleidocranial dysplasia), and taste 2 receptor member 40 (*TAS2R40*; playing a role in the perception of bitterness). Interestingly, some gene families related to animal domestication have appeared as SVs, such as the SLC superfamily of solute carriers and taste 2 receptors.

Similarly, we also identified putative SVs in the Csp-b duck genome assembly by comparison with the CC duck genome, and identified 2694 insertions, 3991 deletions, 609 inversions, and 421 species-specific genes. Functional enrichment among these SV-related genes and species-specific genes from Csp-b ducks was determined using GO analysis and pathway analysis (Table S22). A tatol of 74 functional categories were significantly enriched ($P < 0.05$), and the regulation of small GTPase-mediated signal transduction was ranked as the top category in the GO biological process. We also calculated K_a/K_s ratios by comparing Csp-b ducks to chicken (Figure S9) and Zhedong goose (*A. cygnoides domesticus*) (Figure S10) lineages to find genes accounting for rapid genome evolution. We found that genes with elevated K_a/K_s values in Csp-b ducks were significantly enriched for these functions (FDR < 0.01). Furthermore, these functional GO terms overlapped with the SV-related GO terms by 20.27% (15/74) in Csp-b duck–Zhedong goose pairs and 12.16% (9/74) in Csp-b duck–chicken pairs. We further examined the overlapping GO terms for both pairs, and there were seven categories associated with energy metabolism, nervous system, and signal transduction in the Csp-b duck. We speculate that these seven functional categories might contribute to the duck habitatenvironment-adapted phenotype.

The physiological and genetic basis of crest traits

The crest, which is an interesting phenotypic trait, appears in most bird species worldwide. However, CC ducks are unique duck breeds with bulbous feathers and skin protuberances in China. To fully reveal the physiological basis of crest cushion formation, we investigated the development of the parieto-occipital region of the CC duck during embryonic development by microscopy. The results showed a protuberance at the cranial crest of the E4 duck embryo (**Figure 5**A, Figure S11). Therefore, we speculated that epidermal hyperplasia generated pressure on the adjacent skull cartilage tissue in the fontanelle during the development period, which led to the appearance of perforations in the parieto-occipital region during the cartilage ossification process (Figure S12). Coincidentally, preadipocytes began to differentiate into fat cells. To compensate for the decrease in brain pressure caused by the perforation, different volumes of fat were deposited between the brain and cerebellum (Figure S13). However, spherical feathers are only used to protect against fragile epidermal hyperplasia. Thus, the formation of the crest cushion results from several consecutive coincidences during the development of the skull, scalp, and feathers. The protuberance may be the most fundamental cause of crest formation, and the sub-phenotype of the crest trait was therefore attributed to phenotypic compensation in response to the crest cushion. Furthermore, we found that the inheritance patterns of the crest trait conformed to Mendel's genetic laws in the F_2 generation of 707 CC ducks × CV ducks (crest:crestless = 541:166, $\chi^2_{df=1} = 0.35$).

To identify the genetic basis of crest traits, we performed genome-wide selection tests in CC ducks compared with PK ducks and MDZ, which represent phenotypes for several traits that are relevant for the crest trait of CC ducks. We calculated the global XP-EHH among the CC duck compared with the PK duck and MDZ using a 20-kb sliding window and a shift of 10 kb across the CC duck genome, and 1561 and 1156 putatively selected genomic regions were identified (Figure 5B and C). Additionally, we identified 289 selected regions shared by the two-pair comparison group. In an additional analysis involving the F_{ST} and $\log_2$ θπ ratio using 12 CC ducks and 27 normal ducks, we identified 902 and 980 selective regions, respectively (Figure 5D and E). Combining the results of the selective-sweep analysis of the aforementioned four methods, we identified 26 shared selective regions which spanned 18 candidate genes that we speculated to be associated with crest traits (Table S23). Additional F_{ST} and genetic diversity analysis of the F_2 hybrid population identified 1165 and 997 special selection regions with the top 1% global F_{ST} (Figure 5F) and $\log_2$ θπ ratio values (Figure 5G). Combining the results of between- and within-population selective-sweep analyses, we identified 12 selective windows that may be significantly related to the crest trait (Figure 5H; Table S24). Annotation of the 13 genes putatively influenced by the crest trait revealed functions associated with the sub-phenotypic crest trait.

To fine-map regions identified using selective-sweep methodologies and search for direct evidence of genotype–phenotype associations, we performed genome-wide association study (GWAS) for crest traits with informative phenotypic records. Using a panel of samples from the F_2 hybrid from high-quality SNPs as well as the mixed model, which involved a variance component approach to correct the population structure, we identified two significant signals (harboring 4914 SNPs) that were associated with the crest cushion trait with a threshold of $-\log_{10} P$ value = 8.38 (**Figure 6**A and B). SNPs in the 12 candidate divergent regions (CDRs) associated with crest cushion formation showed extensive linkage disequilibrium (LD). The peak position was located between the *KEL* and *TAS2R40* genes. Furthermore, we found that the genotype frequencies of the related sites in *TAS2R40* and *NANOG* almost separated the crested ducks and normal ducks in the F_2 population (Figure S14). Therefore, we consider that *TAS2R40*, *KEL*, and *NANOG* might be candidate genes for crest cushion formation based on the selective-sweep and GWAS co-localization criteria (Figure 6C–E). To detect the candidate SNPs, we performed Sanger sequencing and genotyping on 30 CC ducks and 75 normal ducks from

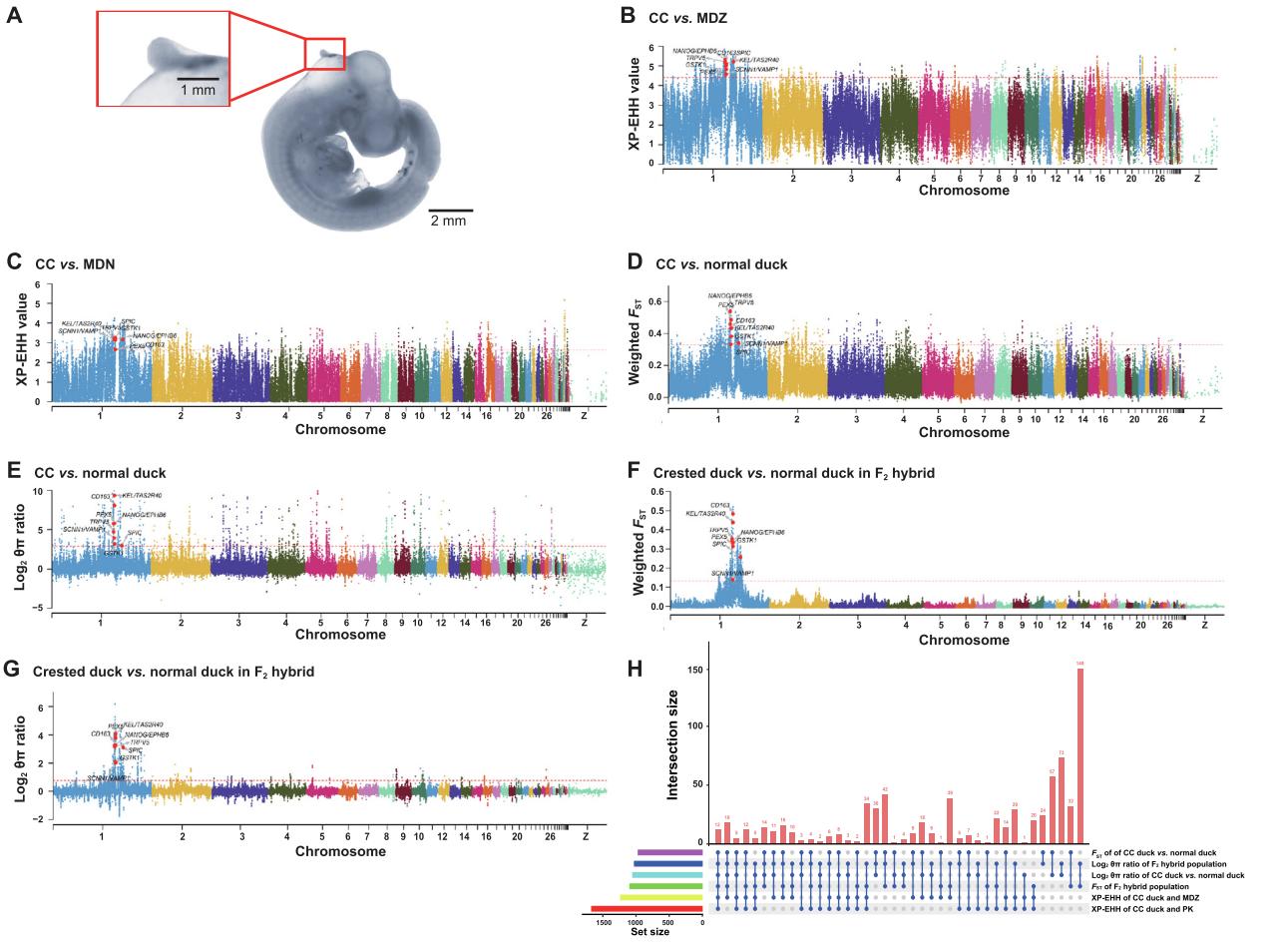

Figure 5 Selective-sweep analysis of the crest cushion of the CC duck

A. The embryo of CC duck and a full image of the crest cushion. XP-EHH values for CC duck compared with MDZ (**B**) and MDN (**C**) ducks. Manhattan plots of F_{ST} (**D**) and $\log_2$ θπ ratio (**E**) of CC duck domestication. Manhattan plots of F_{ST} (**F**) and $\log_2$ θπ ratio (**G**) for selection of crested ducks in F_2 hybrid. **H.** Upset plot showing overlaps between aforementioned selective analysis methods.

three duck breeds. We found that the genotype of the 5′UTR (123272114_c. G78A) of *TAS2R40* (Figure 6F), the first intron (123248845_c. G7127C) of *KEL* (Figure S15), and the fourth exon (120130992_c. G577A_p. V193M and 120131265_c. G850T: p. A284S) of *NANOG* (Figure S16) could separate the CC duck from 15 other non-crested ducks. Among them, only the 123272114_c. G78A of *TAS2R40* showed a 100% frequency of GG genotype in CC ducks, whereas the other two SNP loci showed a percentage of more than 80%. In particular, this SNP exhibited a significant *P* value and could account for 54.68% of the explained phenotypic variation in MLM. Importantly, the tissue expression profile of *TAS2R40* at 56 days of age showed that *TAS2R40* was hardly expressed in the cerebellum, thigh muscle, and breast muscle. The relative expression in the crest cushion and abdominal fat was significantly higher than that in other tissues ($P < 0.01$) (Figure 6G). The results revealed that the mutation in the 5′UTR of *TAS2R40* specifically affected the expression level of *TAS2R40* in the crested tissue of CC ducks. Subsequently, luciferase assay showed that the relative luciferase activity of *TAS2R40* 5′UTR-MT was significantly lower than that of *TAS2R40* 5′UTR-WT ($P < 0.01$) (Figure 6H). A series of results showed that the G > A mutation in the transcription region affected the regulatory effect and reduced its expression in the crested tissue. Combining the aforementioned results, we speculated that the SNP in the 5′UTR of *TAS2R40* was a causative mutation of the crest cushion.

Discussion

Since the first duck draft genome was reported [13], the origin, evolution, domestication, and selection of ducks have been revealed. More importantly, a series of characteristic traits and phenotypes, such as disease resistance, body size, plumage color, and egg color, have been gradually discovered [12,36], providing deep insights into genotype–phenotype associations in animal molecular breeding and germplasm conservation.

During species evolution, directional artificial selection and non-directional natural selection can cause genetic diversity in

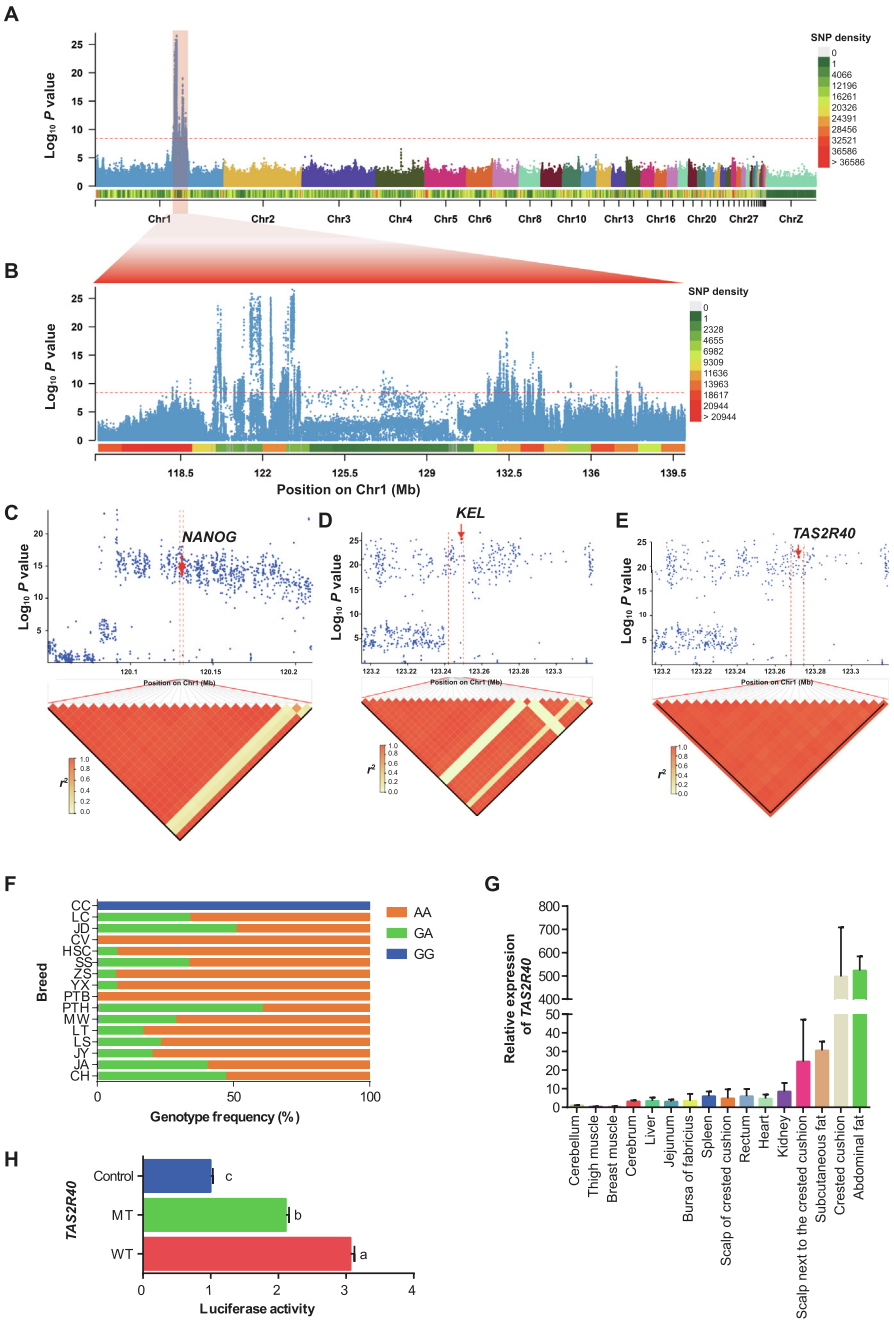

Figure 6 Candidate genes involved in crest cushion formation in crested ducks

A. GWAS of crest traits, including 63 crested ducks and 211 normal ducks. The red horizontal dashed line indicates the Bonferroni significance threshold of 8.39 (0.05/total SNPs) for GWAS. **B.** Magnified view of the peak on Chr1 as shown in (A). **C.**–**E.** Magnified view of 0.2 Mb and 0.1 Mb candidate SNPs for *NANOG* (C), *KEL* (D), and *TAS2R40* (E). The LD heatmap is depicted in red. **F.** Genotype frequency of *TAS2R40* (123272114_c. G78A) in different breeds. LC represents the Lianchen duck; JD represents the Jingding duck; HSC represents the Hese Cai duck; SS represents the Sansui duck; ZS represents the Zhongshanma duck; YX represents the Youxianma duck; PTB represents the Putian white duck; PTH represents the Putian black duck; MW represents the Mawang duck; LT represents the mallard duck; LS represents the Longshencui duck; JY represents the Jingyunma duck; JA represents the Ji'an read feather duck; CH represents the Chaohu duck. **G.** Relative expression of *TAS2R40* in 16 tissues of the CC duck. **H.** Luciferase activities detected after transfection of *TAS2R40* 5′UTR-MT (mutant type) vector and the *TAS2R40* 5′UTR-WT (wild-type) vector. Empty vector was used as a control. Statistical significance is indicated by different lowercase letters. GWAS, genome-wide association analysis.

animals. Adaptive evolution allows animals to acquire certain protective mechanisms that allow the species to continue. Based on the phenotype analysis, we explained the mechanism of crest cushion occurrence at an anatomical level and found that the crest cushion might affect the survival of the CC duck. Theoretically, natural selection promotes the spread of

mutations and removes harmful ones. However, it is not completely effective, and all populations harbor genetic variants with deleterious effects. Human intervention in speciation preservation might maintain the inheritance of harmful mutations and promote the accumulation of beneficial mutations. The results presented herein provide evidence of human intervention leading to genome protection and evolutionary maintenance of species. Considering the SVs, genome evolution-related genes, and gene content enrichment among various birds, there is evidence for genome protection and evolutionary maintenance of species that complement one another. The CC ducks had a greater proportion of genes under adaptive evolution with functions related to tissue repair than the other two ducks.

Crest cushions or crest crowns are conspicuous and diverse features of almost all bird lineages with feather crests and are unique among almost all bird species [37]. The most obvious difference between the CC duck and other existing crested birds is that the crested tissue of the crested duck affects the embryonic development of the crested duck and can even lead to embryonic death. Our results indicate that the crest cushion is caused by the proliferation of relevant cells in the parieto-occipital region during the embryonic stage. This process generates downward pressure, resulting in incomplete closure of the cartilage and, in some crested ducks, likely leading to brain overflow and death as a result of exaggerated crest cushion size. This finding demonstrates that the root cause of crest cushion formation is the rapid proliferation of cells in the parieto-occipital region. Furthermore, we observed that even if some crested duck embryos have a hole in the cartilage, the mortality rate of crested ducks is very low if the scalp heals and adipose tissue compensates for the insufficient brain pressure (**Figure 7**A). Based on the aforementioned results, we propose that the healing of the scalp and the presence of adipose tissue may act as a phenotypic compensation mechanism for crested tissue to reduce the mortality of crested ducks. To reveal the genetic basis, we generated a high-quality chromosome-level CC duck genome. Compared with the recently reported duck genome [12–14,36], the evaluation result of BUSSCO was better than that of other duck genomes. Furthermore, we compared the CC duck genome to other bird genomes and identified some genes related to tumorigenesis. Simultaneously, some immune-related genes in the CC duck genome have also undergone positive selection due to the presence of holes that can cause brain exposure, which is more important for CC duck survival. In addition, we believe that the composition of these phenotypes may be the physiological basis of crest formation under stronger positive conditions. We also identified the genetic basis of crest trait formation and phenotypic composition by inter- and intra-population selective-sweep analyses. We found that 12 CDRs harboring 13 genes were strongly selected in CC ducks. Based on GWAS and experimental evidence, we confirmed that *TAS2R40* might be the most fundamental cause of mortality. We speculate that the 5′UTR mutation of *TAS2R40* may affect the expression of *TAS2R40*, leading to abnormal expansion of certain ectodermal cells in the early embryonic development stage, forming the initial protruding tissue of the crested head, and leading to the occurrence of the crest trait. In addition, ephrin type-b receptor 2 (*EPHB2*), which belongs to the same gene family as ephrin type-b receptor 6 (*EPHB6*) identified here, has proven to be related to the inverse growth of the crest feathers of crested pigeons [23]. Thus, *EPHB6* may control cranial crest feathers to grow clockwise, forming a spherical crested feather phenotype in the CC duck. Thus, the CC duck can form a protective tissue on fragile crested tissue. *NANOG*, which is involved in the development of neural crest stem cells, has been shown to play a role in the pathogenesis of many cancers by regulating cell proliferation, invasion, and metastasis [38,39]. Therefore, we suggest that *NANOG* could be a key gene involved in DNA damage repair and GCR in CC ducks.

Previous studies have shown that all domesticated ducks originated from mallard ducks [12,36]. However, according to the distribution map of the mallard ducks, we found that the mallard ducks exist in two regions of China: the northern and southern groups. However, our data suggest that the CC duck originated from mallard ducks in Zhejiang Province, China, and provide important findings on the history of the CC duck. In recent decades, the CC duck has become endangered, but it has quickly recovered in response to conservation efforts. Our analyses identified 30 candidate genes in the genomic regions under selection in the CC duck domestication process, with most of these genes related to neuron development, response to stress, and response to wounding. Therefore, CC ducks represent a critical example of evolutionary adaptation and genetic compensation.

By comparing the CC duck genome with those of 13 other bird species, we shed new light on how CC ducks likely evolved via the GCR mechanism and propose this breed as a model for studying GCR by natural selection. We found that the four main biological processes were likely co-enriched. The first process involves tumorigenesis and suppression, such as the p53 pathway, the PD-L1 expression and PD-1 checkpoint pathway, and the cellular response to DNA damage stimulus. Interestingly, the p53 pathway is not only enriched for PSGs but also for SV genes and domestication-related genes. The PSG *ATR*, the domestication-related genes *ATM* and *DNM3*, and the pinniped trait candidate gene *NANOG*, as activators of p53 and *GORAB*, participate in the p53 pathway by inhibiting *MDM2* to activate p53. The downstream genes of the p53 pathway, *IGFBP3*, *IGF1*, *FAS*, *PIDD1*, *CASP8*, *BID*, *PERP*, *ZMAT3*, *SIAH1*, and *Apaf1* are involved in the apoptotic process. In addition, *RRM2B* mediates the involvement of *SESN2/3* in DNA repair and damage prevention, whereas *STEAP3* is involved in exosome-mediated secretion (Figure 7B). Based on our observations, we suggest that the root cause of crested head formation may be the short-term rapid proliferation of cranial neural crest cells (similar to local neoplasia). However, with the evolution of the CC duck, the crested duck has evolved a control system that can prevent cells from continuing to proliferate rapidly. Second, some pathways related to tissue repair were enriched, such as cell adhesion molecules and focal adhesions. The genes involved in cell adhesion are also involved in intercellular repair between neuron and Schwan cell, as well as between oligodendrocyte and neuron (growth cone). These genes may control scalp and cartilage healing to prevent encephalocele formation, such as *CADM3*, *CNTN2*, *CNTN1*, *SDC*, *NF186*, *VCAN*, *NFASC*, and *CDH4* (Figure 7C). Third, we identified the genes related to fat synthesis and metabolism. We suspect that the main role of these genes in the brain is participating in the formation of adipose tissue, which compensates for the loss of missing skull intracranial pressure, thus ensuring that crested ducks maintain normal levels of brain pressure.

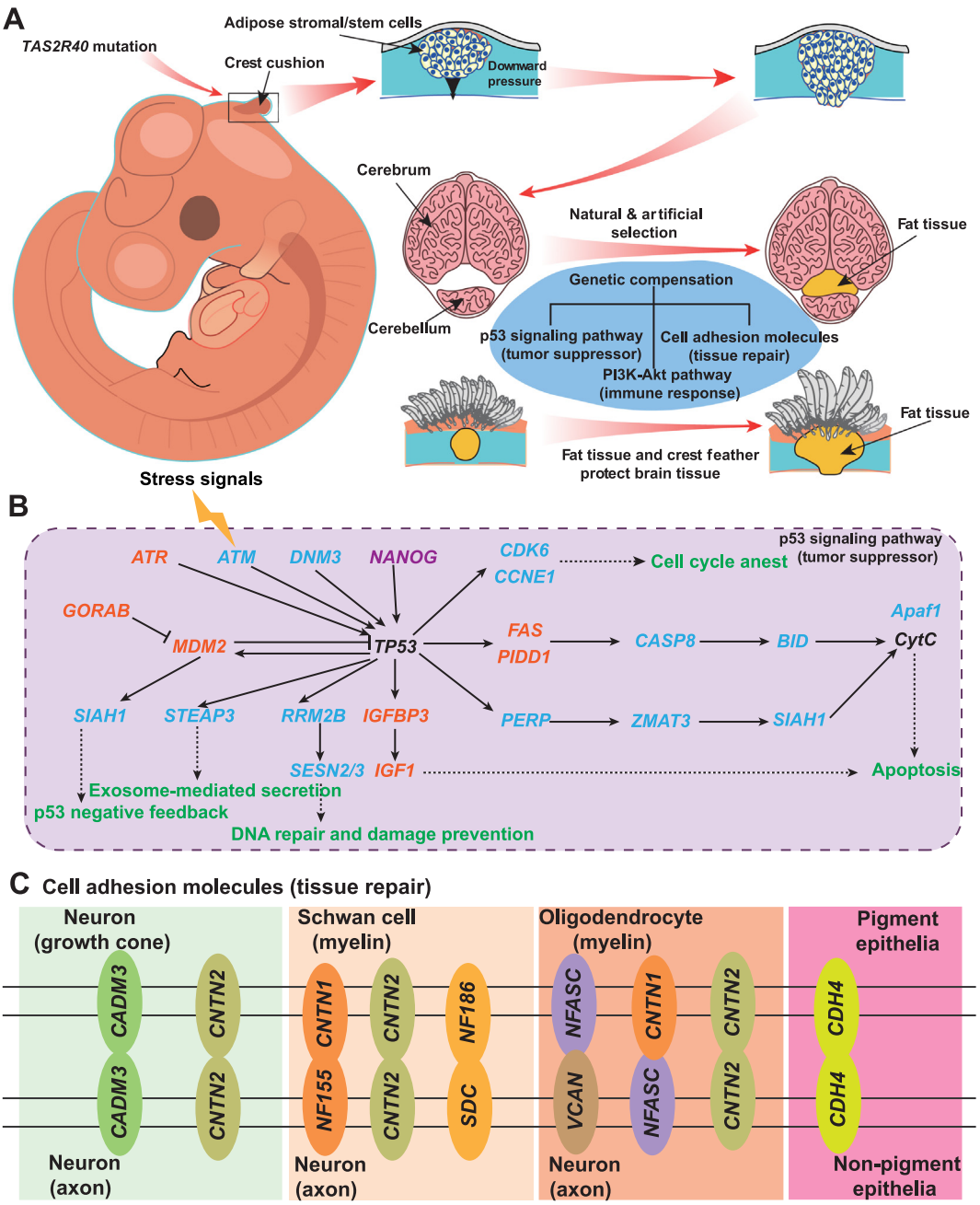

Figure 7 Diagram for crest cushion formation and genetic compensation of CC duck

A. The causative mutation of *TAS2R40* leads to the proliferation of cells in the parieto-occipital region during the embryonic stage. As these cells continue to proliferate, downward pressure is generated on the chondrocyte layer in the parieto-occipital area, forming irregular holes, causing incomplete skull closure, and reducing intracranial pressure. Through a long process of natural and artificial selection, some phenotypic compensation has occurred in the crested tissue. For example, adipose tissue is formed between the cerebrum and the cerebellum to supplement the lost intracranial pressure. Simultaneously, the scalp in the crested area gradually thickens, accompanied by the increase of crested tissue, the change of polarity, and the rapid growth of the top feathers, which protects the leaking brain tissue due to the presence of holes. The most fundamental reason for phenotypic compensation is the selection of pathways, such as cancer suppression (p53 pathway), immunity (PI3K-Akt pathway), and tissue repair (cell adhesion molecules), in the evolution process, which leads to genetic compensation. **B.** Hypothetical model diagram of the genetic compensation mechanism played by the p53 pathway involved in crest cushion formation. PSGs are shown in red, domestication-related genes are shown in blue, and candidate genes for the crest traits are shown in purple. **C.** Hypothetical model diagram of the cell repair pathway involved in phoenix formation.

Fourth, due to the existence of crested tissue, the immune system has also undergone a certain degree of positive selection, such as the immune-related genes enriched in the PI3K-Akt pathway (Figure 7A). In short, the compensatory evolution of a series of genes caused by the occurrence of crest traits has allowed crested ducks to survive and even stabilize the population. Other genes may have evolved due to the accompanying mutations caused by crest traits, incurring GCR and protecting the survival of crested ducks. However, the regulatory relationship of these genes in the mechanism of crest cushion formation remains unclear, and with advances in cell biology, this problem will be gradually solved in the future.

Conclusion

In the present study, we reveal the genetic mechanisms underlying the evolutionary, developmental, and histological origins of the crest trait of CC ducks, and provide insights into the molecular mechanisms of the GCR and its relevance to cancer resistance. The identified genes and their specific mutations provide a starting point for future functional studies of crest cushion development, genetic compensation mechanisms, oncogenesis, and tumor defense.

Materials and methods

Sample preparation and sequencing

A 28-week-old female CC duck from Zhenjiang Tiancheng Agricultural Science and Technology (Zhenjiang, China) was used for genome sequencing and assembly. High-quality genomic DNA was extracted from the blood tissue using a standard phenol/chloroform protocol. A paired-end Illumina sequence library with an insert size of 350 bp and a 10X Genomics linked-read library were constructed and sequenced on the Illumina HiSeq X Ten platform (San Diego, CA). A PacBio library was constructed and sequenced using the PacBio Sequel platform (Menlo Park, CA). RNA-seq libraries for eight tissues (crested tissue, spleen, ovary, liver, duodenum, skin, pectoral, and blood) were constructed and sequenced using Illumina HiSeq4000. Clean reads were assembled using Trinity for gene prediction. In addition, a 28-week-old female Csp-b duck was used for genome sequencing and assembly. Short-insert (250 bp and 350 bp) paired-end libraries and large-insert (2 kb and 5 kb) mate-pair libraries were constructed and sequenced on the Illumina HiSeq4000.

Genome size estimation, assembly, and quality assessment

The genome size of the CC duck genome was estimated based on the *k*-mer distribution using high-quality paired-end reads. The contig assembly of the CC duck was assembled with PacBio reads using FALCON v0.7 [12]. This assembly was polished using Quiver [13] with the default parameters. 10X Genomics was then used to connect contigs to super-scaffolds using FragScaff software [14]. Subsequently, Illumina paired-end reads were used to correct for any errors using Pilon v1.18 [40]. Finally, the scaffolds were anchored and oriented on chromosomes using CHROMONMER v1.07 [18]. A detailed description of the genetic linkage map construction and chromosome anchoring is presented in File S1. To estimate the quality of the final assembly, short paired-end reads were aligned onto the CC duck genome using the Burrows-Wheeler aligner (BWA) with the parameters of '-k 32 -w 10 -B 3 -O 11 -E 4'. BUSCO [16] was used to assess completeness.

Genome annotation

Homology-based and *de novo* predictions were combined to identify repetitive sequences in the CC duck genome. RepeatMasker and RepeatProteinMask (both available from http://www.repeatmasker.org) were used for homologous repeat detection to run against RepBase [41], LTR_FINDER [42], RepeatModeler, and RepeatScout [43] were used to construct a *de novo* repeat library with default settings. Using the *de novo* library, RepeatMasker was run on the CC duck genome. Tandem repeats were identified using TRF v4.07b [44].

Gene prediction was performed using homology-based prediction, *ab initio* prediction, and transcriptome-based prediction. Protein sequences of *A. cygnoides domesticus*, *Aptenodytes forsteri*, *Anas platyrhynchos domestica*, *Coturnix japonica*, *Columba livia*, *E. garzetta*, *G. gallus*, *Homo sapiens*, *Nestor notabilis*, *Struthio camelus*, and *Taeniopygia guttata* were aligned against the CC duck genome using TBLASTN [45]. The blast hits were then conjoined by Solar software [46], and GeneWise [47] was used to predict accurate spliced alignments. For *ab initio* prediction, Augustus [48], Genscan [49], Geneid [50], GlimmerHMM [51], and SNAP [52] were used to predict genes in the repeat-masked genome. RNA-seq data from eight tissues were aligned to the genome using Tophat and Cufflinks [53] to predict gene structures. All predicted genes from the three approaches were integrated using the EvidenceModeler (EVM) [54]. Functional annotation of the predicted genes was carried out using BLASTP against the public databases. To obtain gene functional annotations, KEGG [55], SwissProt [56], and NR databases [57] were used. InterProScan [58] was used to identify domains by searching the InterPro and GO [59] databases.

Comparative genomic analyses

In total, 14 species, including *A. cygnoides domesticus*, *A. forsteri*, *B. regulorum*, *C. japonica*, *C. livia*, CC duck, *E. garzetta*, *G. gallus*, *Gavia stellata*, *N. notabilis*, *O. hoazin*, *P. cristatus*, *S. camelus,* and *T. guttata*, were used for gene family analysis. The longest transcripts of each gene ($>$ 30 amino acids) were retained when a gene had multiple splicing isoforms. 'All-against-all' BLAST v2.2.26 (E-value $\leq$ 1E−7) [45] was used to determine the similarities between the retained genes. OrthoMCL software [22] was used to define the orthologous groups in the aforementioned species with the parameter '-inflation 1.5'. The phylogenetic tree was reconstructed using single-copy orthologs from gene family analysis. Multiple alignments were performed using MUSCLE [60]. The protein alignments were transformed back to coding sequence (CDS) alignments, and then the alignments were concatenated into a super alignment matrix. We constructed a maximum-likelihood phylogenetic tree using RAxML [61]. The mcmctree program from PAML was used for divergence time estimation with eight calibration points from the TimeTree website [62],

and the calibration points are provided in Table S25. We determined the expansion and contraction of orthologous gene families using CAFÉ v1.6 [63] based on a random birth and death model to model gene gain and loss over a phylogeny.

To identify PSGs, all single-copy gene families of five species, including CC duck, *A. cygnoides domesticus*, *G. gallus*, *P. cristatus*, and *A. platyrhynchos domestica*, were used for analysis. Protein-coding sequences were aligned with MUSCLE [64], and the branch-site model of CODEML in PAML was used to identify PSGs by setting the CC duck as the foreground branch. *P* values were calculated using the chi-square test and corrected by the FDR method. Sequence quality and alignment errors have certain influences on the test, so the PSGs with low alignment quality were filtered using the following criteria: (1) FDR > 0.05; (2) presence of gaps near three amino acids around the positively selected sites in the five species. In addition, the kaka_calculator was used to calculate the K_a/K_s ratio [65].

SV detection

We built pairwise local genome alignments between the CC duck and two other duck genome assemblies (*i.e.*, the PK duck and Csp-b duck) using LASTZ v1.04.00 with the parameters 'M = 254, K = 4500, L = 3000, Y = 15000, E = 150, H = 2000, O = 600, and T = 2'. The genomes used for pairwise alignments were soft-masked for repeats using the RepeatMasker software. Then we used "axtChain" to build the co-linear alignment chains and "chainNet" to obtain nets from a set of chains with the default parameters. The "netSyntenic" command was used to add the co-linear information to the nets. The "netToAxt" and "axtSort" were used to convert the net-format to axt-format and change the order of the sequences, respectively. Subsequently, we obtained the best hit for each single location using the utility "axtBest" [66].

SV detection was performed based on the best alignment hits with gapped extensions, which indicated insertions or deletions. In addition, short paired-end reads of the genomes of PK and Csp-b ducks were aligned onto the CC duck genome by BWA [42]. Based on the depth of the reads, we validated our SV detection results. Deletions were defined as gaps with an average depth less than half of the average depth of the whole reference genome and insertions were defined as fragments with an average depth over half of the average depth of the whole assembly. The software source code is available from Li and colleagues [27].

RNA-seq and transcriptomic analysis

Total RNA was extracted from the crest region and adjacent frontal skin tissues of CC ducks and also from the scalps of Cherry Valley ducks using RNAiso Plus reagent (Catalog No. 9109, Takara, Dalian, China) according to the manufacturer's instructions, and 3 μg per sample was used as the input material for RNA sample preparations. The PCR products were purified using an AMPure XP system, and library quality was assessed using an Agilent Bioanalyzer 2100 system. After cluster generation, the library was sequenced using an Illumina HiSeq platform at Novogene Biotechnology (Beijing, China), and 125/150 bp paired-end reads were generated. The quality of the RNA sequences was checked using FastQC, whereas sequence adapters and low-quality reads (read quality < 30) were removed using Trimmomatic v0.36 with parameters 'TRAILING:20' and 'SLIDING WINDOW: 4:15'. The remaining high-quality RNA-seq clean reads were aligned to the corresponding CC duck genome using HISAT2 v2.1.0 with default parameters. FeatureCounts v1.5.0-p3 (parameters: -Q 10 -B -C) was used to count the transcript reads, and StringTie was used to quantify the gene expression levels (fragments per kilobase of transcript per million mapped reads; FPKM) in the detected tissue based on the corresponding transcript annotation. DEGs were identified using negative binomial generalized linear models implemented in DESeq2 v1.20.0. Genes with $P < 0.05$ and $|\log_2$ fold change (FC)$| \geq 1$ were considered DEGs. Hierarchical clustering analysis was performed to determine the variability and repeatability of the samples, and a volcano plot was used to visualize the DEG distribution.

Historical population size estimation

The recent demographic history was inferred from the trends in the *Ne* changes using PopSizeABC v2.1 [67] with default parameters set for the duck population [mutation rate of 7.54×10^{-8} and recombination rate of 1.6×10^{-8}, minor allele count threshold for allele frequency spectrum (AFS) and identity-by-state (IBS) statistics computation = 4, minor allele count threshold for LD statistics computation = 4, and size of each segment = 2,000,000] and 1000 simulated datasets. Summary statistics were extracted using the same parameters, with the tolerance set to 0.05, as recommended.

Alignment and variation calling

A total of 308 samples from GWAS were aligned to the CC duck genome using BWA [68] (settings: mem -t 4 -k 32 -M -R). The sample alignment rates were between 96.00% and 98.00%. The average coverage depth for the reference genome (excluding the *N* region) was between 9.34× and 15.74×, and 4× base coverage (≥ 4) was greater than 82.64%. All the population structure analysis samples were aligned to the CC duck genome using BWA (settings: mem -t 4 -k 32 -M -R), and the sample alignment rate was between 94% and 98.42%. The average coverage depth for the reference genome (excluding the *N* region) was between 6.00× and 17.66×. Variant calling was performed for all samples using the Genome Analysis Toolkit (GATK) v3.7 [69] with the UnifiedGenotyper method. The SNPs were filtered using Perl script. After filtering, the GWAS sample retained 12.6 Mb of SNPs (filter conditions: only two alleles; single-sample quality = 5; single-sample depth = 5–75; total-sample quality = 20; total-sample depth = 308–1,000,000; maximum missing rate of individuals and site = 0.1; and a minor allele frequency = 0.05), and the population genetic analysis retained 5.4 Mb of SNPs (filter conditions: only two alleles; single-sample quality = 5; single-sample depth: 3–75; total-sample quality = 20; total-sample depth: 39–1,000,000; maximum missing rate of individuals and site = 0.1; and a minor allele frequency = 0.05).

Population structure analysis

To clarify the phylogenetic relationship from a genome-wide perspective, an individual-based NJ tree was constructed based

on the p-distance using TreeBeST v1.9.2 [70]; the bootstrap value parameter was 1000. PCA was performed based on all the SNPs using GCTA v1.24.2 [71]. The population genetic structure was examined using an expectation maximization algorithm, as implemented in the program FRAPPE v1.1 [21]. In the population genetic structure analysis, we filtered 5,425,458 SNPs from 36,611,493 SNPs that were filtered by GATK (filter conditions: minor allele frequency = 0.05, maximum missing rate of individuals and site = 0.1, single-sample depth = 3, and single-sample quality = 5). The number of assumed genetic clusters K ranged from 2 to 7, with 10,000 iterations for each run. We compared the patterns of LD using high-quality SNPs. To estimate LD decay, the degree of the LD coefficient (r^2) between pairwise SNPs was calculated using Haploview v4.2, and R v4.1.0 was used to plot LD decay [72]. The program parameters were set as '-n -dprime -minMAF 0.05'. The average r^2 value was calculated for pairwise markers in a 500-kb window and averaged across the whole genome. We found differences in the rate of decay and level of the LD value that reflected variations in the population demographic history and *Ne* among breeds/populations.

We estimated the ancestry of each individual using the genome-wide unlinked SNP dataset, and the model-based assignment software FRAPPE [21] was used to quantify the genome-wide admixture between the wild duck, PK duck, and CC duck populations. FRAPPE was run for each possible group number (K = 2 to 4) with default parameters to estimate the parameter standard errors used to determine the optimal group number (K).

Selective-sweep analysis

To identify the putative selective-sweep regions, the high F_{ST} values [73], high differences in genetic diversity ($\log_2$ θπ ratio), and high XP-EHH values [31] were selected. In brief, we first calculated the F_{ST} and $\log_2$ θπ ratio values in 20-kb sliding windows with 10-kb steps along the autosomes using VCFtools [74]. Then, the F_{ST} values were compared between CC, PK, and Csp-b ducks, as well as between crested and normal ducks in the F_2 population. Subsequently, any windows with fewer than 20 SNPs were filtered out. Moreover, the top 1% of windows or regions with the highest reduction in nucleotide diversity (ROD) values were selected, which represented the extreme ends of the distributions.

GWAS

A case-control GWAS was conducted, including 63 crested ducks [75] and 211 normal ducks (control), involving a total of 12.6 Mb of SNPs. After filtering with PLINK v1.90 [76] ('--geno 0.1 --hwe 1e-05 --maf 0.05 --mind 0.1'), 63 crested ducks and 211 normal ducks with a total of 12.2 Mb of SNPs were used for the subsequent association study. An MLM program, Efficient Mixed-Model Association eXpedited (EMMAX) (beta version) [77], was used to carry out the GWAS. To minimize false positives, the population structure was considered using the top 20 PCA values, which was estimated using PLINK. For the F_2 population, the top 20 PCA values (eigenvectors) were set as fixed effects in the mixed model. The BN kinship matrix was set as a random effect to control for family effects. We defined the whole-genome significance cutoff as the Bonferroni test threshold, which was set as 0.05/total effective SNPs. The GWAS threshold for the crest cushion was 8.38. Manhattan plots and QQ plots of GWAS were produced using the *qqman* package in R v4.1.0 [78].

GO and pathway enrichment analyses using DAVID

The Database for Annotation, Visualization, and Integrated Discovery (DAVID) v6.8 (https://david.ncifcrf.gov) [79] was used to perform GO and KEGG pathway enrichment analyses. The *Bonferroni* method, which is a method of the R/stats package, was used to adjust the P values in GO and KEGG pathway enrichment analyses.

Experimental validation

The genotypes of *TAS2R40*, *KEL*, and *NANOG* among 30 CC ducks and 75 normal ducks from three duck breeds were determined by PCR analysis; the primer sequences (designed by Oligo 6) and annealing temperatures are shown in Table S26. The expression levels of *TAS2R40* in multiple tissues were measured by RT-qPCR (Table S26). Three CC ducks of 56 days of age were slaughtered by stunning and exsanguination. Tissue samples, including the cerebellum, thigh muscle, breast muscle, cerebrum, liver, jejunum, bursa of Fabricius, spleen, the scalp of crested cushion, rectum, heart, kidney, scalp next to the crested cushion, subcutaneous fat, crested cushion, and abdominal fat (50–100 mg), were rapidly collected, snap-frozen in liquid nitrogen, and stored at −80 °C. Glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*) was used as an endogenous control. The collected data were analyzed using the $2^{-\Delta\Delta Ct}$ method. For luciferase activity analysis, wild-type and mutant fragments of the 5′UTR of *TAS2R40* were cloned and inserted between the *Nhe*I and *Xho*I restriction sites of the pGL-Basic 3.0 vector. Empty vector was used as a control. Luciferase activity was measured 36 h after transfection using the dual-luciferase reporter system (Promega, Madison, WI). Firefly luciferase activity was normalized to Renilla luciferase activity.

Ethical statement

All experiments with ducks were performed in accordance with the Regulations on the Administration of Experimental Animals issued by the Ministry of Science and Technology (Beijing, China) in 1988 (last modified in 2001). The experimental protocols were approved by the Animal Care and Use Committee of Yangzhou University, China (Approval No. YZUDWSY2017-11-07). All efforts were made to minimize animal discomfort and suffering.

Data availability

The genome assembly and all of the resequencing data generated in this study have been deposited in the Genome Sequence Archive [80] at the National Genomics Data Center (NGDC), Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (CAS) / China National Center for Bioinformation (CNCB) (GSA: CRA005019), and are publicly accessible at https://ngdc.cncb.ac.cn/gsa. The assembled genome and

gene structures of CC duck and Csp-b duck have been deposited in the Genome Warehouse [81] at the NGDC, BIG, CAS / CNCB (GWH: GWHAAZG00000000 for CC duck, GWHAAZF00000000 for Csp-b duck), and are publicly accessible at https://ngdc.cncb.ac.cn/gwh.

Competing interests

Xiaofang Cao is the current employee of Novogene Co., Ltd. Wangcheng Dai is the current employee of Zhenjiang Tiancheng Agricultural Science and Technology Co., Ltd. All the other authors have declared no competing interests.

CRediT authorship contribution statement

Guobin Chang: Conceptualization, Project administration, Funding acquisition. **Xiaoya Yuan:** Resources, Validation, Resources, Writing – original draft. **Qixin Guo:** Formal analysis, Data curation, Writing – original draft, Writing – review & editing. **Hao Bai:** Formal analysis, Writing – review & editing. **Xiaofang Cao:** Formal analysis, Writing – original draft. **Meng Liu:** Formal analysis, Writing – review & editing. **Zhixiu Wang:** Resources. **Bichun Li:** Writing – review & editing. **Shasha Wang:** Resources. **Yong Jiang:** Resources. **Zhiquan Wang:** Writing – review & editing. **Yang Zhang:** Resources. **Qi Xu:** Investigation. **Qianqian Song:** Resources. **Rui Pan:** Resources. **Lingling Qiu:** Resources. **Tiantian Gu:** Resources. **Xinsheng Wu:** Writing – review & editing. **Yulin Bi:** Writing – review & editing. **Zhengfeng Cao:** Resources. **Yu Zhang:** Resources. **Yang Chen:** Resources. **Hong Li:** Formal analysis. **Jianfeng Liu:** Writing – review & editing. **Wangcheng Dai:** Resources. **Guohong Chen:** Conceptualization, Project administration, Funding acquisition. All authors have read and approved the final manuscript.

Acknowledgments

This work was supported by the China Agriculture Research System (Grant No. CARS-42), the Jiangsu Agricultural Technology System (Grant No. JATS[2020]435), and the Jiangsu Agricultural Science and Technology Innovation Fund (Grant No. CX[18]1004), China. We are deeply grateful to all the donors who participated in this program. We thank Prof. Lizhi Lu from Zhejiang Academy of Agricultural Sciences and Prof. Lujiang Qu from China Agricultural University for providing the next-generation sequencing data of mallard ducks, Prof. Zhuocheng Hou from China Agricultural University for providing the genome annotation file of mallard ducks, Prof. Yunzeng Zhang and Prof. Duonan Yu from Yangzhou University for their suggestions on the data analyses and manuscript writing, and Prof. Zhiqiang Du from Northeast China Agricultural University for helpful suggestions on the F_2 population design. We also thank Qiqi Liang from Novogene Bioinformatics Institute for providing the analysis strategy. Besides, we thank the help form Shenghan Zheng for material collection.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2023.08.002.

ORCID

ORCID 0000-0001-5712-626X (Guobin Chang)
ORCID 0000-0003-0847-2726 (Xiaoya Yuan)
ORCID 0000-0003-3795-2585 (Qixin Guo)
ORCID 0000-0002-5738-4186 (Hao Bai)
ORCID 0000-0002-6972-8687 (Xiaofang Cao)
ORCID 0000-0002-3092-8526 (Meng Liu)
ORCID 0000-0001-5243-735X (Zhixiu Wang)
ORCID 0000-0002-6862-6064 (Bichun Li)
ORCID 0000-0002-2136-6576 (Shasha Wang)
ORCID 0000-0001-5540-377X (Yong Jiang)
ORCID 0000-0002-5813-5368 (Zhiquan Wang)
ORCID 0000-0001-8755-1806 (Yang Zhang)
ORCID 0000-0003-2791-0429 (Qi Xu)
ORCID 0000-0002-3824-7549 (Qianqian Song)
ORCID 0000-0002-9484-8867 (Rui Pan)
ORCID 0000-0002-4228-6540 (Lingling Qiu)
ORCID 0000-0002-4430-9320 (Tiantian Gu)
ORCID 0000-0001-9871-0251 (Xinsheng Wu)
ORCID 0000-0001-8033-9466 (Yulin Bi)
ORCID 0000-0003-3020-5114 (Zhengfeng Cao)
ORCID 0000-0003-3353-4239 (Yu Zhang)
ORCID 0000-0001-9097-5132 (Yang Chen)
ORCID 0000-0003-2193-5721 (Hong Li)
ORCID 0000-0002-5766-7864 (Jianfeng Liu)
ORCID 0000-0002-8703-4734 (Wangcheng Dai)
ORCID 0000-0003-1967-0728 (Guohong Chen)

References

[1] Waddington CH. Canalization of development and genetic assimilation of acquired characters. Nature 1959;183:1654–5.
[2] Grether GF. Environmental change, phenotypic plasticity, and genetic compensation. Am Nat 2005;166:E115–23.
[3] El-Brolosy MA, Stainier DYR. Genetic compensation: a phenomenon in search of mechanisms. PLoS Genet 2017;13:e1006780.
[4] Mather K. Genetical control of stability in development. Heredity 1953;7:297–336.
[5] Rossi A, Kontarakis Z, Gerri C, Nolte H, Holper S, Kruger M, et al. Genetic compensation induced by deleterious mutations but not gene knockdowns. Nature 2015;524:230–3.
[6] Ma Z, Zhu P, Shi H, Guo L, Zhang Q, Chen Y, et al. PTC-bearing mRNA elicits a genetic compensation response via Upf3a and COMPASS components. Nature 2019;568:259–63.
[7] Nedvetzki S, Gonen E, Assayag N, Reich R, Williams RO, Thurmond RL, et al. RHAMM, a receptor for hyaluronan-mediated motility, compensates for CD44 in inflamed CD44-knockout mice: a different interpretation of redundancy. Proc Natl Acad Sci U S A 2004;101:18081–6.
[8] Gao Y, Zhang Y, Zhang D, Dai X, Estelle M, Zhao Y. Auxin binding protein 1 (ABP1) is not required for either auxin signaling or *Arabidopsis* development. Proc Natl Acad Sci U S A 2015;112:2275–80.
[9] Bartels T, Krautwald-Junghanns ME, Portmann SBJ, Kummerfeld N, Sohn HG, Dorsch B. The use of conventional radiography and computer-assisted tomography as instruments for demonstration of gross pathological lesions in the cranium and cerebrum in the crested breed of the domestic duck (*Anas platyrhynchos* f.dom.). Avian Pathol 2000;29:101–8.
[10] Bartels T, Brinkmeier J, Portmann S, Baulain U, Zinke A, Krautwald-Junghanns ME, et al. Magnetic resonance imaging of intracranial tissue accumulations in domestic ducks (*Anas*

platyrhynchos f. dom.) with feather crests. Vet Radiol Ultrasound 2005;42:254–8.
[11] Bartels T, Brinkmeier J, Portmann S, Krautwald-Junghanns ME, Kummerfeld N, Boos A. Osteological investigations of the incidence of cranial alterations in domestic ducks (*Anas platyrhynchos* f. dom.) with feather crests. Ann Anat 2001;183:73–80.
[12] Zhou Z, Li M, Cheng H, Fan W, Yuan Z, Gao Q, et al. An intercross population study reveals genes associated with body size and plumage color in ducks. Nat Commun 2018;9:2648.
[13] Huang YH, Li YR, Burt DW, Chen HL, Zhang Y, Qian WB, et al. The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. Nat Genet 2013;45:776–83.
[14] Li J, Zhang J, Liu J, Zhou Y, Cai C, Xu L, et al. A new duck genome reveals conserved and convergently evolved chromosome architectures of birds and mammals. Gigascience 2021:10:giaa142.
[15] Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods 2016;13:1050–4.
[16] Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 2013;10:563–9.
[17] Adey A, Kitzman JO, Burton JN, Daza R, Kumar A, Christiansen L, et al. *In vitro*, long-range sequence information for *de novo* genome assembly via transposase contiguity. Genome Res 2014;24:2041–9.
[18] Small CM, Bassham S, Catchen J, Amores A, Fuiten AM, Brown RS, et al. The genome of the Gulf pipefish enables understanding of evolutionary innovations. Genome Biol 2016;17:258.
[19] Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;31:3210–2.
[20] Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, et al. *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Res 2010;20:265–72.
[21] Tang H, Peng J, Wang P, Risch NJ. Estimation of individual admixture: analytical and study design considerations. Genet Epidemiol 2005;28:289–301.
[22] Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 2003;13:2178–89.
[23] Shapiro MD, Kronenberg Z, Li C, Domyan ET, Pan H, Campbell M, et al. Genomic diversity and evolution of the head crest in the rock pigeon. Science 2013;339:1063–7.
[24] Lu LZ, Chen Y, Wang Z, Li XF, Chen WH, Tao ZR, et al. The goose genome sequence leads to insights into the evolution of waterfowl and susceptibility to fatty liver. Genome Biol 2015;16:89.
[25] Zhang GJ, Li C, Li QY, Li B, Larkin DM, Lee C, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. Science 2014;346:1311–20.
[26] Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol 2013;30:1987–97.
[27] Chakraborty A, Diefenbacher ME, Mylona A, Kassel O, Behrens A. The E3 ubiquitin ligase Trim7 mediates c-Jun/AP-1 activation by Ras signalling. Nat Commun 2015;6:6782.
[28] Hachem LD, Mothe AJ, Tator CH. The role of TRIM family proteins in the regulation of cancer stem cell self-renewal. Stem Cells 2020;38:187–94.
[29] Liu F, Chen Y, Zhu G, Hysi PG, Wu S, Adhikari K, et al. Meta-analysis of genome-wide association studies identifies 8 novel loci involved in shape variation of human head hair. Hum Mol Genet 2018;27:559–75.
[30] Van Maerken T, Vandesompele J, Rihani A, De Paepe A, Speleman F. Escape from p53-mediated tumor surveillance in neuroblastoma: switching off the p14(ARF)-MDM2-p53 axis. Cell Death Differ 2009;16:1563–72.
[31] Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature 2007;449:913–8.
[32] Lin T, Chao C, Si S, Mazur SJ, Murphy ME, Appella E, et al. p53 induces differentiation of mouse embryonic stem cells by suppressing Nanog expression. Nat Cell Biol 2005;7:165–71.
[33] Li YR, Zheng HC, Luo RB, Wu HL, Zhu HM, Li RQ, et al. Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome *de novo* assembly. Nat Biotechnol 2011;29:723–30.
[34] Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16:284–7.
[35] Supek F, Bosnjak M, Skunca N, Smuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One 2011;6:e21800.
[36] Zhang Z, Jia Y, Almeida P, Mank JE, van Tuinen M, Wang Q, et al. Whole-genome resequencing reveals signatures of selection and timing of duck domestication. Gigascience 2018:7:giy027.
[37] Ng CS, Li WH. Genetic and molecular basis of feather diversity in birds. Genome Biol Evol 2018;10:2572–86.
[38] Lu X, Mazur SJ, Lin T, Appella E, Xu Y. The pluripotency factor nanog promotes breast cancer tumorigenesis and metastasis. Oncogene 2014;33:2655–64.
[39] Huang C, Yoon C, Zhou XH, Zhou YC, Zhou WW, Liu H, et al. ERK1/2-Nanog signaling pathway enhances CD44(+) cancer stem-like cell phenotypes and epithelial-to-mesenchymal transition in head and neck squamous cell carcinomas. Cell Death Dis 2020;11:266.
[40] Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 2014;9:e112963.
[41] Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 2015;6:11.
[42] Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 2007;35:W265–8.
[43] Price AL, Jones NC, Pevzner PA. *De novo* identification of repeat families in large genomes. Bioinformatics 2005;21:i351–8.
[44] Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 1999;27:573–80.
[45] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.
[46] Yu XJ, Zheng HK, Wang J, Wang W, Su B. Detecting lineage-specific adaptive evolution of brain-expressed genes in human using rhesus macaque as outgroup. Genomics 2006;88:745–51.
[47] Birney E, Clamp M, Durbin R. GeneWise and genomewise. Genome Res 2004;14:988–95.
[48] Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res 2005;33:W465–7.
[49] Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol 1997;268:78–94.
[50] Blanco E, Parra G, Guigo R. Using geneid to identify genes. Curr Protoc Bioinformatics 2007;Chapter 4:Unit 4.3.
[51] Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. Bioinformatics 2004;20:2878–9.
[52] Korf I. Gene finding in novel genomes. BMC Bioinformatics 2004;5:59.

[53] Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 2012;7:562–78.
[54] Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol 2008;9:R7.
[55] Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 2000;28:27–30.
[56] Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31:365–70.
[57] Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 2007;35:D61–5.
[58] Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan: protein domains identifier. Nucleic Acids Res 2005;33:W116–20.
[59] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet 2000;25:25–9.
[60] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792–7.
[61] Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 2006;22:2688–90.
[62] Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 1997;13:555–6.
[63] De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study of gene family evolution. Bioinformatics 2006;22:1269–71.
[64] Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 2004;5:1–19.
[65] Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteomics Bioinformatics 2010;8:77–80.
[66] Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, et al. Human–mouse alignments with BLASTZ. Genome Res 2003;13:103–7.
[67] Boitard S, Rodriguez W, Jay F, Mona S, Austerlitz F. Inferring population size history from large samples of genome-wide molecular data – an approximate Bayesian computation approach. PLoS Genet 2016;12:e1005877.
[68] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60.
[69] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297–303.
[70] Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 2009;19:327–35.
[71] Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 2011;88:76–82.
[72] Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21:263–5.
[73] Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. Evolution 1984;38:1358–70.
[74] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics 2011;27:2156–8.
[75] Case LA, Wood BJ, Miller SP. The genetic parameters of feed efficiency and its component traits in the turkey (*Meleagris gallopavo*). Genet Sel Evol 2012;44:2.
[76] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–75.
[77] Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet 2010;42:348–54.
[78] Turner SD. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. Journal of Open Source Software 2018;3:731.
[79] Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009;4:44–57.
[80] Chen T, Chen X, Zhang S, Zhu J, Tang B, Wang A, et al. The Genome Sequence Archive Family: toward explosive data growth and diverse data types. Genomics Proteomics Bioinformatics 2021;19:578–83.
[81] Chen M, Ma Y, Wu S, Zheng X, Kang H, Sang J, et al. Genome Warehouse: a public repository housing genome-scale data. Genomics Proteomics Bioinformatics 2021;19:584–9.