Journal of Animal Science, 2022, **100**, 1–9
https://doi.org/10.1093/jas/skac336
Advance access publication 12 October 2022
Animal Genetics and Genomics

New insights into the novel sequences of the chicken pan-genome by liquid chip

Fei Wang,[†,1] Yingwei Guo,[†,1] Zhenyu Liu,[†,1] Qiao Wang,[‡] Yu Jiang,[†,] and Guiping Zhao[‡,2]

[†]Key Laboratory of Animal Genetics, Breeding and Reproduction of Shaanxi Province, College of Animal Science and Technology, Northwest A&F University, Yangling, Shaanxi, China
[‡]State Key Laboratory of Animal Nutrition, Institute of Animal Sciences, Chinese Academy of Agricultural Sciences, Beijing, China

[1]These authors contributed equally to this study.
[2]Corresponding author: zhaoguiping@caas.cn

Abstract

Increasing evidence indicates that the missing sequences and genes in the chicken reference genome are involved in many crucial biological pathways, including metabolism and immunity. The low detection rate of novel sequences by resequencing hindered the acquisition of these sequences and the exploration of the relationship between new genes and economic traits. To improve the capture ratio of novel sequences, a 48K liquid chip including 25K from the reference sequence and 23K from the novel sequence was designed. The assay was tested on a panel of 218 animals from 5 chicken breeds. The average capture ratio of the reference sequence was 99.55%, and the average sequencing depth of the target sites was approximately 187X, indicating a good performance and successful application of liquid chips in farm animals. For the target region in the novel sequence, the average capture ratio was 33.15% and the average sequencing depth of target sites was approximately 60X, both of which were higher than that of resequencing. However, the different capture ratios and capture regions among varieties and individuals proved the difficulty of capturing these regions with complex structures. After genotyping, GWAS showed variations in novel sequences potentially relevant to immune-related traits. For example, a SNP close to the differentiation of lymphocyte-related gene *IGHV3-23-like* was associated with the H/L ratio. These results suggest that targeted capture sequencing is a preferred method to capture these sequences with complex structures and genes potentially associated with immune-related traits.

Lay Summary

A total of 48K target sites were selected to be placed on the liquid chip, including 23K from the novel sequence of the chicken pan-genome. The high average capture ratio (99.55%) of the reference sequence in five populations indicated the good performance of the liquid chip. For the target region in the novel sequence, the average capture ratio was approximately 33.15% and the average sequencing depth of target sites was approximately 60X, both of which were higher than that of resequencing. However, the capture ratio was different among varieties, ranging from 29.2% (White Leghorn) to 33.4% (B line). GWAS (Genome-wide association study) showed variations in novel sequences potentially related to immune-related traits. For example, an SNP (single nucleotide polymorphism) close to the differentiation of the lymphocyte-related gene *IGHV3-23-like* was associated with the H/L (heterophil/lymphocyte ratio) ratio. Overall, this study not only improved the capture ratio of regions with complex structures in novel sequences but also preliminarily explored the association of variations in these regions with chicken economic traits.

Key words: GWAS, liquid chip, novel sequence, pan-genome

Abbreviations: *ACADM*, acyl-CoA dehydrogenase medium chain; *ADGRB3*, G protein-coupled receptor B3; BW, body weight; *DGAT1*, diacylglycerol o-acyltransferase 1; GBTS, genotyping by target sequencing; GWAS, genome-wide association study; H/L, heterophil/lymphocyte ratio; *HSD17B8*, hydroxysteroid 17-beta dehydrogenase 8; *IGHV*, immunoglobulin heavy chain variable; LD, linkage disequilibrium; MAF, minor allele frequency; *MRPL52*, mitochondrial ribosomal protein L52; NJ, neighbor-joining; PCA, principal component analysis; Q-Q, quantile–quantile; QTL, quantitative trait locus; SNPs, single nucleotide polymorphisms

Introduction

The chicken was the first agricultural animal to have its genome sequenced (International Chicken Genome Sequencing, 2004). After several updates, the macrochromosomes and several microchromosomes exhibit high quality(Groenen et al., 2009). However, special regions, especially in some microchromosomes remain of lower quality, although great efforts toward a complete sequence have been made. Therefore, some studies have hypothesized that chicken and other bird genomes lack some crucial genes when compared with mammals and amphibians(Lovell et al., 2014; Zhang et al., 2014). Previous research has indicated that the missing sequences contain some genes(Botero-Castro et al., 2017; Yin et al., 2019). For example, PacBio sequencing reads of five red jungle fowl cDNAs reconstruct several genes including leptin and mitochondrial ribosomal protein l52 (*MRPL52*), which are absent from the chicken reference genome (GRCg6a, GCA_000000185.5) (Beauclair et al., 2019). In our previous report, 158.98 Mb of novel sequences including 1,335 protein-coding genes, which were not found in GRCg6a, were identified by the construction of the chicken pan-genome(Li et al., 2022). These genes participate in many crucial biological

Received August 26, 2022 Accepted October 11, 2022.

processes, including immune pathways. For example, the novel gene inhibitor of nuclear factor kappa b kinase regulatory subunit gamma (IKKγ) is a subunit of the canonical IκB kinase complex, which is best known as the activator of the transcription factor nuclear factor-κB (NF-κB) (Clark et al., 2013). Another novel gene RELB is a subunit of NF-κB (Vallabhapurapu and Karin, 2009), which regulates both innate and adaptive immune responses as well as the development and maintenance of the cells and tissues that comprise the immune system at multiple steps (Hayden and Ghosh, 2011). However, the detection rate of novel sequences was extremely low, approximately 0.43% among the resequenced data, which is a major obstacle for exploring the function of new genes and their relationship with economic traits. Therefore, an attempt to capture, genotype and subsequently perform genome-wide association study (GWAS) of these regions will assist scientists in obtaining a holistic understanding of these new genes.

Recently, a new genotyping technology called genotyping by target sequencing (GBTS) was developed, and it can accurately capture any position and length of the genome (Burridge et al., 2018).To date, two main methods have been applied in GBTS. One method mainly relied on multiplex Polymerase chain reaction (PCR). In the first round of PCR, target regions were specifically amplified by primers. In the second round of PCR, the sequencing adaptor and indices were added to obtain the library, which was subsequently sequenced and genotyped. Another method relied on probe hybridization with target DNA. After retrieval and amplification, target regions were sequenced and genotyped (Guo et al., 2019). Compared with chip-based genotyping platforms, GBTSs have many advantages, such as ultrahigh throughput, flexibility, and low cost.

Here, we developed a liquid chip based on reference and novel sequences of the pan-genome to capture the target regions by sequencing some Chinese native breeds and commercial chicken lines. To date, no liquid chip has been designed and used for chickens. This study was conducted to systematically compare the capture ratio between reference and novel sequences, and the capture ratio of different breeds. After genotyping, GWAS was used to explore the potential association of variation with traits, especially on novel sequences.

Materials and Methods

All experimental procedures were approved by the Animal Management Committee (in charge of animal welfare issues) of the Institute of Animal Science, Chinese Academy of Agricultural Sciences (IAS-CAAS, Beijing, China) and performed in accordance with the guidelines.

Resequencing data analysis and selection of the first group of candidate target sites

We collected 961 resequencing datasets representing wild and domestic chicken breeds from all over the world. The raw reads were filtered by Fastp to remove adaptor and low-quality sequences (Chen et al., 2018). Then clean reads were mapped to the chicken reference genome using BWA with default parameters. Duplicate reads were excluded using Picard Tools (http://broadinstitute.github.io/picard/). Single nucleotide polymorphism (SNP) calling was carried out by GATK (Van der Auwera et al., 2013), and high-quality SNPs were selected using the following criteria: (1) depth > mean read depth/3 and <mean read depth × 3; (2) QD > 2.0; (3) FS < 60.0; (4) MQ > 40.0; (5) SOR < 3.0; (6) MappingQualityRankSum > −12.5; and (7) ReadPosRankSum > −8.0.

To select highly informative SNPs that were uniformly distributed throughout the genome, we divided the genome into several intervals (length = 40 k). For each genomic interval, we followed the pipeline described in a previous study to sequentially select SNPs, mainly using the greedy algorithm approach (http://www.nist.gov/dads/HTML/greedyalgo.html). Briefly, each SNP was scored depending on its minor allele frequency (MAF), and its proximity to the center of the intervals. The SNP with the maximum score within an interval was considered as the target site and used for the probe design.

Selection of the second group of candidate target sites from chicken pan-genome

In our previous research (Li et al., 2022), we built a chicken pan-genome and acquired 158.98 Mb of nonredundant novel sequences that were missing in the reference genome. Therefore, the second group of candidate SNPs was selected from these novel sequences. We first selected the exon region as the candidate region for targeted amplification. Then, according to the content of direct repeats and G4 motifs in these sequences, we chose regions with low complex structures as the final targeted regions. For contigs without exons, we randomly selected target regions according to the content of complex structures.

Production of probes and sequencing

After the selection of the target sites, probes were designed according to the following criteria: (1) the length of the probe was 100 bp; (2) the GC content was between 30% and 70%; and (3) the number of its homologous regions was less than 10. Due to the high content of repeat sequence in the novel sequences, the number of its homologous regions was relaxed to 300. After high-throughput synthesis, these probes were used for the letter GBTS.

The genomic DNA from Wenchang, Dagu, Douji, White Leghorn, and B lines was isolated from the blood by the standard method of phenol-chloroform extraction. After DNA quality evaluation, the libraries were constructed using the CAGT library Preparation Kit (Compass, Beijing, China) according to the manufacturer's protocol. Next, libraries, probes, and hybridization reagents were mixed and placed on PCR to hybridize and capture target regions. After washing, the target fragments were amplified to enrich these regions. Finally, a dsDNA HS Assay Kit for Qubit (YESEN, Shanghai, China) and qPCR were used to quantify the library concentration and sequencing was performed with PE150 on the DNBSEQ T7 platform (MGI, Shenzhen, China). Sequencing was performed at Beijing Compass Biotechnology Co., Ltd (Beijing, China).

Validation of the liquid chip in four chicken breeds/lines

After sequencing, the data were mapped with the reference genome and novel sequences of the pan-genome. SNP calling was conducted according to the abovementioned pipeline. The sequencing depth was calculated by SAMtools (Danecek et al., 2021). If there were SNPs within the region of upstream or downstream of the target site (±250 bp), we

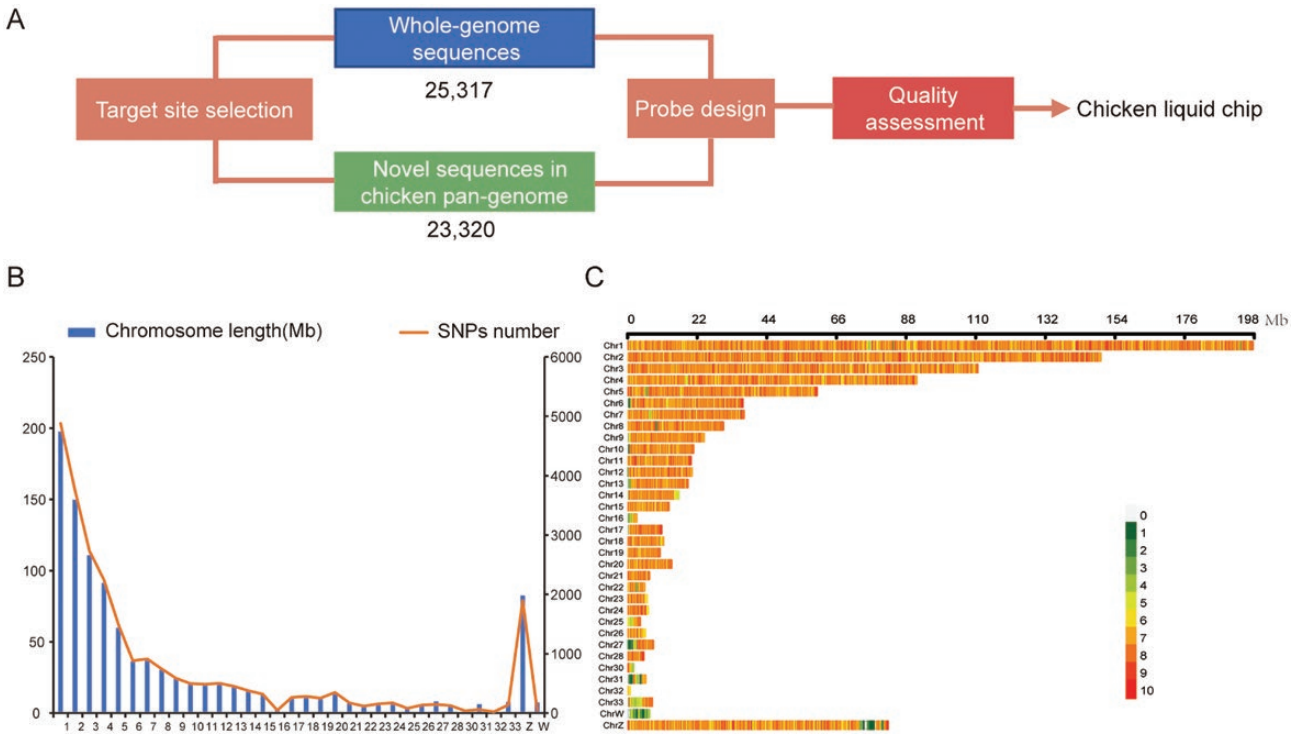

Figure 1. Roadmap and characterization of the target sites on the chicken liquid chip. (a) Design pipeline of the chicken liquid chip; (b) The chromosome wise SNP density of the chicken liquid chip. Chromosome length is shown on the left axis (based on galGal-5) and SNP density is shown on the right axis. (c) Target site density within 500-kb windows throughout the whole genome.

define this region were captured. The capture ratio refers to the ratio of the number of captured regions to the total number of regions, which was the number of target sites. The capture ratio was calculated by bedtools (Quinlan and Hall, 2010). Allele frequency was calculated by PLINK (Purcell et al., 2007).

For population genetic structure analysis, we first implemented principal component analysis (PCA) based on the SmartPCA program (Patterson et al., 2006). Considering that the high linkage disequilibrium (LD) between SNPs may bias the PCA results, SNPs were pruned to obtain independent SNPs. A neighbor-joining (NJ) tree was constructed and visualized with the ggtree package in R (Yu et al., 2017).

Phenotyping and genome-wide association study

The individual body weight (BW) was weighed on Day 42. Other phenotypes, including heterophils and lymphocytes, were counted for all 180 B line chickens, which had free access to feed and water and were managed in the same environment. The measurement methods have been previously detailed by Zhu et al. and Wang et al. (Zhu et al., 2019; Wang et al., 2020a). The phenotypes that did not follow a normal distribution were transformed in R and then the transformed data were used in the following genetic analyses.

Before GWAS analysis, PCA was implemented to assess the population structure. If there was no obvious population stratification, the principal components were not used as covariates in the analysis model. In addition, the genomic relationship matrix was constructed with SNPs using GEMMA and used as a random effect in the later analysis (Zhou and Stephens, 2012).

The mixed linear model was used for GWAS, as implemented in the GEMMA. The basic model formula was: $y = W\alpha + x\beta + u + \varepsilon$, where y is the phenotypic value for every trait; W is a matrix of covariates including sex; α is a vector of corresponding coefficients including the intercept; β is the SNP effect and x is a vector of SNP genotypes; u is a vector of random effects with a covariance structure that follows a normal distribution as $u \sim N(0, KVg)$, where K is genetic relationship matrix calculated by SNP markers; and Vg means polygenic additive variance, and ε is a vector of random errors.

The genome-wide significance P value threshold was calculated using Bonferroni correction with an effective number of independent sites. The number of genome-wide independent markers was calculated using PLINK with the parameter "--indep-pairwise 25 5 0.2". The significance level was set as 9.29E-07 (0.05/53,820). Manhattan and quantile–quantile (Q–Q) plots were visualized using the qqman package in R.

Results

Genome resequencing and novel sequences in the chicken pan-genome supplying the target sites of the liquid chip

Target sites in this liquid chip were selected from two major datasets. As shown in Figure 1a, the first dataset included 961 whole-genome sequences of wild and domestic chicken breeds around the world. The data summary of each breed is provided in Supplementary Table S1. After filtering, a total of 32 Mb SNPs were detected in the overall breeds. The SNPs in the reference sequence were incorporated into the chip panel based on the greed algorithm (GA), which has already been used in assay design for cattle (see Methods; Matukumalli et al., 2009). Finally, 25,317 SNPs were selected to be placed on the chip. The distribution of SNPs on the chromosomes is

shown in Figure 1b, c and Supplementary Table S2. A strong correlation was found between the length of chromosomes and the number of sites selected on each chromosome, indicating that all sites in the panel were uniformly distributed throughout the genome.

The second group of candidate target sites was derived from novel sequences, which were obtained from the chicken pan-genome and were absent from GRCg6a (https://zenodo.org/record/5881830/files/ChickenPangenome.NovelSequences.fa.gz?download=1). According to the tandem repeats and secondary structures of these sequences, 23,320 sites per regions with fewer complex structures were chosen and integrated into the liquid chip. The details of the target region are shown in Supplementary Table S3.

Genotyping performance of the liquid chip

Performance metrics for the liquid chip were tested by genotyping a panel of 218 animals, including 189 animals from the B line, 3 animals from Douji, 7 animals from Dagu, 12 animals from White Leghorn, and 7 animals from Wenchang chickens. The evaluation results are shown below.

Capture ratio

The capture ratio of the target sites on the reference and novel sequences was analyzed. For the target regions in the reference sequences, the number of capture regions ranged from 25,144 to 25,253, and the average capture ratio was 99.55% (25,202/25,317) (Figure 2a), which means that almost all regions were detected. The capture ratio of individuals is listed in Supplementary Table S4.

For the target region on the novel sequences, the number of detected regions ranged from 6,389 to 8,522, and the average capture ratio was 33.15% (7,730/23,320) (Figure 2b), which was lower than that on the reference genome. Our previous research found that the length of novel sequences from assemblies of different breeds was quite varied (Li et al., 2022). Therefore, we speculated that the capture ratio among breeds may be different. As shown in Figure 2c, the average capture ratio for each breed ranged from 29.2% (White Leghorn) to 33.4% (B line). The average capture ratio in White Leghorn was significantly lower than that in other breeds. In addition to the great differences in the capture ratio among breeds, great variation was also detected among individuals within the breed. For individuals in the B line, the difference in the capture ratio can be up to 6% (Supplementary Table S5), which indicates differences in the capture ratio between species and individuals.

Considering the instability of the capture ratio in novel sequences, the capture regions among breeds and individuals may also vary. Therefore, we calculated the capture regions in different breeds and validated the reproduction. Among the 14,427 captured regions, 93.5% were detectable in multiple breeds and 53.4% were shared among these five breeds

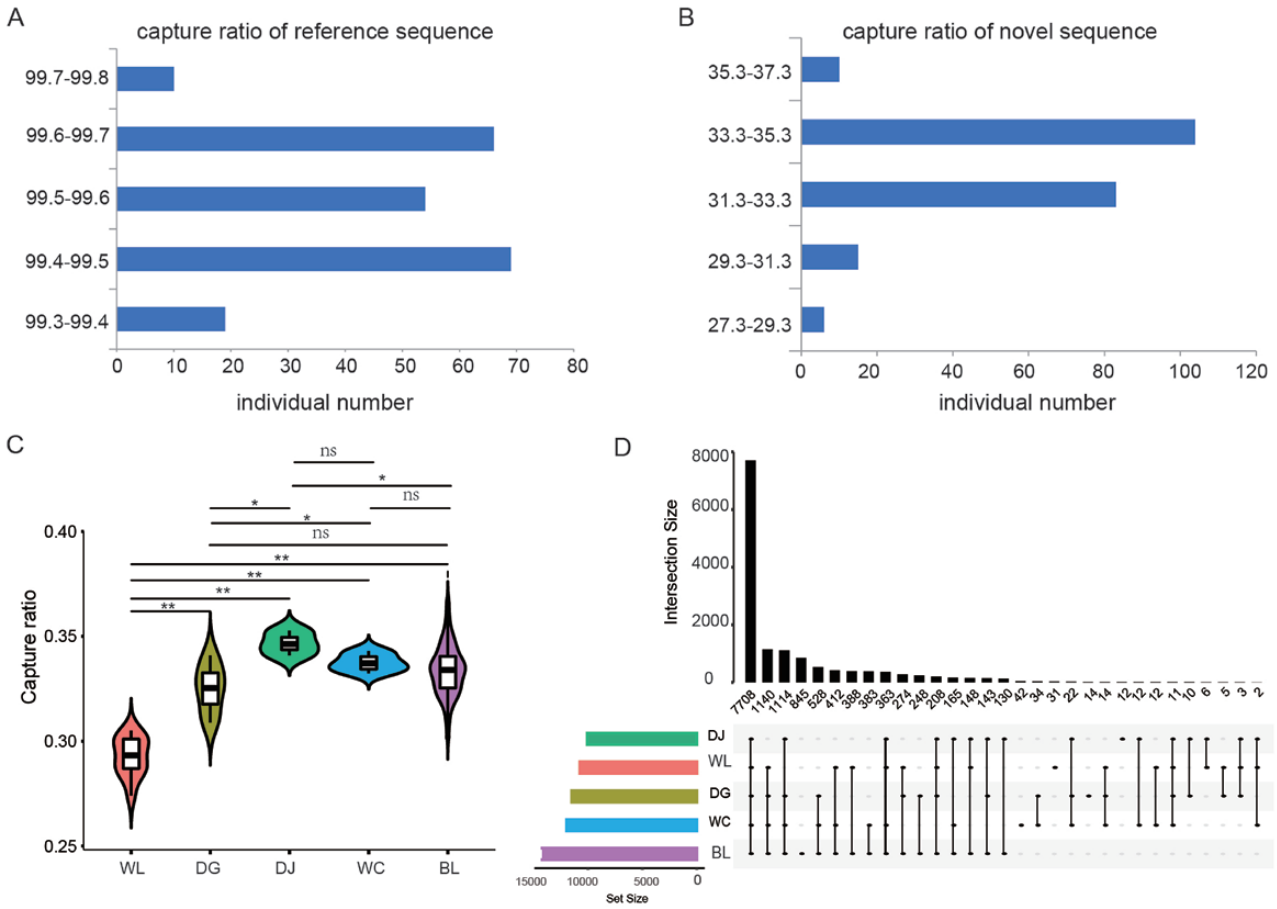

Figure 2. Capture ratio and regions in reference and novel sequences of the liquid chip. (a, b) Distribution of markers on the capture ratio of reference (a) and novel sequences (b) in the test population. (c) Distribution of the capture ratio among different varieties; (d) Number of shared and unique capture regions among different varieties.

(Figure 2d). Furthermore, we compared the capture region of three Wenchang samples whose capture ratios were similar (33.2%, 33.3%, and 33.5%), and 65% were detected in three samples (Supplementary Figure S1). Collectively, these results indicated that there were breed and individual differences in capture ratio and region.

Sequencing depth

In our previous research, the median sequencing depth of the novel sequences was lower than the whole-genome depth, because tandem repeats form noncanonical DNA structures leading to in sufficient DNA sequencing (Guiblet et al., 2018). Therefore, we compared the sequencing depth of reference sequences with that of novel sequences by GBTS. As shown in Figure 3a, the average depth of the target sites in the reference sequences was approximately187X. Within 100 and 200 bp near the target sites, the average depth can still reach more than approximately 80X and 12X, respectively. For the novel sequences, the average sequencing depth of the target sites was only one-third of that in the reference genome (Figure 3b). However, the sequencing depth by GBTS was higher than that by resequencing, which may be a supplementary strategy for acquiring novel sequences and assisting chicken genome assembly.

Variation

After SNP calling, the number and MAF of SNPs were also calculated. A total of 272.3K high-quality SNPs were identified (MAF > 0.05). Of these SNPs, 91.8% were located on the reference genome and evenly occurred on each chromosome (Supplementary Figure S2). Of these SNPs, 78.7% had an MAF > 10% (Figure 3c). Of the polymorphisms, 4.60% were in the 1-kb region upstream of the transcription start site or downstream of the transcription end site, and 41.72% were

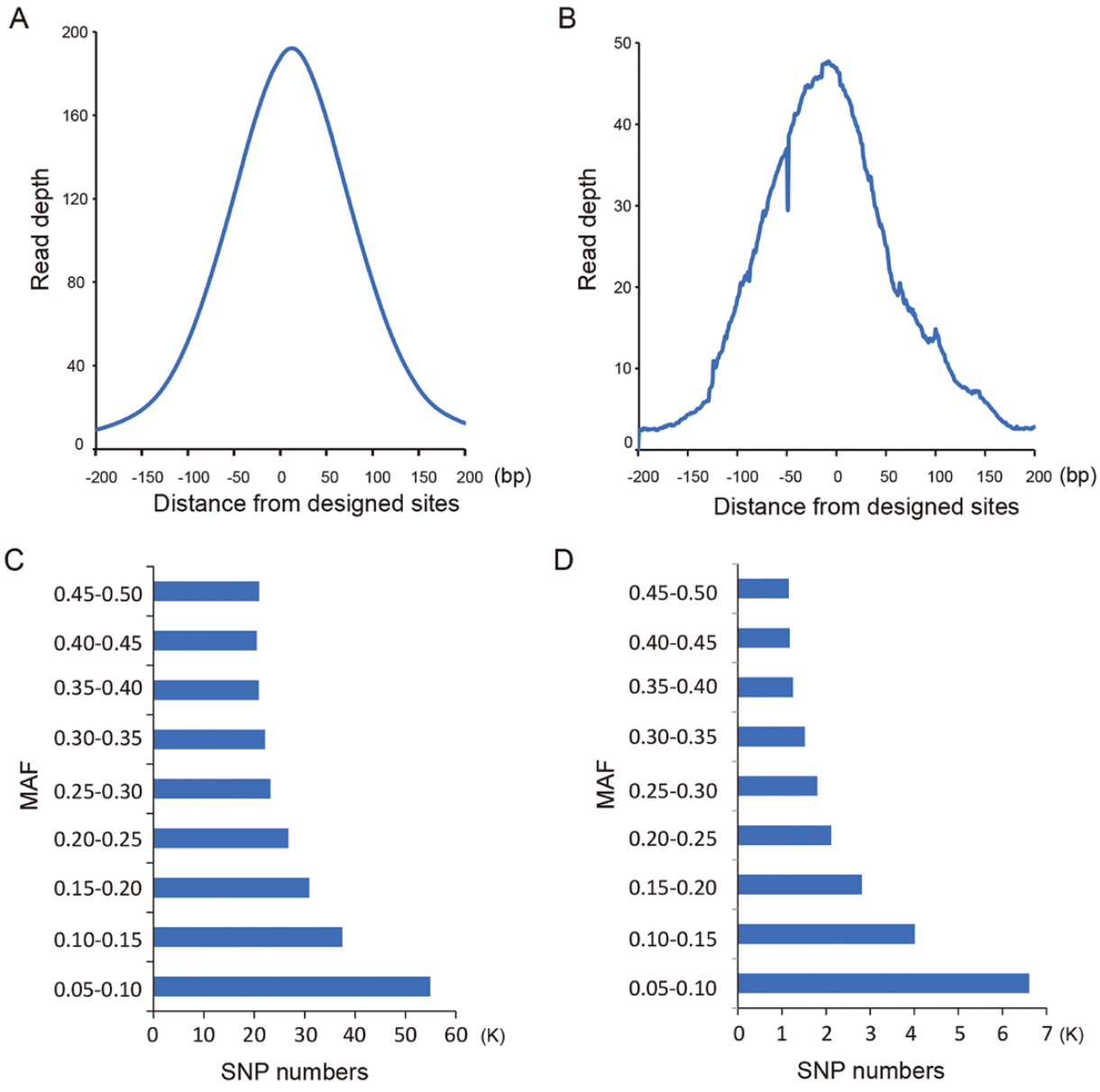

Figure 3. Profile of read depth and SNP number in the test population. Read depth of target sites and their flanking regions in reference (a) and novel sequences (b). (c, d) Distribution of the MAF of the SNPs in reference (c) and novel sequences (d).

in the intergenic region. The remaining polymorphisms were either in the exon or intron, of which only 14.4% were in the exon (Figure S3). According to the annotation of the chicken reference genome, the target polymorphisms covered 34% (24,244) of the protein-coding genes. For the novel sequences, 22.5K variations were identified, 70.6% of which had a MAF > 10% (Figure 3d). Therefore, the ratio (22.5/257.8) of the number of variants in novel sequences to that in the reference genome was lower than the ratio of the length of the novel sequence to that of the reference sequence (0.158/1.1), which may be attributed to the limited capture ratio of the novel sequence.

The ability to detect population stratification

PCA was performed using the genotyped data to investigate the ability of the liquid chip to detect population stratification in the validated samples. As shown in Figure 4a and b, individuals originating from Douji and Wenchang were tightly clustered. The commercial layer of White Leghorn chickens was relatively far away from the Chinese local breeds and commercial broilers. This phylogenetic pattern was also supported by a phylogenetic tree constructed using the weighted NJ method.

GWAS for growth and immune traits

Annotation and enrichment analyses indicated that many novel genes were involved in many crucial biological processes, including metabolism and immune response. To benchmark the association of novel genes with traits, we conducted GWAS on the weight at 42 d, the number of heterophils and lymphocytes, and the H/L ratio. The H/L ratio is a reliable and accurate physiological indicator of the chicken stress response and is related to bacterial killing ability. First, stratified tests were conducted on the B line population. PCA using the first two principal components showed that the B line clustered together and did not form separate clusters (Figure S4). Therefore, population stratification was not accounted for in the GWAS. The analysis revealed 65 SNPs significantly associated with these traits with genome-wide significance (–log10(P) > 4.73) (Figure 5 and Supplementary Table S6). Of these SNPs, 41 were associated with BW at 42 d and mapped to chromosomes 1, 2, 3, 4, 5, 6, 11, 12, and 25. Based on the SNP annotation, 27 of the significant SNPs were located in intergenic regions, and 2 were in exons. For immune traits, 24 genome-wide significant SNPs are located on the reference genome, mainly on chromosomes 1, 3, 4, 11, 17, 23, 24, and 33. Annotation indicated that most of them were located in intergenic and intronic regions. Considering fewer variations on novel sequences, the threshold of significant SNPs dropped to 10^{-4} to reveal putative candidate genes, and five SNPs located in novel sequences were identified. One H/L-ratio-related SNP was in the sequence of Silkies_2#wuji_913#4450#7400, in which a homologous gene *Ighv3-23* was located. *Ighv3-23* is a member of the immunoglobulin heavy chain variable (*IGHV*) gene family, and the protein encoded by this gene is a B-cell receptor. In chickens, this gene was highly expressed in immune-related tissues (Supplementary Figure S5), indicating the possible relationship with the H/L ratio.

Discussion

In our previous research, a chicken pan-genome was constructed from 20 de novo assembled genomes with high sequencing depth, and 1,335 novel genes were identified (Li et al., 2022). These novel genes participate in many biological processes, for example, acyl-CoA dehydrogenase medium chain (*ACADM*) (Gregersen et al., 2001), diacylglycerol o-acyltransferase 1(*DGAT1*) (Liu et al., 2012), and hydroxysteroid 17-beta dehydrogenase 8(*HSD17B8*) (Hiltunen et al., 2019) are involved in fatty acid metabolism and steroid synthesis. IKKγ, RELB, and KH-type splicing regulatory protein(*KHSRP*) play a role in the immune response (Liu et al.,

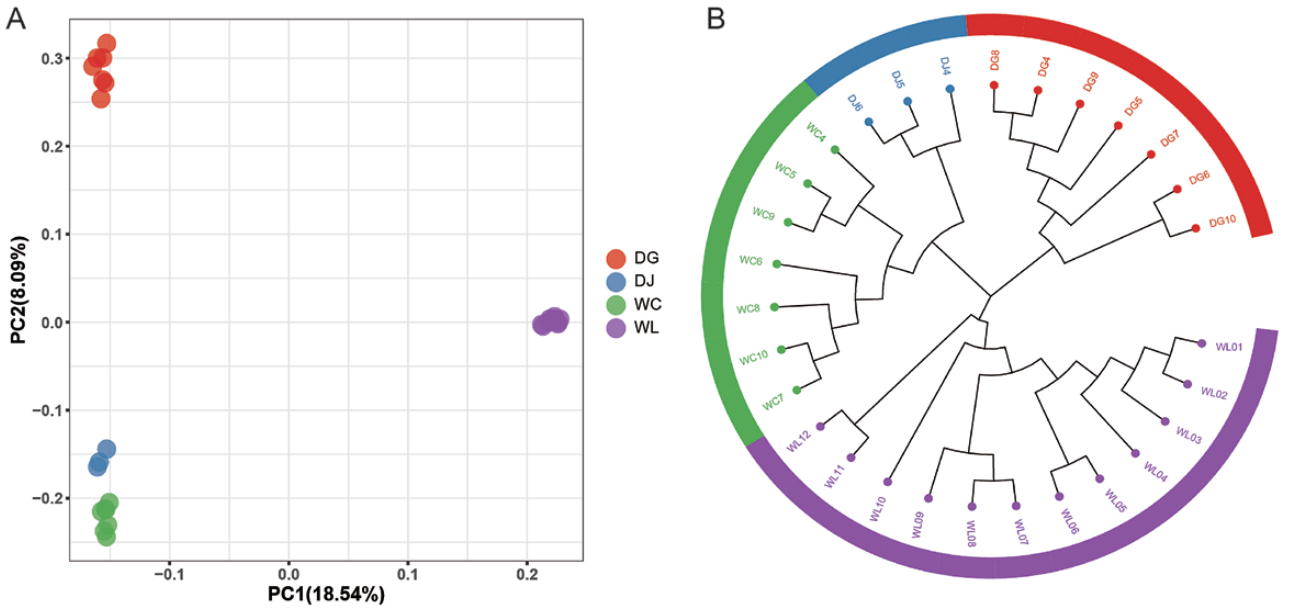

Figure 4. Population stratification identified by the liquid chip. (a) Principal components 1 and 2 for the 5 breeds per lines. (b) NJ tree constructed using p-distances between individuals.

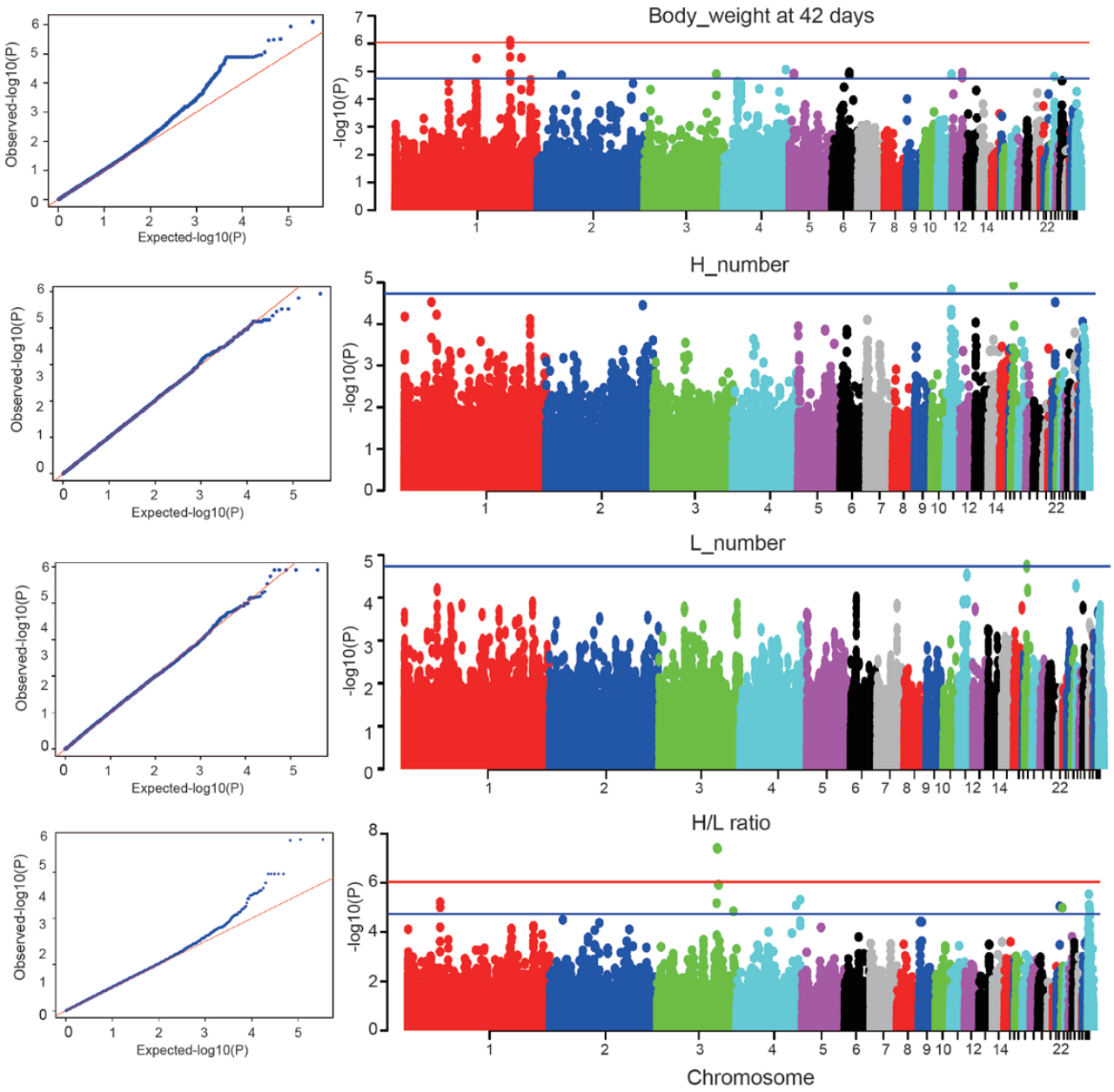

Figure 5. Manhattan and quantile–quantile plots of GWAS for growth and immune traits. Each dot represents an SNP in the dataset. The horizontal red and blue lines indicate the thresholds for genome-wide significance and suggestive significance, respectively.

2015). However, the higher GC content and tandem repeats of these novel genes can form noncanonical DNA structures, such as G-quadruplexes, which result in an extremely low detection rate of the novel sequences among the resequenced data and hinder the exploration of the relationship between novel genes and traits. In this study, although we avoided genome regions with complex structures as much as possible before selecting the target regions and designing probes for the novel sequences, approximately 30% of them were captured in every sample. Two main possible reasons may be attributed to this low capture ratio. One is that the novel sequence acquired by 20 de novo assembled genomes may not be present in every breed. The cattle pan-genome using 12 de novo assemblies has provent that many nonreference sequences are mostly specific to one breed. Pan-genome research on plants also reported this phenomenon (Liu et al., 2020). Another is that the complex structure causes unstable detection. As shown in Supplementary Figure S1, partial capture regions were significantly different among three individuals of the same breed. However, the capture ratio of the novel sequence was much higher than that of resequencing, which indicated the advantage of target sequencing in capturing regions with complex structures, compared with next-generation sequencing. It may be a new method to obtain genome sequences with complex structures and can be used to assist genome assembly. On the other hand, the variety of capture regions among breeds and individuals also highlights the challenge of assembling a complete genome with only one DNA sample using current sequencing technologies.

After GBTS, a total of 272.3K high-quality SNPs were identified, which was beyond the capabilities of the bead

chip. To date, two SNP arrays have been widely used in chicken breed relationship analysis, GWAS, and genomic selection. One is the 600 K SNP array developed by Kranis et al. in 2013 (Kranis et al., 2013). However, this array was designed based on foreign commercial broilers and egg layers, lacking the genomic variation information of Chinese indigenous breeds. Therefore in 2019, Liu et al. developed a new chip that contains the genetic variation of Chinese indigenous breeds (Liu et al., 2019). However, the chip-based genotyping platform has several disadvantages, such as low customization efficiency, higher cost, and less flexibility. To overcome these shortcomings, targeted capture sequencing with ultrahigh-throughput and cost-effectiveness has been developed and is now mainly used in plants (Liu et al., 2022). Our attempt in chickens proved its applicability in animals.

To explore the function of novel genes, we conducted GWAS of growth and immune traits with our liquid chip. These traits are crucial for the broiler industry. In our study, quantitative trait loci (QTLs) affecting BW at 42 d were identified on chromosome 1. Consistently, previous studies also found that QTLs on the end of chromosome 1 were significantly associated with growth traits in different chicken populations (Liu et al., 2008; Sheng et al., 2013; Wang et al., 2020b). For example, researchers identified a region (chromosome 1: 173.5–175.0 Mb) of the chicken genome to be strongly associated with growth traits. Zhang et al. performed a GWAS in a Gushi-Anka F2 chicken population and mapped the major QTLs of BW at 169.4 Mb on chromosome 1 (Zhang et al., 2021). This indicated the reliability of our results and the efficacy of the liquid chip for GWAS. For immune traits, SNPs on adhesion G protein-coupled receptor B3 (*ADGRB3*) were associated with the H/L ratio, and this gene was associated with fecal egg counts in sheep (Becker et al., 2022). For novel sequences, no significant SNPs were associated with BW or immune-related traits at the genome-wide level. Considering the limited SNPs in the novel sequences, we lowered the threshold and found some candidate SNPs. For example, one SNP in the novel sequences was related to the H/L ratio. This SNP was proximal to the *IGHV3-23-like* gene, which is related to the differentiation of lymphocytes and exhibits high expression in immune tissues (Bomben et al., 2010; Dal-Bo et al., 2011). Due to limited phenotypes and the capture ratio of novel sequences, exploration of the correlation between these new genes and more traits will be needed in the future.

Conclusions

In summary, compared with resequencing, the capture ratio of novel sequences in pan-genome was greatly improved by the designed liquid chip. The variation in the capture region among species and individuals showed us that multiple sequencing samples may be necessary for complete chicken genome assembly with current sequencing technologies. Using GWAS, a few SNPs in novel sequences were found to be potentially associated with growth and immune-related traits.

Supplementary Data

Supplementary data are available at *Journal of Animal Science* online.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 32072708), Natural Science Basic Research Program of Shaanxi (2022JQ-171).

Conflict of Interest Statement

The authors declare no real or perceived conflicts of interest.

Literature Cited

Beauclair, L., C. Ramé, P. Arensburger, B. Piégu, F. Guillou, J. Dupont, and Y. Bigot. 2019. Sequence properties of certain GC rich avian genes, their origins and absence from genome assemblies: case studies. *BMC Genomics* 20:734. doi:10.1186/s12864-019-6131-1

Becker, G. M., J. M. Burke, R. M. Lewis, J. E. Miller, J. L. M. Morgan, B. D. Rosen, C. P. Van Tassell, D. R. Notter, and B. M. Murdoch. 2022. Variants within genes EDIL3 and ADGRB3 are associated with divergent fecal egg counts in Katahdin sheep at weaning. *Front. Genet*. 13:817319. doi:10.3389/fgene.2022.817319

Bomben, R., M. Dal-Bo, D. Benedetti, D. Capello, F. Forconi, D. Marconi, F. Bertoni, R. Maffei, L. Laurenti, D. Rossi, et al. 2010. Expression of mutated IGHV3-23 genes in chronic lymphocytic leukemia identifies a disease subset with peculiar clinical and biological features. *Clin. Cancer Res*. 16:620–628. doi:10.1158/1078-0432.Ccr-09-1638

Botero-Castro, F., E. Figuet, M. K. Tilak, B. Nabholz, and N. Galtier. 2017. Avian genomes revisited: hidden genes uncovered and the rates versus traits paradox in birds. *Mol. Biol. Evol*. 34:3123–3131. doi:10.1093/molbev/msx236

Burridge, A. J., P. A. Wilkinson, M. O. Winfield, G. L. A. Barker, A. M. Allen, J. A. Coghill, C. Waterfall, and K. J. Edwards. 2018. Conversion of array-based single nucleotide polymorphic markers for use in targeted genotyping by sequencing in hexaploid wheat (Triticum aestivum). *Plant Biotechnol. J*. 16:867–876. doi:10.1111/pbi.12834

Chen, S. F., Y. Q. Zhou, Y. R. Chen, and J. Gu. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34:884–890. doi:10.1093/bioinformatics/bty560

Clark, K., S. Nanda, and P. Cohen. 2013. Molecular control of the NEMO family of ubiquitin-binding proteins. *Nat. Rev. Mol. Cell Biol*. 14:673–685. doi:10.1038/nrm3644

Dal-Bo, M., I. Del Giudice, R. Bomben, D. Capello, F. Bertoni, F. Forconi, L. Laurenti, D. Rossi, A. Zucchetto, G. Pozzato, et al. 2011. B-cell receptor, clinical course and prognosis in chronic lymphocytic leukaemia: the growing saga of the IGHV3 subgroup gene usage. *Br. J. Haematol*. 153:3–14. doi:10.1111/j.1365-2141.2010.08440.x

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, et al. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* 10. doi:10.1093/gigascience/giab008

Gregersen, N., B. S. Andresen, M. Corydon, T. J. Corydon, R. K. J. Olsen, L. Bolund, and P. Bross. 2001. Mutation analysis in mitochondrial fatty acid oxidation defects: exemplified by acyl-CoA dehydrogenase deficiencies, with special focus on genotype-phenotype relationship. *Hum. Mutat*. 18:169–189. doi:10.1002/humu.1174

Groenen, M. A., P. Wahlberg, M. Foglio, H. H. Cheng, H. J. Megens, R. P. Crooijmans, F. Besnier, M. Lathrop, W. M. Muir, G. K. Wong, et al. 2009. A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. *Genome Res*. 19:510–519. doi:10.1101/gr.086538.108

Guiblet, W. M., M. A. Cremona, M. Cechova, R. S. Harris, I. Kejnovska, E. Kejnovsky, K. Eckert, F. Chiaromonte, and K. D. Makova. 2018. Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate. *Genome Res*. 28:1767–1778. doi:10.1101/gr.241257.118

Guo, Z. F., H. W. Wang, J. J. Tao, Y. H. Ren, C. Xu, K. S. Wu, C. Zou, J. N. Zhang, and Y. B. Xu. 2019. Development of multiple SNP marker panels affordable to breeders through genotyping by target sequencing (GBTS) in maize. *Mol. Breed.* 39:37. doi:10.1007/s11032-019-0940-4

Hayden, M. S., and S. Ghosh. 2011. NF-kappa B in immunobiology. *Cell Res.* 21:223–244. doi:10.1038/cr.2011.13

Hiltunen, J. K., A. J. Kastaniotis, K. J. Autio, G. Y. Jiang, Z. J. Chen, and T. Glumoff. 2019. 17B-hydroxysteroid dehydrogenases as acyl thioester metabolizing enzymes. *Mol. Cell. Endocrinol.* 489:107–118. doi:10.1016/j.mce.2018.11.012

International Chicken Genome Sequencing, C. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432:695–716. doi:10.1038/nature03154

Kranis, A., A. A. Gheyas, C. Boschiero, F. Turner, L. Yu, S. Smith, R. Talbot, A. Pirani, F. Brew, P. Kaiser, et al. 2013. Development of a high density 600K SNP genotyping array for chicken. *BMC Genomics* 14:59. doi:10.1186/1471-2164-14-59

Li, M., C. Sun, N. Xu, P. Bian, X. Tian, X. Wang, Y. Wang, X. Jia, R. Heller, M. Wang, et al. 2022. De novo assembly of 20 chicken genomes reveals the undetectable phenomenon for thousands of core genes on microchromosomes and subtelomeric regions. *Mol. Biol. Evol.* 39. doi:10.1093/molbev/msac066

Liu, A. L., Y. F. Li, W. B. Qi, X. L. Ma, K. X. Yu, B. Huang, M. Liao, F. Li, J. Pan, and M. X. Song. 2015. Comparative analysis of selected innate immune-related genes following infection of immortal DF-1 cells with highly pathogenic (H5N1) and low pathogenic (H9N2) avian influenza viruses. *Virus Genes* 50:189–199. doi:10.1007/s11262-014-1151-z

Liu, Q., R. M. P. Siloto, R. Lehner, S. J. Stone, and R. J. Weselake. 2012. Acyl-CoA:diacylglycerol acyltransferase: molecular biology, biochemistry and biotechnology. *Prog. Lipid Res.* 51:350–377. doi:10.1016/j.plipres.2012.06.001

Liu, R., S. Xing, J. Wang, M. Zheng, H. Cui, R. Crooijmans, Q. Li, G. Zhao, and J. Wen. 2019. A new chicken 55K SNP genotyping array. *BMC Genomics* 20:410. doi:10.1186/s12864-019-5736-8

Liu, X., H. Zhang, H. Li, N. Li, Y. Zhang, Q. Zhang, S. Wang, Q. Wang, and H. Wang. 2008. Fine-mapping quantitative trait loci for body weight and abdominal fat traits: effects of marker density and sample size. *Poult. Sci.* 87:1314–1319. doi:10.3382/ps.2007-00512

Liu, Y., S. Liu, Z. Zhang, L. Ni, X. Chen, Y. Ge, G. Zhou, and Z. Tian. 2022. GenoBaits Soy40K: a highly flexible and low-cost SNP array for soybean studies. *Sci. China Life Sci.* 65:1898–1901. doi:10.1007/s11427-022-2130-8

Liu, Y. C., H. L. Du, P. C. Li, Y. T. Shen, H. Peng, S. L. Liu, G. A. Zhou, H. K. Zhang, Z. Liu, M. Shi, et al. 2020. Pan-genome of wild and cultivated soybeans. *Cell* 182:162. doi:10.1016/j.cell.2020.05.023

Lovell, P. V., M. Wirthlin, L. Wilhelm, P. Minx, N. H. Lazar, L. Carbone, W. C. Warren, and C. V. Mello. 2014. Conserved syntenic clusters of protein coding genes are missing in birds. *Genome Biol.* 15:565. doi:10.1186/s13059-014-0565-1

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. Smith, T. S. Sonstegard, et al. 2009. Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One* 4:e5350. doi:10.1371/journal.pone.0005350

Patterson, N., A. L. Price, and D. Reich. 2006. Population structure and eigenanalysis. *PLoS Genet.* 2:e190. doi:10.1371/journal.pgen.0020190

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81:559–575. doi:10.1086/519795

Quinlan, A. R., and I. M. Hall. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842. doi:10.1093/bioinformatics/btq033

Sheng, Z. Y., M. E. Pettersson, X. X. Hu, C. L. Luo, H. Qu, D. M. Shu, X. Shen, O. Carlborg, and N. Li. 2013. Genetic dissection of growth traits in a Chinese indigenous x commercial broiler chicken cross. *BMC Genomics* 14:151. doi:10.1186/1471-2164-14-151

Vallabhapurapu, S., and M. Karin. 2009. Regulation and function of NF-kappaB transcription factors in the immune system. *Annu. Rev. Immunol.* 27:693–733. doi:10.1146/annurev.immunol.021908.132641

Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, et al. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 43:11.10.11–11.10.33. doi:10.1002/0471250953.bi1110s43

Wang, Y. Z., L. N. Bu, X. M. Cao, H. Qu, C. Y. Zhang, J. L. Ren, Z. L. Huang, Y. Q. Zhao, C. L. Luo, X. X. Hu, et al. 2020b. Genetic dissection of growth traits in a unique chicken advanced intercross line. *Front. Genet.* 11:894. doi:10.3389/fgene.2020.00894

Wang, J., B. Zhu, J. Wen, Q. Li, and G. Zhao. 2020a. Genome-wide association study and pathway analysis for Heterophil/Lymphocyte (H/L) ratio in chicken. *Genes (Basel)* 11:1005. doi:10.3390/genes11091005

Yin, Z. T., F. Zhu, F. B. Lin, T. Jia, Z. Wang, D. T. Sun, G. S. Li, C. L. Zhang, J. Smith, N. Yang, et al. 2019. Revisiting avian 'missing' genes from de novo assembled transcripts. *BMC Genomics* 20:4. doi:10.1186/s12864-018-5407-1

Yu, G. C., D. K. Smith, H. C. Zhu, Y. Guan, and T. T. Y. Lam. 2017. GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* 8:28–36. doi:10.1111/2041-210x.12628

Zhang, G. J., C. Li, Q. Y. Li, B. Li, D. M. Larkin, C. Lee, J. F. Storz, A. Antunes, M. J. Greenwold, R. W. Meredith, et al. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* 346:1311–1320. doi:10.1126/science.1251385

Zhang, Y., Y. Wang, Y. Li, J. Wu, X. Wang, C. Bian, Y. Tian, G. Sun, R. Han, X. Liu, et al. 2021. Genome-wide association study reveals the genetic determinism of growth traits in a Gushi-Anka F2 chicken population. *Heredity (Edinb)* 126:293–307. doi:10.1038/s41437-020-00365-x

Zhou, X., and M. Stephens. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44:821–824. doi:10.1038/ng.2310

Zhu, B., Q. Li, R. Liu, M. Zheng, J. Wen, and G. Zhao. 2019. Genome-wide association study of H/L Traits in Chicken. *Animals (Basel)* 9:260. doi:10.3390/ani9050260