bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Large scale loss-of-function mutations during chicken evolution and domestication

Siwen Wu[1*], Kun Wang[2*], Xuehai Ge[2*], Tengfei Dou[2*], Sisi Yuan[1], Shixiong Yan[2], Zhiqiang Xu[2], Yong Liu[2], Zonghui Jian[2], Jingying Zhao[2], Rouhan Zhao[2], Xiannian Zi[2], Dahai Gu[2], Lixian Liu[2], Qihua Li[2], Dong-Dong Wu[3,4], Junjing Jia[2#], Changrong Ge[2#], Zhengchang Su[1#]

[1]Department of Bioinformatics and Genomics, the University of North Carolina at Charlotte, Charlotte, NC 28078, USA

[2]Faculty of Animal Science and Technology, Yunnan Agricultural University, Kunming, Yunnan, China

[3]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China

[4]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China

[*]These authors contributed equally.

[#]These authors jointly supervised this work and should be addressed to junjingli2009@hotmail.com (JJ), gcrzal@126.com (CG) and zcsu@uncc.edu (ZS).

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Abstract

Despite recent progresses, the driving force of evolution and domestication of chickens remains poorly understood. To fill this gap, we recently sequenced and assembled genomes of four distinct indigenous chickens from Yunnan, China. Unexpectedly, we found large numbers of pseudogenes which have lost their functions and are fixed in their corresponding populations, and we also found highly variable proteomes in the genomes of the four indigenous chickens as well as the sequenced wild red jungle fowl (RJF) (GRCg6a). Although the four indigenous chicken breeds are closely related to the *G. g. spadiceous* subspecies, for the first time, we found that the RJF (GRCg6a) is of the *G. g. bankiva* origin. Thus, the five chicken share the most recent common ancestor (MRCA) before subspeciation. Our results support a scenario that the MRCA of the four indigenous chickens and the RJF possessed at least 21,972 genes, of which 7,993 are dispensable. Each chicken has lost functions of thousands of the dispensable genes during their evolution and domestication via complete gene loss and pseudogenization. The occurring pattens of completely lost genes and pseudogenes segregate the chickens as their phylogenetic tree does. Therefore, loss-of-function mutations might play important roles in chicken evolution and domestication.

Keywords

Chicken; pseudogenization; loss-of-function; domestication; evolution.

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Introduction

As the first non-mammalian vertebrate sequenced [1], chicken (*Gallus gallus*) provides us with most protein sources in our daily life and also is an important model organism to study development and immunity of vertebrates [2]. Since the first release of the draft genome of a red jungle fowl (RJF) [1], its assembly quality has been greatly improved (galgal5 and GRCg6a) [3, 4]. More recently, the Vertebrate Genomes Project (VGP) assembled pseudo-haplotype genomes (GRCg7b and GRCg7w) of a hybrid individual from a broiler mother and a layer father using long sequencing reads and multiple scaffolding data [5, 6]. Studies based on these assemblies have provided insightful understandings of not only the domestication and evolution processes, but also the genetic basis of selected traits of domesticated chickens. However, earlier studies were limited in revealing driving forces of chicken domestication and evolution due to low-quality genome assemblies and insufficient gene annotations. Consequently, contradictory conclusions have been drawn. For example, on the one hand, it was reported that the chicken genome has undergone a large number of segmental deletions [1], resulting in a substantial reduction of genome sizes and a large number of concomitant gene loss; and therefore, chicken might have fewer genes than other tetrapods [7]. On the other hand, it was concluded that selection for loss-of-function mutations had no prominent role in chicken domestication [8, 9]. However, accumulating genomic evidence supports loss-of-function mutations as a major driving force for the evolution (for a review, see [10-12]) of animals [13-17] and plants [18] as well as for the domestication (for a review, see [19, 20]) of many farm animals [21] and crops [22, 23].

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

It has been shown that RJF subspecies *G. g. Spadiceus* is the primary ancestor of domestic chickens (*Gallus gallus domesticus*) all over the world [24]. *G. g. Spadiceus* diverged from other RJF subspecies *G. g. murphy*, *G. g. jabouillei*, *G. g. gallus* and *G. g. bankiva* 50,000-500,000 years ago [24], substantially earlier than the advent of chicken domestication [25]. Indigenous chickens in Yunnan, a southwestern province in China, are among the earliest domesticated birds, and they are formed by less-intensive traditional family-based artificial selection in villages in isolated mountainous areas since 2,000–6,000 BC [25]. It has been estimated that indigenous chickens in Yunnan share the most recent common ancestor (MRCA) with wild *G. g. Spadiceus* less than 8,000 years ago [24]. Although domestic indigenous chicken such as those from southeast Asian and commercial chickens such as white leghorn may have substantial introgression from other RJF subspecies, indigenous chickens in Yunnan have minimal (~4%) introgression only from *G. g. jabouille* [24]. Thus, indigenous chickens in Yunnan are good candidates to investigate the driving force of evolution and domestication of indigenous chickens.

We therefore recently sequenced and assembled genomes at chromosome-level of four indigenous chicken breeds in Yunnan [26]. These chicken breeds include Daweishan chicken with a miniature body size, Hu chicken with a large body size and stout legs, Piao chicken with a rumpless trait and Wuding chicken good at running. Using an annotation pipeline that combines homology-based and RNA-seq-based methods, we found that the Daweishan, Hu, Piao and Wuding chicken genomes encoded 17,718, 17,497, 17,711 and 17,646 protein-coding genes, respectively [27]. Of these genes in the four genomes, a total of 1,420 are not seen in the annotations of the RJF (GRCg6a), the broiler (GRCg7b) and the layer (GRCg7w) assemblies,

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

we thus refer them to as newly annotated genes (NAGs) [27]. Unexpectedly, we also identified a large number of pseudogenes in the Daweishan (747), Hu (606), Piao (682) and Wuding (667) chicken genomes [27]. Interestingly, most of the NAGs also are either encoded or become pseudogenes in the GRCg6a, GRCg7b, and GRCga7w assemblies. We therefore increase the numbers of both annotated genes and pseudogenes in GRCg6a (18,463 and 542), GRCg7b (19,002 and 474) and GRCg7w (18,978 and 435) [27]. In addition to the varying numbers of genes and pseudogenes, each pair of chicken genomes share 81%-92% of their genes, which diverge only 7,000-500,000 years ago. This is in stark contrast to the observation that humans and chimpanzees share 98% of their genes, which split at least 6-7 million years ago [28]. To understand the underlying reasons for such high variation in gene and pseudogene compositions in the chickens, we analyzed the occurring patterns and evolutionary behaviors of the pseudogenes as well as presence and absence variation of genes in the four indigenous chicken genomes and the RJF genome. We did not include the commercial chickens in this analysis as they may possess mosaic genomes of different ancestries inherited from multiple RJF subspecies [24, 29]. Our results suggest that loss-of-function mutations via pseudogenization (contain a premature stop codon or an open-reading-frame (ORF) shift mutation) and complete gene loss (losing all features of a gene beyond to be detected even as a pseudogene) might play critical roles in chicken domestication and evolution.

Results

Chicken genomes harbor highly varying sets of protein-coding genes and pseudogenes

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Although the Daweishan (17,718), Hu (17,497), Piao (17,711) and Wuding (17,646) chicken genomes harbor quite similar numbers of protein-coding genes [27], they shared only 15,050 genes (Figure 1a), comprising only 84.9%-86.0% of their genes. Interestingly, the indigenous chicken genomes encode 745-966 fewer genes than the RJF genome (GRCg6a) (18,463). Consequently, the indigenous chicken genomes share only 13,979 genes with the RJF genome (Figure 1b), comprising only 75.7%-79.9% of their genes. Moreover, each pair of these five chicken genomes share only 84%-92% of their genes (Figure 1c), even though they only diverged 7,000-500,000 years ago [25]. These results indicate that the indigenous chicken breeds and RJF have undergone more dramatic changes in their gene compositions in the last 7,000-500,000 years than have humans and chimpanzees (share 98% of their genes) in the last 6-7 million years [28].

Moreover, we identified a larger number of pseudogenes in each of the four indigenous chicken genomes (606-747) (Table 1) based on three sources [27]. Most (486-622 or 80.2%-83.5%) of the pseudogenes in the indigenous chicken genomes were predicted based on homology to annotated genes in GRCg6a (322-395), GRCg7b (313-385) and GRCg7w (311-398) (Tables 1 and S1-S4). In other words, some functional genes in the reference genomes become pseudogenes in an indigenous chicken genome due to at least a pseudogenization mutation (premature nonsense mutation or open-reading frame (ORF) shift mutation). A small portion (24-35 or 3.6%-5.8%) were predicted based on homology to the 1,420 NAGs, i.e., some functional NAGs in indigenous genomes become pseudogenes in other indigenous genomes. The remaining 83-94 (12.2%-14.0%) were predicted based on homology to previously annotated pseudogenes in GRCg6a (49-57), GRCg7b (55-64) and GRCg7w (51-58). In other

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

words, some pseudogenes in reference genomes also are pseudogenes in indigenous chicken genomes. Most pseudogenes (576-713, 94.9-96.0%) in each indigenous chicken genome are transcribed in multiple tissues (Tables S5-S8), suggesting that their regulatory systems might be still at least partially functional, and thus they might arise quite recently. Furthermore, based on pseudogenization mutations of the 1,420 NAGs in the reference genomes, we increased the numbers of annotated pseudogenes in GRCg6a (from 262 to 542) (Table 1). Notably, the indigenous chicken genomes harbor 64-205 more pseudogenes than GRCg6a, which partially explains why the former genomes harbor fewer genes than the latter genome (17,497-17,716 vs 18,463). The indigenous chickens share 142 pseudogenes (Figure 1d) among themselves, and 33 pseudogenes with the RJF (Figure 1e). Each pair of the chicken share 8%-48% of their pseudogenes (Figure 1f). In total, we found 1,995 pseudogenes that appeared in at least one of the five chicken genomes (Table S9). In summary, these results suggest a highly complex picture of gene presence and absence as well as pseudogenization in various chicken breeds.

Vast majority of pseudogenes are unprocessed and unitary

Most (91.6%-92.8%, 556-684) of the pseudogenes in each indigenous chicken genome are unprocessed (Table 1), i.e., they arose due to direct pseudogenization mutations, while the remaining small portion (7.2-8.4%, 48-63) are processed, i.e., they arose due to retrotransposition followed by pseudogenization mutations [30]. Unprocessed pseudogenization mutations could occur after a duplication event to eliminate a redundant copy [30]. However, we failed to find an intact paralog for most (88.2%-90.3%) of the unprocessed pseudogenes in the same genomes (Tables S1-S4), suggesting that most of the unprocessed pseudogenes are not related to gene duplications, and thus are unitary

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

pseudogenes [31]. There are a total of 1,814 unprocessed pseudogenes in the five chicken genomes (Table S10). The indigenous chickens share 129 unprocessed pseudogenes among themselves (Figure S1a), and 22 with the RJF (Figure S1b). Interestingly, compared to the cases in indigenous chickens, a smaller proportion of the pseudogenes in RJF (85.2%, 462) are unprocessed. However, the number (80) of processed pseudogenes in RJF is similar to those (63, 50, 55 and 48) in the indigenous chickens (Table 1). In the following analyses, we will focus on the unprocessed pseudogenes.

Pseudogenization mutations are biased to the two ends of parental genes

To see whether the arise of the unprocessed pseudogenes in the indigenous chickens is under natural/artificial selection, selectively neutral or a result of random genetic drift, we compared the distribution of their first pseudogenization mutation sites along the CDSs of their parental genes in relevant genome(s) with the distribution of the synonymous mutation sites along the CDSs of true genes. As shown in Figure 2a, synonymous mutations in true genes are largely uniformly distributed along the CDSs as expected for neutral mutations, except at the two ends, where the density decreases, consistent with an earlier report in chickens [32]. The reduced synonymous mutation rates at the two ends suggests that the two ends of CDSs are generally under purifying selection, suggesting that the two ends might harbor functional elements not related to their amino acid coding functions, such as transcriptional and post-transcriptional regulatory elements [33]. Interestingly, the synonymous mutations in the pseudogenes are also largely uniformly distributed along the CDSs including the two ends (Figure 2a), indicating that purifying selection on the two ends of pseudogenes is relaxed. Thus, transcriptional and post-

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

transcriptional regulatory elements at the two ends of pseudogenes might have been deteriorated since their pseudogenization.

By stark contrast, the first pseudogenization mutations in the four indigenous chickens are strongly biased to the two ends of parental CDSs (Figure 2a), consistent with earlier reports in chickens [32] and humans [34]. Specifically, 22.2%, 64.2% and 13.6% of first loss-of-function mutations occur at the first 10%, the middle 80% and the last 10% lengths of the CDSs. Almost all the pseudogenes have their first (Figure 2b) and last (Figure 2c) coding nucleotides aligned with those of parental CDSs, indicating that the biased pseudogenization mutations to the 5'- and 3'-ends are not due to incorrect predictions of the two ends of the pseudogenes. These results strongly suggest that the biased first pseudogenization mutations towards the two ends of pseudogenes are under positive selection.

Biased pseudogenization mutations to the two ends of parental CDSs might facilitate loss of functions of genes

To see whether the first pseudogenization mutation along the parental CDSs result in loss of function of the genes, we compared the evolutionary pressures on true genes with that on pseudogenes in the four indigenous chickens using their ratio of the number of nonsynonymous mutations over the number of synonymous mutations (dN/dS). As shown in Figure 2d, the pseudogenes have significantly higher dN/dS values than the true genes (p<6.13e-295). This is also true when the pseudogenes with the first pseudogenization mutation occurring in the first 10% (p<7.87e-46), the middle 80% (p<1.23e-230) or the last 10% (p<1.93e-30) of CDSs are compared with the true genes (Wilcoxon rank-sum test). These results suggest that at least most pseudogenes are no longer under purifying selection, and thus might lose gene functions.

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Moreover, the pseudogenes with the first pseudogenization mutation occurring in the first 10% of CDSs and occurring in the last 10% of CDSs have similar dN/dS values (p=0.46), and both have significantly lower dN/dS values than the pseudogenes with the first pseudogenization mutation occurring in the middle 80% of CDSs (Figure 2d, p<9.24e-6 and p<7.75e-5, respectively, Wilcoxon rank-sum test). The underlying cause is not clear to us but might be due to our finding that the two ends of CDSs were under purifying selection before pseudogenization events occurred (Figure 2a).

Clearly, the closer a pseudogenization mutation to the 5'-end of a CDS, the larger portion of the peptide chain is affected and the more likely a pseudogenization mutation occurs. Loss of function of a gene could also occur when critical amino acids at the N-terminus of the protein or regulatory DNA elements at the 3'-end of the CDS are disrupted by a pseudogenization mutation. For the former possibility, we noted that the identity of amino acids at the C-terminus of proteins are elevated (Figure 2a), indicating that the C-terminus may harbor critical amino acids. For the later possibility, as the 3'-UTRs of genes often harbor miRNA binding sites for post-transcriptional regulation [35], we hypothesize that 3'-ends of CDSs may also contain miRNA binding sites, and disruption of such sites in either 3'-ends of CDSs or 3'-UTRs may have functional consequence. To test this, we scanned the pseudogenes' and parental genes' CDSs in the four indigenous chickens and their 1,000 bp downstream sequences as putative 3'-UTRs for potential miRNA binding sites. As shown in Figure 2e, putative 3'-UTRs and 3'-ends of both parental genes and pseudogenes have higher density of putative miRNA binding sites than their upstream coding regions, consistent with previously reports [35], suggesting that putative 3'-UTRs and 3'-ends indeed tend to contain a miRNA binding sites.

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Interestingly, pseudogenes have a fewer number of miRNA binding sites in their 3'-ends and 3'-UTRs than do parental genes (Figure 2e), suggesting that pseudogenization mutations might disrupt the miRNA binding sites in the 3'-end of CDSs.

The functions of parental genes of most pseudogenes are lost in the indigenous chicken genomes

To see whether alternative isoforms of the pseudogenes in the four indigenous chickens could skip the exons harboring the pseudogenization mutations, we assembled transcripts of all the transcribed pseudogenes in each indigenous chicken genomes (Table 1). We found that most transcribed pseudogenes had only one type of transcript containing the pseudogenization mutations, while for those that had more than one isoform, very few of them had transcripts that skipped the pseudogenization mutations (Tables S11-S14). For example, in Daweishan chicken, only 139 (20.32%) of the 684 unprocessed pseudogenes (Table 1) have alternative splicing transcripts, and only one of them has transcripts that skip the pseudogenization mutation (Table S11). In Hu chicken, only 137 (24.64%) of the 556 unprocessed pseudogenes (Table 1) have alternative splicing transcripts, and none of them has transcripts that skip the pseudogenization mutations (Table S12). In Piao chicken, only 131 (20.89%) of the 556 unprocessed pseudogenes (Table 1) have alternative splicing transcripts, and only two of them have transcripts that skip the pseudogenization mutations (Table S13). In Wuding chicken, only 139 (22.46%) of the 619 unprocessed pseudogenes (Table 1) have alternative splicing transcripts, and none of them has transcripts that skip the pseudogenization mutations (Table S14). These results suggest that almost all the pseudogenes in the four indigenous chickens did not skip the exons harboring the pseudogenization mutations, and that the functions of

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

parental genes cannot be rescued by alternative isoforms of the pseudogenes. Moreover, as we indicted earlier, most of the unprocessed pseudogenes do not have a functional copy in the same genomes, thus, the functions of their parental genes might be lost in the indigenous chickens.

The GCRg6a assembly might be of an individual of *G. g. bankiva* origin

Although the RJF *G. g. spadiceus* subspecies is believed to be the major ancestor of domestic chickens all over the world [24], no high-quality genome of a *G. g. spadiceus* individual has yet been available. Thus, we would compare the gene compositions in the four indigenous chicken genomes with that of the RJF genome (GRCg6a). To infer the subspecies origin of the RJF individual and the indigenous chickens belonging to, we performed a principal component analysis (PCA) on the SNPs profiles of the RJF individual and populations of the five RJF subspecies (*G. g. Spadiceus, G. g. murphy*, *G. g. jabouillei*, *G. g. gallus* and *G. g. bankiva*) as well as of the four indigenous chickens (Methods and Materials). As expected, individuals of the four indigenous chicken breeds form a compact cluster with those of the *G. g. spadiceus* subspecies (Figure 3a), indicating that the four indigenous chicken breeds are indeed derived from *G. g. spadiceus*. Constituent with a previous report [24], individuals of *G. g. murphy* form a widely spread cluster that cannot be separated from the compact cluster formed by individuals of *G. g. jabouillei* (Figure 3a), suggesting the diversity of the individuals of *G. g. murphy* and possible admixture with *G. g. jabouillei.* Individuals of *G. g. gallus* form a cluster in between the one formed by individuals of *G. g. jabouillei* and the one formed by individuals of *G. g. bankiva.* Consistent with the previous report [24], individuals of *G. g. bankiva* form a cluster that is farthest away from those formed by other subspecies and the indigenous chickens (Figure 3a),

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

indicating that *G. g. bankiva* diverged earliest from the other subspecies. Interestingly, the sequenced RJF (GRCg6a) is separated far away from the cluster formed by the indigenous chickens and *G. g. spadiceus* individuals, and is closest to the cluster formed by *G. g. bankiva* individuals (Figure 3a), suggesting that it might be of *G. g. bankiva* origin. We also analyzed the genetic structures of the chickens (Methods and Materials). In agreement with the PCA result, GRCg6a has highly similar genetic structure to the individuals of *G. g bankiva* (Figures 3b-3d). Both *G. g. murphy* and *G. g. spadiceous* have diverse genetic structures (Figures 3b-3d) due to their broader geographic origins as previously indicted [24]. Hu and Piao chickens have mosaic genetic structures, while Daweishan and Wuding chickens have quite uniform genetic structures (Figures 3b-3d). The four indigenous chicken breeds have large genetic admixture from *G. g. spadiceous* but little from the other subspecies (Figures 3b-3d). These results further strengthen our conclusion that the four indigenous chicken breeds might be mainly derived from *G. g. spadiceous*, while the GRCg6a assembly belongs to an individual of the *G. g. bankiva* origin. The latter conclusion might not be surprising given the fact that the sequenced RJF individual was from of the UCD0001 line that was originated from a wild population from Malaysia [1], where *G. g. bankiva* inhabits [24].

Indigenous chickens and the RJF have lost thousands of genes since their divergence

Based on our aforementioned results, it is reasonable to assume that the indigenous chickens and the RJF share a MRCA A1 before subspeciation, and the indigenous chickens share a MRCA A2 of *G. g. spadiceus* after subspeciation. There are two possible scenarios that the indigenous chickens and the RJF could be derived from the two MRCAs. In one scenario, a MRCA possessed at least the union of genes in the derived chickens plus functional versions of the intersection of

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

unprocessed pseudogenes of all the derived chickens, and the derived chickens selectively lost genes via two unnecessarily exclusive forms of loss-of-function mutations, i.e., complete gene loss and pseudogenization, during the evolution and domestication processes. In the case of the four indigenous chickens, as illustrated in Figure 4a, their MRCA A2 would possess 20,760 (20,631 genes (Figure 1a) + 129 unprocessed pseudogenes (Figure S1a)) genes, and Daweishan, Hu, Piao and Wuding chickens would lose functions of 3,042, 3,263, 3,049 and 3,114 genes, respectively, though pseudogenization (684, 556, 627 and 619) and complete gene loss (2,358, 2,707, 2,422 and 2,495) during their evolution and domestication processes. In the case of the four indigenous chickens and the RJF, their MRCA A1 would possess 21,972 (21,947 genes (Figure 1b) + 25 unprocessed pseudogenes (Figure S1b)) genes, and the RJF and MRCA A2 would lose functions of 3,509 and 1,212 genes, respectively, though pseudogenization (462 and 0) and complete gene loss (3,047 and 1,212) during the subspeciation and evolution processes (Figure 4a). Moreover, from MRCA A1, Daweishan, Hu, Piao and Wuding chicken would lose function of additional 1,212 genes (Figure 4a). This explanation is in agreement with the earlier finding that chicken genome has undergone a large number of segmental deletions, resulting in a substantial reduction of genome sizes and the number of genes [1]. In the other scenario, the MRCA would possesses at most the intersection of genes in the derived chickens, and the derived chickens selectively gain genes during the evolution and domestication processes. In the case of the four indigenous chickens and the RJF, their MRCA A1 would possess at most 13,979 genes, and RJF, Daweishan, Hu, Piao and Wuding chickens would have gained 4,484, 3,739, 3,618, 3,732 and 3,667 genes, respectively, since their divergence (Figure 1b). However,

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

this scenario is unlikely since there is no evidence of a large-scale introgression in the RJF and the indigenous chickens (Figure 1b).

Assuming MRCA A1 of the RJF and four indigenous chickens possessed 21,972 genes (Figure 4a), then, as the five chickens share 13,979 core genes (Figure 1b), the remaining 7,993 would lose functions though complete gene loss or pseudogenization in at least one of the five chickens. Specifically, of the 7,993 dispensable genes, 1,583 (19.8%) are pseudogenized in at least one of the five chickens, while the remaining 6,410 (80.2%) are not pseudogenized in any of the five chickens, but are completely lost in at least one of the five chickens, and we refer them as missing genes for the convenience of discussion.

Loss-of-function mutations affect an array of important biological pathways of chickens

We analyzed the biological functions of the 7,993 dispensable genes in MRCA A1 that are either completely lost (6,410) or pseudogenized (1,583) in at least one of the five indigenous chicken genomes (Figure 4b). To this end, we performed a two-way hierarchical clustering on the 7,993 dispensable genes and the five chickens based on the occurring patterns of these dispensable genes in the five chicken genomes based on Euclid distances using the UPGMA method. The dispensable genes form distinct clusters along the clustering hierarchy (Figure 4b). Based on the distinct features of clusters, we divided them into 31 exclusive clusters as described in Table S15. Although only 1,567 (19.6%) of the dispensable genes have GO term assignments to their functional parental genes in GRCg6a/GRCg7b/GRCg7w or humans, most clusters (27/31, 87.1%) containing genes related to important biological pathways (Figure 4b, Table S15). For instance, cluster 29 containing 241 genes that are completely lost or pseudogenized in Daweishan chicken but functional in other four genomes are involved in 20 pathways for metabolism

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(thiamin metabolism), signal transduction (PI3 kinase pathway, p53 pathway feedback loops 2, etc.), cell growth (EGF receptor signaling pathway, VEGF signaling pathway, PDGF signaling pathway, TGF-beta signaling pathway, FGF signaling pathway), neuronal function (nicotinic acetylcholine receptor signaling pathway, metabotropic glutamate receptor group II pathway, etc.), immunity (Inflammation mediated by chemokine and cytokine signaling pathway), and cardiovascular function (blood coagulation, angiogenesis). Cluster 30 containing 1,071 genes that are functional in all the four indigenous chicken genes but are completely lost or pseudogenized in the RJF are involved in 27 pathways for metabolism (ATP synthesis, de novo purine biosynthesis, glycolysis,), signal transduction (FAS signaling pathway, heterotrimeric G-protein signaling pathway-Gi alpha, etc.), cardiovascular functions (hypoxia response via HIF activation, angiogenesis), reproduction (gonadotropin-releasing hormone receptor pathway), Immunity (T cell activation, interleukin signaling pathway, etc.), cell growth (TGF-beta signaling pathway, EGF receptor signaling pathway, PDGF signaling pathway, etc.), neuronal functions (nicotinic acetylcholine receptor signaling pathway). Cluster 26 containing 217 genes that are completely lost or pseudogenized in Wuding chicken but are functional in other four chickens are involved in 17 pathways for metabolism (glutamine glutamate conversion, mannose metabolism, ATP synthesis), neuronal functions (GABA-B receptor II signaling), Immunity (toll receptor signaling pathway), cell growth (EGF receptor signaling pathway, Endothelin signaling pathway, FGF signaling pathway), reproduction (gonadotropin-releasing hormone receptor pathway), immunity (inflammation mediated by chemokine and cytokine signaling pathway), signal transduction (heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha mediated pathway, etc.), and cardiovascular function (angiogenesis). Cluster 1 containing 1,316

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

genes that are functional in the RJF but completely lost or pseudogenized in all the four indigenous chickens are involved in 26 pathways for Signal transduction (cadherin signaling pathway, Wnt signaling pathway, etc.), neuronal function (alpha adrenergic receptor signaling pathway), cardiovascular function (Hypoxia response via HIF activation), immunity (interferon-gamma signaling pathway, B cell activation, etc.), metabolism (de novo purine biosynthesis, oxidative stress response, etc.), cell growth (FGF signaling pathway, EGF receptor signaling pathway, PDGF signaling pathway), and reproduction (gonadotropin-releasing hormone receptor pathway). Cluster 27 containing 221 genes that are completely lost or pseudogenized in Piao chicken but functional in other four chickens are involved in 17 pathways for metabolism (glutamine glutamate conversion, mannose metabolism, ATP synthesis), neuronal functions (GABA-B receptor II signaling), immunity (toll receptor signaling pathway, inflammation mediated by chemokine and cytokine signaling pathway), signal transduction (heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha mediated pathway), cell growth (EGF receptor signaling pathway, FGF signaling pathway, etc.), reproduction (gonadotropin-releasing hormone receptor pathway), and cardiovascular function (angiogenesis, endothelin signaling pathway). Taken together, the importance of these affected pathways suggests that loss-of-function mutations might have shaped the traits of the indigenous chickens for them to adapt to domesticated conditions and for the RJF to adapt to its unique ecological niche.

Pseudogenes and functional versions of missing genes are preferentially located on micro-chromosomes and have high G/C contents

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

We analyzed the distributions of the 1,583 pseudogenes and the 6,410 missing genes in the chromosomes in each indigenous chicken genome. For missing genes in a chicken genome, we used its functional copy in another chicken for the analysis. Both the pseudogenes and functional versions of the missing genes are located in almost all the chromosomes in each of the four indigenous chickens (Figures 5a, 5b). However, the micro-chromosomes (chr14-chr39) and unplaced contigs have higher densities of both the pseudogenes and functional versions of the missing genes, harboring more than a third of the pseudogenes (39.4%-41.3%, Figure 5c) and more than half of functional versions of the missing genes (51.5%-54.6%, Figure 5d) while only comprising 13.5%-14.4% of the genomes. Likewise, both the ratio of the number of pseudogenes/the number of genes (Figure 5e) and ratios of the number of functional versions of missing genes/the number of genes (Figure 5f) tend to be higher on the micro-chromosomes and unplaced contigs. Both the pseudogenes and functional versions of the missing genes (in other chickens) have a significantly higher G/C contents than true genes for each of the four indigenous chicken genomes (Figure 5g). Interestingly, the pseudogenes exhibit even significantly higher G/C contents than functional versions of the missing genes for all the four chicken genomes except for the Hu chicken genome. The same results were seen when the analyses were done separately on macro-chromosomes (chr1-chr5 and chrZ) (Figure S2a), middle-chromosomes (chr6-chr13 and chrW) (Figure S2b) and micro-chromosomes (chr14-chr39) (Figure S2c) to account for their different G/C contents [26]. It is not clear to us why functional copies of the missing genes have elevated G/C contents compared to true genes in the chicken genome. However, the elevated G/C contents in the pseudogenes compared with those in true genes and functional versions of the missing genes might be due to G/C-biased

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

gene conversion during miotic recombination and DNA repairing [36] after purifying selection on the pseudogenes were relaxed.

Most pseudogenes arise recently and are fixed in respective populations

The vast majority (93%-97%) of the pseudogenes in each of the four indigenous chicken genomes share more than 95% of sequence identity with their parent genes (Figures 6a-6d), indicating that they arose quite recently. Only a small portion (0%-1%) with less than 80% of sequence identity with their functional parental genes arose a relatively long time ago. To see whether the first pseudogenization mutation along a pseudogene in the four indigenous chicken genomes are fixed or not in their respective populations, we computed the frequencies of the mutated alleles in the populations of the four breeds based on DNA re-sequencing reads (Materials and methods). As shown in Figures 6e-6h, most (~75%) of the mutations in the pseudogenes in each chicken genome are fixed or nearly fixed (allele frequency > 80%) in their respective populations. The high probability of fixation of the pseudogenes suggests that they might be under positive selection. A few examples of fixed or nearly fixed unitary pseudogenes in indigenous chicken populations are shown in Figure S3.

Most missing genes have residual sequences in the indigenous chicken genomes but lose gene features in respective populations

To see whether the missing genes in the assembled genomes also are completely lost in their respective populations, we mapped re-sequencing DNA short reads from the individual chickens of a breed population to the functional versions of the missing genes in the corresponding assembly (Materials and methods). As shown in Figures 6i-6l, in each of the four indigenous chicken genomes, vast majority (99.9%) of the missing genes still have residual

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

sequences in the individual genomes, covering up to 90% of their functional versions, while only few missing genes lack detectable residual sequences in the populations. These results strongly suggest that most of the missing genes were once functional in the ancestors, but lost gene features beyond recognitions recently. The missing genes might have lost function earlier than the pseudogenes as the former have lost the features to be predicted as even pseudogenes. Moreover, most (99%) of the missing genes in assembled indigenous chicken genome have a missing frequency > 99% in their respective populations, and only few are detected in all the re-sequenced individuals in respective populations (Figures 6m-6p). The high missing rate of the missing genes in the relevant populations suggests that loss-of-function (null) mutations are fixed in the populations, and thus might be under positive selection. Two examples of fixed loss-of-function mutations of missing genes in indigenous chicken populations are shown in Figures S4a and S4b. In both cases, the residual sequences cover different parts of the functional versions of the missing genes in different breeds, with a missing match rate > 11% and all containing gaps (Figure S4).

Occurring patterns of loss-of-function mutations reflect evolutionary history of the chickens

The patterns of loss-of-function mutations of genes in the chicken genomes might provide hints to mutation orders during chicken evolution and domestication. Most pseudogenes in the GRCg6a genome are completely lost in the four indigenous chicken genomes (Figure 4b), suggesting that the indigenous chickens completely lost these genes after their divergence from the MRCA A1, during their domestication process and/or subspeciation of *G.g. spadiceous*, while GRCg6a is still in the process of completely losing these genes. In contrast, most pseudogenes in the four indigenous chicken genomes have intact copies in GRCg6a (Figure 4b),

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

suggesting that these GRCg6a genes might lose their functions in the indigenous chickens after their separation from MRCA A1, during the domestication process and/or subspeciation of *G.g. spadiceous*, and might be in the process of being completely lost.

To see whether the occurring patterns of the complete gene loss and pseudogenization in the five chicken genomes reflect their evolutionary relationships, we constructed a neighbor-joining (NJ) phylogenetic tree using the occurring patterns in the five chicken genomes of the 7,993 MRCA A1 genes that are either completely lost or pseudogenized in at least one of the five indigenous chicken genomes. As shown in Figure 7a, consistent with the UPGMA tree rooted with the RJF (Figure 4b), Wuding and Piao chickens form a clade that is joined by Daweishan chicken, and the resulting cluster is joined by Hu chicken. The tree is also consistent with the NJ tree constructed using 6,744 essential protein-coding genes in the five chicken genomes and quail (*Coturnix jcponica*) with quail as the root (Figure 7b). Therefore, the occurring patterns of complete gene loss and pseudogenization in the five chickens segregate them in the way by their evolutionary relationships, and thus, reflect their evolutionary relationships. This result is in contrast to the earlier reports that loss-of-function mutations failed to segregate between wild and domestic chickens, and thus selection for loss-of-function mutations had a little role in chicken domestication [8, 9].

Discussion

We find larger numbers of pseudogenes in the four indigenous chicken genomes as well as in the RJF genome (GRCg6a). Most of the pseudogenes in each chicken genome are unprocessed and unitary, while only a small number of them are processed. This finding is consistent with

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

the previous results [37, 38], presumably because the chicken's LINE1 like CR1 (L1) elements lack retro transposase activity [37, 38]. However, the large number of unprocessed pseudogenes that we found in each chicken genome is in stark contrast to the findings in humans and mice. For example, a previous study found that only a few dozen unprocessed pseudogenes were found in human population [39, 40], not to mentioning in a human individual. In a more recent study [41], 165 and 303 unprocessed pseudogenes were found in large mouse and human populations, respectively. However, these numbers are still much smaller than those (1,995) that we found in only five chicken genomes. Thus, we observed a larger scale of unprocessed pseudogenization in chickens than in mice and humans.

Our results strongly suggest that most of the pseudogenes lose their protein-coding functions. First, the pseudogenes have elevated G/C contents and higher dN/dS ratios compared with true genes, no matter where the first pseudogenizations occur along the CDSs of their parental genes, indicating that they are no longer under purifying selection. Second, in true genes, synonymous mutations are largely uniformly distributed along the CDSs, but decrease at the two-ends of the CDSs, suggesting that both ends are under purifying selection. However, such purifying selection is relaxed on the two ends of pseudogenes. Third, although most pseudogenes have transcripts, very few have isoforms that skip the exon with the first pseudogenization mutations. Finally, there are two scenarios for a pseudogene to arises: 1) when the function of the gene is no longer needed; and 2) after a gene duplication event, removal of a redundant copy is beneficial. We found that most of the pseudogenes in the chickens are unitary, and thus, were not related to gene duplications. Therefore, functions of most of parental genes are lost in the genomes that harbor the pseudogenes.

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Moreover, we find that the compositions of protein-coding genes in the five chicken genomes are highly variable even in the indigenous chicken genomes that harbor similar number of genes. For example, even though the Daweishan and Piao chicken genomes encode almost the same number of genes (17,718 vs 17,711), they share only 91% of their genes even though they have diverged for only a few thousand years [24]. These results are in stark contrast with the observation that humans and chimpanzees share almost the same sets (98%) of genes although they split at least 6-7 million years ago [28]. Both the unexpectedly large number of pseudogenes and the highly variable compositions of the proteomes in the five chicken genomes strongly suggest that chickens have undergone dramatic changes in their proteomes in the last 7,000-500,000 years of evolution and domestication.

We confirm that the four indigenous chickens from Yunnan province are mainly derived from the *G. g. spadiceous* subspecies, and infer for the first time that the sequenced RJF (GRCg6a) is of *G. g. bankiva* origin. Thus, the indigenous chickens and the RJF might share their MRCA A1 ~500,000 years ago, before their subspeciation [24]. There are two possible scenarios that the highly variable proteomes in the five chicken genomes could arise: 1) their MRCA A1 only possessed the intersection of their genes, and each derived chicken selectively gained thousands of new genes; and 2) their MRCA A1 harbored the union of their genes plus functional versions of the intersection of their unprocessed unitary pseudogenes genes, and each derived chicken selectively lost thousands of genes. Our results are against scenario 1 but favor scenario 2. First, the RJF and the four indigenous chicken have little gene introgression from other RJF subspecies (Figures 3b-3d). Thus, it is unlikely that the five chickens have gain large numbers of genes from other subspecies. Second, the hundreds of unprocessed unitary

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

pseudogenes in each chicken genome arose quite recently, and are still in the process of being completely lost, and they might be once functional in recent ancestors. Third, although all the missing genes in a genome have lost all gene features, most of them have residual sequences left in the genomes, strongly suggesting that they were also once functional in recent ancestors. The missing genes might have lost functions earlier than the pseudogenes, because the former have lost the gene features, while the latter still possess few to be recognized as pseudogenes. In addition, it has been shown that chicken genome has undergone a large-scale of segmental deletions, resulting in a substantial reduction of the number of genes [1].

Although 7,993 of the 21,947 genes estimated in the MRCA A1 of the RJF and the four indigenous chicken genomes are dispensable, many of them are involved in important biological processes. Thus, their selective retainment or loss in a genome might be beneficial to the chicken in its unique natural or domestic conditions. More specifically, although the thousands of genes in the RJF genome that are either completely lost or pseudogenized in the four indigenous chicken genomes might be essential for RJF to live in its unique ecological niche, loss-of-function mutations of these genes in the indigenous chickens might be beneficial for them to live in domestic conditions. Similarly, although the thousands of genes in the four indigenous chicken genomes that are either completely lost or pseudogenized in the RJF genome might be essential for the indigenous chickens to live in their domestic conditions, loss-of-function mutations of these genes in the RJF might be beneficial for it to live in its unique ecological niche.

Our results strongly suggest that loss-of-function mutations via complete gene loss and pseudogenization are a result of natural/artificial selection. First, unlike synonymous mutations

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

along the true genes and pseudogenes, which are largely uniformly distributed along the CDSs as expected for neutral mutations, the first pseudogenization mutations in pseudogenes are strongly biased to the two ends of parental CDSs, particularly, the 5'-ends. Such biases would facilitate eliminating the functions of parental genes. It is well-known that a promoter can extend into the 5'-end of the CDS, thus mutations in the region may disrupt the promoter of the gene [42]. Moreover, the closer a pseudogenization mutation is toward the 5'-end of the CDS, the greater impact of the mutation could have on the gene function and the more likely the gene would lose its function. Although pseudogenizations at the 3'-ends of CDSs can potentially produce at least partially functional proteins, this is unlikely for at least most of the pseudogenes that we found in the indigenous chicken genomes. This is because dN/dS ratios for pseudogenes with the first pseudogenization sites occurring in the last 10% and in the first 10% of the CDSs are not significantly different, but both are significantly higher than those for true genes. In other words, pseudogenizations in the 3'-ends of CDSs can be as effective as pseudogenizations in the other parts of the CDSs to eliminate the functions of genes. We found that 3'-ends of CDSs might harbor miRNA binding sites, and pseudogenization could disrupt such binding sites, which might change post-transcriptional regulation, and thus, the functions of genes.

Second, most pseudogenization mutations are fixed in the indigenous chicken populations (Figures 6e~6h), and thus the mutations are likely under positive selection, although the selection on the other parts of the pseudogenes is relaxed. This is possible since most of the pseudogenes arose quite recently as indicated by their high sequence identity with normal copies (Figures 6a~6d). Thus, unlike completely lost gene, pseudogenes have not had

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

enough time to be fully degraded after they are no longer under negative selection pressure. Of course, with time the pseudogenes without any other functions will be eventually degraded beyond recognition. Third, most of the missing genes in each assembled indigenous chicken genomes also are missing in the corresponding population (Figures 6m~6p), i.e., the null alleles are fixed, and thus are likely under positive selection. Finally, the occurring patterns of the 7,993 dispensable genes in the MRCA A1 segregate the four indigenous chickens and the RJF (Figure 7a) in the exactly same way as the phylogenetic tree of the chickens constructed using more than 6,000 essential avian protein-coding genes (Figure 7b). Taken together, these results strongly suggests that loss-of-function mutations via pseudogenization and complete loss of thousands of genes in RJF and the indigenous chickens since their divergence play critical roles in chicken evolution and domestication. This conclusion is in contrast to an earlier report that loss-of-function mutations play a little role in chicken domestication [8]. Complete gene loss and pseudogenization are not unnecessarily exclusive forms of loss-of-function mutations. One a gene is pseudogenized, it will be rapidly degraded as purifying selection on the pseudogene is relaxed (Figure 2d).

It is worth pointing out that although it has been shown that deleterious mutations might play roles in the domestication of plants [43, 44] and animals [45, 46], lost-of-function mutations are not necessarily deleterious. In fact, it has been well documented that loss of certain genes might be the results of adaptation of birds for flight [17, 47-49], of beef cattle for meat production [21], and of humans for new abilities [39]. It has been proposed that loss-of-function mutations may be important factors in rapid evolution as occurred during domestication—the “less is more” hypothesis [20], which has since gained substantial evidence

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

supports [10, 11, 19, 50-57], including the data that we present in the current study. Thus, the earlier conclusion that fixation of null alleles is not a common mechanism for phenotypic evolution in chicken domestication [8, 9] might be incorrectly drawn because of the low quality of earlier chicken genome assemblies, leading to the failure to detect inactivating mutations such as large scale pseudogenizations and high variation of gene presence and absence [58].

Materials and Methods

Chicken population: The GRCg6a and quail (*coturnix japonica*) genomes and annotation files were downloaded from the NCBI Genbank with accession numbers GCF_000002315.6 and GCF_001577835.2, respectively. Our previously assembled four indigenous chicken genomes were downloaded from the NCBI Genbank with the BioProject number PRJNA865263. All the Illumina short DNA sequencing reads and RNA-seq reads of different tissues of the four indigenous chickens were downloaded from the NCBI SRA database with accession number PRJNA865247.

We downloaded the re-sequencing data of different RJF subspecies from the ChickenSD database (http://bigd.big.ac.cn/chickensd/) with accession numbers listed in [24]. All the re-sequencing data of the indigenous chickens were downloaded from the NCBI SRA database with the accession number PRJNA893352. The sequences of 8,338 essential avian proteins were obtained from the BUSCO aves_odb10 database [59].

Protein-coding gene and pseudogene annotation: We used a combination of homology-based and RNA-seq-based method to annotate the protein coding genes and pseudogenes as previously described [27].

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Single nucleotide variants calling: We mapped short DNA reads from each individual chicken to the GRCg7b reference genome using Bowtie (2.4.1), and called SNVs and small indels in each individual chicken using GATK (4.1.6) [60].

Calculation of alle frequencies of pseudogenes: We computed allele frequencies of the first pseudogenization mutation of each pseudogene in each chicken breeds using GATK (4.1.6) [60] based on call SNVs and indels.

Neighbor-joining tree construction: We mapped the 8,338 essential avian proteins to each of the five chicken's CDSs as well as the quail's CDSs using blastx (2.11.0) [61]. We selected the 6,744 genes with greater than 70% sequence identity with the essential avian proteins in each of the six genomes to construct a neighbor-joining tree. Since it is hard to make multiple alignments for very long sequences, we evenly divided the genes in each bird into 68 groups (each contains about 100 genes). We then aligned sequences of the same group in the six birds using Clustal Omega (1.2.4) [62]. We finally concatenated the 68 multiple alignments with a fixed order and constructed a consensus neighbor-joining trees with 1,000 rounds of bootstrapping using Phylip (3.697) [63]. The 6,744 essential avian genes that we used to construct the tree is listed in Table S16.

PCA and population structure analysis: We used the SNPs called in each individual chicken of each population to perform the PCA and population genetic structure analysis. PCA was performed using PLINK (1.90) [64] with the default settings, and population genetic structure analysis was inferred using ADMIXTURE (1.3.0) [65] with K=2, 3, …, 15 using the default settings.

Prediction of miRNA binding sites: For each pair of pseudogene and its parental gene, we scanned their CDSs and 1,000 bp downstream sequences as putative 3'-UTRs for miRNA binding

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

sites using RNAhybrid (2.1.2) [66]. The miRNAs predicted in the genome harboring the pseudogene are used as the database for the scanning. We consider the putative binding sites with a p-value<0.05.

Calculation of gene body coverage ratio and missing rates of missing genes: For each assembled indigenous chicken breed genome, we collected functional version (reference genes) of its missing genes from either the RJF genome (GRCg6a) or any of the other three indigenous chicken genomes. We mapped the re-sequencing short reads of each individual chicken of each breed (n= 25, 10, 23 and 23 for Daweishan, Hu, Piao and Wuding, respectively) to the reference genes for the breed using Bowtie (2.4.1) [67] allowing no mismatch and gaps. For each missing gene in the assembled genome of a breed, we computed the gene body coverage ratio as the average length of the reference gene body covered by reads among all the individuals of the breed over the length of the reference gene body. We also computed missing rate for each missing gene in the assembled genome of a breed as the ratio of the number of individuals whose re-sequencing reads cannot fully cover the reference gene body over the number of total individuals of the breed.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (U2002205 and U1702232), Yunling Scholar Training Program of Yunnan Province (2014NO48), Yunling Industry and Technology Leading Talent Training Program of Yunnan Province (YNWR-CYJS-2015-027), Natural Science Foundation of Yunnan Province (2019IC008 and 2016ZA008), and Department of Bioinformatics and Genomics of the University of North Carolina at Charlotte.

Author contributions

JJ, CG and ZS supervised and conceived the project; KW, XG, TD, SY^2, ZX, YL, ZJ, JZ, RZ, XZ, DG, LL, QL and DW collected tissue samples and conducted molecular biology experiments; SW and SY^1 assembled and corrected the genomes; SW and ZS performed data analysis; and SW and ZS wrote the manuscript.

Data availability

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

The annotation code and pipeline description are available at https://github.com/zhengchangsulab/A-genome-assebmly-and-annotation-pipeline.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. *Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.* Nature, 2004. **432**(7018): p. 695-716.
2. Burt, D.W., *The chicken genome and the developmental biologist.* Mech Dev, 2004. **121**(9): p. 1129-35.
3. Schmid, M., et al., *Third Report on Chicken Genes and Chromosomes 2015.* Cytogenet Genome Res, 2015. **145**(2): p. 78-179.
4. Warren, W.C., et al., *A New Chicken Genome Assembly Provides Insight into Avian Genome Structure.* G3 (Bethesda), 2017. **7**(1): p. 109-117.
5. Rhie, A., et al., *Towards complete and error-free genome assemblies of all vertebrate species.* Nature, 2021. **592**(7856): p. 737-746.
6. Smith, J., et al., *Fourth Report on Chicken Genes and Chromosomes 2022.* Cytogenet Genome Res, 2022. **162**(8-9): p. 405-528.
7. Lovell, P.V., et al., *Conserved syntenic clusters of protein coding genes are missing in birds.* Genome Biol, 2014. **15**(12): p. 565.
8. Rubin, C.J., et al., *Whole-genome resequencing reveals loci under selection during chicken domestication.* Nature., 2010. **464**(7288): p. 587-91. doi: 10.1038/nature08832. Epub 2010 Mar 10.
9. Wong, G.K., et al., *A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms.* Nature., 2004. **432**(7018): p. 717-22.
10. Albalat, R. and C. Cañestro, *Evolution by gene loss.* Nat Rev Genet, 2016. **17**(7): p. 379-91.
11. Murray, A.W., *Can gene-inactivating mutations lead to evolutionary novelty?* Curr Biol, 2020. **30**(10): p. R465-r471.
12. Monroe, J.G., et al., *The population genomics of adaptive loss of function.* Heredity (Edinb), 2021. **126**(3): p. 383-395.
13. Steinworth, B.M., M.Q. Martindale, and J.F. Ryan, *Gene Loss may have Shaped the Cnidarian and Bilaterian Hox and ParaHox Complement.* Genome Biol Evol, 2023. **15**(1).
14. Zhao, X.W., et al., *Massive Loss of Transcription Factors Promotes the Initial Diversification of Placental Mammals.* Int J Mol Sci, 2022. **23**(17).
15. Zheng, Z., et al., *Gene losses may contribute to subterranean adaptations in naked mole-rat and blind mole-rat.* BMC Biol, 2022. **20**(1): p. 44.
16. Arunkumar, R., et al., *Natural selection has driven the recurrent loss of an immunity gene that protects Drosophila against a major natural parasite.* Proc Natl Acad Sci U S A, 2023. **120**(33): p. e2211019120.
17. Haimson, B., et al., *Natural loss of function of ephrin-B3 shapes spinal flight circuitry in birds.* Sci Adv, 2021. **7**(24).

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

18. Caseys, C., *Loss of Function, a Strategy for Adaptation in Arabidopsis.* Plant Cell, 2019. **31**(5): p. 935.

19. Xu, Y.C. and Y.L. Guo, *Less Is More, Natural Loss-of-Function Mutation Is a Strategy for Adaptation.* Plant Commun, 2020. **1**(6): p. 100103.

20. Olson, M.V., *When less is more: gene loss as an engine of evolutionary change.* Am J Hum Genet, 1999. **64**(1): p. 18-23.

21. Grobet, L., et al., *Molecular definition of an allelic series of mutations disrupting the myostatin function and causing double-muscling in cattle.* Mamm Genome, 1998. **9**(3): p. 210-3.

22. Wei, S., et al., *A loss-of-function mutant allele of a glycosyl hydrolase gene has been co-opted for seed weight control during soybean domestication.* J Integr Plant Biol, 2023. **65**(11): p. 2469-2489.

23. Torkamaneh, D., et al., *Identification of candidate domestication-related genes with a systematic survey of loss-of-function mutations.* Plant J, 2018. **96**(6): p. 1218-1227.

24. Wang, M.S., et al., *863 genomes reveal the origin and domestication of chicken.* Cell Res, 2020. **30**(8): p. 693-701.

25. Miao, Y.W., et al., *Chicken domestication: an updated perspective based on mitochondrial genomes.* Heredity (Edinb), 2013. **110**(3): p. 277-82.

26. Wu, S., et al., *High quality assemblies of four indigenous chicken genomes and related functional data resources.* Sci Data, 2024. **11**(1): p. 300.

27. Siwen Wu, T.D., Sisi Yuan, Shixiong Yan, Zhiqiang Xu, Yong Liu, Zonghui Jian, Jingying Zhao, Rouhan Zhao, Xiannian Zi, Dahai Gu, Lixian Liu, Qihua Li, Dong-Dong Wu, Zhengchang Su, Junjing Jia, Changrong Ge, Kun Wang, *Annotations of four high-quality indigenous chicken genomes identify more than one thousand missing genes in sub-telomeric regions with high G/C contents.* BioRxiv, 2024.

28. Consortium, C.S.a.A., *Initial sequence of the chimpanzee genome and comparison with the human genome.* Nature, 2005. **437**(7055): p. 69-87.

29. Guo, Y., et al., *Researching on the fine structure and admixture of the worldwide chicken population reveal connections between populations and important events in breeding history*. Evol Appl, 2022. **15**(4): p. 553-564.

30. Hurles, M., *Gene duplication: the genomic trade in spare parts.* PLoS Biol, 2004. **2**(7): p. E206.

31. Zhang, Z.D., et al., *Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates.* Genome Biol, 2010. **11**(3): p. R26.

32. Derks, M.F.L., et al., *A survey of functional genomic variation in domesticated chickens.* Genet Sel Evol, 2018. **50**(1): p. 17.

33. Parmley, J.L. and L.D. Hurst, *How do synonymous mutations affect fitness?* Bioessays, 2007. **29**(6): p. 515-9.

34. Ng, P.C., et al., *Genetic variation in an individual human exome.* PLoS Genet, 2008. **4**(8): p. e1000160.

35. Jonas, S. and E. Izaurralde, *Towards a molecular understanding of microRNA-mediated gene silencing.* Nat Rev Genet, 2015. **16**(7): p. 421-33.

36. Duret, L. and N. Galtier, *Biased gene conversion and the evolution of mammalian genomic landscapes.* Annu Rev Genomics Hum Genet, 2009. **10**: p. 285-311.

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

37. *Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.* Nature., 2004. **432**(7018): p. 695-716.
38. Yu, Z., et al., *Analysis of the role of retrotransposition in gene evolution in vertebrates.* BMC Bioinformatics, 2007. **8**: p. 308.
39. Wang, X., W.E. Grus, and J. Zhang, *Gene losses during human origins.* PLoS Biol, 2006. **4**(3): p. e52.
40. MacArthur, D.G., et al., *A systematic survey of loss-of-function variants in human protein-coding genes.* Science, 2012. **335**(6070): p. 823-8.
41. Sisu, C., et al., *Transcriptional activity and strain-specific history of mouse pseudogenes.* Nat Commun, 2020. **11**(1): p. 3695.
42. Aguezzoul, M., A. Andrieux, and E. Denarier, *Overlap of promoter and coding sequences in the mouse STOP gene (Mtap6).* Genomics, 2003. **81**(6): p. 623-7.
43. Renaut, S. and L.H. Rieseberg, *The Accumulation of Deleterious Mutations as a Consequence of Domestication and Improvement in Sunflowers and Other Compositae Crops.* Mol Biol Evol, 2015. **32**(9): p. 2273-83.
44. Lu, J., et al., *The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication.* Trends Genet, 2006. **22**(3): p. 126-31.
45. Cruz, F., C. Vilà, and M.T. Webster, *The legacy of domestication: accumulation of deleterious mutations in the dog genome.* Mol Biol Evol, 2008. **25**(11): p. 2331-6.
46. Xie, X., et al., *Accumulation of deleterious mutations in the domestic yak genome.* Anim Genet, 2018. **49**(5): p. 384-392.
47. Daković, N., et al., *The loss of adipokine genes in the chicken genome and implications for insulin metabolism.* Mol Biol Evol, 2014. **31**(10): p. 2637-46.
48. Mello, C.V. and P.V. Lovell, *Avian genomics lends insights into endocrine function in birds.* Gen Comp Endocrinol, 2018. **256**: p. 123-129.
49. Krchlíková, V., et al., *Repeated MDA5 Gene Loss in Birds: An Evolutionary Perspective.* Viruses, 2021. **13**(11).
50. Satoh, T., *Bird evolution by insulin resistance.* Trends Endocrinol Metab, 2021. **32**(10): p. 803-813.
51. Salve, B.G., A.M. Kurian, and N. Vijay, *Concurrent loss of ciliary genes WDR93 and CFAP46 in phylogenetically distant birds.* R Soc Open Sci, 2023. **10**(8): p. 230801.
52. Hecker, N., V. Sharma, and M. Hiller, *Convergent gene losses illuminate metabolic and physiological changes in herbivores and carnivores.* Proc Natl Acad Sci U S A, 2019. **116**(8): p. 3036-3041.
53. Jongepier, E., et al., *Convergent Loss of Chemoreceptors across Independent Origins of Slave-Making in Ants.* Mol Biol Evol, 2022. **39**(1).
54. Monroe, J.G., et al., *Drought adaptation in Arabidopsis thaliana by extensive genetic loss-of-function.* Elife, 2018. **7**.
55. Behe, M.J., *Experimental evolution, loss-of-function mutations, and "the first rule of adaptive evolution".* Q Rev Biol, 2010. **85**(4): p. 419-45.
56. Farkas, Z., et al., *Gene loss and compensatory evolution promotes the emergence of morphological novelties in budding yeast.* Nat Ecol Evol, 2022. **6**(6): p. 763-773.
57. Oh, H.J., et al., *Loss of gene function and evolution of human phenotypes.* BMB Rep, 2015. **48**(7): p. 373-9.
58. Andersson, L., *Molecular consequences of animal breeding.* Curr Opin Genet Dev, 2013. **23**(3): p. 295-301.

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

59. Simão, F.A., et al., *BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.* Bioinformatics, 2015. **31**(19): p. 3210-2.

60. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Res, 2010. **20**(9): p. 1297-303.

61. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

62. Sievers, F. and D.G. Higgins, *Clustal Omega for making accurate alignments of many protein sequences.* Protein Sci, 2018. **27**(1): p. 135-145.

63. Felsenstein, J., *Comparative methods with sampling error and within-species variation: contrasts revisited and revised.* Am Nat, 2008. **171**(6): p. 713-25.

64. Chen, Z.L., et al., *A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides.* Nat Commun, 2019. **10**(1): p. 3404.

65. Alexander, D.H., J. Novembre, and K. Lange, *Fast model-based estimation of ancestry in unrelated individuals.* Genome Res, 2009. **19**(9): p. 1655-64.

66. Krüger, J. and M. Rehmsmeier, *RNAhybrid: microRNA target prediction easy, fast and flexible.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W451-4.

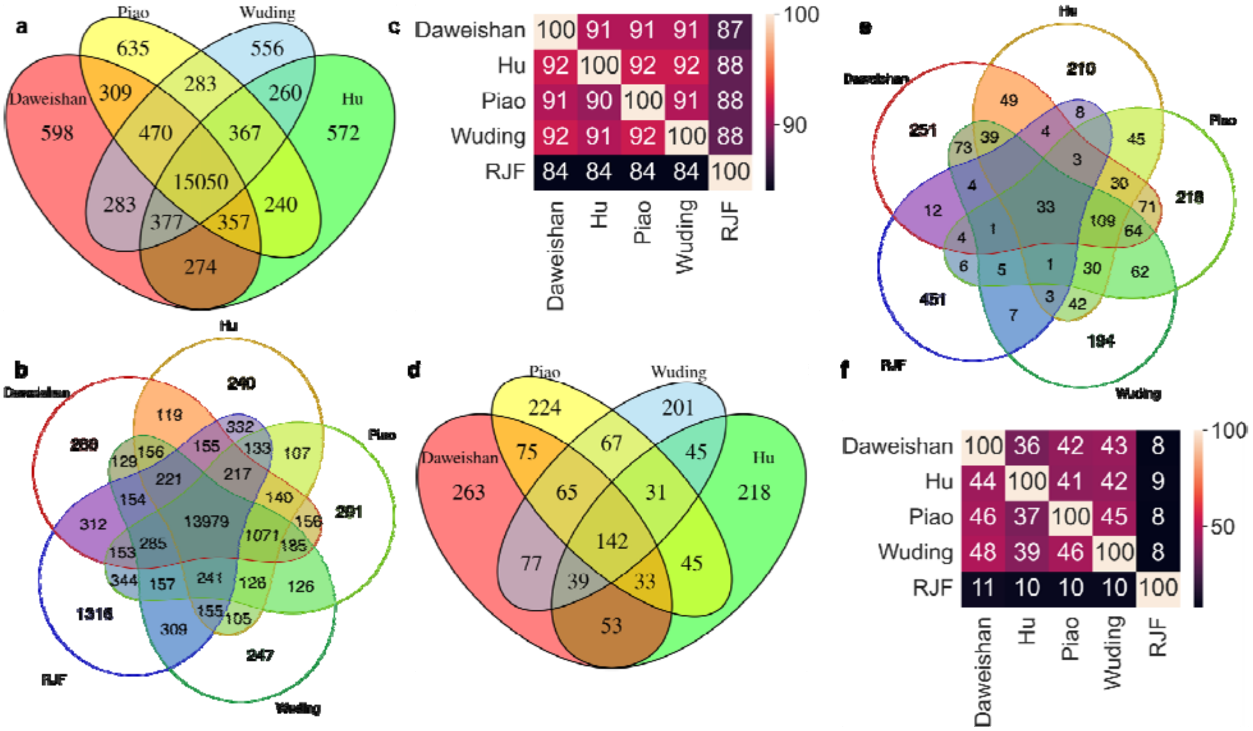
67. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Table 1: Summary of annotated pseudogenes in the four indigenous chicken genomes in comparison with those in GRCg6a

Chicken	Homology-based method		RNA-seq-based method	# Originally annotated pseudogenes	# Pseudogenes added	Final pseudogenes		# Trancribed pseudogenes
	Gene-supported	Pseudogene-supported	# NAG-supported pseudogenes					
	# Pseudogenes	# Pseudogenes				# Processed	# Unprocessed	
Daweishan	622	94	31	-	-	63	684	713
Hu	486	85	35	-	-	50	556	576
Piao	564	83	35	-	-	55	627	655
Wuding	557	86	24	-	-	48	619	633
RJF(GRCg6a)	-	-	-	262	280	80	462	-

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



	Daweishan	Hu	Piao	Wuding	RJF
Daweishan	100	91	91	91	87
Hu	92	100	92	92	88
Piao	91	90	100	91	88
Wuding	92	91	92	100	88
RJF	84	84	84	84	100

	Daweishan	Hu	Piao	Wuding	RJF
Daweishan	100	36	42	43	8
Hu	44	100	41	42	9
Piao	46	37	100	45	8
Wuding	48	39	46	100	8
RJF	11	10	10	10	100

Figure 1. Comparison of protein-coding genes and pseudogenes among the five chicken breeds. **a.** Venn diagram of the protein-coding genes of the four indigenous chickens. **b.** Venn diagram of the protein-coding genes of the five chickens. **c.** Comparison of the protein-coding genes among each pair of the five chickens. **d.** Venn diagram of the pseudogenes of the four indigenous chickens. **e.** Venn diagram of the pseudogenes of the five chickens. **f.** Comparison of the pseudogenes among each pair of the five chickens.

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

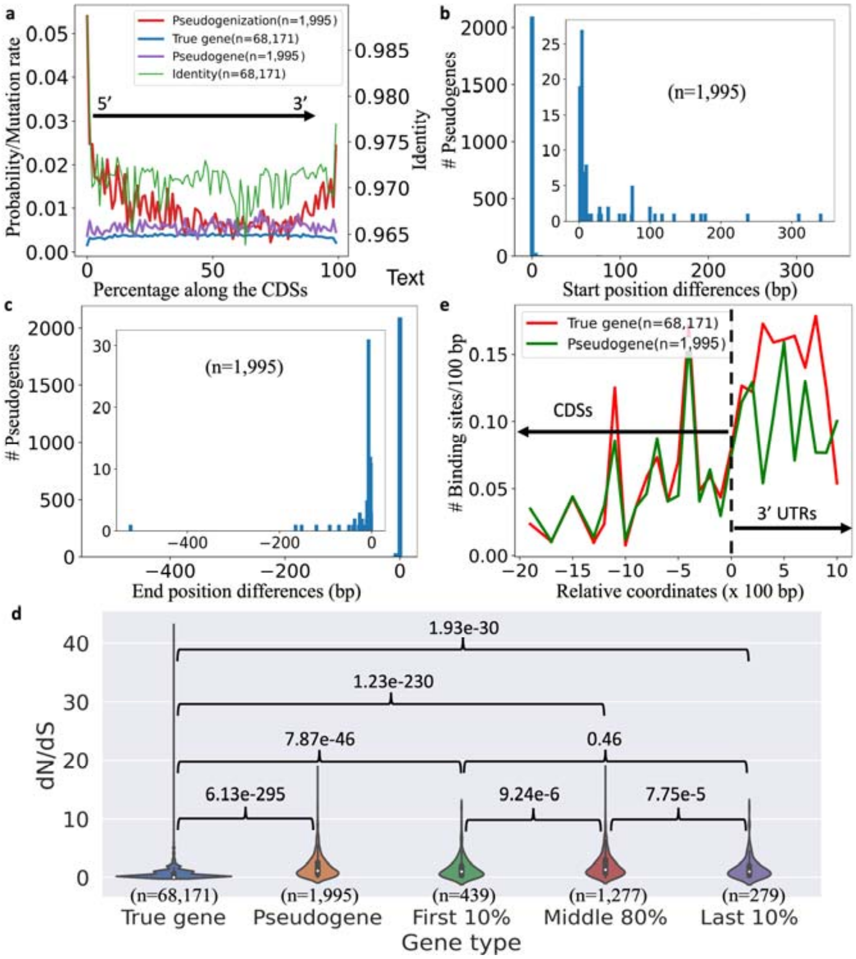


Figure 2. Pseudogenization mutations tend to occur at the two ends of CDSs. **a.** Probability of first pseudogenization mutations (red line) in 100 evenly divided CDS segments from the 5′-ends to the 3′-ends of the parental genes of the pseudogenes in the four chickens, mean rates of synonymous mutations in 100 evenly divided CDS segments from the 5′-ends to the 3′-ends of the true genes (blue line) and pseudogenes (purple) in the four chickens, and mean identity of the true genes in 100 evenly divided CDS segments from the 5′-ends to the 3′-ends of the genes (green line). **b.** Start position of the "CDS" of the pseudogenes in the four chickens with respect to the nucleotide positions of their parental genes starting with 0 with the downstream positions being positive integers. **c.** End positions of the "CDS" of the pseudogenes in the four chickens with respect to the nucleotide positions of their parental genes ending with 0 with the upstream positions being negative integers. **d.** Violin plots of the dN/dS values of all true genes, all pseudogenes, pseudogenes with the first pseudogenization occurring in the first 10%, the middle 80% and the last 10% of the CDSs in the four indigenous chickens. **e.** Number of predicted miRNA binding sites per 100pb along the CDSs and 3′ UTRs of the true genes (red line) and pseudogenes (green line). In the figure, '0' represents the end positions of the CDSs, the positive numbers represent the relative positions of 1,000 bp sequences downstream of the end of CDSs, and the negative numbers represent the relative positions of the CDSs with respect to the ends of CDSs.

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

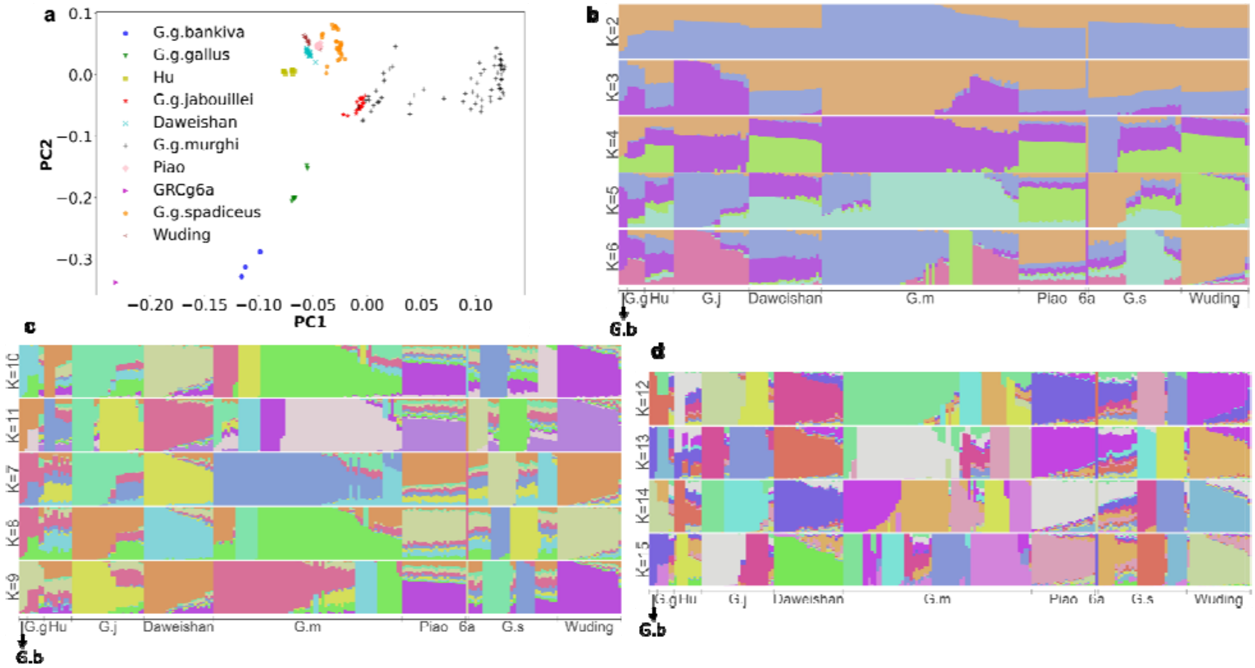
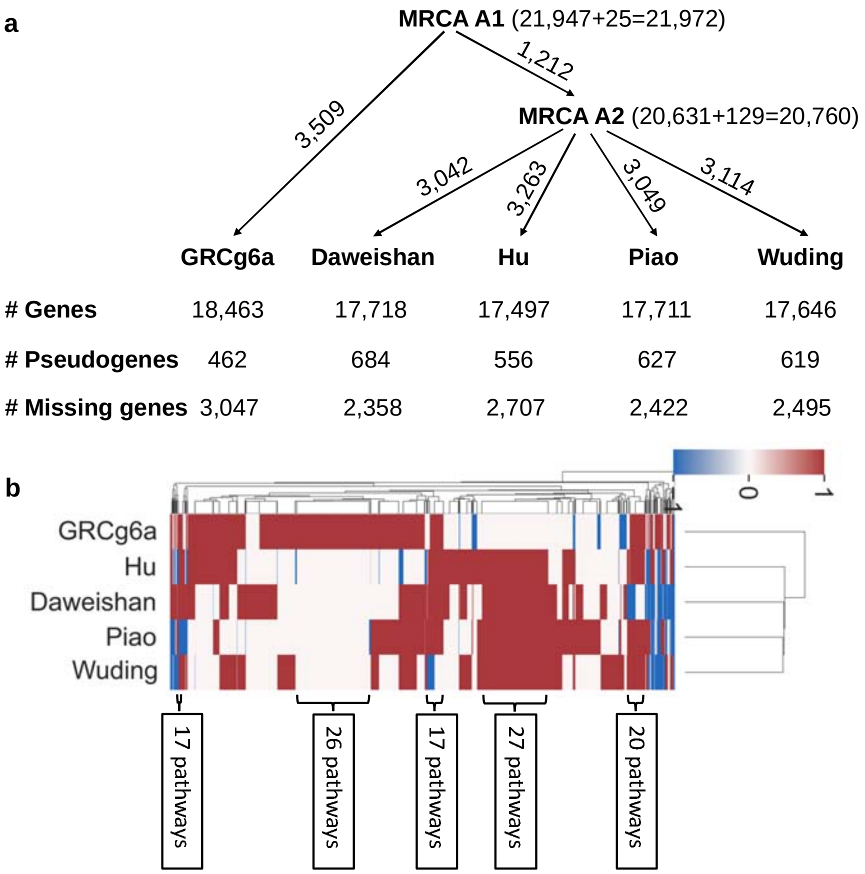


Figure 3. Analysis of frequency spectrums of SNPs. a. Principal component analysis of the RJF subspecies, indigenous chickens and the RJF individual (GRCg6a) based on their SNP profiles**. b-d.** Genetic structures of the RJF subspecies, indigenous chickens and the RJF individual (GRCg6a) estimated using the ADMIXTURE program for K=2, 3, …, 15.

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



	GRCg6a	Daweishan	Hu	Piao	Wuding
# Genes	18,463	17,718	17,497	17,711	17,646
# Pseudogenes	462	684	556	627	619
# Missing genes	3,047	2,358	2,707	2,422	2,495

Figure 4. Evolutionary pattern of the five chickens. **a.** A hypothetical scenario for loss-of-functions of the five chickens since their divergence from MRCA A1 and A2. **b.** Heatmap of two-way hierarchical clustering of the 7,993 dispensable genes in MRCA A1 that are either completely lost or pseudogenized in at least one of the five indigenous chicken genomes based on their appearance as an intact form (1, brown), absence (0, white) and as a pseudogenized form (-1, blue) in the five chicken genomes.

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

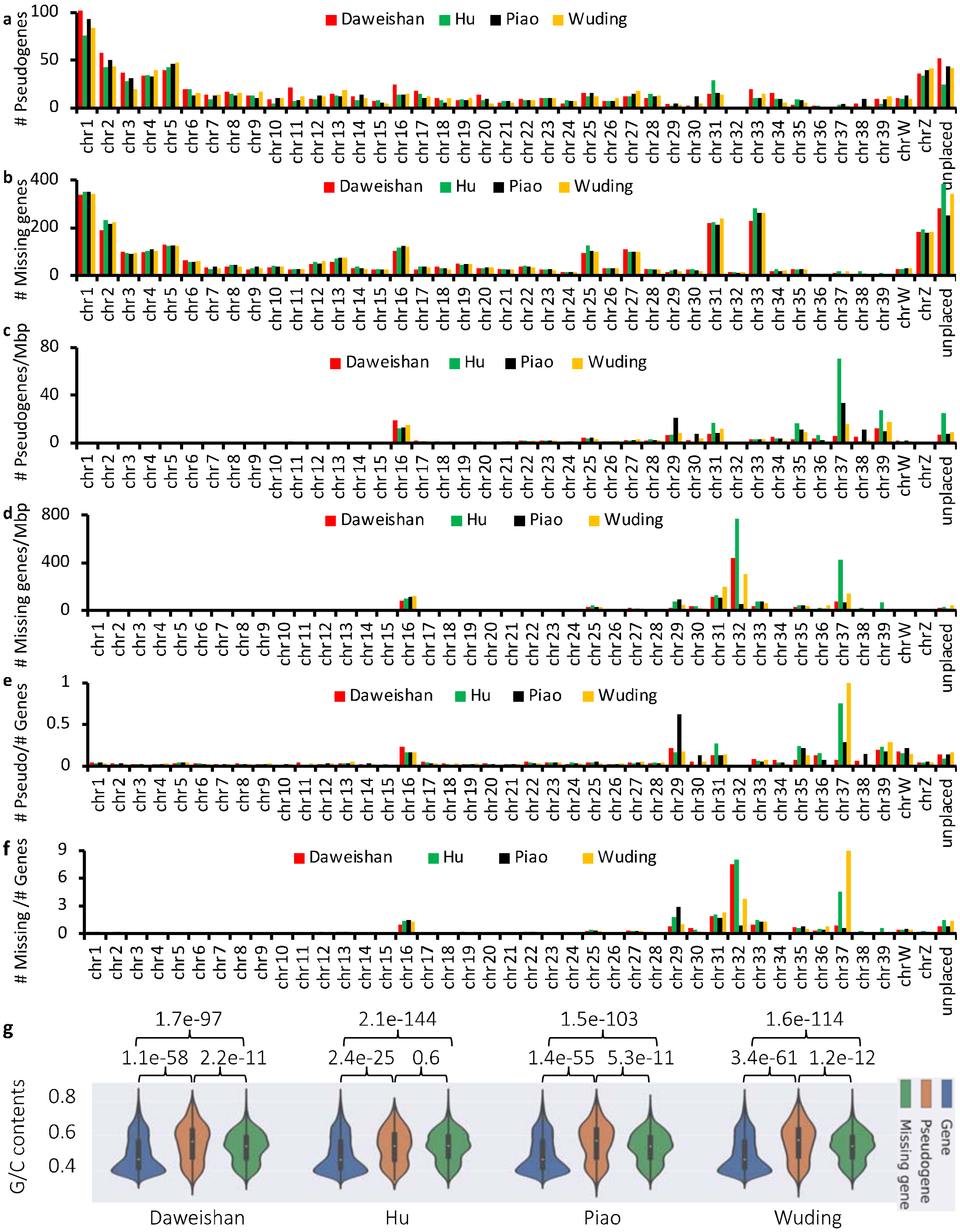
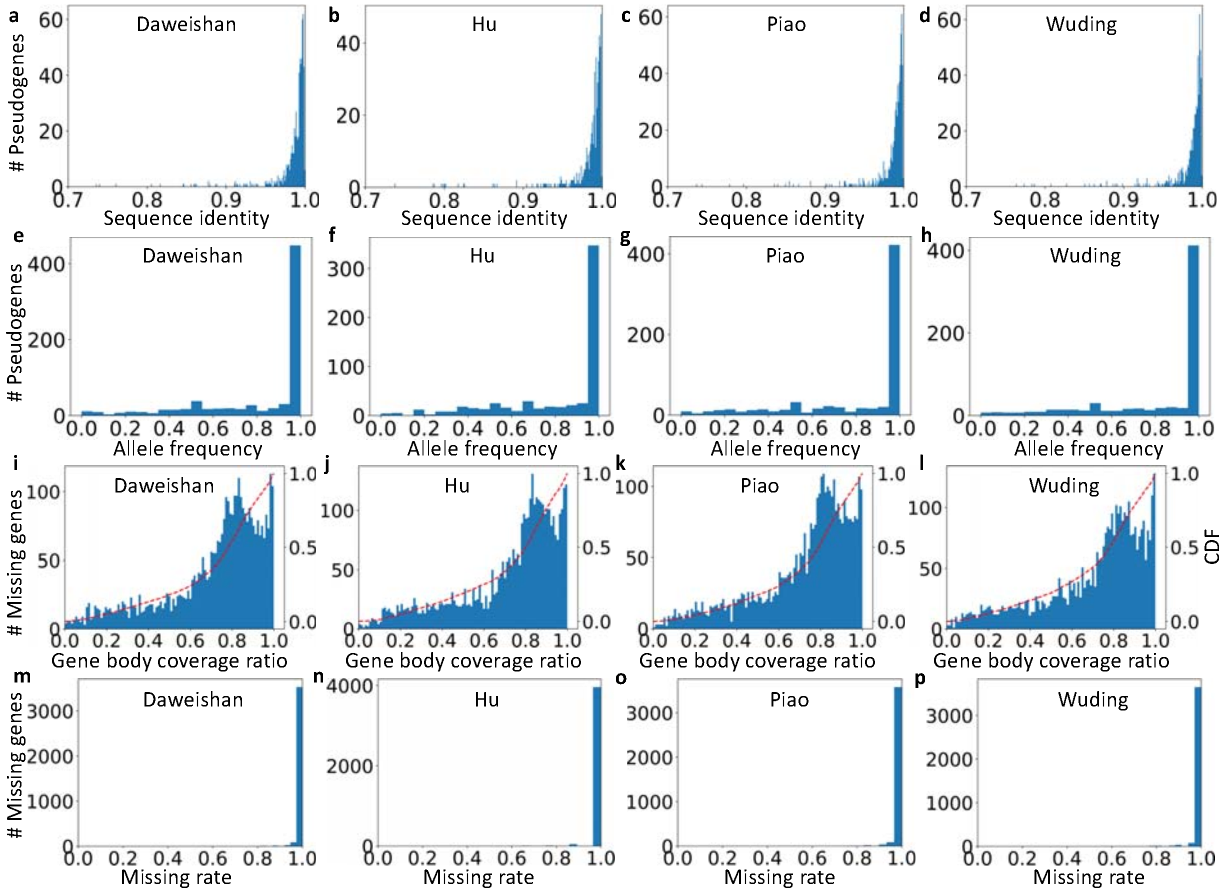


Figure 5. Distribution of pseudogenes and missing genes on each chromosome of the four indigenous chicken genomes. **a, b.** Number of pseudogenes (a) and missing genes (b) on each chromosome of the chicken genomes. **c, d.** Density of pseudogenes (c) and missing genes (d) on each chromosome of the chicken genomes. **e, f.** Ratio of number of pseudogenes/genes (e) and mising genes/genes (f) on each chromosome of the chicken genomes. **g.** Comparison of G/C contents of true genes, pseudogenes and missing genes in the chicken genomes. Statistical tests were done using one-tailed t-test.

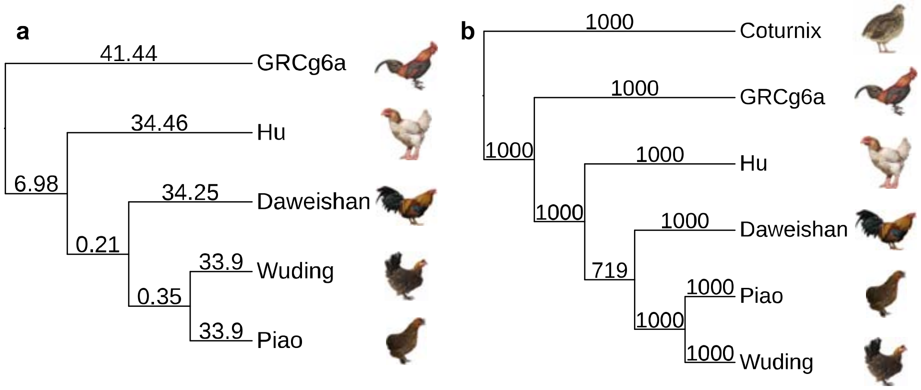
bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



838

839 **Figure 6.** Most loss-of-function mutations are fixed in the indigenous chicken populations. **a-d.** Number of pseudogenes in each of the indigenous chicken genomes with the indicated identity with their parental genes. **e-h.** Number of pseudogenes with the indicated pseudogenization rate in the chicken populations. **i-l.** Number of missing genes in each of the indigenous chicken genomes with the indicated reads coverage on their functional versions. The dashed red lines are the CDFs of coverage ratios. **m-p.** Number of missing genes with the indicated missing rate in the chicken populations.

bioRxiv preprint doi: https://doi.org/10.1101/2024.05.21.595210; this version posted May 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



846
847 **Figure 7.** Evolutionary relationships of the five chickens. **a.** Neighbor-joining phylogenetic tree of the five chickens, constructed using the occurring patterns of the 7,993 dispensable genes in their genomes. The numbers on the branches are Euclid distance between the pattern vectors. **b.** Neighbor-joining phylogenetic tree of the five chickens, constructed using the 6,744 essential protein-coding genes in their genomes and the quail genome. The numbers on the nodes are bootstrapping value for 1,000 repeats.