Recovering the ‘missing’ avian genes using multi-omics data

Recovering the 'missing' avian genes using multi-omics data

Zhong-Tao Yin[1], Jacqueline Smith[2], Zhuo-Cheng Hou[1,*]

Affiliation

1. National Engineering Laboratory for Animal Breeding and Key Laboratory of Animal Genetics, Breeding and Reproduction, MARA; College of Animal Science and Technology, China Agricultural University, No. 2 Yuanmingyuan West Rd, Beijing 100193, China
2. The Roslin Institute & R(D)SVS, University of Edinburgh, Easter Bush, Midlothian, EH25 9RG, UK

Email addresses

Co-authors: yinzhngtao@cau.edu.cn (Z.T.Y.)

Co-authors: jacqueline.smith@roslin.ed.ac.uk (JS)

Corresponding author: zchou@cau.edu.cn (Z.C.H.)

Gene gain and loss are common events in the evolution of species, especially for birds, which have evolved many unique characteristics such as feathers, wings, and flight capabilities, strong and lightweight skeletons, toothless beaks, high metabolic rates, and heat absorption sex, and unique respiratory and excretory systems [Kennedy and Vevers, 1976, Blomme et al., 2006]. The release of the first chicken genome provided the basis for systematic analysis of the similarities and differences between vertebrate and avian genomes [International Chicken Genome Sequencing et al., 2004]. In comparison with other amniotes, bird genomes are more compact, and this difference may be related to the overall smaller cell size [Hughes and Hughes, 1995, Hughes and Friedman, 2008]. The reductions in genome size may be the result of the loss of noncoding DNA sequences, with bird genomes having less repetitive DNA, fewer pseudogenes, and shorter introns than mammalian genomes [Hillier et al., 2004, Hughes and Piontkivska, 2005]. Importantly, the evolution of avian genomes also appears to involve the loss of protein-coding genes, as the total number of uniquely identified avian-coding genes is much smaller than in other tetrapods (i.e., 23,294 in humans, GRCh38.p14; 19,404 in lizards, AnoCar2.0; 17,007 in chickens, GRCg7b). Paralog analysis revealed a higher overall incidence of gene families with fewer members in birds compared to other vertebrates [Hughes and Friedman, 2008]. Likewise, birds have a high rate of chromosomal rearrangements compared to other organisms, all of which may result in the deletion of protein-coding genes [Backstrom et al., 2010]. In recent years, the genomes of a large number of birds and lizards have been assembled and annotated, including zebra finches [Warren et al., 2010], chickens [Hillier et al., 2004], turkeys [Dalloul et al., 2010] and duck [Zhu et al., 2021]. Moreover, large-scale bird genome projects [Jarvis et al., 2014, Zhang et al., 2014], and chicken pan-genomes [Wang et al., 2021, Li et al., 2022] have also generated considerable genomic data. These large comparative genomic datasets identified hundreds of lost genomic-blocks in the bird genomes, and also suggested that hundreds of genes are missing in birds [Lovell et al., 2014, Zhang et al., 2014].

The missing genes seem to be directly related to the unique physiological phenomena of birds. Several functionally important genes in mammals are supposed

‘missing’ in chickens and have caused long-debated questions in bird biology. Spurious discovery of the missing/hidden genes in the bird genome has continued for decades. Previously, *BGN* [Blaschke et al., 1996], *CORO1A* [Xavier et al., 2008], *MAPK3* [Lemoine et al., 2009], *MMP14* [Simsa et al., 2007], *TBX6* [Lardelli et al., 2003, Ahn et al., 2012], *TSSK4* [Shang et al., 2013] and five adipokine genes [Dakovic et al., 2014] were reported to be missing in birds, however, several long-debated genes including *TNF-alpha*, and *leptin* have been cloned in birds [Prokop et al., 2014, Seroussi et al., 2016, Rohde et al., 2018]. This hide-and-seek game still continues, and does not appear to be ending anytime soon [Elleder and Kaspers, 2019]. Here we summarize recent efforts using multi-omics data to probe those genes missing/hidden in avian genomes.

Reconstruction of missing genes in the chicken genome

While the hypothesis of missing genes in birds has been proposed for decades, researchers have found that some of the missing genes were, in fact, present in chickens or other birds. In the presence of large gaps and imperfect gene annotation in the genome, the *de novo* assembly of gene sequences using RNA-seq is considered to be an efficient way to identify unannotated genes in the genome. Attempts that only used a few tissues/organs have identified many missing genes in birds [Hron et al., 2015, Bornelov et al., 2017, Botero-Castro et al., 2017]. Recently, we used the raw data from 26 chicken tissues downloaded from the GenBank database to assemble and obtain 2,048,631 transcripts and identified 589 missing genes in birds [Yin et al., 2019].

At the same time, the continuity and integrity of chicken genome assemblies have been rapidly improving. The chicken genome released in 2017 was assembled by third-generation sequencing technology, and the number of annotated genes increased significantly (2,768 noncoding and 1,911 protein-coding genes) [Warren et al., 2017]. In the Gallus_gallus-5.0 genome, 442 (77.41%, from a total of 571) genes thought to be missing in chickens (in Lovell et al., 2014 see Table S1 and Table S6, plus select entries in Table S4 and Table S18) were annotated, indicating that there is no systematic deletion of genes in birds. With the development of sequencing and hybrid assembly technology, the genomes of different chicken breeds continue to be assembled and

another 136 missing genes were further annotated in our recently assembled *Silkie* genome (unpublished). To date, it has now been shown that 528 (92.47%) genes that were thought to be missing, actually exist in chickens. This has been made possible by exploiting a large amount of multi-omics data available in chicken and has led to the revelation of genes with important functions such as *TNF-α* and *Leptin* [Seroussi et al., 2016, Rohde et al., 2018]. Recent large-scale chicken pan-genome data have also identified thousands of genes that are not presented in the current chicken reference genome [Li et al., 2022].

Reconstruction of missing genes from other birds

In addition to chicken, researchers have reconstructed many genes thought to be missing from other birds. We collected data from various important tissues from duck (24), pigeon (11), goose (8), and zebra finch (22) [Yin et al., 2019], and an avian transcriptomic database containing a total of 9,296,247 transcripts was constructed by *de novo* transcriptome assembly. From this, we identified several genes in duck (583), pigeon (558), goose (537), and zebra finch (543) from 806 genes that were thought to be missing in birds (in Lovell et al., 2014 see Table S1 and Zhang et al., 2014 see Table S10). Only 135 genes were not found in this bird transcriptome database. The number of missing genes reconstructed in different birds by *de novo* assembly of large transcriptome data is similar, indicating that these genes thought to be missing exist across different bird species.

In recent years, duck functional genomics has developed rapidly. We have assembled the Mallard, Pekin duck, and Shaoxing laying duck genomes using a combination of third-generation sequencing, Bionano, and Hi-C sequencing technologies. These have proved to be a rich source of genetic information [Zhu et al., 2021]. In the Mallard duck the CAU_wild 1.0 genome has 1,872 more protein-coding genes annotated than the previous CAU 1.0 genome, including 89 genes previously thought to be missing in birds. Among these 89 genes, 5 genes have become pseudogenes, losing part of their gene function, 3 genes have been annotated as lncRNAs, and the remaining 81 genes remain as protein-coding genes. In addition, 240

genes were annotated as paralogous genes and 108 genes had similar segments in the genome. Mining large multi-omics data assemblies and annotations now reveals that only 10 genes (from a total of 806 missing genes), to date, have not been reconstructed in birds, with the rest of the genes thought to be missing in birds having been shown to actually exist. The recovered gene list is shown in Supplementary Table 1.

Development of new methods to identify more missing genes

Summarizing the characteristics of these reconstructed missing genes in birds and the reasons why they are thought to be missing can provide insights and methods for us to identify more missing genes. First, these reconstructed gene sequences have high GC content and length in many birds. The GC content of most of these 'missing' genes is more than 60%, and few genes even have over 80% (the median GC content of the chicken genome is 42.22% and the median GC content of the duck genome is 41.99%) [Hron et al., 2015, Bornelov et al., 2017, Botero-Castro et al., 2017, Yin et al., 2019]. At the same time, the multi-tissue transcriptome expression profiles of birds showed that most of the reconstructed genes usually have strong tissue-specific expression. These genes are generally expressed predominantly in one tissue and are rarely expressed in the other tissues [Yin et al., 2019]. High-throughput transcriptome-based assembly approaches have limitations for fully recovering missing genes due to technical factors such as the PCR amplification bias against GC-rich fragments [Beauclair et al., 2019]. Expression patterns, i.e. tissue-specific expression patterns, and low expression, also limit the ability for full transcriptome assembly. Now, the third-generation sequencing technologies, which have less GC bias, such as single-molecule real-time (SMRT) and nanopore sequencing technologies, can obtain full-length transcripts directly, without assembly [Yin et al., 2019; Kuo et al. 2020]. The missing genes will continue to be discovered with the accumulation of full-length transcriptome data from more avian tissues from different physiological conditions.

Furthermore, the missing genes annotated in the chicken and duck genomes are mainly distributed on the micro-chromosomes, the ends of the chromosomes, and within regions showing a high content of tandem repeats clustering with non-canonical

DNA structures. [Zhu et al., 2021, Li et al., 2022]. Long repetitive regions [Treangen and Salzberg, 2011], regions of high GC content [Chen et al., 2013], telomeric regions, fragmented micro-chromosomes [O'Connor et al., 2019], and adaptive assembly strategies have always proved problematic for enabling complete bird genome assembly. To fully resolve the whole chicken gene sets, a Telomere-to-Telomere (T2T) genome is necessary. The recently completed human T2T genome has now paved the way for the finished bird genome assembly [Miga et al., 2020, Hoyt et al., 2022, Mao and Zhang, 2022, Nurk et al., 2022]. Ultra-long ONT sequencing, high-precision HiFi sequencing data, multi-type auxiliary assembly data, and hybrid assembly using multiple strategies will greatly promote the quality of bird genome assembly [Sohn and Nam, 2018]. For large presence/absence variations within species, we can enrich genomic information by constructing high-quality multi-breed pan-genomes [Vernikos et al., 2015]. The Bird 10,000 Genomes (B10K) Project [Zhang et al., 2015] has generated insightful results and the future bird T2T genome and pan-genome will undoubtedly reveal more genes. This complete gene map of birds will be critical for the further understanding of the biology and evolution of birds.

Finally, precise genome annotation will also provide the necessary sequence and structural information for mining more genes in birds. Annotation errors are unavoidable in genome annotation using automated processes, especially for some protein-coding genes that cannot be annotated in complex and high GC regions [Salzberg et al., 2019]. While applying full-length transcriptomic data for genome annotation [Nudelman et al., 2018, Wang et al., 2019; Kuo et al., 2020], the use of novel annotation methods developed based on machine learning can further improve the accuracy of annotation [Mahood et al., 2020, Stiehler et al., 2020]. More accurate manual annotation of important genome regions is also necessary for novel gene identification [Dunn et al., 2019]. It can be seen that, with the continuous development of omics technology and analysis methods, the genome information will be more complete, the annotation will be more accurate, and the genes that were previously thought to be missing in birds will continue to be discovered.

169
170

Statement of Ethics

All experiments with birds were performed under the guidance of ethical regulations from the Animal Care and Use Committee of China Agricultural University, Beijing, China.

Conflict of Interest Statement

The authors declare no competing interests.

Funding Sources

The work was supported by the National Waterfowl-Industry Technology Research System (CARS-42), the National Nature Science Foundation of China (31972525, 31572388), Beijing Joint Research Program for Germplasm Innovation and New Variety Breeding (G20220628007), Beijing Municipal Science & Technology Commission (Z211100004621007).

Author Contributions

Z.C.H designed the study. Z.T.Y collected the samples and performed the analyses of the identification of the missing genes in birds. Z.C.H, J.S. and Z.T.Y wrote and revised the paper.

Data Availability Statement

Data have been submitted to the public databases under the following accession numbers: The raw data for assembling the transcriptome database of chicken, duck, goose, pigeon, and zebra finch were deposited in Sequence Read Archive (SRA) database under the accession number SRP141084. The Mallard genome is stored in NCBI under accession number PRJNA554956. The raw data for the Silkie genome assembly can be found in the SRA database under the accession number PRJNA805080 (Unpublished data).

References

Ahn D, You KH, Kim CH. Evolution of the tbx6/16 subfamily genes in vertebrates: insights from zebrafish. Mol Biol Evol. 2012;29(12):3959-83.

Backstrom N, Forstmeier W, Schielzeth H, Mellenius H, Nam K, Bolund E, et al. The recombination landscape of the zebra finch Taeniopygia guttata genome. Genome Res. 2010;20(4):485-95.

Beauclair L, Rame C, Arensburger P, Piegu B, Guillou F, Dupont J, et al. Sequence properties of certain GC rich avian genes, their origins and absence from genome assemblies: case studies. BMC Genomics. 2019;20(1):734.

Blaschke UK, Hedbom E, Bruckner P. Distinct isoforms of chicken decorin contain either one or two dermatan sulfate chains. J Biol Chem. 1996;271(48):30347-53.

Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y. The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol. 2006;7(5):R43.

Bornelov S, Seroussi E, Yosefi S, Pendavis K, Burgess SC, Grabherr M, et al. Correspondence on Lovell et al.: identification of chicken genes previously assumed to be evolutionarily lost. Genome Biol. 2017;18(1):112.

Botero-Castro F, Figuet E, Tilak MK, Nabholz B, Galtier N. Avian Genomes Revisited: Hidden Genes Uncovered and the Rates versus Traits Paradox in Birds. Mol Biol Evol. 2017;34(12):3123-31.

Chen YC, Liu T, Yu CH, Chiang TY, Hwang CC. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. PLoS One. 2013;8(4):e62856.

Dakovic N, Terezol M, Pitel F, Maillard V, Elis S, Leroux S, et al. The loss of adipokine genes in the chicken genome and implications for insulin metabolism. Mol Biol Evol. 2014;31(10):2637-46.

Dalloul RA, Long JA, Zimin AV, Aslam L, Beal K, Blomberg Le A, et al. Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis. PLoS Biol. 2010;8(9).

Dunn NA, Unni DR, Diesh C, Munoz-Torres M, Harris NL, Yao E, et al. Apollo: Democratizing genome annotation. PLoS Comput Biol. 2019;15(2):e1006790.

Elleder D, Kaspers B. After TNF-alpha, still playing hide-and-seek with chicken genes. Poult Sci. 2019;98(10):4373-4.

Hillier LW MW, Birney E, Warren WC, Consortium. atmotCG. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 2004;432(7018):695-716.

Hoyt SJ, Storer JM, Hartley GA, Grady PGS, Gershman A, de Lima LG, et al. From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. Science. 2022;376(6588):eabk3112.

Hron T, Pajer P, Paces J, Bartunek P, Elleder D. Hidden genes in birds. Genome Biol. 2015;16:164.

Hughes AL, Friedman R. Genome size reduction in the chicken has involved massive loss of ancestral protein-coding genes. Mol Biol Evol. 2008;25(12):2681-8.

Hughes AL, Hughes MK. Small genomes for better flyers. Nature. 1995;377(6548):391.

Hughes AL, Piontkivska H. DNA repeat arrays in chicken and human genomes and the adaptive evolution of avian genome size. BMC Evol Biol. 2005;5:12.

International Chicken Genome Sequencing C. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 2004;432(7018):695-716.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science. 2014;346(6215):1320-31.

Kennedy GY, Vevers HG. A survey of avian eggshell pigments. Comp Biochem Physiol B. 1976;55(1):117-23.

Kuo RI, Cheng Y, Zhang R, Brown JWS, Smith J, Archibald AL, Burt DW. Illuminating the dark side of the human transcriptome with long read transcript sequencing. BMC Genomics. 2020 Oct 30;21(1):751.

Lardelli M. The evolutionary relationships of zebrafish genes tbx6, tbx16/spadetail and mga. Dev Genes Evol. 2003;213(10):519-22.

Lemoine M, Dupont J, Guillory V, Tesseraud S, Blesbois E. Potential involvement of several signaling pathways in initiation of the chicken acrosome reaction. Biol Reprod. 2009;81(4):657-65.

Li M, Sun C, Xu N, Bian P, Tian X, Wang X, et al. De Novo Assembly of 20 Chicken Genomes Reveals the Undetectable Phenomenon for Thousands of Core Genes on Microchromosomes and Subtelomeric Regions. Mol Biol Evol. 2022;39(4).

Lovell PV, Wirthlin M, Wilhelm L, Minx P, Lazar NH, Carbone L, et al. Conserved syntenic clusters of protein coding genes are missing in birds. Genome Biol. 2014;15(12):565.

Mahood EH, Kruse LH, Moghe GD. Machine learning: A powerful tool for gene function prediction in plants. Appl Plant Sci. 2020;8(7):e11376.

Mao Y, Zhang G. A complete, telomere-to-telomere human genome sequence presents new opportunities for evolutionary genomics. Nat Methods. 2022;19(6):635-8.

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, et al. Telomere-to-telomere assembly of a complete human X chromosome. Nature. 2020;585(7823):79-84.

Nudelman G, Frasca A, Kent B, Sadler KC, Sealfon SC, Walsh MJ, et al. High resolution annotation of zebrafish transcriptome using long-read sequencing. Genome Res. 2018;28(9):1415-25.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376(6588):44-53.

O'Connor RE, Kiazim L, Skinner B, Fonseka G, Joseph S, Jennings R, et al. Patterns of microchromosome organization remain highly conserved throughout avian evolution. Chromosoma. 2019;128(1):21-9.

Prokop JW, Schmidt C, Gasper D, Duff RJ, Milsted A, Ohkubo T, et al. Discovery of the elusive leptin in birds: identification of several 'missing links' in the evolution of leptin and its receptor. PLoS One. 2014;9(3):e92751.

Rohde F, Schusser B, Hron T, Farkasova H, Plachy J, Hartle S, et al. Characterization of Chicken Tumor Necrosis Factor-alpha, a Long Missed Cytokine in Birds. Front Immunol. 2018;9:605.

Salzberg SL. Next-generation genome annotation: we still struggle to get it right. Genome Biol. 2019;20(1):92.

Seroussi E, Cinnamon Y, Yosefi S, Genin O, Smith JG, Rafati N, et al. Identification of the Long-Sought Leptin in Chicken and Duck: Expression Pattern of the Highly GC-Rich Avian leptin Fits an Autocrine/Paracrine Rather Than Endocrine Function. Endocrinology. 2016;157(2):737-51.

Shang P, Hoogerbrugge J, Baarends WM, Grootegoed JA. Evolution of testis-specific kinases TSSK1B and TSSK2 in primates. Andrology. 2013;1(1):160-8.

Simsa S, Genina O, Ornan EM. Matrix metalloproteinase expression and localization in turkey (Meleagris gallopavo) during the endochondral ossification process. J Anim Sci. 2007;85(6):1393-401.

Sohn JI, Nam JW. The present and future of de novo whole-genome assembly. Brief Bioinform.

2018;19(1):23-40.

Stiehler F, Steinborn M, Scholz S, Dey D, Weber APM, Denton AK. Helixer: Cross-species gene annotation of large eukaryotic genomes using deep learning. Bioinformatics. 2020.

Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2011;13(1):36-46.

Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. Curr Opin Microbiol. 2015;23:148-54.

Wang K, Hu H, Tian Y, Li J, Scheben A, Zhang C, et al. The Chicken Pan-Genome Reveals Gene Content Variation and a Promoter Region Deletion in IGF2BP1 Affecting Body Size. Mol Biol Evol. 2021;38(11):5066-81.

Wang X, You X, Langer JD, Hou J, Rupprecht F, Vlatkovic I, et al. Full-length transcriptome reconstruction reveals a large diversity of RNA and protein isoforms in rat hippocampus. Nat Commun. 2019;10(1):5009.

Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Kunstner A, et al. The genome of a songbird. Nature. 2010;464(7289):757-62.

Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, et al. A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. G3 (Bethesda). 2017;7(1):109-17.

Xavier CP, Eichinger L, Fernandez MP, Morgan RO, Clemen CS. Evolutionary and functional diversity of coronin proteins. Subcell Biochem. 2008;48:98-109.

Yin Z, Zhang F, Smith J, Kuo R, Hou ZC. Full-length transcriptome sequencing from multiple tissues of duck, Anas platyrhynchos. Sci Data. 2019;6(1):275.

Yin ZT, Zhu F, Lin FB, Jia T, Wang Z, Sun DT, et al. Revisiting avian 'missing' genes from de novo assembled transcripts. BMC Genomics. 2019;20(1):4.

Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. Science. 2014;346(6215):1311-20.

Zhang G, Rahbek C, Graves GR, Lei F, Jarvis ED, Gilbert MT. Genomics: Bird sequencing project takes off. Nature. 2015;522(7554):34.

Zhu F, Yin ZT, Wang Z, Smith J, Zhang F, Martin F, et al. Three chromosome-level duck genome assemblies provide insights into genomic variation during domestication. Nat Commun. 2021;12(1):5932.

Supplementary Table 1. The recovered 'missing genes' from birds

Previously reported gene list	No. of recovered missing genes	Gene Symbols
Evidence supporting the missing genes in five birds [Yin et al., 2019]	446	*ABCD1 ATL3 ADCY4 AIF1 ANKRD23 AP2A1 APLP1 ARAF ARHGEF25 ATP6AP1 AVIL AVPR2 BBS1 BCAT2 BGN BRSK1 CA11 CALM3 CD37 CD97 CDC42EP5 CDH24 CDKN2D CEBPE CORO1A CPT1C CYP2F1 DNAJC4 DOC2A DOCK6 DUSP9 EGLN2 EPN1 ERF FAM120C FBXL19 FERMT3 FGD1 FOXA3 FOXH1 GDI1 GLS2 GMPR2 GPR173 GSK3A HAS1 HIF3A HIGD1C HOMEZ HOOK2 HSPB6 IGLON5 IRAK1 IRF2BP1 IRF9 ITPKC KANK2 KCNAB3 KCNK4 KCNK6 KCNN4 KDELR1 KIFC2 KIRREL2 KLC2 KMT5C KREMEN2 LOC100555519 MACROD1 MAMSTR MAP3K10 MAP4K1 MAPK3 MARCH9 METTL21B MMP14* MRPL52* MYH14 NFATC4 NR1H2 NTN5 ORAI3 PACS1 PALM3 PARP2 PIM2 PLXNA3 PLXNB3 PNCK PODNL1 POU5F1 PPFIA3 PPP1R3E PPP6R1 PRDX2 PRKACA PRRG2 PRSS8* PRX PSPN PTGIR PTPRH RAB1B RAB3D RBCK1 RCN3* RGL3 RNF181 RPS6KA4 RRAS S1PR5 SERPINI2 SLC22A17 SLC6A16 SLC6A8 SLC7A7 SNX15 SPRED3 STX4 SYN1 SYNGR4 SYT5 TBX6 TEAD2 TEP1 TIMM29 TIMP1 TMC4 TMEM150A TMEM91 TRMT1* TRPT1 TSPAN31* TSSK4 TUBB4A TULP2 USP11 WRAP53 YIPF2 YPEL3 ZNF784 ZSWIM4 ASF1B ATP2B3 BCL7C CCDC88B CHMP4A CNTD2 DENND1C DNAH2 EFS ELK1 EMP3 EPS8L1 ESRRA FAM98C FBRS FGF21 FKBPL FUZ GAPDHS GNG8 GRAMD1A HSD17B14 IRF3 JPH4 JUNB KCND1 KCNJ14 KMT2B LRP10 LRRC8E MMP25 MYL6B NPHS1* NTF4 NUCB1 PDZD4 PELI3 PLCB3* PLEKHA4 PLPPR2* PRDX5 PRR14 PRRC2A PRSS53 PSMB11 PSME1 RAB2B RASGRP4 REC8 REM2 RPGRIP1 SLC25A23SLC26A10 SLC44A4 SLC7A8 SPHK2 SPIB STX10 STXBP2 SYP TBC1D10B TCF19 TRIB3 TRPM4 TTC5 TTC9C TTYH1 UBL4A YIF1A ZNF385A NFE2 OR52N4 CGREF1 OPA3 SCAND1 RCE1 FIS1 PIP4P1 SOAT2 SIX5 MPDU1 EPSTI1 PLOD3 OR52N2 ACP4 GRWD1 PRODH2 CIB4 GPR137 ABHD16B SMG9 RCOR2 IDO1 KDM6B ITGB7 PAX8 MAP3K11 TMEM147 SCYL1 PSRC1 FAM131C CADM4 B4GALNT2 BAZ2A PCMTD2 ZNF583 PHF23 ANKRD45 GATA1 IGHV1-2 IGHV4-34 ABCB5 DDX39A PELP1 PRKCSH GTF2IRD2 MED25 TMEM30B S100A5 IKBKG ABHD16A PPP1R12C GRIPAP1 BRMS1 AQP6 RBM42 HCFC1 WDR6 ARHGAP11B MBD1 OR7C1 DBP CHD3 AKR1C3 KCTD13 COX4I2 ARG1 CABP5 NELFE OR1C1 CPA3 THAP8 RASIP1 ZNF653 NANOS2 ECT2L TIMM17B LMTK3 ALAS2 EHBP1L1 SLC52A1 B3GAT3 EPOR PXDNL INO80E EFEMP2 SLC25A18 RNF31 OR52B4 PIWIL4 KHNYN BTBD18 OLIG1 PRLH STK19 SLC43A1 ADH4 KRI1 CD2BP2 LYL1 GPR108 GNB2 PRMT5 AXL CYB561D1 MEGF8 TSSK1B CALML6 SPNS1 ARR3 STAC2 RELA SOWAHB PGAM2 OXA1L CHRNB1 TP53 THOC6 SH2D6 GPR45 CIART ACY3 AIF1L PHF1 MUS81 OR11H6 NOSIP ARHGAP27 SRRM2 LSR TNS2 CXXC1 CUTA SRSF9 CDIPT ARL4D CHRM1 PSMA8 SART1 CFL1 PLSCR3 OR10V1*

		TBC1D10C ALKBH7 DNAJC30 RNF183 DCAF11 ACAP1 NOVA2 RMND5B TNFAIP8L2 NCKAP5L BEST2 ENTHD1 ABT1 SLC12A6 KLC3 CLIC1 ABHD4 GDPD3 DVL2 FAM89B ZDHHC11 COX7A1 PLD2 METTL21A TAOK2 MORC1 UXT SHANK1 OR51A7 NDRG2 CNOT3 XRCC1 HUWE1 HIGD1B RARRES3 RNASEH2A TBC1D17 DGKA DNAH11 CLN3 NXPH3 NUMBL HOXD1 OR2B3 TNNT1 ZNF500 PSMB8 AKR1C1 WAS CNFN CCDC22 ARHGEF1 POU2F2 WDR74 CIC PPP1R13L FXR2 ACIN1 EMC9 RFX1 ZFPL1 B4GALNT1 SLC35E3 PHKG2 GNL1 OR52A5 EML3 TSC22D4 DUSP2 LIN37 BAG6 BCL6B ANXA9 MZT2A ARRB2 IER2 TOX4 CDK20 FAM151A SLC1A5 OR10G6 RASGRP2 MMP12 SLC25A45 OR5L1 TFPT C10ORF131 C11ORF58 ENSG00000238163 ENSG00000250246 ENSG00000255613 U82695.1
Newly annotated missing genes in Mallard genome [Zhu et al., 2021]	89	*ZBTB39 GPR182 OLFM2 HIPK4 HSPA12B WNT1 ADCY6 HOXC6 GPAA1 KLHL33 GABBR1 SYT3 NR4A1 SOX12 SLC44A2 USP39 GUCY2D FMNL3 PRPH SHMT2 CCDC65 CACNB3 DCC KANSL2 MPZ OPLAH SGCA ILF3 CYP27B1 OSBPL7 FARSA TBX21 ERBB3 TBKBP1 ALDOA PPP5C MYL6 POLI TARBP2 COL5A3 RAB27B STAT6 SLC6A8 RAB5B PDLIM2 ATG4D PACS1 JOSD2 BLVRB RCE1 YIF1B ANKRD39 SAE1 ESYT1 FLOT1 KRI1 LETMD1 METTL1 TSFM RPS26 STAC3 PPOX TMEM150A ARHGAP9 UBA1 TINF2 TECR B9D2 DDIT3 FKBP11 TMEM88 RBCK1 DAZAP2 DCTN2 UBL5 TMEM147 TMEM205 STAP2 STARD6 RLN3 IRF3 RABGGTA CCDC68 PDZD4 C12ORF44 C18ORF54 C1ORF192 C2ORF68 CCDC88B*
Newly annotated missing genes in chicken genome (Gallus_gallus-5.0) [Warren et al., 2017]	240	*ASNA1 ATP5B AVIL B4GALT3 CALR CDK2 CFAP126 (aka C1orf192) COPZ1 ECSIT ERBB3 ESYT1 EXOSC5 FARSA FBXL12 FBXW9 GPR182 HNRNPUL1 HOXC6 IKZF4 KCNH2 KLHL33 MIP NOS3 PRIM1 PRPH RASAL3 RDH5 RPS26 S100A10 SCNM1 SDHC SEMA4C SMARCC2 SOX12 STAP2 TARBP2 TBX21 TGFB1 TNPO2 TTC9C YIF1B ZBTB39 ADCY6 ADGRL1 aka LPHN1 AKT2 ANKRD39 APEX1 ARF3 ATAT1 ATG4D BBS1 BCAP31 C12orf44 aka ATG101 C19orf52 CACNA1A CACNB3 CACNG7 CACNG8 CAMSAP3 CCDC120 CCDC130 CCDC65 CCDC97 CCNT1 CHD8 CLASRP CLEC17A CLPP CSAD CSRNP2 CYTH2 DAZAP2 DCTN2 DDIT3 DNAJB1 DNM2 DPF1 ETFB EXOSC4 FKBP11 FLOT1 FMNL3 FUS FUZ GABBR1 GATA1 GEMIN7 GNG3 GNL3L GPAA1 GPKOW GRIK5 GTF2F1 HCFC1 HDAC6 HIF3A HOOK2 HSPBP1 ILF3 IPO4 JOSD2 JUNB KANSL2 KCNA7 KEAP1 KHSRP KMT5C KRI1 L1CAM LDLR LENG8 LETMD1 LIN7B LRRC4B LSMD1 aka NAA38 MAP2K7 MAP4K1 MARK2 MARS MBOAT7 METTL21B METTL3 MMP14 MPZ MRPL52 MYBPC2 NDUFB7 NKPD1 NOSIP NR4A1 NRXN2 OS9 OTUD5 PIH1D1 PLD3 POU2F2 POU6F1 PPOX PPP1R10 PPP1R12C PPP1R9B PPP4C PPP5C PRKCG PRMT5 PRPF31 PRX PSMD8 QPRT RAB4B RABGGTA REC8 RELB RENBP RING1 RNASEH2A RNF31 RUVBL2 SAE1 SCAF1 SCN1B SETD1A SHANK1 SHMT2 SIPA1L3 SLC11A2 SLC17A7 SLC35A2 SLC39A7 SLC44A2 SMC1A SMG9 SNRNP70 SPINT2 SPRYD3 SPTBN4 SRPK3 SSR4 STAC3 STAT6 STIP1 STRN4 TECR TFCP2 TFE3 TFPT TINF2 TNNT1 TRMT112 TRPT1 TSFM TSPYL2 UBA1 UBL4A UBL5 USP39*

		WNT1 ZNF385A ZNF653 ZNF668 ZNF865 AGAP2 ARL2 ASPDH AXL BLVRB C19orf53 C6orf136 CATSPERB CDK16 CYP27B1 DCAF11 FAM50A FKBP2 FLRT1 GPT GRM6 GUCY2D KDMB KIF5A LENG1 LMTK3 LTBP4 NOP9 OPLAH OSGEP PHF8 PIP4K2C PPP1R18 PQBP1 PRKCSH PSENEN RCE1 SRCAP STX1B THOC6 TRAPPC1 UXTWDR45XAB2
Newly annotated missing genes in chicken genome (Silkie)	136	*ZBTB12 ZNF865 GABBR1 LENG9 PPP1R10 BAG6 FLOT1 ABCF1 C6orf136 LSM2 MRPS18B VARS ATAT1 FKBPL ERBB3 DNM2 OSBPL7 MARK2 PACS1 RASAL3 ILF3 RUVBL2 ESYT1 FARSA ZBTB39 RFX1 POLD1 CSAD UBA1 OPLAH KEAP1 PRPF31 KANSL2 ALDOA XAB2 U2AF2 SPRYD3 WNT1 MAP2K7 PRIM1 POU6F1 B4GALNT1 KLHL33 KHSRP PPP1R37 RASIP1 STX1B ADCK5 PORCN STAT6 OSGEP GPR182 TARBP2 WDR45 PPP5C RDH5 MAP4K1 TBKBP1 CCDC65 RBM23 PPP1R12C CACNG7 TRMT1 ARF3 TECR USP39 RNF31 SMG9 PSMB5 DCTN2 SLC35A2 CLPP PRKCSH OS9 SYMPK RPL13A SNRNP70 CCDC22 EXOSC5 NPHS1 METTL1 DTX3 RNASEH2A MPZ PABPN1 NOSIP RPL18 CARM1 ASPDH LETMD1 TGFB1 STAC3 SGCA SOX12 SYP GTF2F1 BCAP31 TTC5 GPAA1 CCDC130 MED25 KRI1 EMC4 RENBP TMEM147 ANKRD39 ATP6AP1 QPRT RCE1 PIH1D1 LIN37 BAX WRAP53 CYB5D1 PSMD8 BSCL2 RBCK1 TMEM150A TSPAN31 CCDC97 MBOAT7 DAZAP2 ALKBH7 THOC6 FUZ PPP1R3E XH1 CCDC106 TFPT UXT BBS1 RLN3 TSFM DDIT3 CD2BP2 NOP9 TNF-α LEPTIN*

*: The gene has also been recovered from chicken RNA-Seq data in recent studies [Hron et al., 2015, Bornelov et al., 2017, Botero-Castro et al., 2017].