The chicken pan-genome reveals gene content variation and a promoter region deletion in *IGF2BP1* affecting body size

Kejun Wang[1,2†], Haifei Hu[3†] ,Yadong Tian[1,2], Jingyi Li[4], Armin Scheben[5], Chenxi Zhang[1,2], Yiyi Li[1,2], Junfeng Wu[1,2], Lan Yang[1,2], Xuewei Fan[1,2], Guirong Sun[1,2], Donghua Li[1,2], Yanhua Zhang[1,2], Ruili Han[1,2], Ruirui Jiang[1,2], Hetian Huang[1,2], Fengbin Yan[2], Yanbin Wang[2], Zhuanjian Li[1,2], Guoxi Li[1,2], Xiaojun Liu[1,2], Wenting Li[1,2]*, David Edwards[3]*, Xiangtao Kang[1,2]*

[1]College of Animal Science and Technology, Henan Agricultural University, Zhengzhou 450046, China

[2]Henan Key laboratory for innovation and utilization of chicken germplasm resources,Zhengzhou, 450046, China

[3]School of Biological Sciences and Institute of Agriculture, University of Western Australia, Crawley, 6009 WA, Australia

[4]Key Laboratory of Agricultural Animal Genetics, Breeding and Reproduction of Ministry of Education, College of Animal Science and Technology, Huazhong Agricultural University, 430070 Wuhan, Hubei, China

[5]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA

[†]These authors contributed equally to this work.

[*]Correspondence and requests for materials should be addressed to W. L. (email:liwenting_5959@hotmail.com), D.E. (email: dave.edwards@uwa.edu.au) and X.K. (email:xtkang2001@263.net).

Abstract

Domestication and breeding have reshaped the genomic architecture of chicken, but the retention and loss of genomic elements during these evolutionary processes remain unclear. We present the first chicken pan-genome constructed using 664 individuals, which identified an additional ~66.5 Mb sequences that are absent from the reference genome (GRCg6a). The constructed pan-genome encoded 20,491 predicated protein-coding genes, of which higher expression level are observed in conserved genes relative to dispensable genes. Presence/absence variation (PAV) analyses demonstrated that gene PAV in chicken was shaped by selection, genetic drift, and hybridization. PAV-based GWAS identified numerous candidate mutations related to growth, carcass composition, meat quality, or physiological traits. Among them, a deletion in the promoter region of *IGF2BP1* affecting chicken body size is reported, which is supported by functional studies and extra samples. This is the first time to report the causal variant of chicken body size QTL located at chromosome 27 which was repeatedly reported. Therefore, the chicken pan-genome is a useful resource for biological discovery and breeding. It improves our understanding of chicken genome diversity and provides materials to unveil the evolution history of chicken domestication.

Introduction

Chicken (*Gallus gallus*) is the most abundant domesticated animal in the world. The publication of the chicken genome in 2004 (Hillier, et al. 2004) paved the way to identify the QTLs or QTGs involved in economically important traits, dissect the evolutionary processes of domestication, and understand the genetic basis of distinct phenotypes differentiating domesticated chickens and their wild relatives. Recently, the domestic chicken *Gallus gallus domesticus* was reported to have been domesticated from one subspecies of red jungle fowl, *Gallus gallus spadiceus* (Wang, Thakur, et al. 2020b). Nevertheless, subspecies of *Gallus gallus* and other jungle fowls can introgress with *Gallus gallus domesticus* and these interspecies hybridizations have affected the

genetic content of the species during evolution (Barton 2001; DESTA 2019; Lawal, et al. 2020; Wang, Thakur, et al. 2020b). Traits such as yellow skin, pencilled feathers and the spotted comb of domesticated chickens are likely the result of introgressions from *Gallus sonneratii*, *Gallus varius* and *Gallus lafayettii* (Morejohn 1968b; Eriksson, et al. 2008; Fallahshahroudi, et al. 2019). Hybridizations leading to fertile offspring have been documented between *Gallus* species (DANFORTH 1958; Morejohn 1968a). These indicate that *Gallus gallus domesticus* is an admixed species, not only derived from red jungle fowl (Wang, Thakur, et al. 2020a). A recent study also found different genome sizes between red jungle fowl and domestic chicken lineages (Piegu, et al. 2020). Moreover, growing evidence suggests that structural variations are present in a substantial proportion of the genomes of many animals (Bickhart and Liu 2014), including human (Sherman, et al. 2019), pig (Zhao, et al. 2016; Li, Chen, et al. 2017; Tian, et al. 2020), salmon (Bertolotti, et al. 2020), and chicken (Kerstens, et al. 2011; Seol, et al. 2019). A range of phenotypes in chicken was reported to be determined by structural variations, such as feathered legs (Li, Lee, et al. 2020), crest (Li, et al. 2021), blue egg shell (Wang, et al. 2013), muffs and beard (Guo, et al. 2016), comb (Wright, et al. 2009; Imsland, et al. 2012), and fibromelanosis (Dorshorst, et al. 2011). The current chicken reference genome (GRCg6a) is derived from a single red jungle fowl individual. This reference therefore cannot fully capture the genetic diversity of domesticated chickens, and may be unable to reveal the genetic basis of some phenotypes. Recently, an increasing number of reports for pan-genomes in human (Sherman, et al. 2019), pig (Tian, et al. 2020), goat (Li, et al. 2019), and also in plants (Bayer, et al. 2020), have focused on capturing genetic variations between different individuals within the species. The pan-genome represents the gene set of the species rather than a representative individual, which can uncover the genetic diversity and resolve structural variations that are missed by studies using a single reference genome. The pan-genome can also provide a straightforward way to detect presence/absence variations (PAV) and explore the distributions of these variants at the population level.

Body size is an important quantitative trait that has been intensively selected during

chicken improvement and possibly associated with genome structural variations. One of the well-known candidate genes linked to body size is insulin-like growth factor 2 mRNA-binding protein 1 (*IGF2BP1*). *IGF2BP1* can regulate cell proliferation, differentiation, morphology, and metabolism through regulating mRNA localization, stability, and translation of targeted genes (Stohr, et al. 2012; Bell, et al. 2013). In recent studies, *IGF2BP1* was reported as N6-methyladenosine (m6A) readers to regulate the above functions (Huang, et al. 2018; Zhu, et al. 2020; Zhang, et al. 2021). Knockout of *IGF2BP1* in mouse led to mild active colitis, mild-to-moderate active enteritis, and decreasing of barrier function and body weight (Singh, et al. 2020). Dwarfism and impaired gut development were also observed in IGF2BP1-deficient mice (Hansen, et al. 2004). Evidence from genome-wide association studies (GWAS) and QTL mapping revealed that the genomic regions upstream of *IGF2BP1* were significantly associated with body weight, head weight, gizzard weight, chest width, leg weight, and wing weight in chicken and duck (Sheng, et al. 2013; Ma, et al. 2018; Zhou, et al. 2018; Wang, Bu, et al. 2020; Wang, Cao, et al. 2020; Zhang, et al. 2020). However, the causal variations of *IGF2BP1* that responsible for body sizes in chicken and duck remain unclear.

Here, we constructed the first chicken pan-genome and comprehensively investigate PAV using this pan-genome, revealing changes in allele frequencies associated with chicken evolution. We found that deletions in the promoter region of *IGF2BP1* can increase transcriptional activity and gene expression, regulating the body size in commercial chickens. Dissection of the causal variation of *IGF2BP1* associated with body size can accelerate the breeding process for high growth rate chickens using marker-assisted selection. These findings will improve our understanding of changes in chicken gene content during domestication and breeding and help to design highly productive chicken breeds in the future

Results

Pan-genome construction of chicken

We constructed the first *Gallus gallus* pan-genome using an iterative mapping and

assembly approach based on the chicken reference genome GRCg6a assembly. A set of whole-genome sequencing data including 664 individuals was used in the pan-genome construction, which contains 5 *Gallus gallus* wild subspecies, 28 native breeds (indigenous chicken breeds raised by farmers that did not experience intense artificial selection) and 4 commercial breeds (Supplementary Table S1; Figure 1a).

The *Gallus gallus* pan-genome identified an additional ~66.5 Mb sequences that are absent from the reference genome (GRCg6a), encoding an additional 4,063 high-confidence genes (Supplementary Table S2-S3). Of these, 49% (1,976 genes) non-reference genes are only present in a small proportion of chickens (Figure 1b). Together, the chicken pan-genome, including reference and non-reference sequences, consists of 1131.9 Mb and contains 20,941 predicted protein-coding genes. A total of 81 RNA-seq datasets from 27 tissues (including digestive, respiratory, kinetic, urinary, reproductive, endocrine, circulatory, nervous, immune, epithelium, and connective system) were used to investigate the gene expression (Supplementary Table S4). We observed an average normalized transcript per million (TPM) abundance greater than 1 for 90.6% of the autosomal genes in the reference genome and 19.4% of the non-reference genes. This pattern is similar to those found in other pan-genome studies (Zhao, et al. 2018; Gao, et al. 2019), which showed that genes in the reference genome generally have higher expression than genes in the non-reference contigs (Supplementary Figure S1a).

Discovery of gene Presence/Absence Variation (PAV)

After sample selection (see Methods and Supplementary Figure S2), a total of 268 individuals with average sequence depth larger than 10x based on pan-genome estimation were available for gene PAV detection, including 6 wild, 217 native and 45 commercial individuals (Supplementary Table S1).

We categorized genes in the chicken pan-genome according to their gene presence frequencies. 15,205 (76.32%) core genes are shared by 268 individuals. 4,738 genes are variable including 391 softcore, 2,351 shell and 1,976 cloud genes, which are present in more than 99%, 1-99%, and less than 1% of all individuals, respectively (Figure 1b). The chicken pan-genome showed a moderate core gene content (76.32%) compared to

that of human (96.88%) (Duan, et al. 2019), mussel (69.2%) (Gerdol, et al. 2020) , and plants (35 ~81%) (Gao, et al. 2019). Gene Ontology (GO) enrichment results of each cluster of variable genes are presented in Supplementary Table S5. Variable genes were enriched in the function associated with reproduction, nutrient absorption, metabolic and biosynthetic process (Figure 1c). RNA-seq analysis revealed that the expression level of flexible genes (shell and cloud genes) was significantly lower than that of conserved genes (core and softcore genes) (Supplementary Figure S1b). No apparent difference of expression was identified between conserved genes in the reference and non-reference sequences, but the expression of conserved genes was significantly higher than that of flexible genes in both reference and non-reference sequences (Supplementary Figure S1c). Pan-genome modelling revealed a closed pan-genome with an estimated total of 19,190 genes (genes on sex chromosomes were excluded because the gene content was different between chromosomes Z and W) (Figure 1d). This suggests the chicken pan-genome assembled using our selected 268 individuals included all or nearly all of the *Gallus gallus* gene content.

Gene PAV shaped by selection, genetic drift and hybridization

We observed a broad gene PAV distribution within different groups, with substantial variation in the native chickens and wild relatives (red jungle fowls) (Figure 2a). PAV-based PCA and phylogenetic analysis also showed high diversity among wild relatives and native chickens, while commercial broiler and layer clustered together (Figure 2b-c). Moreover, two clades of commercial chickens were further separated into two groups, meat-production (two broiler breeds: BRA and BRB) and egg-production (two layer breeds: BL and WL). These differences between commercial and native or wild chickens are likely due to selection, but the genetic drift and other factors can not be ruled out. Therefore, we further investigated whether selection, genetic drift, and hybridization can alter gene PAV content since SNP-based allelic frequencies can be shaped by selection, genetic drift, and hybridization (Edwards 2008).

We analyzed the pool sequencing data of 'Virginia body weight lines' and compared the gene PAV content between high weight selected (HWS) lines and low

weight selected (LWS) lines which were divergently selected from the same founder White Plymouth Rock population (Lillie, et al. 2018). Two lines had been suffered from intensive bidirectional selection for 8-week body weight, between which about 15-fold phenotypic difference presented. PAV-based PCA and phylogenetic analysis showed two distinct clusters were consistent with selected lines (Supplementary Figure S3a-3b). This suggests that gene PAV can be shaped by intensively artificial selection. We have further compared the frequencies of gene PAV between HWS and LWS and identified the candidate genes related to the intensive bidirectional selection for body weight. Twenty-four genes were found to be completely absent in HWS and present in LWS or entirely present in HWS and absent in LWS (Supplementary Figure S3c). The well studied *SH3 domain containing ring finger 2* (*SH3RF2*, *ENSGAT00000090177*) gene, regulating appetite and affecting body weight, was also identified as one of these genes that is completely absent in HWS but present in LWS (Rubin, et al. 2010; Jing, et al. 2020).

We further investigated whether gene PAV within the chicken population can be affected by genetic drift or hybridization. Firstly, we studied conserved populations of varying size. The subpopulations GS1, GS2, and GS3 are from the Gushi chicken populations, of which GS1 (n=30) and GS2 (n=30) were sampled from a small conserved population in 2010 and 2019 respectively, while GS3 (n=30) was sampled from a large conserved population in 2019. The subpopulations XB1 (n=30) and XB2 (n=30), are the Xichuan black-bone chicken populations, which were sampled from a large conserved population in 2010 and 2019, respectively (Supplementary Note**;** Supplementary Figure S4a). We did not observe the change of PAV content during short period (less than 9 years), whatever in a small or big conservation population, by comparing XB1 and XB2, or GS1 and GS2. However, we observed an apparent division between GS1 or GS2 and GS3 based on the results of PCA and phylogenetic analysis (Supplementary Figure S4b-d). We also found a significant reduction of genetic diversity in GS1 and GS2 in comparison with GS3 based on SNP heterozygosity and allelic richness analysis (observed heterozygosity: 0.18 in GS1, 0.17 in GS2, 0.23 in

GS3; allelic richness: 0.53 in GS1, 0.47 in GS2, 0.75 in GS3). These results are consistent with significant differences in gene PAV content (Supplementary Figure S4b-d) and previous studies showing that small conserved populations suffer from genetic drift after long periods of isolation which leads to a reduction of genetic diversity (Whitlock 2000). Based on the above evidence, we compared the gene PAV frequencies between GS3 and GS1+GS2 to investigate gene PAV involving genetic drift by long periods of isolation. According to Fisher's exact test (FDR < 0.001) (Gao, et al. 2019), we only identified six genes that were significantly different in frequencies between GS3 and GS1+GS2 (Supplementary Figure S4e), and none of these genes has a clear functional annotation (annotated as proteins without known function). Of these, four gene PAVs were fixed or nearly fixed in GS1+GS2, that consistent with the reduction of genetic diversity of GS1 and GS2. Secondly, we compared the gene PAV between Gushi chickens and the Gushi×Anak F2 population (Supplementary Figure S5a) and identified a clear divergence between Gushi breeds and the F2 population according to the PAV distribution (Supplementary Figure S5b). PAV-based PCA and the phylogenetic analysis also revealed that Gushi pure breeds and hybrid population fall into two distinct clades. These results suggest a relatively larger effect of hybridization on gene content, which is also significantly more extensive than that from genetic drift by comparing their clustering (Supplementary Figure S5c-d).

Change of gene PAV frequency during breeding

Gene PAV can be shaped by domestication and improvement; therefore, PAV within populations can also be applied to track the evolutionary history of a species (Gao, et al. 2019; Guo, et al. 2020). By comparing the gene presence frequency between the commercial and native breeds, we identified 30 significantly increased genes and 83 significantly decreased genes associated with post-domestication breeding improvement (Supplementary Table S6; Supplementary Figure S6). Of these, 10 significant genes (7 increased and 3 decreased) are located on the reference genome. We observed that two uncharacterized genes (*PanGallus_Gene02610* and *ENSGALT00000098327*) are lost in modern breeds. We also observed four immune-

related genes significantly decreased during improvement, including a class I histocompatibility antigen (*ENSGALT00000081489*), a B-cell differentiation antigen CD72-like (*PanGallus_Gene00218*), a T-cell differentiation antigen CD6 (*PanGallus_Gene04583*), and an Immunoglobulin G-binding protein A (*PanGallus_Gene03891*).

Tibetan chicken living at the Tibetan Plateau shows the environmental adaptation to high altitudes, particularly to the hypoxic environment (Wang, et al. 2015). Therefore, we compared the gene PAV frequencies between Tibetan chicken and other lower land indigenous chicken to identify candidate genes associated with the environmental adaptation to high altitude (Supplementary Table S7). A total of 121 genes showing significant difference in PAV frequencies were identified, of which frequencies of 118 genes were significantly increased in Tibetan chicken. *Vasodilator-stimulated phosphoprotein* (*VASP*, *ENSGALT00000100137*) was found to have a high presence frequency (0.906) in Tibetan chicken compared to other lower land native chickens (0.476). *VASP* has been reported to protect endothelial barrier function during hypoxia (Schmit, et al. 2012). Vasculature of *VASP* deletion mouse exhibited patterning defects and lacks structural integrity, leading to edema and hemorrhaging (Furman, et al. 2007). This evidence suggests *VASP* is likely to play an essential role in vasculature function and structure in a hypoxic environment. *Transitional endoplasmic reticulum ATPase* gene (*ENSGALT00000056168*) was nearly completely lost in Tibetan chicken (frequency is 0.093), while had moderate frequency in other lower land native chickens (0.568). Previous studies revealed that transitional endoplasmic reticulum ATPase activity is significantly inhibited during hypoxia in rat and western painted turtles (Henrich and Buckler 2013; Smith, et al. 2015). This suggests that the absence of *transitional endoplasmic reticulum ATPase* gene is potentially associated with the adaptation to hypoxic environment.

Change of PAV frequency in promoter regions during breeding

Most PAV analysis in previous pan-genome studies has focused on the protein coding regions. However, further investigations of the roles of regulatory regions are also

required since they can affect gene expression and phenotype (Van Laere, et al. 2003; Swinnen, et al. 2019). Similarity between orthologous promoters drastically decreased when distance was longer than 2 kb from the gene transcription start site (TSS) (Keightley, et al. 2005). Therefore, promoter regions are generally anchored within the 2 kb upstream genomic region of the TSS (Farre, et al. 2007; Abe and Gemmell 2014). In this study, the promoter region was defined as the 3 kb upstream genomic region from TSS to maximize the captured promotor regions. In order to detect smaller PAV in the promoter region, we divided each of the promoter region into three windows: 0-1 kb, 1-2 kb, and 2-3 kb upstream of the gene and investigated the frequencies of PAV in each window (Figure 3). We observed that frequencies of 143 PAVs in the 0-1 kb region of commercial chickens were significantly different from that of native chickens, which contains 117 increased and 26 decreased. In the same comparison, the frequencies of 80 PAVs differed significantly in the 1-2 kb regions with 56 increased and 24 decreased, and 78 PAVs differed in the 2-3 kb regions with 55 increased and 23 decreased (Figure 3a-c; Supplementary Table S8). We found 12 genes in the olfactory receptor gene family that showed reduced presence frequency in the promoter regions of commercial chickens relative to native chickens (Supplementary Figure S7a). We also observed that the presence frequencies of the promoter region of 9 immunoglobulin related genes were altered during improvement (Supplementary Figure S7b). Genes with significantly altered PAVs frequencies in promoter regions during breeding were enriched mainly in the GO terms of modulation by virus of host process, cyclin-dependent protein kinase holoenzyme complex, and p53 binding (Supplementary Figure S8).

Interestingly, we found two PAVs located at both 1-2 kb and 2-3 kb upstream region of *IGF2BP1* gene respectively, which their presence frequencies are significantly less in commercial chickens than native chickens (Figure 3b-c). A high loss rate was observed in commercial breeds compared to native breeds, with a 1-2 kb promoter region presence frequency of 0.04 in commercial breeds and 0.83 in native breeds (Supplementary Table S8). Similarly, the 2-3 kb promoter region presence frequency

was 0.04 in commercial breeds and 0.89 in native breeds (Figure 3b-c).

PAV-based GWAS on promoter regions

To uncover traits determined by promoter region PAV, we further conducted PAV-based GWAS on the promoter regions using the Gushi×Anak population with 204 F2 individuals (Figure 3d-f). Anak chicken is a commercial broiler breed from Israel, while Gushi is an indigenous chicken of China that did not experience from an intensive selection. We identified 56 association events for 0-1 kb promoter regions, 61 for 1-2 kb promoter regions and 78 for 2-3 kb promoter regions (Supplementary Table S8). These association events are involved in 81 traits, including body size, growth, carcass, meat quality, and physiological traits (Supplementary Note). For example, the PAV for 2-3 kb upstream region of *ENSGALG00000052768* (low-density lipoprotein receptor precursor, *LDLR*) was functionally related to serum CREA (creatinine) level. *ENSGALG00000051173* (olfactory receptor 14C36-like) was found to be associated with ileum length (IL), jejunum length (JL) and cecum length (CL). We also found that the promoter region PAV of immune-related genes showed associations with production traits. For instance, *ENSGALG00000054397* (class I histocompatibility antigen, F10 alpha chain-like isoform X1) was associated with breast bone length (BBL12) and *ENSGALG00000050329* (class I histocompatibility antigen, F10 alpha chain-like isoform X1) was correlated with body weight at birth (BBW). *ENSGALG00000051088* (Gallus gallus class I histocompatibility antigen, F10 alpha chain-like) was linked with BBL12 and body slanting length at 12 weeks (BSL12) (Supplementary Table S8). Immunoglobulin related genes were also identified to correlate with production traits. *ENSGALG00000049846* (immunoglobulin-like receptor CHIR2D-751 precursor) was associated with breast muscle weight (BMW) and the ratio of head weight to body weight at 12 weeks (HR1). *ENSGALG00000045164* (leukocyte immunoglobulin-like receptor subfamily A member 2) was associated with breast muscle weight (BMW) and shank girth (SG8). *ENSGALG00000050779* (immunoglobulin superfamily member 1) was linked with six carcass composition traits, and *ENSGALG00000050638* (immunoglobulin-like receptor CHIR2D-878 precursor) was associated with shank

length (SL12) (Supplementary Table S8).

As expected, we also found that the promoter region of *IGF2BP1* was associated with growth traits, including claw weight (CW1), the ratio of claw weight to body weight (CR), double pinion weight (DPW) and semi-evisceration weight (SEW), based on PAV-based GWAS of both 1-2 kb and 2-3 kb upstream regions (Figure 3e-f; Supplementary Table S8). The most significant association was identified between *IGF2BP1* and CW1 (p = 1.92E-07) based on 1-2 kb PAV-based GWAS (Figure 3h-i).

Dissection of the structure and function of *IGF2BP1* promoter region

To dissect the structure of *IGF2BP1* promoter region, we comprehensively analyzed the results of PCR and WGS read mapping. Three alleles of *IGF2BP1* promoter region were identified, which were defined as W (wild type), L1 allele (3.2 kb deletion at GRCg6a chr27:6082202-6085435), and L2 allele (1.5 kb deletion at chr27:6083984-6085538) (Figure 4a; Supplementary Figure S9a). We conducted allele specific PCR to genotype these three alleles in wild, native and commercial chickens (Figure 4a-b). As expected, the W allele was dominant in native breeds and wild relatives. In contrast, all the two commercial broiler breeds and commercial crossed chickens mainly had absence variant (L1 or L2 alleles). Absence variant (L1 or L2) was also dominant in commercial layer breeds, except for White Leghorn breeds. This result is consistent with the distribution of 1-2 kb and 2-3 kb upstream region PAV frequency for the *IGF2BP1* promoter region, which showed that commercial breeds were almost uniform for the mutant absence variant. We also compared the promoter region of *IGF2BP1* between high weight selection (HWS) lines and low weight selections (LWS) line using their pool sequencing data (Lillie, et al. 2018). We found that L1 was fixed in high body weight lines, whereas W was fixed in the low body weight lines, including the relaxed selection lines (Supplementary Figure S9b). It implies that W and L1 alleles have been selected to be fixed at an earlier time, before the divergence of relaxed selection lines.

Via the single genotype marker association analysis, the associations between the L1 allele and the body size, body weight or carcass composition still hold true when enlarging the sample size of Gushi×Anak F2 population to 734 (Figure 5;

Supplementary Figure S10). The associated traits include claw weight (CW1, CR), shank length (SL12), breast bone length (BBL8 and BBL12), wing weight (DPW), evisceration weight (EW and SEW), head weight (HW1), carcass weight (CW), leg weight (LW and LMW), pelvis breadth (PB12), shank girth (SG12 and SG8), body slanting length (BSL8 and BSL12), gizzard weight (GW), body weight (BWHR, BW6 and BW10), and growth rate (GR0_4). In those traits, the L1L1 genotype was always linked to better performance of production (such as body size, carcass weight and body weight) than the WW genotype (Figure 5; Supplementary Figure S10). The significant association between the *IGF2BP1* genotype and body size confirms the PAV-based GWAS results in promoter regions (Figure 3). Of these, associations are most significant in traits CW1 (p = 2.32E-14) and CR (p = 3.70E-12), which account for 4.01% and 3.85% of the phenotypic variations, respectively. Interestingly, we observed a larger effect of L1 in females relative to males, which explained 11.5% in females and 7.3% in males of the phenotypic variations for CW1 trait (Figure 5a). Besides, a female phenotype variation of 8.2% for DPW and 6.2% for SL12 was explained by L1. These associations are also consistent with the chicken and duck SNP-based GWAS results, which indicated that SNPs located near *IGF2BP1* were associated with body weight, head weight, gizzard weight, chest width, leg weight, and wing weight (Ma, et al. 2018; Zhou, et al. 2018; Wang, Bu, et al. 2020; Wang, Cao, et al. 2020; Zhang, et al. 2020). Unexpectedly, L2 allele was not found in the F2 population.

To further verify the molecular effects of the deletions, luciferase expression levels were investigated to represent the transcriptional activity through transfecting three kinds of recombinant plasmids (pGL3-L1, pGL3-L2, and pGL3-W) into chicken DF-1 cells (Figure 6a). Before performing the luciferase activity experiment, we screened the genome region which inserted the pGL3 construct, and confirmed that we did not find any difference except the L1 and L2 deletion. Therefore, the activity difference among the three constructs was derived from the deletions. The transcriptional activities of these two deletions (L1 and L2) were significantly higher than that of wild type (W). Further, the activity of the L1 genotype is also higher than that of L2 (Figure 6a).

Subsequently, we compared the mRNA expression level between the L1L1 (Ross 308) and WW genotypes (Gushi chicken). The expression of *IGF2BP1* mRNA in WW genotype is significantly lower than that in the L1L1 genotype in almost all investigated tissues at 6 weeks of age (Figure 6b). To reduce the difference in the genetic background among individuals with different genotypes and investigate the effect of deletions more accurately, we performed cross-breeding between chickens with the same heterozygous genotypes (L1WxL1W and L2WxL2W) to generate L1L1, L2L2 and WW genotype chicken with half-sib or full-sib relationship, and then compared the expressions of *IGF2BP1*. In spleen and duodenum tissues at 3 weeks of age, we observed higher expressions in L1L1 and L2L2 than WW genotype, while L1L1 also showed a higher value than L2L2 (Figure 6c). This is completely consistent with the result of transcriptional activity (Figure 6a). We also observed three conserved elements located in L1 deletion based on 77 vertebrates basewise PhyloP conservation score (https://hgdownload.soe.ucsc.edu/goldenPath/galGal6/phastCons77way/), of which one conserved element located in L2 deletion (Figure 4a). These suggest the functional importance of three conserved elements which possibly regulating *IGF2BP1* expression.

Investigation of the genomic regions flanking the deletion

Since *IGF2BP1* was also reported as the body size candidate gene using SNP-based studies (Sheng, et al. 2013; Ma, et al. 2018; Wang, Bu, et al. 2020; Wang, Cao, et al. 2020; Zhang, et al. 2020), we explored the SNPs within the region from 10 kb upstream of L1 deletion and 10 kb downstream in order to test if any of the signal driving SNPs in previous studies could be the causal. SNP calling was done using the same 664 individual sequencing data for building the pan-genome; however, 210 individuals were excluded for the low quality of SNP calling or their heterozygous genotype. The remaining individuals include 325 WW, 117 L1L1, and 12 L2L2 samples. We searched for SNPs that associated with the deletions in three different ways, L1L1 vs. WW, L2L2 vs. WW, and L1L1 + L2L2 vs. WW. Altogether, five associated SNPs were detected. Among them, the highest PhyloP conservation score is 0.97, and that SNP (chr27:

6087849) is not within a conservation element. The other four SNPs have negative conservation scores. This implies that none of these five associated SNPs are highly conserved, which supports that the deletion is likely to be the only functional mutation within this region.

Discussion

Construction of the first chicken pan-genome and dissection of genetic changes in the chicken population

Here, we constructed the first pan-genome of chicken, capturing ~66.5 Mb novel sequences that are absent from the reference genome (GRCg6a). Similar novel additional pan-genome sequences were captured in pig (Tian, et al. 2020) (~72.5 Mb), human (Sherman, et al. 2019) (~296.5 Mb), and plants (Yao, et al. 2015; Golicz, Bayer, et al. 2016; Montenegro, et al. 2017) (15.8 Mb ~ 350 Mb). Absent sequences from the reference genome were predicted to encode additional 4,063 high-confidence genes. We also found about one-third of the gene PAV is variable among the 268 individuals used for PAV calling. This highlights the heterogenicity of genetic makeup among chicken breeds and shows a potential utility for further breeding (Gao, et al. 2019).

We observed that red jungle fowls and native chickens contained most of the genetic diversity of chickens, while limited genetic diversity was found in commercial chickens (Figure 2). This result is consistent with known reductions in genetic diversity of modern livestock compared to their wild ancestors (Malomane, et al. 2019; Frantz, et al. 2020). Similarly, peach (Guo, et al. 2020), chickpea (Varshney, et al. 2019) and tomato (Gao, et al. 2019) pan-genome studies found that their wild relatives and landraces are more genetically diverse compared with modern cultivars. We also found that intra-species gene content variation can be affected by selection, genetic drift, or hybridization (Supplementary Figure S3-S5). We proposed that the reduction of genetic diversity in commercial chickens might occur due to intensive artificial selection during breeding, but other factors can not be ruled out, such as genetic drift.

PAVs are associated with physiological traits and the presence frequency of immune related loci was reduced during modern chicken breeding

We found that the promoter region PAV of genes showed associations with physiological related traits, such as *LDLR* and olfactory receptors (Supplementary Table S8). Lipid accumulation can enhance *LDLR* expression leading to an increase of serum creatinine (Sun, et al. 2013; Zhang, et al. 2016). LDLR knockout mouse and rat showed substantial increases in plasma creatinine (Bisgaard, et al. 2016; Sithu, et al. 2017). Variation in the promoter region of *LDLR* may reduce its expression and further upregulate serum creatinine level. Olfactory receptors were first discovered in the olfactory epithelium, functioning in odorant recognition involving various physiological behaviors, such as food choice and intake. However, recent studies indicate that these genes are also expressed in the intestinal tract (Priori, et al. 2015; Kim, et al. 2017; Kotlo, et al. 2020), and olfactory receptors play a role in intestinal inflammatory reaction (Kotlo, et al. 2020), secretion (Kim, et al. 2017) and microbiota metabolites (Priori, et al. 2015). We also found olfactory receptor 14C36-like gene associated with ileum length (IL), jejunum length (JL) and cecum length (CL) (Supplementary Table S8). It is thus possible that olfactory receptors are involved in feed digestion and conversion via regulation of intestine development and thus were under selection during modern breeding (Supplementary Figure S7a).

We observed the presence frequency of immune related gene or promoter region (including MHC and immunoglobulin) decreased in commercial chicken compared with the native breed. Of these, some immune gene PAVs showed significant association with production traits (Supplementary Table S8). This is consistent with a previous report that a high immune response is negatively correlated with chicken egg production and body weight (Warner, et al. 1987). MHC genes are involved in immune recognition and susceptibility to infectious disease (Sommer 2005). There is a possible genetic linkage between MHC genes and growth or reproduction genes (Warner, et al. 1987). Another possible explanation is that increased productivity may also increase the metabolic burden of immune gene maintenance in modern breeds. A trade-off might occur between the conservation of production-related genes and the loss of immune-related genes due to human selection for desirable production traits (van der Most, et

al. 2011).

***IGF2BP1* deletion is the causal variant for a major QTL associated with body size**

Many QTGs or QTLs associated with chicken growth traits have been identified, of which loci located at chromosomes 27, 4 and 1 have the largest impact on growth in chicken (Sheng, et al. 2013; Ma, et al. 2018; Wang, Bu, et al. 2020; Wang, Cao, et al. 2020; Zhang, et al. 2020). To our knowledge, the study in 2003 was the first time to report that a large QTL region located between 4.0 Mb and 6.1 Mb in chromosome 27 was associated with chicken body size (Kerje, et al. 2003). After that, many studies identified the chicken growth trait QTL in chromosome 27, including the gene *IGF2BP1* by SNP-based GWAS (Sheng, et al. 2013; Ma, et al. 2018; Wang, Bu, et al. 2020; Wang, Cao, et al. 2020). Our previous GWAS also revealed a signal peak correlated to body size trait, which was located at the genomic upstream of *IGF2BP1* (Zhang, et al. 2020). GWAS in duck also revealed SNPs located at the genomic upstream region of *IGF2BP1* that showed significant association with body size traits, while a higher expression level of *IGF2BP1* is correlated to better performance (Zhou, et al. 2018). Altogether, *IGF2BP1* is a potential major gene associated with body size traits, but the causal variant regulating these traits has not been reported previously.

In this study, using a genotype-phenotype association, we found two mutant alleles in the *IGF2BP1* promoter region that contributed to larger body size. We also observed a stronger association in females than males (Figure 5; Supplementary Figure S10). We compared the phenotypes among L1W, L1L1, and WW chickens to estimate the inheritance mode of the deletions. Taking the CW1 trait as an example, we found no significant difference in CW1 between L1W and L1L1 ($p = 0.68$) in males, while both are significantly heavier than WW (L1W vs WW, $p = 1.26E-3$, L1L1 vs WW, $p= 2.60E-5$). We inferred that there is a possible dominant effect of L1 against W in males. In females, however, we found no significant difference between WW and L1W ($p = 0.42$), while L1L1 are significantly heavier than L1W ($p = 5.35E-7$) and WW ($p = 4.0E-6$). There is a possible recessive effect of L1 against W in females. One possible reason is that this autosomal deletion locus shows sex-influenced inheritance, with a dominant

effect in males and a recessive effect in females. There may be a putative binding site of androgen-mediated transcription factor located on this deletion region. We also found three conserved elements based on 77 vertebrates basewise PhyloP conservation score, suggesting a putative regulatory function (Figure 4a). These deletions in the promoter region may increase *IGF2BP1* expression by upregulating its transcriptional activity (Figure 6). Further studies are required to elucidate the upstream regulatory pathway.

Together with our GWAS analysis, the mutant genotype is associated with higher expression of *IGF2BP1* and improved productivity traits (Figure 5; Figure 6). Our findings are consistent with findings that higher expression of *IGF2BP1* is linked to the larger body size in duck (Zhou, et al. 2018). Although the *IGF2BP1* mutation only explains a moderate 2-4% of phenotypic variation, this is in fact a substantial effect for a complex quantitative trait like body size. For instance, in humans two key variants for lean body mass explained 0.23% and 0.16% of the variance (Zillikens, et al. 2017) and ~50 variants for height only explain ~5% of the variance (Yang, et al. 2010). After examining the flanking regions of the deletion, the only five SNPs correlated to the deletion showed extremely low conservation scores implying that the deletion is the unique functional variant in this region. Based on this combined evidence, we propose that the deletion in the *IGF2BP1* promoter region is the causal variant for the QTL located at chromosome 27 that was previously reported to be related to body size in chicken.

Conclusion

Collectively, this first chicken pan-genome provides a foundation for future chicken population genetics and evolutionary genomics studies. PAV analysis offers an opportunity to uncover genomic architecture and identify the change of gene content during domestication and improvement, helping the designing of future chicken breeds with desired traits. We dissect the causal variant of one of the major QTL contributing to body size in chicken using PAV-based GWAS. The deletions that we found can be applied as markers for breeding programs using marker-assisted selection. As pan-genomic studies become more common, PAV-based GWAS will provide a powerful

complement to SNP-based GWAS for identifying functional variants of economically or evolutionary important traits.

Materials and Methods

Genomic sequencing of chicken

A total of 868 individuals were used in this study, of which 664 were used to construct the chicken pan-genome (Supplementary Table S1). We downloaded 509 accessions, published in recent genome resequencing studies (Fan, et al. 2013; Wang, et al. 2015; Ulfah, et al. 2016; Li, Che, et al. 2017; Lawal, et al. 2018; Qanbari, et al. 2019; Huang, et al. 2020; Wang, Thakur, et al. 2020b), from the National Center for Biotechnology Information (NCBI) Sequence Read Archive database (Supplementary Table S1). Sequencing data of 150 Henan indigenous chickens and 204 Gushi×Anak F2 individuals were generated in this study, and further data for an additional 5 Xichuan black-bone chickens were generated in our previous study (Li, Sun, et al. 2020). Genomic DNA was extracted from chicken blood using Qiagen DNeasy Kit. Paired-end libraries with ~500 bp insert size were constructed and then subjected to sequencing using the BGISEQ-500 platform to generate paired-end 150 bp reads (BGI Genomics Co., Ltd. and Beijing Fuyu Biotechnology Co., Ltd, China). We also downloaded 10 pool sequencing data, including 5 HWS and 5 LWS pool data from the NCBI database using project number PRJNA516366 (Lillie, et al. 2018).

Pan-genome construction and annotation

Raw reads were processed to remove low quality reads and generate adaptor free clean reads using Trimmomatic (v0.36) (Bolger, et al. 2014). The pan-genome was constructed by a reference based iterative mapping and assembly approach using the GRCg6a assembly as a starting reference genome (Golicz, Batley, et al. 2016; Golicz, Bayer, et al. 2016). The reference-based iterative mapping and assembly approach (Golicz, Batley, et al. 2016; Golicz, Bayer, et al. 2016) was first applied in a pan-genome study of the crop *B.oleracea*. This approach allows using sequencing data from a large range of individuals from different populations to construct a pan-genome. Briefly, clean reads were mapped to the reference genome (Ensemble Gallus_gallus.GRCg6a.dna.toplevel.fa) using bowtie2 (v2.3.5.1) (Langmead and Salzberg 2012). Unmapped reads were extracted using SAMtools and then assembled

using MaSuRCA v3.3.1 (Zimin, et al. 2013). After pan-genome construction, newly assembled contigs of non-reference sequences with length larger than 500 bp were kept. Contaminant sequences were filtered by the following two steps. Firstly, contigs were aligned using blastn v2.9.0 (Camacho, et al. 2009) against the NT database (v5, 07-03-2019) of contaminant taxid groups, which includes archaea, viruses, bacteria, fungi and Viridiplantae. Secondly, the remaining contigs were classified and filtered using Kraken2 (v 2.0.9-beta) using the kraken2-microbial database, which consists of archaea, bacteria, fungi, protozoa, viral and human sequences. (https://lomanlab.github.io/mockcommunity/mc_databases.html) (Wood, et al. 2019). The unclassified contigs were defined as contamination-free. The final contamination-free non-reference sequences and the reference Gallus/GRCg6a genome were merged to generate the chicken pan-genome.

A custom repeat library was constructed by scanning the final non-reference sequence using RepeatModeler (v1.0.11) (Flynn, et al. 2020). A custom repeat library and the RepBase database (downloaded in June 2019) of vertebrates were used to detect the repeat sequences with RepeatMasker (v4.0.8) (Tarailo-Graovac and Chen 2009). The MAKER2 annotation pipeline was used to obtain a set of high-confidence annotation based on RNA-seq evidence, homologous protein evidence and *ab initio* gene prediction evidence(Holt and Yandell 2011). RNA-seq evidence was generated using Hisat2-Stringtie pipeline (Pertea, et al. 2016) with published data from available tissues (Supplementary Table S2). Protein sequences of chicken, human and other mammals and vertebrates were collected from the Uniprot database (https://www.uniprot.org/). *Ab initio* gene prediction was implemented using SNAP (Korf 2004) and Augustus (Stanke, et al. 2006) with the 'chicken' model selected. Finally, redundant assembled protein sequences were filtered with CD-HIT (Fu, et al. 2012) (-c 0.9 -n 5 -M 16000 -T 18) with the threshold of 90% similarity.

PAV calling

Gene PAV was determined based on the cumulative coverage of exons of each gene. The longest transcripts were retrieved as the gene body to avoid redundant gene counts.

If at least two reads covered more than 5% cumulative coverage of all exons, this gene was defined as present in an individual. Otherwise, it was defined as absent (Golicz, Bayer, et al. 2016). Clean reads were aligned to the pan-genome using BWA-MEM (v0.7.17) (Li and Durbin 2009) with default parameters and the sequences depth of each sample was captured using Mosdepth package (v0.2.5)(Pedersen and Quinlan 2018). High-depth sequencing data (>30x) is preferable to increase the robustness of PAV analysis; however, it is not economical to sequence large samples numbers at this depth. Low-depth data (<15x) is a viable and more economical means to carry out PAV analysis in large sets of diverse samples (Gao, et al. 2019; Sherman, et al. 2019; Jayakodi, et al. 2020). To estimate the impact of the sequencing depth on gene PAV calling, we extracted reads from reference genome individual with varying depths of sequences to determine the minimum sequence depth required to call a confident gene PAV. An average sequence depth of 10x was considered as the threshold for including a sample since this threshold is estimated to allow a 99.94% recovery rate of gene PAV (Gao, et al. 2019) (Supplementary Figure S2a). We also performed additional simulation analysis using sequencing data of random seven breeds and found 98.4%~99.5% of pan-genome genes can be called when the average sequence depth reaching 10x (Supplementary Figure S2b). Thus, to get a high confident PAV matrix, individuals with an average depth above 10x were kept to perform gene PAV calling. Additionally, the sequencing data of red jungle fowls in Thailand were reported to be contaminated by domestic chicken sequences and were removed for the PAV calling (Qanbari, et al. 2019; Wang, Thakur, et al. 2020b).

PAV calling for promoter region was performed using the same method of gene PAV calling that is described above but based on the gene promoter regions. We divided the promoter region into three 1 kb windows based on the distance to the transcription start site. The three blocks were in the 0-1 kb, 1-2 kb and 2-3 kb regions upstream to the transcription start site of genes in the reference genome. A PAV was considered as present if more than 50% cumulative coverage with at least two reads was identified, otherwise, it was considered absent (Golicz, Bayer, et al. 2016).

PAV analysis

The gene PAV matrix was subjected to population genetic analysis. Principal component analysis and neighbour-joining phylogenetic analysis were conducted using TASSEL5 (Bradbury, et al. 2007). To identify the PAV with frequency significantly changed during improvement, the PAV frequency of each gene was compared between the native breeds and commercial breeds. Fisher's exact test was employed to identify significant PAV with false discovery rate (FDR) 0.001 (Gao, et al. 2019). Significantly increased genes were defined as genes having a significantly higher frequency in the commercial breeds than the native breeds. Inversely, we consider genes with a significantly lower frequency as significantly decreased genes. To identify the promoter region with significantly changed during chicken improvement, PAV patterns were also analyzed using the same method as gene PAV frequency calculation.

PAV-based GWAS

PAV-based genome-wide association study (GWAS) was also implemented to identify the candidate genes associated with 151 traits in a GushixAnak F2 mapping population with 204 individuals. To reduce bias, gene PAVs were removed if they were located on sex chromosomes or showed a minor allele frequency less than 0.05. A general linear model (GLM) was employed for association analysis using TASSEL5 (Bradbury, et al. 2007), with sex and the first five PCA eigenvectors defined as fixed effects. A Bonferroni test was used to define the genome-wide significant (0.05/number of loci) or suggestive (0.1/number of loci) cut-off threshold.

GO annotation

Functional annotation of the pangenome was performed using command line Blast2GO (Conesa et al., 2005) v2.5. The pan-genome genes were aligned to the proteins in the Uniref90 database (downloaded on Sep 2019) using BLASTP (Camacho et al., 2009), and only alignments with E-values < $1x10^{-5}$ were used. Then, the BLAST results were reformatted to satisfy Blast2GO naming requirements. Gene ontology annotation of the variable genes was conducted by the R package topGO (Alexa, et al. 2006) using Fisher's exact test with the approach 'elim' used to correct for multiple comparisons.

Genotyping of *IGF2BP1* PAV and association analysis

Three primers, including one forward and two reverse primers, were designed based on the sequence of the *IGF2BP1* promoter region (Figure 4a). One pair of primer, *Asp-F* and *Asp-R*, was used for genotyping L1 and W allele, while another pair, *2k-F* and *Asp-R*, was used to genotype L2 and W (Figure 4a; Supplementary Table S8). PCR reaction was conducted as described below: 5 pmol of each primer, 100 ng of genomic DNA, 2 ul 10 x PCR buffer (Takara), 100 uM dNTP mixture and 1 ul Taq polymerase. Association analysis of the validation population was conducted between genotypes (L1L1, L1W and WW) in *IGF2BP1* PAV and 151 traits (Supplementary Note) in F2 population with 734 individuals using GLM as described as above. The value of marker R squared was used to explain the phenotype variation of *IGF2BP* loci, as computed from the marker sum of squares (SS) after fitting all other model terms divided by the total SS (Bradbury, et al. 2007).

Functional assay of *IGF2BP1* promoter region and *IGF2BP1* expression

Three kinds of *IGF2BP1* promoter region (L1, L2 and W) were cloned into pGL3-Basic luciferase vector (Promega) using *Clone-F* and *Clone-R* primer (Supplementary Table S9). All recombinant plasmids, together with the pRL-TK plasmid (Promega), were transfected into DF-1 (chicken fibroblast cell) cell line. After 48 hours, the transcriptional activity was investigated by the Dual-Luciferase Reporter Assay System (Promega). Quantitative PCR was conducted to investigate the mRNA level of *IGF2BP1* using primer *IGF2BP1-qF* and *IGF2BP1-qR* (Supplementary Table S9). The relative expression level of *IGF2BP1* was normalized by GAPDH using the $2^{-\triangle\triangle ct}$ method.

Investigating the flanking SNPs of deletions

The deletion and its flanking regions (chr27:6072202-6095435) were analyzed by GATK (v3.8) pipeline (McKenna, et al. 2010) using the same 664 individuals for building the pan-genome. Genotypes of the *IGF2BP1* deletion of each sample were determined by the GATK results and manually checking the alignments by IGV (version 2.4.3). Then samples with the same genotypes were grouped together, and the

SNP associated with the *IGF2BP1* genotypes was defined as 1) significant in the Chi-squared test, 2) the mutant allele of the SNP has an allelic frequency higher than 0.8 in the deletion group, and 3) the allelic frequency difference between the two compared groups greater than 0.5. Three different deletion groups were used in three different comparisons. They are L1L1 group, L2L2 group, and L1L1 + L2L2 group, all compared with WW group, respectively.

Availability of Data and Materials

All the sequence data generated in this study have been deposited in the National Genomics Data center (https://bigd.big.ac.cn) with the accession codes PRJCA004227 and PRJCA004441. Downloaded sequence data used in this study were presented in Supplementary Table S1. The chicken pan-genome and relevant data are available at the DRYAD database (https://doi.org/10.5061/dryad.7pvmcvds1)

Author contribution

K.W., H.H. and W. L. designed analysis, performed analysis and wrote manuscript; W.L., C.Z., Y.L., J.W., L.Y., and X.F. performed the wet-lab experiment; X.K., Y.T., G.S., D.L., Y.Z., R.H., R.J., F.Y., Y.W., Z.L., G.L, X.L. contributed to sample collection and construction of F2 resource population. J. L. and A.S. assisted with data analysis and manuscript revision. W. L., D.E. and X. K. conceived research designed analysis and revised manuscript.

Acknowledgements and funding information

We thank Leif Andersson for comments on an earlier version of this manuscript. We thank Longxian Zhang and Jiangying Huang for help on computing resource. This work was supported by the Program for Innovation Research Team of the Ministry of Education (IRT16R23), National Natural Science Foundation of China (31902144) and the Scientific Studio of Zhongyuan Scholars (30601985). H.H. thanks the China Scholarship Council for supporting his studies at the University of Western Australia.

Ethics declarations

Ethics approval for this study was obtained from Henan Agricultural University.

Competing interests

692 The authors declare that they have no competing interests.

693 **References**

694 Abe H, Gemmell NJ. 2014. Abundance, arrangement, and function of sequence motifs in the chicken
695 promoters. BMC Genomics 15:900.
696 Alexa A, Rahnenfuhrer J, Lengauer T. 2006. Improved scoring of functional groups from gene expression
697 data by decorrelating GO graph structure. Bioinformatics 22:1600-1607.
698 Barton NH. 2001. The role of hybridization in evolution. Mol Ecol 10:551-568.
699 Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. 2020. Plant pan-genomes are the new reference.
700 Nature Plants 6:914-920.
701 Bell JL, Wachter K, Muhleck B, Pazaitis N, Kohn M, Lederer M, Huttelmaier S. 2013. Insulin-like
702 growth factor 2 mRNA-binding proteins (IGF2BPs): post-transcriptional drivers of cancer progression?
703 Cell Mol Life Sci 70:2657-2675.
704 Bertolotti AC, Layer RM, Gundappa MK, Gallagher MD, Pehlivanoglu E, Nome T, Robledo D, Kent
705 MP, Rosaeg LL, Holen MM, et al. 2020. The structural variation landscape in 492 Atlantic salmon
706 genomes. Nature Communications 11.
707 Bickhart DM, Liu GE. 2014. The challenges and importance of structural variation detection in livestock.
708 Frontiers in Genetics 5.
709 Bisgaard LS, Bosteen MH, Fink LN, Sorensen CM, Rosendahl A, Mogensen CK, Rasmussen SE, Rolin
710 B, Nielsen LB, Pedersen TX. 2016. Liraglutide Reduces Both Atherosclerosis and Kidney Inflammation
711 in Moderately Uremic LDLr-/- Mice. PLoS One 11:e0168396.
712 Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data.
713 Bioinformatics 30:2114-2120.
714 Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. TASSEL: software
715 for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635.
716 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+:
717 architecture and applications. BMC Bioinformatics 10:421.
718 DANFORTH CH. 1958. GALLUS SONNERATI AND THE DOMESTIC FOWL. Journal of Heredity
719 49:167-170.
720 DESTA TT. 2019. Phenotypic characteristic of junglefowl and chicken. World's poultry science journal
721 2019 v.75 no.1:pp. 69-82.
722 Dorshorst B, Molin AM, Rubin CJ, Johansson AM, Stromstedt L, Pham MH, Chen CF, Hallbook F,
723 Ashwell C, Andersson L. 2011. A complex genomic rearrangement involving the endothelin 3 locus
724 causes dermal hyperpigmentation in the chicken. PLoS Genet 7:e1002412.
725 Duan Z, Qiao Y, Lu J, Lu H, Zhang W, Yan F, Sun C, Hu Z, Zhang Z, Li G, et al. 2019. HUPAN: a pan-
726 genome analysis pipeline for human genomes. Genome Biol 20:149.
727 Edwards AW. 2008. G. H. Hardy (1908) and Hardy-Weinberg equilibrium. Genetics 179:1143-1150.
728 Eriksson J, Larson G, Gunnarsson U, Bed'hom B, Tixier-Boichard M, Stromstedt L, Wright D, Jungerius
729 A, Vereijken A, Randi E, et al. 2008. Identification of the yellow skin gene reveals a hybrid origin of the
730 domestic chicken. PLoS Genet 4:e1000010.
731 Fallahshahroudi A, Sorato E, Altimiras J, Jensen P. 2019. The Domestic BCO2 Allele Buffers Low-
732 Carotenoid Diets in Chickens: Possible Fitness Increase Through Species Hybridization. Genetics
733 212:1445-1452.
734 Fan WL, Ng CS, Chen CF, Lu MY, Chen YH, Liu CJ, Wu SM, Chen CK, Chen JJ, Mao CT, et al. 2013.

Genome-wide patterns of genetic variation in two domestic chickens. Genome Biol Evol 5:1376-1392.
Farre D, Bellora N, Mularoni L, Messeguer X, Alba MM. 2007. Housekeeping genes tend to show reduced upstream sequence conservation. Genome Biol 8:R140.
Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A 117:9451-9457.
Frantz LAF, Bradley DG, Larson G, Orlando L. 2020. Animal domestication in the era of ancient genomics. Nat Rev Genet 21:449-460.
Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150-3152.
Furman C, Sieminski AL, Kwiatkowski AV, Rubinson DA, Vasile E, Bronson RT, Fassler R, Gertler FB. 2007. Ena/VASP is required for endothelial barrier function in vivo. Journal of Cell Biology 179:761-775.
Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, Burzynski-Chang EA, Fish TL, Stromberg KA, Sacks GL, et al. 2019. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat Genet 51:1044-1051.
Gerdol M, Moreira R, Cruz F, Gomez-Garrido J, Vlasova A, Rosani U, Venier P, Naranjo-Ortiz MA, Murgarella M, Greco S, et al. 2020. Massive gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel. Genome Biology 21.
Golicz AA, Batley J, Edwards D. 2016. Towards plant pangenomics. Plant Biotechnol J 14:1099-1105.
Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, Chan CK, Severn-Ellis A, McCombie WR, Parkin IA, et al. 2016. The pangenome of an agronomically important crop plant Brassica oleracea. Nat Commun 7:13390.
Guo J, Cao K, Deng C, Li Y, Zhu G, Fang W, Chen C, Wang X, Wu J, Guan L, et al. 2020. An integrated peach genome structural variation map uncovers genes associated with fruit traits. Genome Biol 21:258.
Guo Y, Gu X, Sheng Z, Wang Y, Luo C, Liu R, Qu H, Shu D, Wen J, Crooijmans RP, et al. 2016. A Complex Structural Variation on Chromosome 27 Leads to the Ectopic Expression of HOXB8 and the Muffs and Beard Phenotype in Chickens. PLoS Genet 12:e1006071.
Hansen TV, Hammer NA, Nielsen J, Madsen M, Dalbaeck C, Wewer UM, Christiansen J, Nielsen FC. 2004. Dwarfism and impaired gut development in insulin-like growth factor II mRNA-binding protein 1-deficient mice. Mol Cell Biol 24:4448-4464.
Henrich M, Buckler KJ. 2013. Cytosolic calcium regulation in rat afferent vagal neurons during anoxia. Cell Calcium 54:416-427.
Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MAM, Delany ME, et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695-716.
Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12:491.
Huang H, Weng H, Sun W, Qin X, Shi H, Wu H, Zhao BS, Mesquita A, Liu C, Yuan CL, et al. 2018. Recognition of RNA N(6)-methyladenosine by IGF2BP proteins enhances mRNA stability and translation. Nat Cell Biol 20:285-295.
Huang X, Otecko NO, Peng M, Weng Z, Li W, Chen J, Zhong M, Zhong F, Jin S, Geng Z, et al. 2020. Genome-wide genetic structure and selection signatures for color in 10 traditional Chinese yellow-feathered chicken breeds. BMC Genomics 21:316.

779 Imsland F, Feng C, Boije H, Bed'hom B, Fillon V, Dorshorst B, Rubin CJ, Liu R, Gao Y, Gu X, et al.
780 2012. The Rose-comb mutation in chickens constitutes a structural rearrangement causing both altered
781 comb morphology and defective sperm motility. PLoS Genet 8:e1002775.
782 Jayakodi M, Padmarasu S, Haberer G, Bonthala VS, Gundlach H, Monat C, Lux T, Kamal N, Lang D,
783 Himmelbach A, et al. 2020. The barley pan-genome reveals the hidden legacy of mutation breeding.
784 Nature.
785 Jing Z, Wang X, Cheng Y, Wei C, Hou D, Li T, Li W, Han R, Li H, Sun G, et al. 2020. Detection of CNV
786 in the SH3RF2 gene and its effects on growth and carcass traits in chickens. BMC Genet 21:22.
787 Keightley PD, Lercher MJ, Eyre-Walker A. 2005. Evidence for widespread degradation of gene control
788 regions in hominid genomes. PLoS Biol 3:e42.
789 Kerje S, Carlborg O, Jacobsson L, Schutz K, Hartmann C, Jensen P, Andersson L. 2003. The twofold
790 difference in adult size between the red junglefowl and White Leghorn chickens is largely explained by
791 a limited number of QTLs. Anim Genet 34:264-274.
792 Kerstens HHD, Crooijmans RPMA, Dibbits BW, Vereijken A, Okimoto R, Groenen MAM. 2011.
793 Structural variation in the chicken genome identified by paired-end next-generation DNA sequencing of
794 reduced representation libraries. BMC Genomics 12.
795 Kim KS, Lee IS, Kim KH, Park J, Kim Y, Choi JH, Choi JS, Jang HJ. 2017. Activation of intestinal
796 olfactory receptor stimulates glucagon-like peptide-1 secretion in enteroendocrine cells and attenuates
797 hyperglycemia in type 2 diabetic mice. Sci Rep 7:13978.
798 Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5:59.
799 Kotlo K, Anbazhagan AN, Priyamvada S, Jayawardena D, Kumar A, Chen Y, Xia Y, Finn PW, Perkins
800 DL, Dudeja PK, et al. 2020. The olfactory G protein-coupled receptor (Olfr-78/OR51E2) modulates the
801 intestinal response to colitis. Am J Physiol Cell Physiol 318:C502-C513.
802 Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357-359.
803 Lawal RA, Al-Atiyat RM, Aljumaah RS, Silva P, Mwacharo JM, Hanotte O. 2018. Whole-Genome
804 Resequencing of Red Junglefowl and Indigenous Village Chicken Reveal New Insights on the Genome
805 Dynamics of the Species. Front Genet 9:264.
806 Lawal RA, Martin SH, Vanmechelen K, Vereijken A, Silva P, Al-Atiyat RM, Aljumaah RS, Mwacharo
807 JM, Wu DD, Zhang YP, et al. 2020. The wild species genome ancestry of domestic chickens. BMC Biol
808 18:13.
809 Li D, Che T, Chen B, Tian S, Zhou X, Zhang G, Li M, Gaur U, Li Y, Luo M, et al. 2017. Genomic data
810 for 78 chickens from 14 populations. Gigascience 6:1-5.
811 Li D, Sun G, Zhang M, Cao Y, Zhang C, Fu Y, Li F, Li G, Jiang R, Han R, et al. 2020. Breeding history
812 and candidate genes responsible for black skin of Xichuan black-bone chicken. BMC Genomics 21:511.
813 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.
814 Bioinformatics 25:1754-1760.
815 Li J, Lee M, Davis BW, Lamichhaney S, Dorshorst BJ, Siegel PB, Andersson L. 2020. Mutations
816 Upstream of the TBX5 and PITX1 Transcription Factor Genes Are Associated with Feathered Legs in
817 the Domestic Chicken. Mol Biol Evol 37:2477-2486.
818 Li J, Lee MO, Davis BW, Wu P, Hsieh Li SM, Chuong CM, Andersson L. 2021. The crest phenotype in
819 domestic chicken is caused by a 197 bp duplication in the intron of HOXC10. G3 (Bethesda) 11.
820 Li M, Chen L, Tian S, Lin Y, Tang Q, Zhou X, Li D, Yeung CKL, Che T, Jin L, et al. 2017. Comprehensive
821 variation discovery and recovery of missing sequence in the pig genome using multiple de novo
822 assemblies. Genome Res 27:865-874.

Li R, Fu W, Su R, Tian X, Du D, Zhao Y, Zheng Z, Chen Q, Gao S, Cai Y, et al. 2019. Towards the Complete Goat Pan-Genome by Recovering Missing Genomic Segments From the Reference Genome. Front Genet 10:1169.
Lillie M, Sheng ZY, Honaker CF, Andersson L, Siegel PB, Carlborg O. 2018. Genomic signatures of 60 years of bidirectional selection for 8-week body weight in chickens. Poult Sci 97:781-790.
Ma M, Shen M, Qu L, Dou T, Guo J, Hu Y, Lu J, Li Y, Wang X, Wang K. 2018. Genome-wide association study for carcase traits in spent hens at 72 weeks old. Italian Journal of Animal Science:1-6.
Malomane DK, Simianer H, Weigend A, Reimer C, Schmitt AO, Weigend S. 2019. The SYNBREED chicken diversity panel: a global resource to assess chicken diversity at high genomic resolution. BMC Genomics 20:345.
McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297-1303.
Montenegro JD, Golicz AA, Bayer PE, Hurgobin B, Lee H, Chan CK, Visendi P, Lai K, Dolezel J, Batley J, et al. 2017. The pangenome of hexaploid bread wheat. Plant J 90:1007-1013.
Morejohn GV. 1968a. Breakdown of Isolation Mechanisms in Two Species of Captive Junglefowl (Gallus Gallus and Gallus Sonneratii). Evolution 22:576-582.
Morejohn GV. 1968b. Study of Plumage of the Four Species of the Genus Gallus. The Condor 70:56-65.
Pedersen BS, Quinlan AR. 2018. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics 34:867-868.
Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc 11:1650-1667.
Piegu B, Arensburger P, Beauclair L, Chabault M, Raynaud E, Coustham V, Brard S, Guizard S, Burlot T, Le Bihan-Duval E, et al. 2020. Variations in genome size between wild and domesticated lineages of fowls belonging to the Gallus gallus species. Genomics 112:1660-1673.
Priori D, Colombo M, Clavenzani P, Jansman AJ, Lalles JP, Trevisi P, Bosi P. 2015. The Olfactory Receptor OR51E1 Is Present along the Gastrointestinal Tract of Pigs, Co-Localizes with Enteroendocrine Cells and Is Modulated by Intestinal Microbiota. PLoS One 10:e0129501.
Qanbari S, Rubin CJ, Maqbool K, Weigend S, Weigend A, Geibel J, Kerje S, Wurmser C, Peterson AT, Brisbin IL, Jr., et al. 2019. Genetics of adaptation in modern chicken. PLoS Genet 15:e1007989.
Rubin CJ, Zody MC, Eriksson J, Meadows JR, Sherwood E, Webster MT, Jiang L, Ingman M, Sharpe T, Ka S, et al. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464:587-591.
Schmit MA, Mirakaj V, Stangassinger M, Konig K, Kohler D, Rosenberger P. 2012. Vasodilator Phosphostimulated Protein (VASP) Protects Endothelial Barrier Function During Hypoxia. Inflammation 35:566-573.
Seol D, Ko BJ, Kim B, Chai HH, Lim D, Kim H. 2019. Identification of Copy Number Variation in Domestic Chicken Using Whole-Genome Sequencing Reveals Evidence of Selection in the Genome. Animals 9.
Sheng Z, Pettersson ME, Hu X, Luo C, Qu H, Shu D, Shen X, Carlborg O, Li N. 2013. Genetic dissection of growth traits in a Chinese indigenous x commercial broiler chicken cross. BMC Genomics 14:151.
Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, Boorgula MP, Chavan S, Vergara C, Ortega VE, et al. 2019. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat Genet 51:30-35.

Singh V, Gowda CP, Singh V, Ganapathy AS, Karamchandani DM, Eshelman MA, Yochum GS, Nighot P, Spiegelman VS. 2020. The mRNA-binding protein IGF2BP1 maintains intestinal barrier function by up-regulating occludin expression. Journal of Biological Chemistry 295:8602-8612.
Sithu SD, Malovichko MV, Riggs KA, Wickramasinghe NS, Winner MG, Agarwal A, Hamed-Berair RE, Kalani A, Riggs DW, Bhatnagar A, et al. 2017. Atherogenesis and metabolic dysregulation in LDL receptor-knockout rats. JCI Insight 2.
Smith RW, Cash P, Hogg DW, Buck LT. 2015. Proteomic changes in the brain of the western painted turtle (Chrysemys picta bellii) during exposure to anoxia. Proteomics 15:1587-1597.
Sommer S. 2005. The importance of immune gene variability (MHC) in evolutionary ecology and conservation. Front Zool 2:16.
Stanke M, Schoffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7:62.
Stohr N, Kohn M, Lederer M, Glass M, Reinke C, Singer RH, Huttelmaier S. 2012. IGF2BP1 promotes cell migration by regulating MK5 and PTEN signaling. Genes Dev 26:176-189.
Sun H, Yuan Y, Sun ZL. 2013. Cholesterol Contributes to Diabetic Nephropathy through SCAP-SREBP-2 Pathway. Int J Endocrinol 2013:592576.
Swinnen G, Goossens A, Pauwels L. 2019. Lessons from Domestication: Targeting Cis-Regulatory Elements for Crop Improvement. Trends Plant Sci 24:1065.
Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics Chapter 4:Unit 4 10.
Tian X, Li R, Fu W, Li Y, Wang X, Li M, Du D, Tang Q, Cai Y, Long Y, et al. 2020. Building a sequence map of the pig pan-genome from multiple de novo assemblies and Hi-C data. Sci China Life Sci 63:750-763.
Ulfah M, Kawahara-Miki R, Farajalllah A, Muladno M, Dorshorst B, Martin A, Kono T. 2016. Genetic features of red and green junglefowls and relationship with Indonesian native chickens Sumatera and Kedu Hitam. BMC Genomics 17:320.
van der Most PJ, de Jong B, Parmentier HK, Verhulst S. 2011. Trade-off between growth and immune function: a meta-analysis of selection experiments. Functional Ecology 25:74-80.
Van Laere AS, Nguyen M, Braunschweig M, Nezer C, Collette C, Moreau L, Archibald AL, Haley CS, Buys N, Tally M, et al. 2003. A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. Nature 425:832-836.
Varshney RK, Thudi M, Roorkiwal M, He W, Upadhyaya HD, Yang W, Bajaj P, Cubry P, Rathore A, Jian J, et al. 2019. Resequencing of 429 chickpea accessions from 45 countries provides insights into genome diversity, domestication and agronomic traits. Nat Genet 51:857-864.
Wang M-S, Thakur M, Peng M-S, Jiang Y, Frantz LAF, Li M, Zhang J-J, Wang S, Peters J, Otecko NO, et al. 2020a. 863 genomes reveal the origin and domestication of chicken. Cell Research 30:693-701.
Wang MS, Li Y, Peng MS, Zhong L, Wang ZJ, Li QY, Tu XL, Dong Y, Zhu CL, Wang L, et al. 2015. Genomic Analyses Reveal Potential Independent Adaptation to High Altitude in Tibetan Chickens. Mol Biol Evol 32:1880-1889.
Wang MS, Thakur M, Peng MS, Jiang Y, Frantz LAF, Li M, Zhang JJ, Wang S, Peters J, Otecko NO, et al. 2020b. 863 genomes reveal the origin and domestication of chicken. Cell Res 30:693-701.
Wang Y, Bu L, Cao X, Qu H, Zhang C, Ren J, Huang Z, Zhao Y, Luo C, Hu X, et al. 2020. Genetic Dissection of Growth Traits in a Unique Chicken Advanced Intercross Line. Front Genet 11:894.
Wang Y, Cao X, Luo C, Sheng Z, Zhang C, Bian C, Feng C, Li J, Gao F, Zhao Y, et al. 2020. Multiple

ancestral haplotypes harboring regulatory mutations cumulatively contribute to a QTL affecting chicken growth traits. Commun Biol 3:472.
Wang Z, Qu L, Yao J, Yang X, Li G, Zhang Y, Li J, Wang X, Bai J, Xu G, et al. 2013. An EAV-HP insertion in 5' Flanking region of SLCO1B3 causes blue eggshell in the chicken. PLoS Genet 9:e1003183.
Warner C, Meeker D, Rothschild M. (2.092 co-authors). 1987. Genetic control of immune responsiveness: a review of its use as a tool for selection for disease resistance. Journal of animal science 64:394-406.
Whitlock MC. 2000. Fixation of new alleles and the extinction of small populations: drift load, beneficial alleles, and sexual selection. Evolution 54:1855-1861.
Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. Genome Biol 20:257.
Wright D, Boije H, Meadows JR, Bed'hom B, Gourichon D, Vieaud A, Tixier-Boichard M, Rubin CJ, Imsland F, Hallbook F, et al. 2009. Copy number variation in intron 1 of SOX5 causes the Pea-comb phenotype in chickens. PLoS Genet 5:e1000512.
Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, et al. 2010. Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565-569.
Yao W, Li G, Zhao H, Wang G, Lian X, Xie W. 2015. Exploring the rice dispensable genome using a metagenome-like assembly strategy. Genome Biol 16:187.
Zhang L, Wan YC, Zhang ZH, Jiang Y, Gu ZY, Ma XL, Nie SP, Yang J, Lang JH, Cheng WJ, et al. 2021. IGF2BP1 overexpression stabilizes PEG10 mRNA in an m6A-dependent manner and promotes endometrial cancer progression. Theranostics 11:1100-1114.
Zhang Y, Ma KL, Ruan XZ, Liu BC. 2016. Dysregulation of the Low-Density Lipoprotein Receptor Pathway Is Involved in Lipid Disorder-Mediated Organ Injury. Int J Biol Sci 12:569-579.
Zhang Y, Wang Y, Li Y, Wu J, Wang X, Bian C, Tian Y, Sun G, Han R, Liu X, et al. 2020. Genome-wide association study reveals the genetic determinism of growth traits in a Gushi-Anka F2 chicken population. Heredity (Edinb).
Zhao PJ, Li JH, Kang HM, Wang HF, Fan ZY, Yin ZJ, Wang JF, Zhang Q, Wang ZQ, Liu JF. 2016. Structural Variant Detection by Large-scale Sequencing Reveals New Evolutionary Evidence on Breed Divergence between Chinese and European Pigs. Scientific Reports 6.
Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, Zhan Q, Lu Y, Zhang L, Huang T, et al. 2018. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat Genet 50:278-284.
Zhou Z, Li M, Cheng H, Fan W, Yuan Z, Gao Q, Xu Y, Guo Z, Zhang Y, Hu J, et al. 2018. An intercross population study reveals genes associated with body size and plumage color in ducks. Nat Commun 9:2648.
Zhu S, Wang JZ, Chen, He YT, Meng N, Chen M, Lu RX, Chen XH, Zhang XL, Yan GR. 2020. An oncopeptide regulates m(6)A recognition by the m(6)A reader IGF2BP1 and tumorigenesis. Nat Commun 11:1685.
Zillikens MC, Demissie S, Hsu YH, Yerges-Armstrong LM, Chou WC, Stolk L, Livshits G, Broer L, Johnson T, Koller DL, et al. 2017. Large meta-analysis of genome-wide association studies identifies five loci for lean body mass. Nat Commun 8:80.
Zimin AV, Marcais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. Bioinformatics 29:2669-2677.

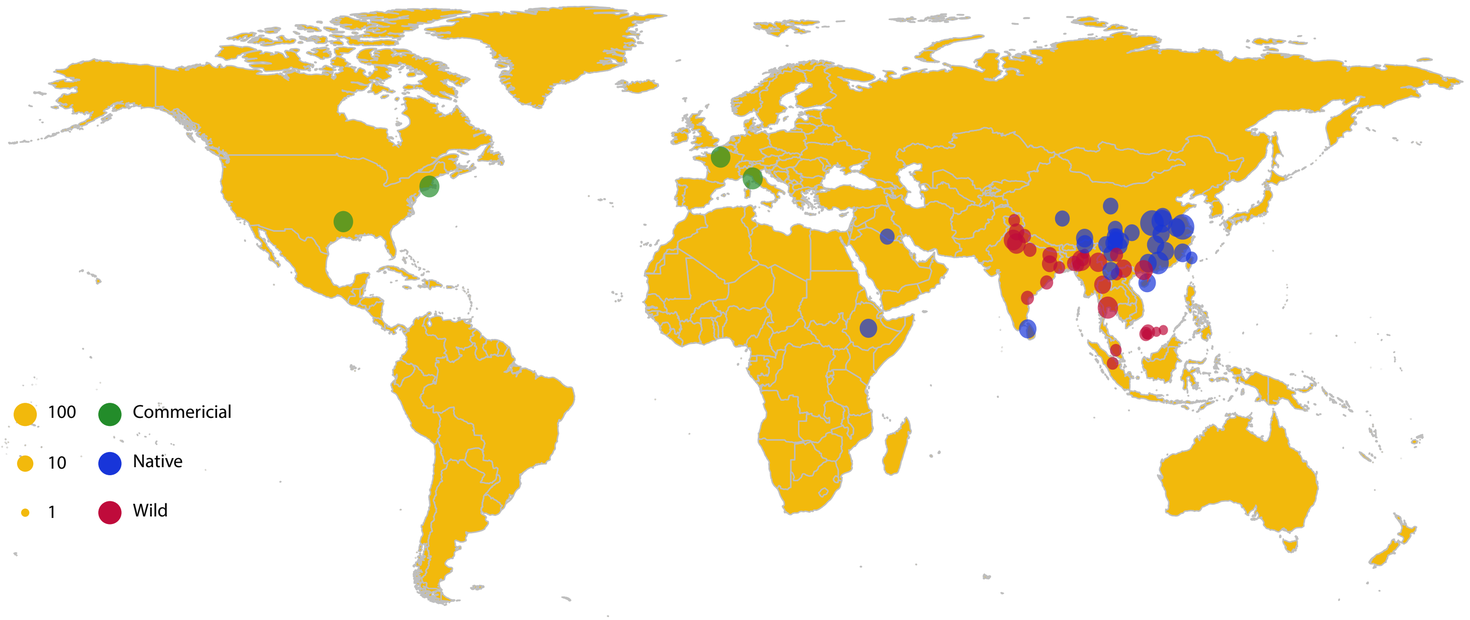
Figure legends

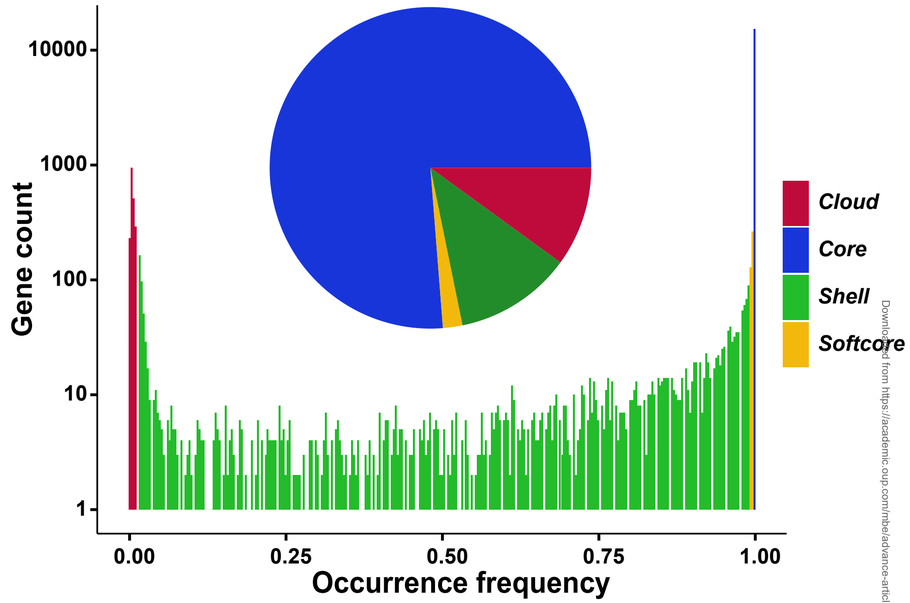
954 **Figure 1. Pan-genome of chicken. a** Geographical distribution of samples used for
955 pan-genome construction. **b** Pan-genome gene classification. **c** Word cloud of the Gene
956 Ontology (GO) enrichment of biological process for variable genes. **d** Pan-genome
957 modelling. The pan-genome modelling shows no more dramatic increases when the
958 number of accession genomes is over 220, indicating that selected individuals were
959 sufficient to capture the majority of PAVs within Gallus gallus. Upper and lower lines
960 represent the pan-genome number and core-genome number, respectively.
961 **Figure 2. Distribution of gene PAV. a** The heatmap shows the PAV of variable genes
962 within wild relatives, native breeds and commercial breeds. **b** The principal component
963 analysis of chicken breeds based on gene PAV. Wild: wild relatives (red jungle fowls);
964 Native, native breeds; commercial breeds consist of two broiler breeds (BRA and BRB)
965 and two layer breeds (BL and WL). **c** Neighbor-Joining phylogenetic tree constructed
966 based on gene PAV matrix.
967 **Figure 3. Change of PAV frequency in promoter region during breeding and PAV-**
968 **based GWAS. a**, **b**, **c**, Scatter plots showing gene occurrence frequencies in Native
969 breeds and Com (commercial) breeds for 0-1 kb (**a**), 1-2 kb (**b**) and 2-3 kb (**c**) upstream
970 promoter regions, respectively. **d**, **e**, **f**, Manhattan plots showing significant promoter
971 region PAVs associated with 151 traits for 0-1 kb (**d**), 1-2 kb (**e**), and 2-3 kb (**f**) upstream
972 promoter regions. All association analysis result was plotted according to the physical
973 location and p-value, with each dot representing an association analysis result. The
974 upper and lower dashed line represents the significant and suggestive thresholds,
975 respectively. CW1, claw weight; CR , the ratio of claw weight to body weight; DPW,
976 double pinion weight; SEW, semi-evisceration weight.
977 **Figure 4. Structure and frequency of the three alleles in *IGF2BP1* promoter region.**
978 **a** Genomic structure of three alleles in *IGF2BP1* promoter region in relation to
979 evolutionarily conserved elements (77 vertebrates basewise PhyloP conservation score).
980 Variant alleles in the promoter region of *IGF2BP1* include wild type (W) and two
981 mutant alleles (L1 and L2). The conserved elements are indicated by red arrows. *Asp-*
982 *F*, *2k-F* and *Asp-R* are the PCR primers for the identification of the allelic type. **b** Allelic

frequency of *IGF2BP1* promoter region in the validated population by allelic-specific PCR genotyping. PCR product sizes of W, L1 and L2 are 2345 bp, 290 bp and 791 bp, respectively. The gel shows the six genotypes derived from the combinations of the three alleles.

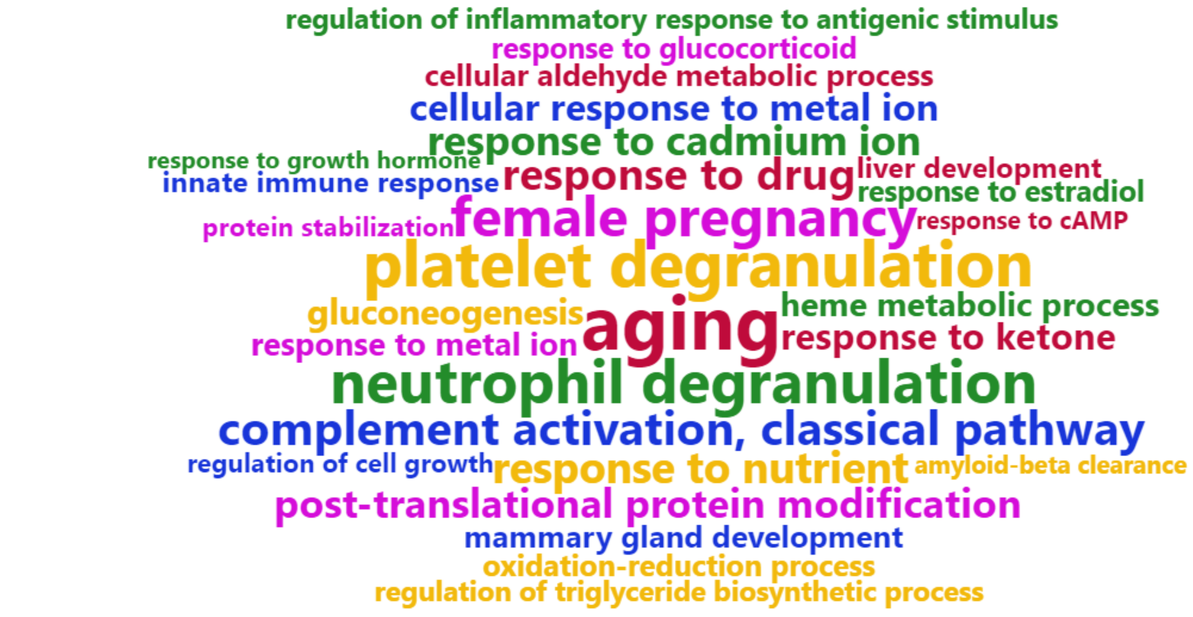
Figure 5. Single-marker genotype association of *IGF2BP1* promoter region in the validated Gushi×Anak F2 population with 734 individuals. Eight representative association events were included and others were showed in Supplementary Figure S10. The number in the bracket is the proportion of phenotype variance explained by *IGF2BP1* loci. CW1, claw weight; CR, the ratio of claw weight to body weight. SL12, shank length; BBL12, breast bone length; DPW, double pinion weight; SEW, semi-evisceration weight; CW, carcass weight; LW, leg weight. All traits were phenotyped at 12 weeks of age.

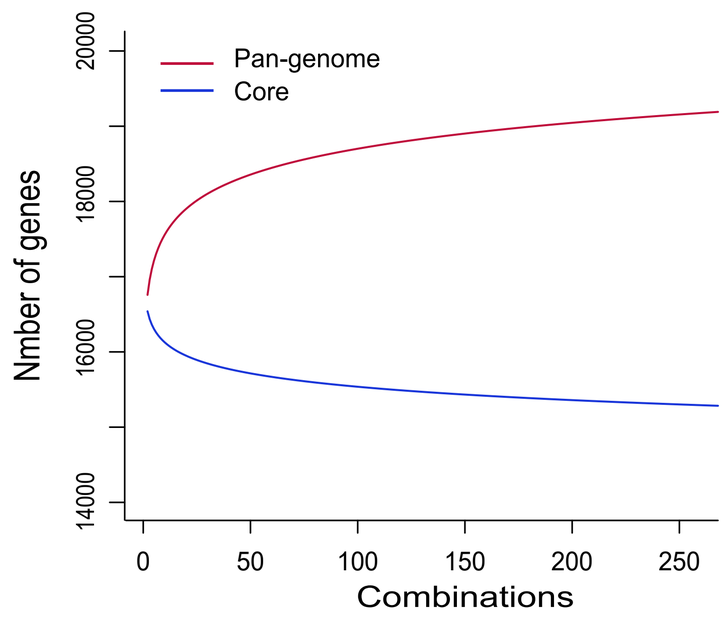
Figure 6. Comparison of transcriptional activity and expression among three *IGF2BP1* genotypes. a Comparison of transcriptional activity among different *IGF2BP1* promoter region in chicken DF-1 cells. Left shows the constructions of the inserted fragment into the pGL3-Basic plasmid. Significance of two-tailed Student's *t*-test: **, $p < 0.01$; ***, $p < 0.001$. **b** Comparison of mRNA expression of *IGF2BP1* between L1L1 (Ross 308) and WW (Gushi) chickens in five tissues at 6 weeks of age. Breast, breast muscle; Leg, leg muscle. P-values were calculated using a two-tailed Student's *t*-test. **c** Comparison of mRNA expression of *IGF2BP1* between L1L1, L2L2 and WW in an *IGF2BP1* genotype segregating population at 3 weeks of age.

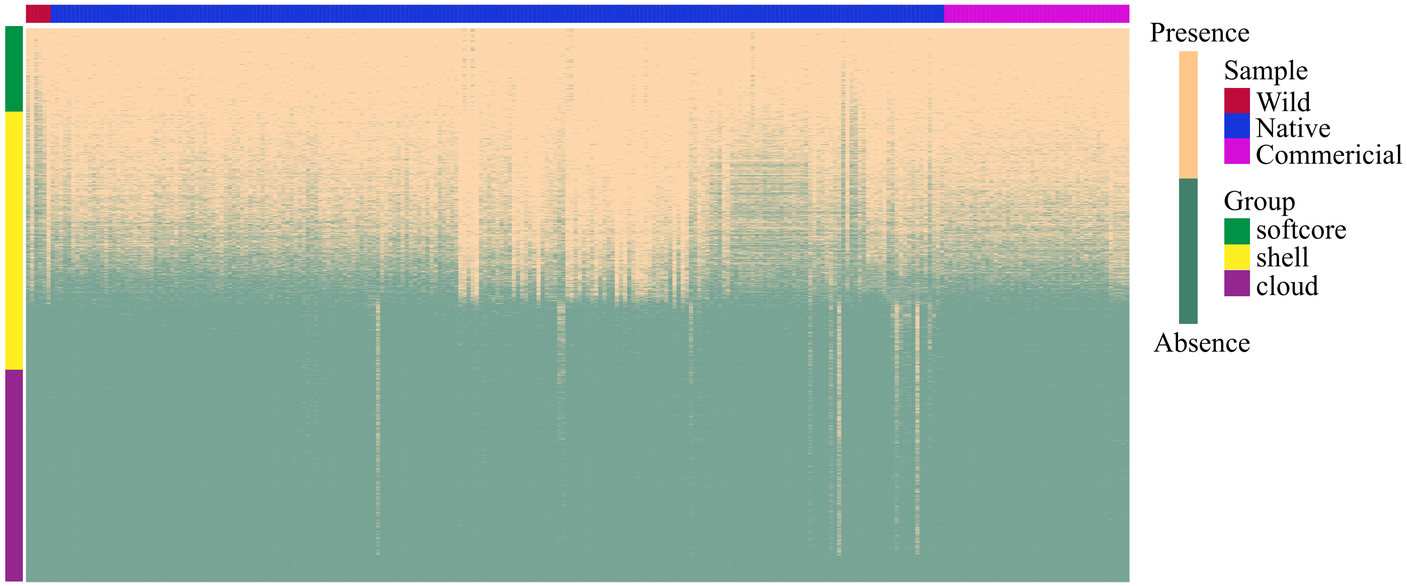


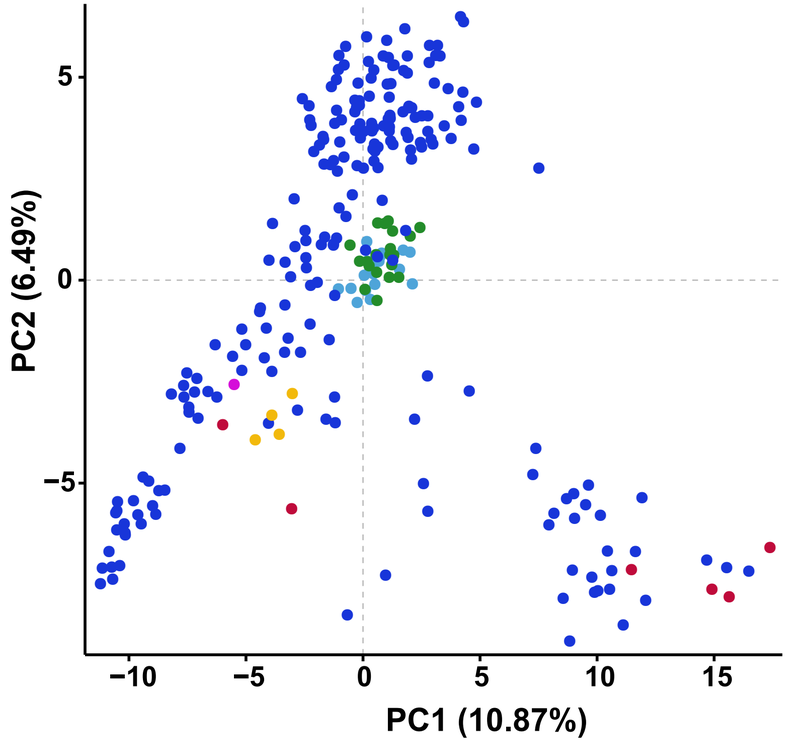


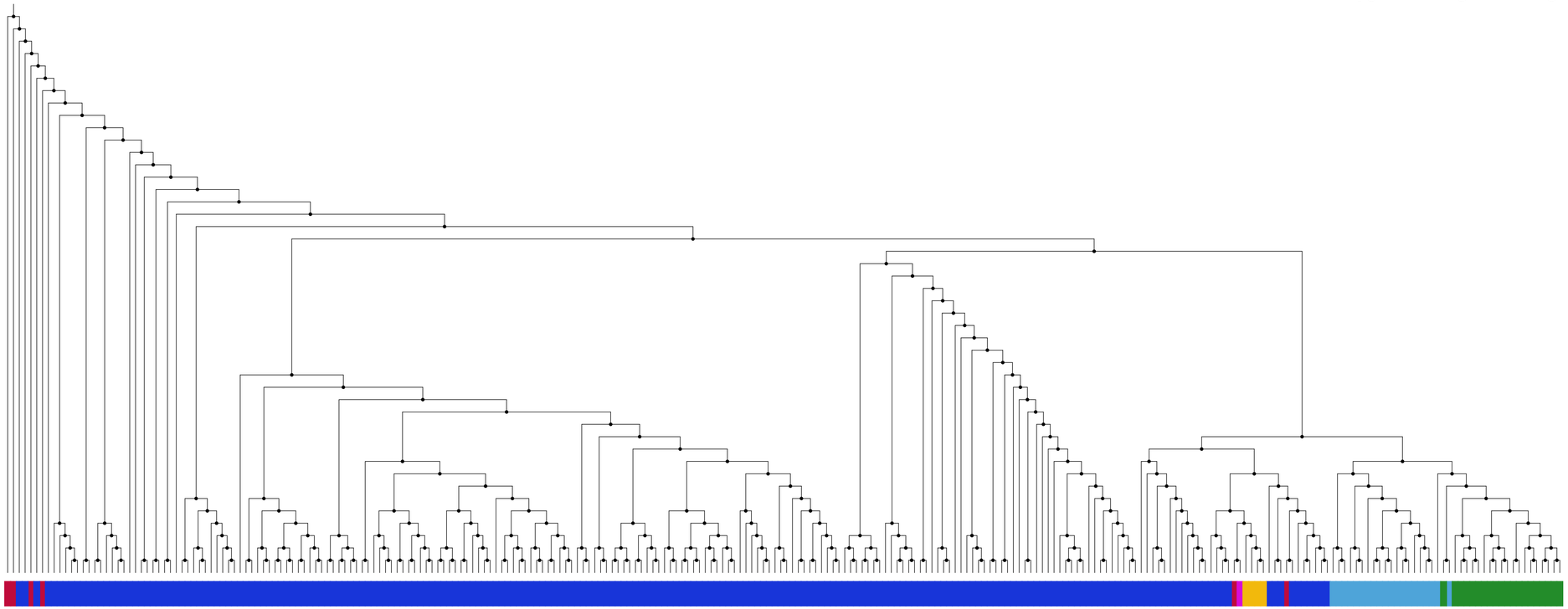
	Core	Softcore	Shell	Cloud
Ref	15042	300	518	0
Non-ref	163	91	1833	1976
Total	15205	391	2351	1976

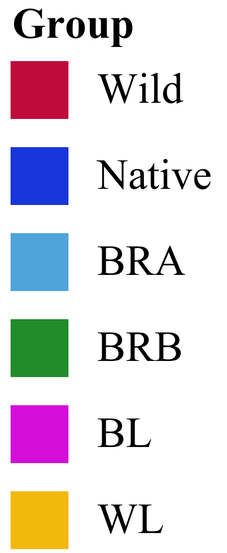


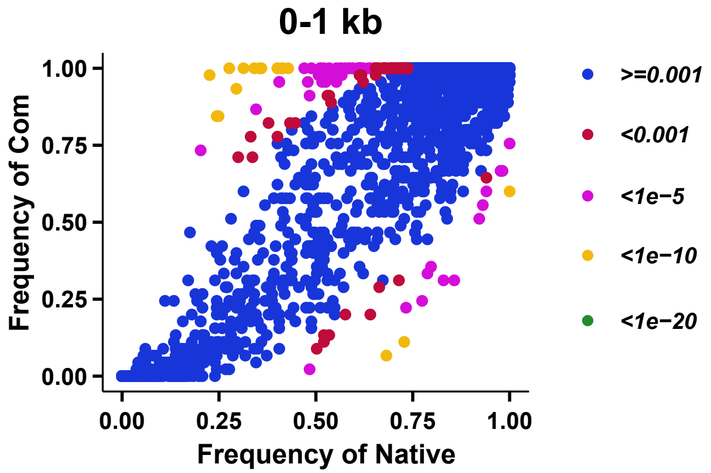


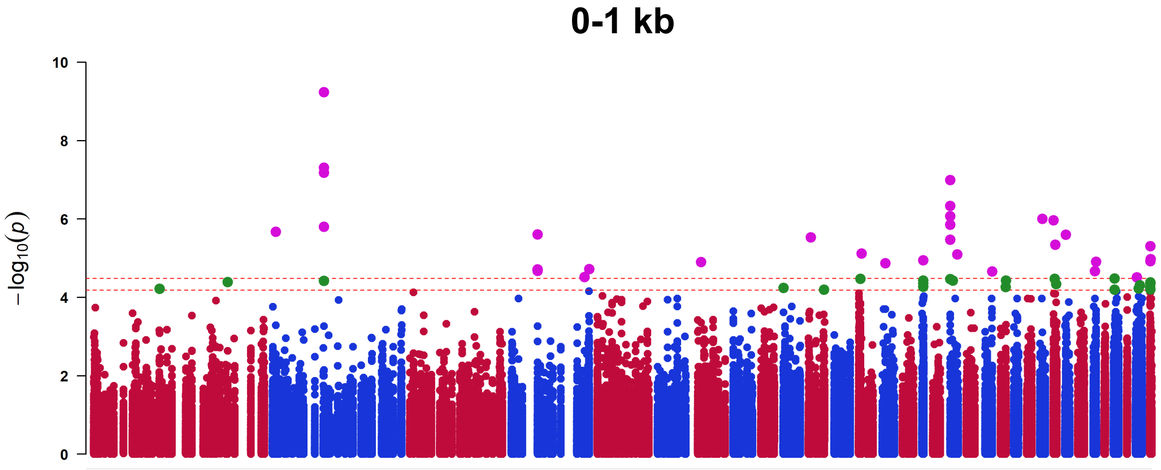


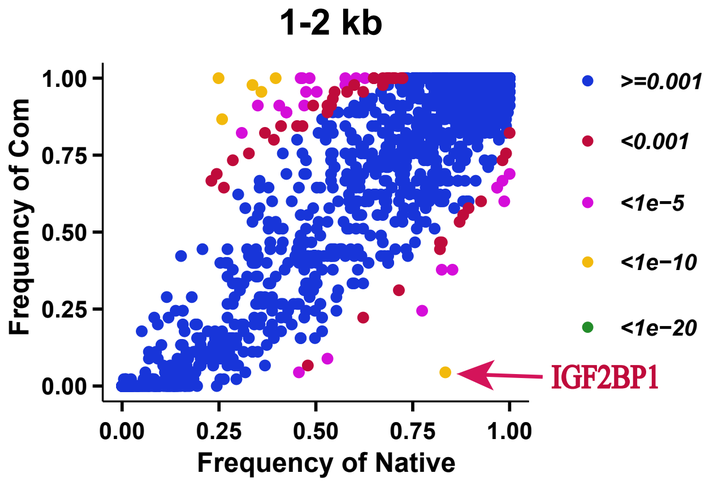


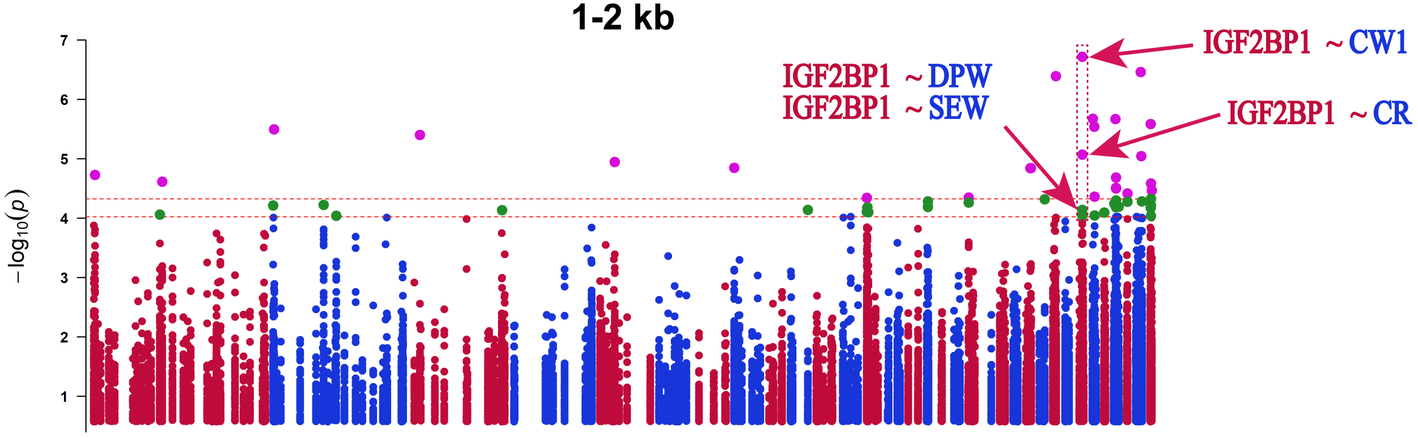


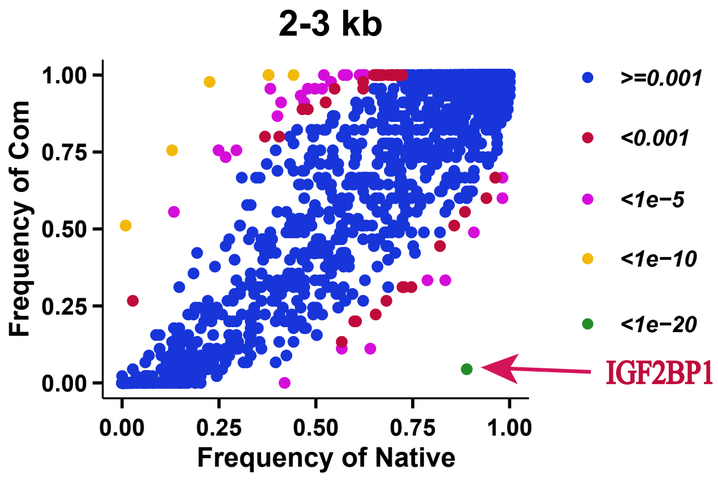


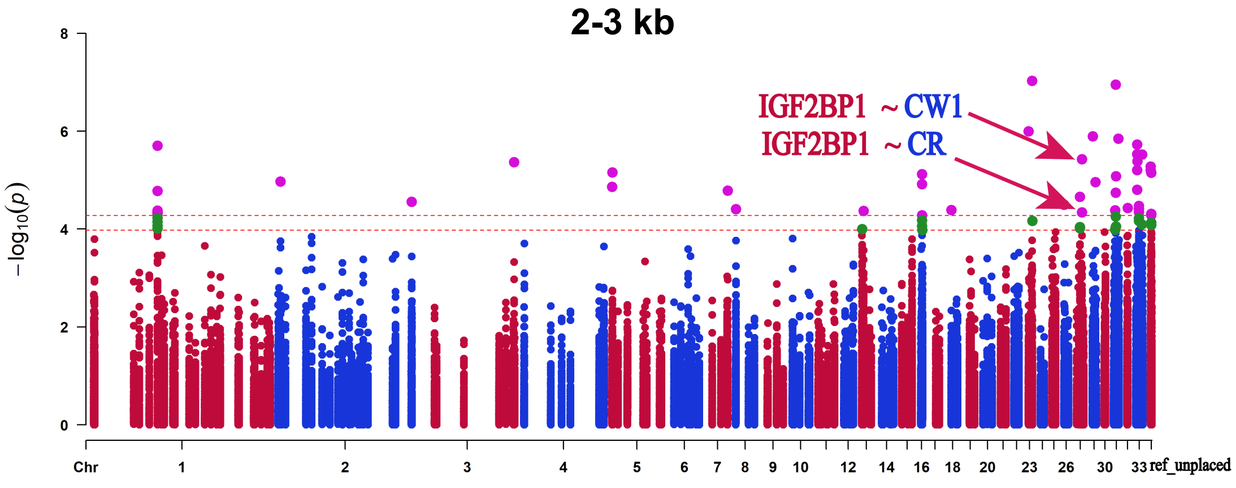


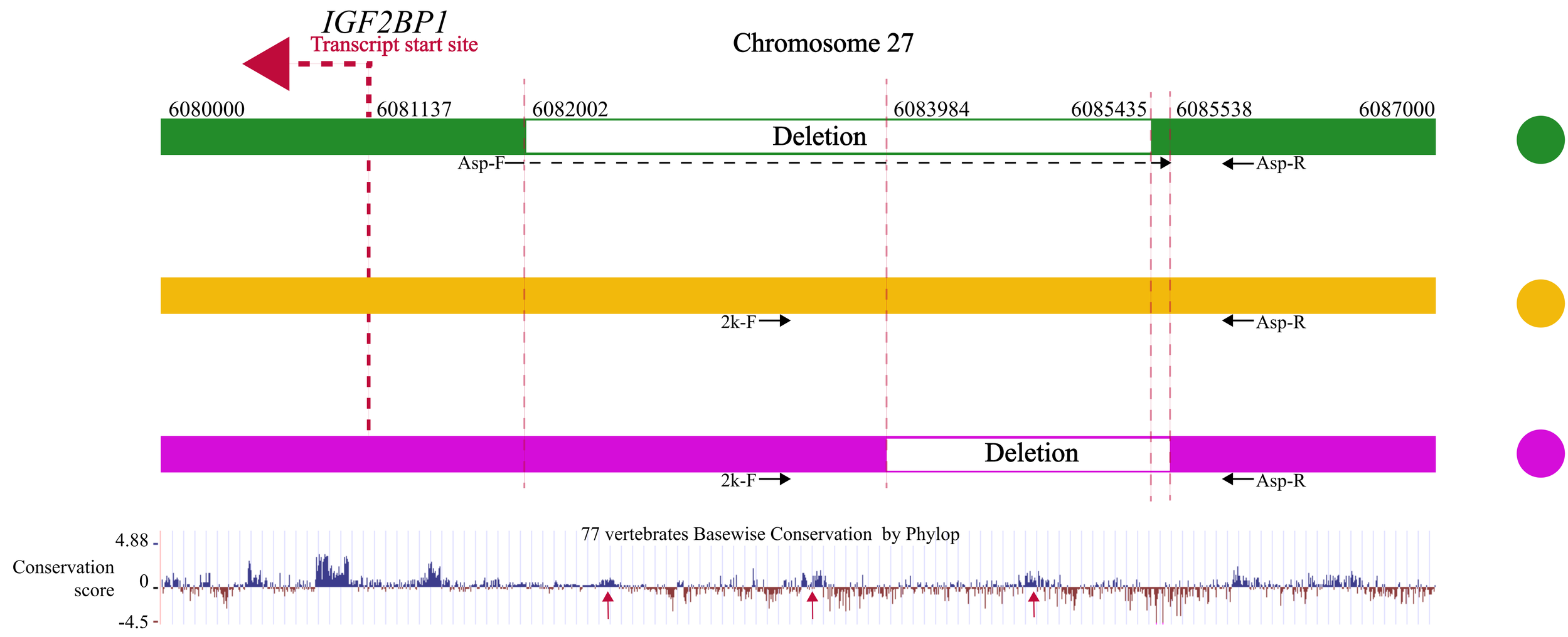


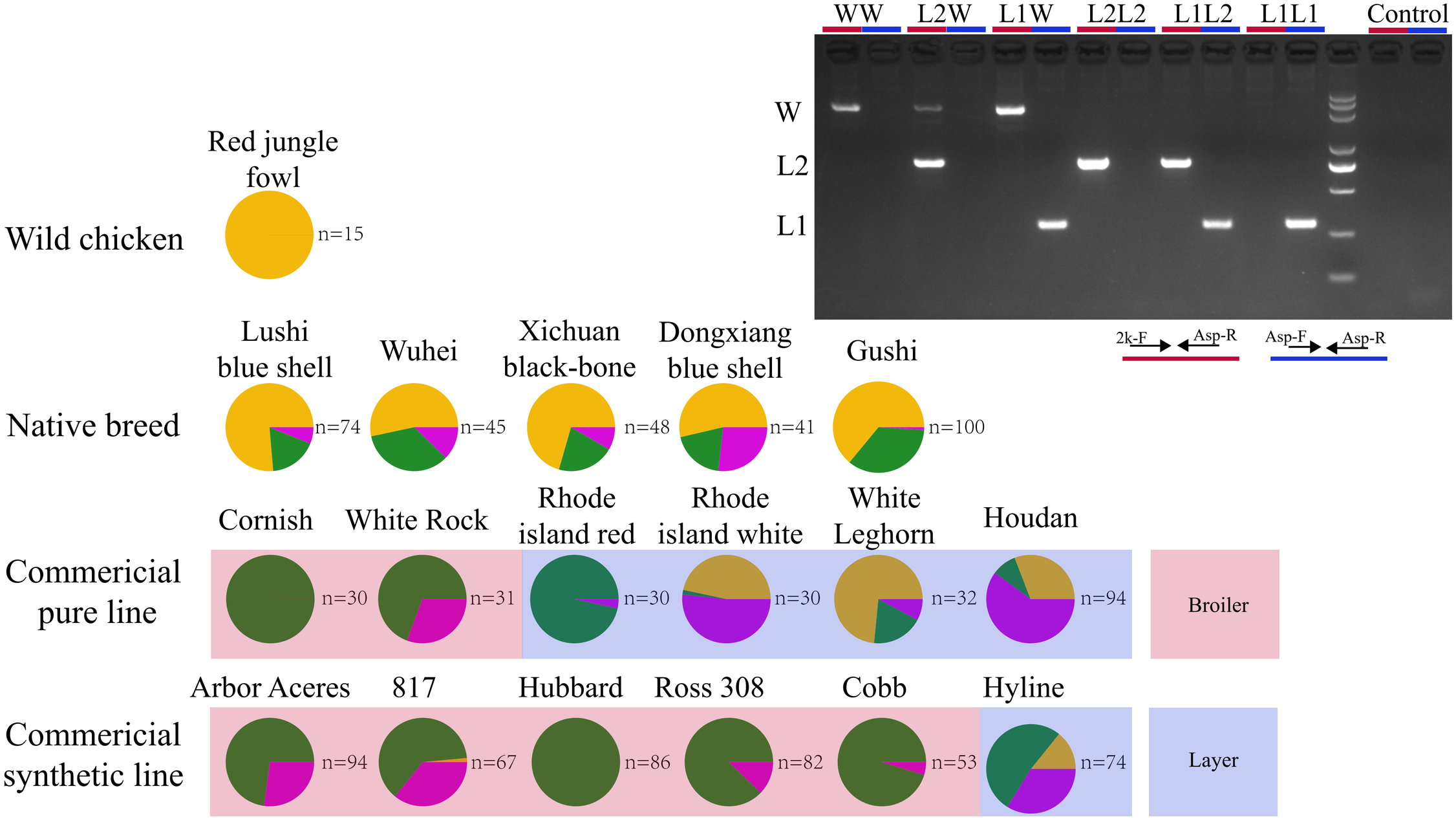


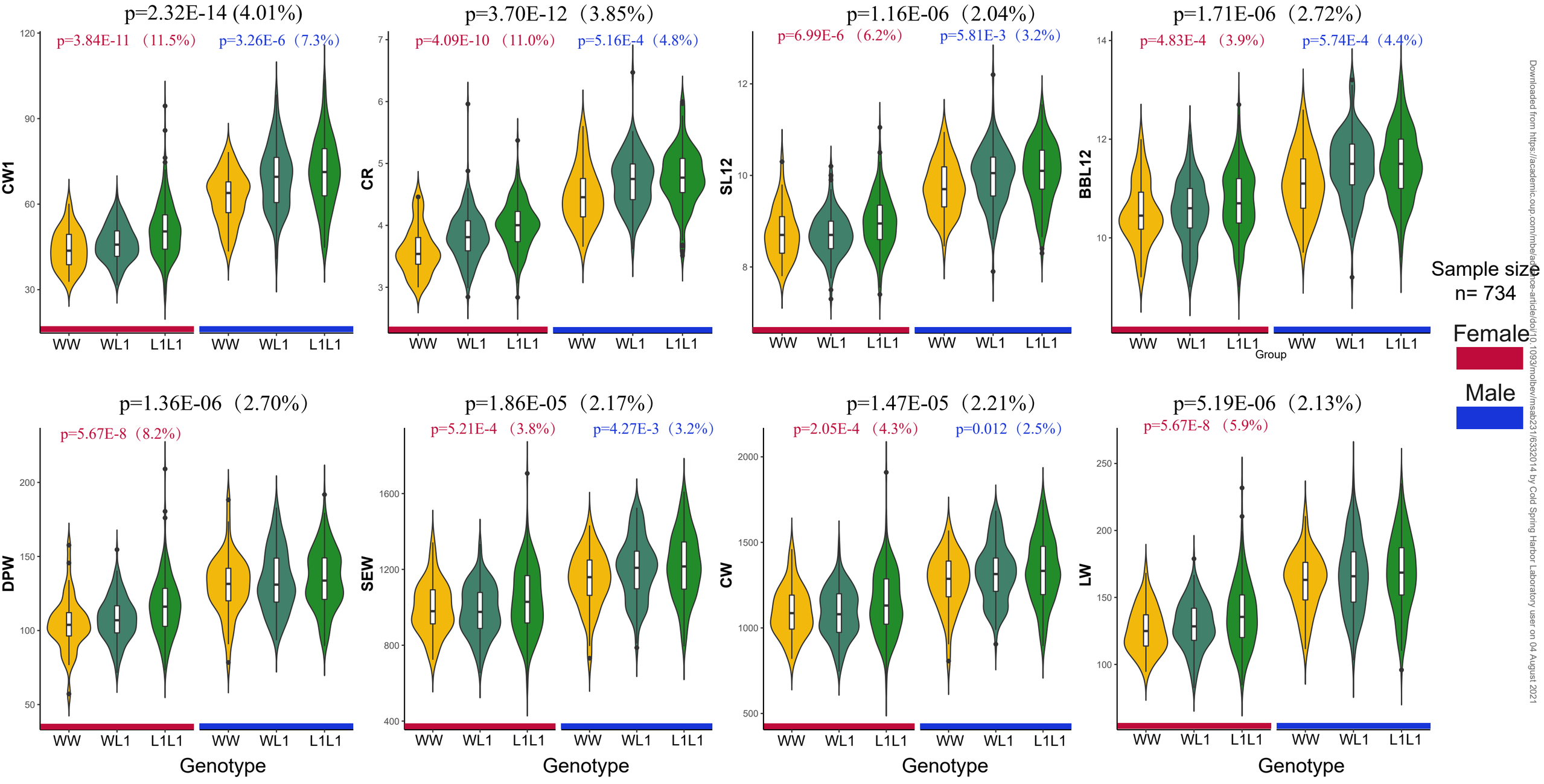




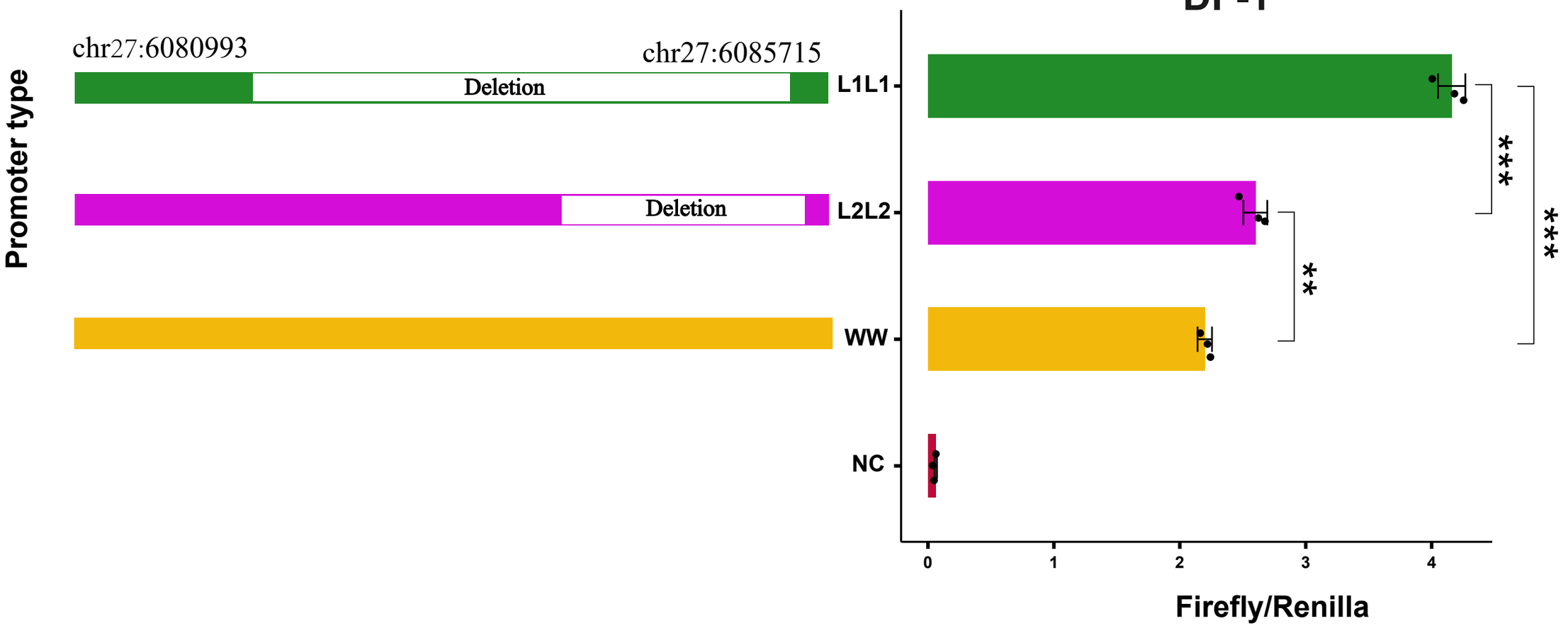




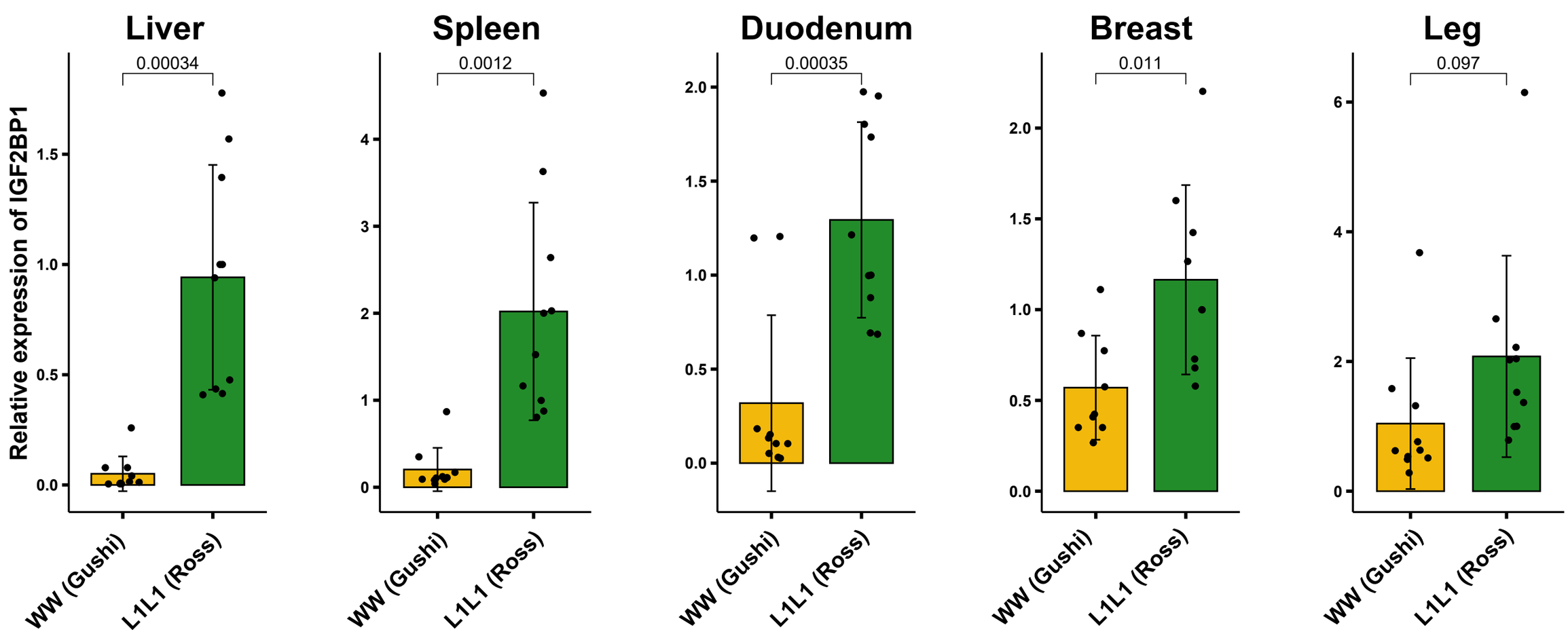




a



b



C