


DATABASE

Open Access



Galbase: a comprehensive repository for integrating chicken multi-omics data

Weiwei Fu^{1†}, Rui Wang^{1†}, Naiyi Xu^{1†}, Jinxin Wang¹, Ran Li¹, Hojjat Asadollahpour Nanaei¹, Qinghua Nie², Xin Zhao³, Jianlin Han^{4,5}, Ning Yang⁶ and Yu Jiang^{1,7*} 

Abstract

Background: Multi-omics data can provide a stereoscopic view to explore potential causal variations and genes, as well as underlying genetic mechanisms of complex traits. However, for many non-mammalian species, including chickens, these resources are poorly integrated and reused, greatly limiting genetic research and breeding processes of the species.

Results: Here, we constructed Galbase, an easily accessible repository that integrates public chicken multi-omics data from 928 re-sequenced genomes, 429 transcriptomes, 379 epigenomes, 15,275 QTL entries, and 7,526 associations. A total of 21.67 million SNPs, 2.71 million InDels, and 488,583 cis-regulatory elements were included. Galbase allows users to retrieve genomic variations in geographical maps, gene expression profiling in heatmaps, and epigenomic signals in peak patterns. It also provides modules for batch annotation of genes, regions, and loci based on multi-layered omics data. Additionally, a series of convenient tools, including the UCSC Genome Browser, WashU Epigenome Browser, BLAT, BLAST, and LiftOver, were also integrated to facilitate search, visualization, and analysis of sequence features.

Conclusion: Galbase grants new opportunities to research communities to undertake in-depth functional genomic studies on chicken. All features of Galbase make it a useful resource to identify genetic variations responsible for chicken complex traits. Galbase is publicly available at <http://animal.nwsuaf.edu.cn/ChickenVar>.

Keywords: Chicken, Omics data, Complex traits, Database

Background

Technological advances and low sequencing costs have led to the generation and accumulation of a large amount of omics data, which are conducive to understand the genetic architecture of complex traits in farm animals and poultry. The availability of multi-omics data in large cohorts provides opportunities

to systematically quantify the genetic control of DNA methylation, genetic and epigenetic regulations of gene expression, and their effects on complex traits. In addition to providing meat and eggs to consumers, chickens are widely used in developmental biology, medical research, and phenotypic evolution studies as a model organism. Modern chickens were domesticated from their wild ancestor, the red jungle fowl (RJF) [1]. Subsequently, chickens have been developed into different breeds exhibiting remarkable differences in morphology, behavior, physiology, and adaptation [2]. Up to now, most genetic studies on chickens were conducted on one type of omics data with a single analysis method, such as differential expression analysis [3], genome-wide association studies (GWAS) [4, 5], and

[†]Weiwei Fu, Rui Wang and Naiyi Xu contributed equally to this work.

*Correspondence: yu.jiang@nwsuaf.edu.cn

¹ Key Laboratory of Animal Genetics, Breeding and Reproduction of Shaanxi Province, College of Animal Science and Technology, Northwest A&F University, Yangling 712100, China

Full list of author information is available at the end of the article



selective sweep analysis [6, 7]. However, these studies only described the characteristics of data from one level and could not detect credible candidate variations and genes based on limited evidence. A few studies have attempted to reveal the genetic basis for complex traits of chickens using multi-omics datasets, such as for small body size of Yuanbao chicken [8], silky-feather [9], and skin pigmentation phenotype [10] of Silkie chicken, and blue eggshell phenotype of Dongxiang chicken [11]. These studies have narrowed the large candidate lists by screening the overlapping regions based on the evidence from different layers of multi-omics data and experimental verification, demonstrating the need for integrated system-level approaches for analyzing multi-omics data for complex traits. Therefore, developing a comprehensive multi-omics database can help users prioritize the candidate variations and genes, select credible genes for follow-up experimental verification, and apply them to chicken molecular breeding programs.

Currently, several chicken genomics and functional genomics databases have been developed. For instance, GEISHA is a repository of the metadata for genes expressed during the first-six-days of chicken embryo development identified by in situ hybridization analysis [12]. Chickspress is a database to collect gene expression of mRNAs, miRNAs, proteins, and peptides based on different tissue types at different developmental stages [13]. ChickenSD provides whole-genome SNPs from 863 re-sequenced chicken genomes [1]. Apart from these dedicated databases, some generic databases, including EBI/EVA [14], SNPchiMp v.3 [15], Expression Atlas [16], ECRbase [17], Animal QTLdb [18], and GWAS Atlas [19], have also collected different aspects of chicken data. Despite being very useful, these existing databases only focus on one specific omics area and provide a limited amount of data, making them unsuitable for inferring and studying complex traits.

To fill this gap, we developed Galbase, a multi-omics repository to integrate reference genomes, annotations, high-quality genetic variants, transcriptomes, histone modifications, open chromatin regions, variant-trait associations, and QTLs for chicken genetic research and breeding. We used a uniform pipeline to identify SNPs and InDels; evaluated the genetic diversity of each breed; introduced a tissue-specific index, tau, to help determine tissue expression patterns and tissue-specific genes; and integrated the WashU Epigenome Browser to visualize gene regulation patterns. With its rich annotations and functionalities, Galbase will serve as an important public resource to narrow the range of trait candidate genes and

to mine for functional variations by selecting the intersections of evidence from different genomic levels.

Construction and content

Resequencing data alignment and variant calling

We integrated whole-genome sequencing data from a total of 928 chicken samples of different breeds, indigenous populations, and ecotypes based on previously published studies [1, 20, 21] [Date of data collection: April, 2021]. The data-set contained 778 domestic chickens (including 47 breeds) and 150 RJFs (including five RJF subspecies), totaling 7,279 Gb of sequencing data (Additional file 1: Table S1). We first trimmed low-quality reads and adapter sequences using Trimmomatic v0.36 [22]. Clean reads were then aligned against the GRCg6a chicken reference genome [GCF_000002315.6] using the default parameters of BWA-MEM v0.7.15 [23]. The BAM files were processed using Picard v2.1, including position sorting, sample merging, and marking duplicates and removal. The Genome Analysis Toolkit (GATK) v3.7 [24] HaplotypeCaller and GenotypeGVCFs algorithms were used to call SNPs and InDels according to previously published methods [25, 26]. We next filtered SNPs/InDels using the parameters “DP < TotalReadDepth/3 || DP > TotalReadDepth*3 || QD < 2.0 || QUAL < 30.0 || MQ < 40.0 || FS > 60.0 || SOR > 3.0 || ReadPosRankSum < -8.0 || MQRankSum < -12.5” / “DP < TotalReadDepth/3 || DP > TotalReadDepth*3 || QD < 2.0 || QUAL < 30.0 || MQ < 40.0 || FS > 200.0 || SOR > 10.0 || ReadPosRankSum < -20.0 || MQRankSum < -12.5”. We only retained the 1–30 bp indels. Finally, we downloaded a gff3 file for GRCg6a from NCBI and annotated these variations by using SnpEff v.4.3 [27]. The allele frequency of each chicken group and minor allele frequency (MAF) of all chickens were calculated with VCFtools [28] and PLINK [29], respectively.

Genetic diversity and inbreeding coefficient

The nucleotide diversity of each group was estimated by VCFtools, using the parameters “-window-pi 50,000 -window-pi-step 25,000”. The ROH (Runs of homozygosity) of each group was calculated with PLINK parameters “-homozyg-window-snp 50 -homozyg-snp 50 -homozyg-kb 500 -homozyg-density 50 -homozyg-window-missing 5 -homozyg-window-threshold 0.05 -homozyg-window-het 3”. We then calculated genomic inbreeding coefficients (F_{ROH}) by the formula: $F_{ROH} = L_{ROH} / L_{total}$ [30, 31]. To avoid the bias due to sample size variation, we reduced the sample size of each group down to five, following random sampling 10 times, then compared with the calculation results of the original samples.

RNA-seq data processing

We downloaded 429 RNA-seq datasets covering 44 tissues of chicken (Additional file 2: Table S2) from NCBI Sequence Read Archive (SRA) [Date of data collection: March, 2022]. The raw data were pre-processed to filter low-quality reads and remove adaptor sequences by using Trimmomatic v0.36 [22]. Clean reads were then aligned against the chicken GRCg6a reference genome with STAR v2.5.1 [32], and unmapped reads were extracted to perform the second alignment by HISAT2 v2.0.3-beta [33] to improve their utilization [34]. Each bam file was then merged by the Picard tool (v2.1.1). The Transcripts Per Million (TPM) values were computed by StringTie v1.3.4 [35]. The tissue-specific index, tau [36], was calculated by an in-house python script. For some specific experimental designed groups derived from the same project, we performed differential expression analysis using DESeq2 [37].

Epigenome data processing

All available chicken epigenomic data [Date of data collection: March, 2022] were downloaded from NCBI SRA, including histone ChIP-seq and ATAC-seq (Additional file 3: Table S3). Trim Galore was used to trim adaptor and low-quality bases, then cleaned reads were aligned against the chicken GRCg6a reference genome using Bowtie2 v2.2.8 [38]. Low-quality and multiple-mapping reads were removed by using samtools [39] with the option “-q 20”. Reads mapping to mitochondrial DNA were also removed for subsequent analysis. The coverage of reads was calculated by the subroutine “bamcoverage” of deepTools [40]. MACS2 v2.1.1 [41] was used to call peaks with the option “-q 0.05”. The enriched peaks were defined as potential regulatory regions. We first classified the epigenome data into different categories according to the experiment type, such as the H3K4me3 marked as active promoters, H3K27ac marked as active promoters and enhancers, H3K4me1 marked as enhancers and other distal regulatory elements, H3K27me3 marked as repressed transcription, CTCF marked to maintain genome 3D structure, and ATAC-seq marked as open chromatin regions. Next, we overlapped the enriched regions of repeated samples using the BEDtools intersect function, then merged the different tissues using the BEDtools merge function. The total number and total length of the set are used to define the potential regulatory elements of each classification. Finally, we combined all the categories to count the total number and length of regulatory elements in the whole chicken genome.

Phenotypic information collection

We collected chicken phenotypic traits from Animal QTLdb [18], GWAS Atlas [19], and previously reported

studies (Additional file 4: Table S4). For Animal QTLdb, we retained entries containing both chromosomal locations and trait names. For variant-trait associations, we collected data from GWAS Atlas and published literature resources. Since GWAS Atlas only collected 10 chicken publications, we re-searched literature in the NCBI PubMed database using “chicken” and “GWAS” as keywords [Date of data collection: April, 2022]. We manually curated 7,526 associations and 192 traits from 62 publications (Additional file 4: Table S4). We used the chicken GRCg6a reference genome as the reference physical map to convert all phenotypic data to this version using liftOver.

Projection of GRCg6a genomic features to the GRCg7b reference genome

We performed pairwise alignment from GRCg7b [GCF_016699485.2] to GRCg6a by using minimap2, converted the results to liftOver chain files, and then configured an online coordinate conversion. In order to be compatible with the latest chicken reference genome (GRCg7b), we also transformed the coordinates of all multi-omics genomic features to the newer assembly by using the LiftOver tool [42] with the default parameters, which will allow users to retrieve multi-omics data through the two most recently assembled genome assemblies (GRCg6a and GRCg7b).

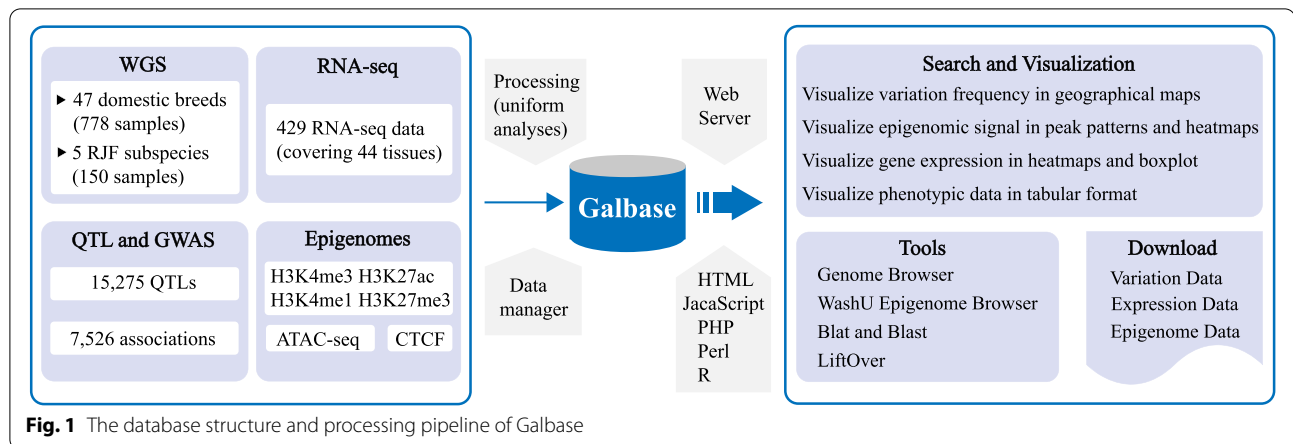
Database implementation

Galbase was built in a LAMP (Linux + Apache + MySQL + PHP) environment, and implemented by a model-view controller (MVC) design pattern based on CodeIgniter framework. The web interfaces were written by PHP, HTML, CSS, JavaScript language, jQuery, Bootstrap, and the open-sourced ECharts framework [43]. The UCSC Genome Browser [42], WashU Epigenome Browser [44], BLAST [45], BLAT, and liftOver [42] functions were also built and adjusted to adapt to our database. We used in-house Perl scripts to process all multi-omics data and manage this large volume of information through MySQL database. The website has been tested in mainstream browsers such as Internet Explorer (version ≥ 9), Firefox, Google Chrome, and Safari.

Utility and discussion

Omics data collection and statistics

Galbase collected the public chicken multi-omics data of 928 re-sequenced genomes (Additional file 1: Table S1), 429 transcriptomes (Additional file 2: Table S2), 379 epigenomes (Additional file 3: Table S3), 15,275 QTL entries (Data from Animal QTLdb), and 7,526 associations (Additional file 4: Table S4). After applying quality



control and standardized data processing procedures, these datasets have been converted to usable table information in the MySQL database (Fig. 1).

The final list of variations included 21,672,487 SNPs and 2,708,244 InDels, which were annotated into 25 consequence types (Additional file 4: Table S5 and S6). We estimated π and F_{ROH} values to evaluate the genomic diversity of each breed (Fig. 2 and Additional file 4: Fig. S1). We found higher level of nucleotide diversity (Fig. 2a and Additional file 4: Fig. S1a) and relatively lower level of inbreeding coefficients (Fig. 2b and Additional file 4: Fig. S1b) in RJFs than domestic chickens, representing a higher level of genetic diversity in wild populations. In domestic chickens, breeds in Southwest Asia, South Asia, Southeast Asia, and Southern China had higher level of nucleotide diversity (Fig. 2a and Additional file 4: Fig. S1a) but lower level of inbreeding (Fig. 2b and Additional file 4: Fig. S1b) than those in Northern China, due possibly to more frequent gene flow between RJFs and sympatric domestic chickens [1]. Some highly selected breeds and native breeds, including White Leghorn, Commercial broilers, Araucana Blue-shelled chicken, Dongxiang Blue-shelled chicken, Daweishan Mini chicken, Yuanbao chicken, Guangxi chicken, Miyi chicken, Jiangxi Silkies, Anyi Gray chicken, Langshan chicken, Shouguang chicken, and Gushi chicken, showed higher levels of inbreeding (Fig. 2b and Additional file 4: Fig. S1b), indicating potential risks of their extinction following inbreeding depression and thus calling for conservation measures [31, 46].

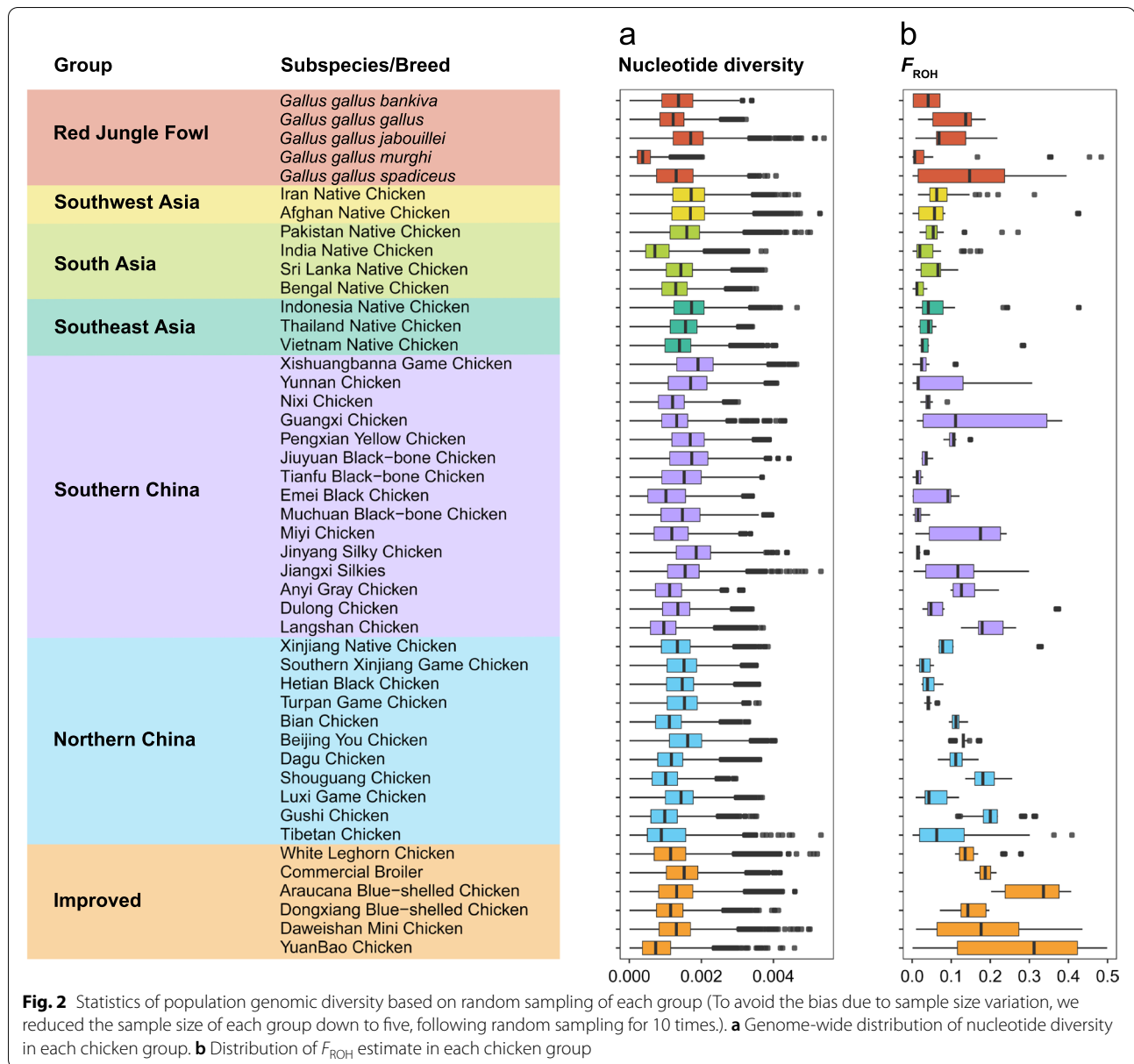
The transcriptome data covers 44 tissues at different developmental stages. We added some valuable groups to provide information for chicken trait studies, such as high and low altitude samples, slow and fast-growing muscles, as well as normal and frizzle shaped feathers. We calculated TPM values and tissue specific index (τ) for 23,729 genes. The gene ontology (GO) analysis validated tissue-specific genes as being involved in

the known tissue-relevant biological processes (Additional file 4: Table S7). We also performed differential expression analysis for each experimental designed group from the same project and provided a list of differentially expressed genes. To better interpret the changes in expression levels, we collected and included four histone modification marks (H3K4me3, H3K27ac, H3K4me1, and H3K27me3) and one transcription factor, CCCTC-binding factor (CTCF), based on ChIP-seq, as well as one open chromatin marker, based on ATAC-seq, to identify cis-regulatory elements. A total of 488,583 cis-regulatory elements were identified, accounting for 49.37% of the whole genome size (Additional file 4: Table S8).

The phenotypic data were collected from AnimalQTLdb, GWAS Atlas, and public literature resources (Additional file 4: Table S4). We mapped reported genes and positional information for all collected phenotypic data to the chicken GRCg6a genome. The data includes 609 different chicken traits which were divided into 15 major categories (Additional file 4: Table S9). We found that the reported traits were mainly concentrated in the “Growth Related Traits”, “Egg Related Traits”, “Exterior Features”, “Behavior Related Traits”, and “Feeding Related Traits” categories, which is consistent with the mainstream research on chickens.

Database characteristics

Galbase comprises a data storage warehouse through MySQL, a user search engine by CodeIgniter, and a set of tools for analysis and visualization. We categorized the chicken multi-omics data into five main retrieval functionalities: (i) variation module; (ii) expression module; (iii) epigenomic module; (iv) phenotypic module; (v) batch annotation; and (vi) a series of useful tools. Each



module has its own page, and features are linked to each other by gene symbols and chromosomal locations.

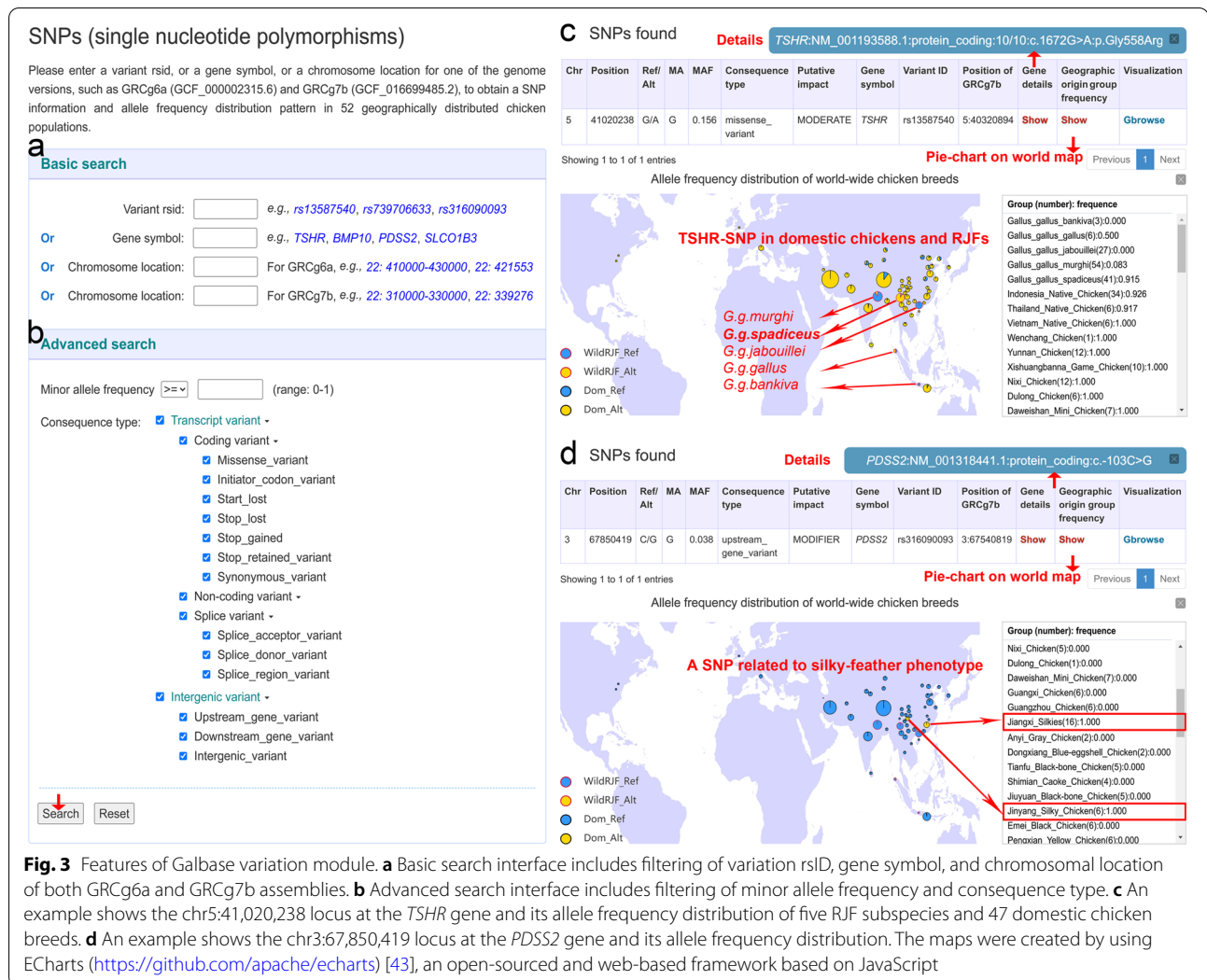
Variation module

The “Variation module” was designed to dynamically retrieve relevant SNP and InDel information in a tabular format or in a genome browser interface. Simply by specifying a variant rsID, a gene symbol, or a chromosomal location (Fig. 3a), users can easily obtain the results of a query for all annotated variation information, including chromosome, position, reference/alternative alleles, MAF, consequence type, variant ID, and allele frequency (Fig. 3c and 3d). Gbrowse (integrated from UCSC Genome

Browser, here we named it “Gbrowse”) was linked to this page to help view other sequence features (Fig. 3c and 3d). More filtering parameters can be set to obtain those variations that fall in different gene bodies or non-protein-coding fraction (Fig. 3b). Moreover, a geographical map showing the allele frequency of five RJF subspecies and 47 chicken breeds can also be displayed, accompanied by each search (Fig. 3c and 3d). This function can help users to identify the breed- or trait-associated variants.

Expression module

The “Expression module” displays gene expression profiles in three ways. The first displays the gene expression



matrix by heatmap and also incorporates a tau value (Fig. 4a), which enables users to easily distinguish gene expression patterns and tissues with specific or high abundance expression. The other two ways display RNA-seq data in Gbrowse. The “expreBar” track (Fig. 4b) can be linked to a more detailed boxplot display page (Fig. 4c) by clicking one gene symbol, where users can view the expression levels of all samples (Fig. 4c). The “RNA-seqReadsCoverage” track, displays normalized read coverage depth by converting BAM files to 1 × sequencing depth, so that users can compare the expression levels between different samples (Fig. 4d). We also performed differential expression analysis by DESeq2 [37] for some specific experimental designed groups, and provide upregulated and downregulated gene lists. Users can download expression matrices and differential gene lists in a CSV format or plain-text files for further analysis.

Epigenomics module

Cis-regulatory elements can help elucidate the causes of complex traits and altered expression levels. Galbase encompasses data relating to four histone modifications: H3K4me3 (active promoters), H3K27ac (active promoters and enhancers), H3K4me1 (enhancers and other distal regulatory elements), H3K27me3 (repressed transcription); one transcription factor, CCCTC-binding factor (CTCF), contributing to 3D genome organization; and one open chromatin marker based on ATAC-seq to identify regulatory regions. The epigenomics metadata, including sample information, experiment type, and epigenetic mark, can be displayed by heatmap views or typical ‘wiggle’ views in WashU Epigenome Browser (Fig. 5b). The epigenomics data can also be retrieved by a table browser. Both regulatory peaks and read coverage (normalized 1 × sequencing depth) can be visualized in Gbrowse (Fig. 5c). Tracks can be sorted, organized,

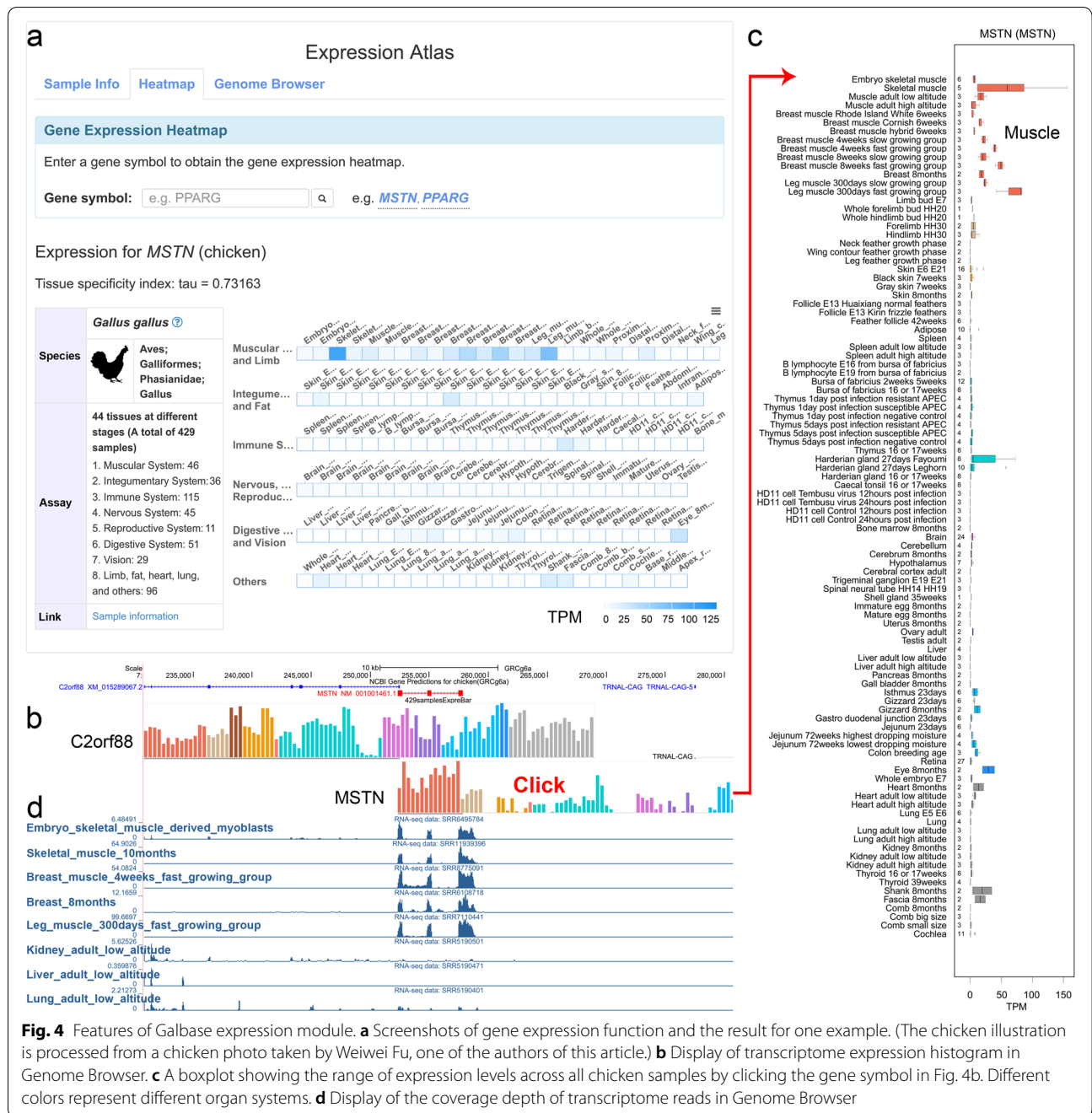


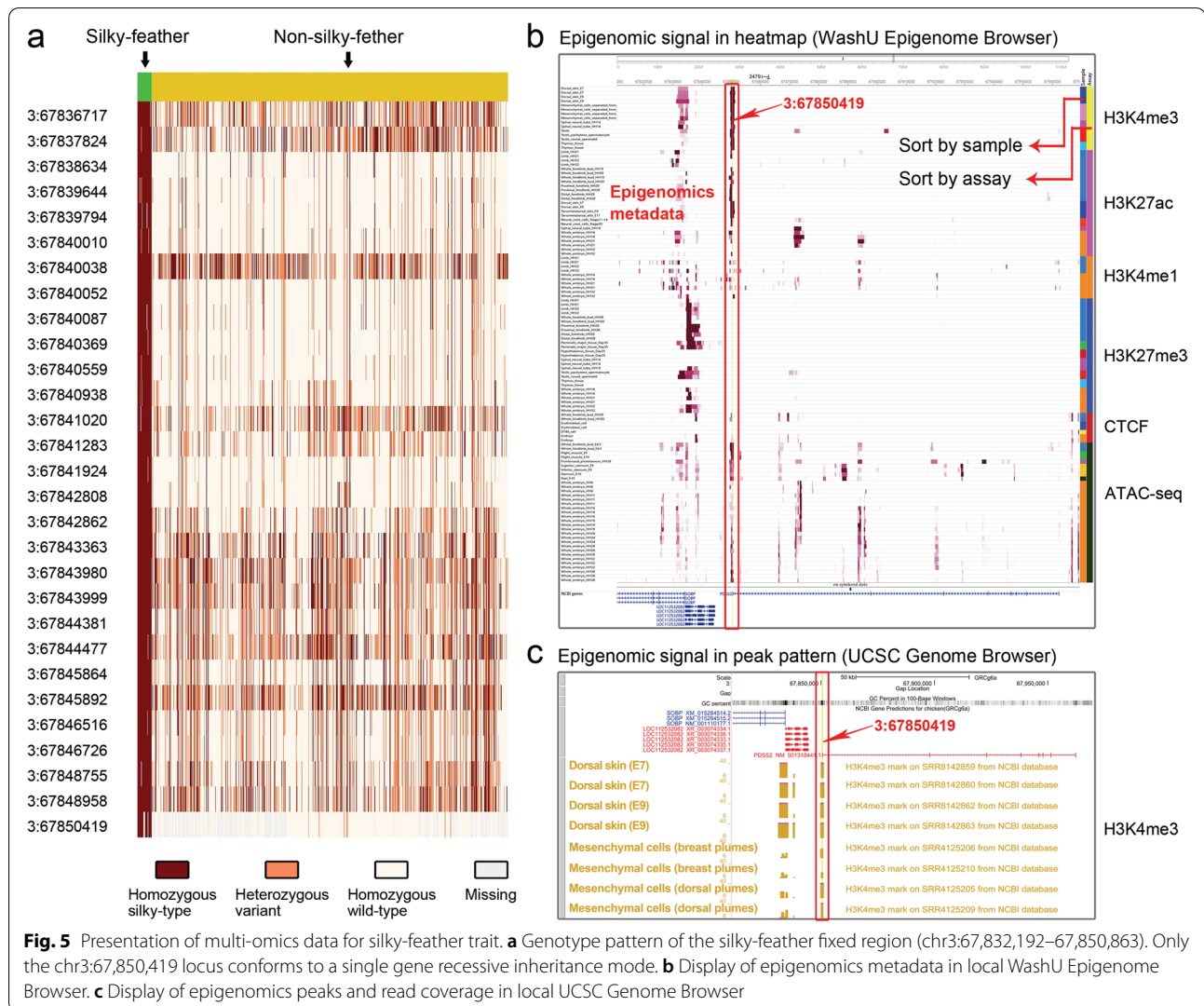
Fig. 4 Features of Galbase expression module. **a** Screenshots of gene expression function and the result for one example. (The chicken illustration is processed from a chicken photo taken by Weiwei Fu, one of the authors of this article.) **b** Display of transcriptome expression histogram in Genome Browser. **c** A boxplot showing the range of expression levels across all chicken samples by clicking the gene symbol in Fig. 4b. Different colors represent different organ systems. **d** Display of the coverage depth of transcriptome reads in Genome Browser

dragged, and printed in a PDF format based on user preferences.

Phenotypic module

The “Phenotypic module” contains 15,275 QTL entries and 7,526 variant-trait associations manually curated from AnimalQTLdb [18], GWAS Atlas [19], and literature resources. To unify different chicken traits, we divided the trait entities

into 15 major categories using the standard classification of chicken QTLdb (Additional file 4: Table S9). We provided various ways to browse and retrieve the phenotypic data: search by gene symbol, search by variant rsid, find QTLs or associations by genome location, and find associated genes by a trait name or a keyword. In order to be compatible with the latest version of the chicken reference genome, we transformed the coordinates to the newest genome assembly by using the LiftOver tool [42] with the default parameters.



Batch annotation

This module was designed on the home page to realize full database retrieval. We integrated all aforementioned multi-omics data to annotate genes, genomic regions, and loci in batches, which can improve the accuracy and reliability of screening, and help users to better analyze and judge the function of genes. Users can enter a candidate list of genes or positions to view all available datasets: (i) SNPs; (ii) InDels; (iii) Epigenomics; (iv) Expression; (v) QTL; (vi) GWAS; (vii) Gene Ontology; and (viii) KEGG pathway. The search results are presented in a tabular format and can be downloaded in a CSV format for downstream analyses.

A series of useful tools

In order to facilitate the usage of the database and to compensate for the lack of tools in the latest chicken GRCg7b reference genome, we have introduced and

built several commonly used tools. The currently available tools included two genome browsers (the UCSC Genome Browser [42] and WashU Epigenome Browser [44]), a BLAST server [45], a BLAT server, and a liftOver tool [42]. Users can use the UCSC Genome Browser to visualize the multi-omics data in a global view. Currently, 1196 tracks for the GRCg6a assembly and 388 tracks for GRCg7b assembly have been released. Users can view SNPs, InDels, gene expression, epigenomics signal, and phenotypic data by searching for a gene symbol or a genomic region. The WashU Epigenome Browser was designed to display epigenomic data specially and we configured the epigenomic data to the Browser. The BLAST and BLAT servers target GRCg6a and GRCg7b genome assemblies, which allows researchers to perform sequence alignment, locate the position of the sequence on the genome, and infer the sequence function. The liftOver tool can offer an online coordinate conversion from

GRCg6a to GRCg7b and chain files can be downloaded to support the liftOver server version.

Example applications and discussion

Establishing a systematic multi-omics database is critical to streamline all mega-datasets to provide an easy access for different users in the field of animal genetics and breeding, however, such databases are very limited in domestic animals. Here, we use chicken as a paradigm for archive, analysis, and visualization of multi-omics data on genome-wide SNPs, InDels, expression, epigenomics, GWAS, and QTL. Compared with other specialized databases, including GEISHA [12], Chickspress [13], and ChickenSD [1], Galbase excels in the following two aspects:

First, Galbase provides variants and their allele frequency for both wild and domestic chickens in nearly 1000 genomes, which can help investigation of the population history of chickens. For instance, a previous study reported a missense mutation in the *TSHR* gene (GRCg6a; chr5:41,020,238 G/A; TSHR-Gly558Arg) to be a domestication locus since it was nearly fixed in all domestic chickens [2]. However, subsequent studies found that the frequency of TSHR-558Arg mutation in European archaeological chickens sharply increased only in the last 1000 years [47], while this allele also had a very high frequency in the ancestor of domestic chickens [1], *Gallus gallus spadiceus*, suggesting this mutation may not be a domestication locus following the complex domestication history of all chickens. By querying our database, users can easily obtain the frequency distribution pattern of domestic chickens and RJFs (Fig. 3c). This intuitive display of geographic allele frequency can help quickly verify hypothesized population history.

Secondly, Galbase provides multi-omics data to help comprehensively judge the potential causal variation of chicken complex traits. For instance, the silky-feather phenotype is controlled by a single recessive gene [9, 48]. We calculated the F_{ST} values of silky-feather and non-silky-feather groups through the vcf file that was downloaded from our database. We selected highly differentiated loci ($F_{ST} > 0.4$) (Additional file 4: Table S10) in the previously reported fine mapping interval of the silky-feather phenotype (GRCg6a; chr3:67,832,192–67,850,863) [9], and presented its genotype patterns (Fig. 5a). We found that only the chr3:67,850,419 locus conformed to the single gene recessive inheritance mode, that is, the homozygous silky-type did not exist in the non-silky-feather group. By querying the epigenomics data in our database, we found that chr3:67,850,419

was located in a region showing strongly-enriched signals for H3K4me3, H3K27ac, and ATAC-seq (Fig. 5b and 5c), which was consistent with the published experimental results showing that the chr3:67,850,419 locus leads to the silky-feather phenotype by affecting promoter activity [9]. This exemplifies the use of multi-omics resources obtained from Galbase to reveal complex traits and reduce the verification work of downstream experiments. In addition, Galbase provides a variety of downloadable forms of expression data. Users can easily screen tissue-specific genes according to tau index or filter differentially expressed genes according to the result of DEseq2, which makes it more convenient to investigate tissue traits.

Data management plan and future update

We will continue to incorporate newly released chicken multi-omics data, and provide dedicated tools required to explore and visualize these data. In order to connect and integrate these external resources more quickly, we will develop an automatic interface to download daily published data, upload it to the supercomputer platform for quality control, processing and analysis when the sample size gets larger than 100 individuals, and finally process the offline data into the website format. For the analysis of re-sequenced genomes, which require a lot of computational resources, we will maintain a major update every year. Our plan for the next phase is to use deep learning algorithms to integrate multi-omics data, which will provide a comprehensive insight from genotype to phenotype, so as to better evaluate and mine heterogeneous multi-omics information.

Conclusions

We present a comprehensive chicken multi-omics database, named Galbase, for the identification of credible candidate genes and loci from different omics layers. To make the data easily accessible, and the usage of the information more effective and flexible, Galbase offers several convenient modules and tools to retrieve, present, and analyze abundant genetic variants, transcriptional, epigenomic, and phenotypic data, which can help uncover the biomechanisms behind complex traits. Galbase provides the largest integrated chicken data repository to date and will help provide new functional insights from genomic data thus offering great promise for the chicken research community.

Abbreviations

QTL: Quantitative Trait Locus; RJF: Red jungle fowl; GWAS: Genome-wide association studies; MAF: Minor allele frequency; SRA: Sequence Read Archive;

TPM: Transcripts Per Million; MVC: Model-view controller; Tau: Tissue specific index; CTCF: CTCF-binding factor.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-022-08598-2>.

Additional file 1:

Additional file 2:

Additional file 3:

Additional file 4:

Acknowledgements

We thank the High-Performance Computing platform of Northwest A&F University for providing computing resources. We are grateful for the advice provided by Hans H. Cheng about the Galbase construction, development and update.

Authors' contributions

Y.J. and W.F. conceived and designed the project; W.F. drafted the manuscript; W.F. and R.W. constructed the database; N.X., J.W., R.L., and H.A.N. collected and analyzed the datasets; Q.N., X.Z., J.H., N.Y., and Y.J. interpreted the analysis results and revised the manuscript. All authors read, commented on, and approved the manuscript.

Funding

This work was supported by grants from the National Natural Science Foundation of China to Y.J. (U21A20247, 31822052), and China Postdoctoral Science Foundation to W.F. (No. 2021M702690). The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

Galbase is freely available at <http://animal.nwsuaf.edu.cn/ChickenVar>. All data used in this study are available from NCBI SRA and PubMed database. The accession numbers analyzed in this study are listed in Additional file 1: Table S1, Additional file 2: Table S2, Additional file 3: Table S3, and Additional file 4: Table S4. And two chicken genome assemblies can be accessed on NCBI Assembly database via the accession numbers: GRCg6a (GCF_000002315.6) and GRCg7b (GCF_016699485.2).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Key Laboratory of Animal Genetics, Breeding and Reproduction of Shaanxi Province, College of Animal Science and Technology, Northwest A&F University, Yangling 712100, China. ²Department of Animal Genetics, Breeding and Reproduction, College of Animal Science, South China Agricultural University, Guangzhou 510642, Guangdong, China. ³Department of Animal Science, McGill University, Montreal, Québec, Canada. ⁴CAAS-ILRI Joint Laboratory On Livestock and Forage Genetic Resources, Institute of Animal Science, Chinese Academy of Agricultural Sciences (CAAS), Beijing, China. ⁵Livestock Genetics Program, International Livestock Research Institute (ILRI), Nairobi, Kenya. ⁶National Engineering Laboratory for Animal Breeding and Key Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture and Rural Affairs, China Agricultural University, Beijing 100193, China. ⁷Center

for Functional Genomics, Institute of Future Agriculture, Northwest A&F University, Yangling, China.

Received: 10 February 2022 Accepted: 28 April 2022

Published online: 12 May 2022

References

- Wang MS, Thakur M, Peng MS, Jiang Y, Frantz LAF, Li M, et al. 863 genomes reveal the origin and domestication of chicken. *Cell Res.* 2020;30(8):693–701. <https://doi.org/10.1038/s41422-020-0349-y>.
- Rubin CJ, Zody MC, Eriksson J, Meadows JR, Sherwood E, Webster MT, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature.* 2010;464(7288):587–91. <https://doi.org/10.1038/nature08832>.
- Zhang J, Kaiser MG, Deist MS, Gallardo RA, Bunn DA, Kelly TR, et al. Transcriptome Analysis in Spleen Reveals Differential Regulation of Response to Newcastle Disease Virus in Two Chicken Lines. *Sci Rep.* 2018;8(1):1278. <https://doi.org/10.1038/s41598-018-19754-8>.
- Pértille F, Moreira GC, Zanella R, Nunes JR, Boschiero C, Rovadoscki GA, et al. Genome-wide association study for performance traits in chickens using genotype by sequencing approach. *Sci Rep.* 2017;7:41748. <https://doi.org/10.1038/srep41748>.
- Raeesi V, Ehsani A, Torshizi RV, Sargolzaei M, Masoudi AA, Dideban R. Genome-wide association study of cell-mediated immune response in chicken. *J Anim Breed Genet.* 2017;134(5):405–11. <https://doi.org/10.1111/jbg.12265>.
- Elferink MG, Megens HJ, Vereijken A, Hu X, Crooijmans RP, Groenen MA. Signatures of selection in the genomes of commercial and non-commercial chicken breeds. *PLoS ONE.* 2012;7(2):e32720. <https://doi.org/10.1371/journal.pone.0032720>.
- Qanbari S, Rubin CJ, Maqbool K, Weigend S, Weigend A, Geibel J, et al. Genetics of adaptation in modern chicken. *PLoS Genet.* 2019;15(4):e1007989. <https://doi.org/10.1371/journal.pgen.1007989>.
- Wang MS, Huo YX, Li Y, Otecko NO, Su LY, Xu HB, et al. Comparative population genomics reveals genetic basis underlying body size of domestic chickens. *J Mol Cell Biol.* 2016;8(6):542–52. <https://doi.org/10.1093/jmcb/mjw044>.
- Feng C, Gao Y, Dorshorst B, Song C, Gu X, Li Q, et al. A cis-regulatory mutation of PDSS2 causes silky-feather in chickens. *PLoS Genet.* 2014;10(8):e1004576. <https://doi.org/10.1371/journal.pgen.1004576>.
- Dorshorst B, Molin AM, Rubin CJ, Johansson AM, Strömstedt L, Pham MH, et al. A complex genomic rearrangement involving the endothelin 3 locus causes dermal hyperpigmentation in the chicken. *PLoS Genet.* 2011;7(12):e1002412. <https://doi.org/10.1371/journal.pgen.1002412>.
- Wang Z, Qu L, Yao J, Yang X, Li G, Zhang Y, et al. An EAV-HP insertion in 5' Flanking region of SLCO1B3 causes blue eggshell in the chicken. *PLoS Genet.* 2013;9(1):e1003183. <https://doi.org/10.1371/journal.pgen.1003183>.
- Antin PB, Yatskievych TA, Davey S, Darnell DK. GEISHA: an evolving gene expression resource for the chicken embryo. *Nucleic Acids Res.* 2014;42(Database issue):D933–937. <https://doi.org/10.1093/nar/gkt962>.
- McCarthy FM, Pendarvis K, Cooksey AM, Gresham CR, Bomhoff M, Davey S, et al. Chickspress: a resource for chicken gene expression. *Database (Oxford).* 2019;2019. <https://doi.org/10.1093/database/baz058>.
- Cook CE, Bergman MT, Finn RD, Cochrane G, Birney E, Apweiler R. The European Bioinformatics Institute in 2016: Data growth and integration. *Nucleic Acids Res.* 2016;44(D1):D20–26. <https://doi.org/10.1093/nar/gkv1352>.
- Nicolazzi EL, Picciolini M, Strozzi F, Schnabel RD, Lawley C, Pirani A, et al. SNPchipMp: a database to disentangle the SNPchip jungle in bovine livestock. *BMC Genomics.* 2014;15:123. <https://doi.org/10.1186/1471-2164-15-123>.
- Papathodorou I, Moreno P, Manning J, Fuentes AM, George N, Fexova S, et al. Expression Atlas update: from tissues to single cells. *Nucleic Acids Res.* 2020;48(D1):D77–83. <https://doi.org/10.1093/nar/gkz947>.
- Loots G, Ovcharenko I. ECRbase: database of evolutionary conserved regions, promoters, and transcription factor binding sites in vertebrate genomes. *Bioinformatics.* 2007;23(1):122–4. <https://doi.org/10.1093/bioinformatics/btl546>.

18. Hu ZL, Park CA, Reecy JM. Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrdB. *Nucleic Acids Res.* 2019;47(D1):D701–10. <https://doi.org/10.1093/nar/gky1084>.
19. Tian D, Wang P, Tang B, Teng X, Li C, Liu X, et al. GWAS Atlas: a curated resource of genome-wide variant-trait associations in plants and animals. *Nucleic Acids Res.* 2020;48(D1):D927–32. <https://doi.org/10.1093/nar/gkz828>.
20. Ye S, Gao N, Zheng R, Chen Z, Teng J, Yuan X, et al. Strategies for Obtaining and Pruning Imputed Whole-Genome Sequence Data for Genomic Prediction. *Front Genet.* 2019;10:673. <https://doi.org/10.3389/fgene.2019.00673>.
21. Ulfah M, Kawahara-Miki R, Farajallah A, Muladno M, Dorshorst B, Martin A, et al. Genetic features of red and green junglefowls and relationship with Indonesian native chickens Sumatera and Kedu Hitam. *BMC Genomics.* 2016;17:320. <https://doi.org/10.1186/s12864-016-2652-z>.
22. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30(15):2114–20. <https://doi.org/10.1093/bioinformatics/btu170>.
23. Li H. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics.* 2012;28(14):1838–44. <https://doi.org/10.1093/bioinformatics/bts280>.
24. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297–303. <https://doi.org/10.1101/gr.107524.110>.
25. Fu W, Wang R, Yu J, Hu D, Cai Y, Shao J, et al. GGVD: A goat genome variation database for tracking the dynamic evolutionary process of selective signatures and ancient introgressions. *J Genet Genomics.* 2021;48(3):248–56. <https://doi.org/10.1016/j.jgg.2021.03.003>.
26. Chen N, Fu W, Zhao J, Shen J, Chen Q, Zheng Z, et al. BGVD: An Integrated Database for Bovine Sequencing Variations and Selective Signatures. *Genomics Proteomics Bioinformatics.* 2020;18(2):186–93. <https://doi.org/10.1016/j.gpb.2019.03.007>.
27. Cingolani P, Platts A, Le Wang L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin).* 2012;6(2):80–92. <https://doi.org/10.4161/fly.19695>.
28. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics.* 2011;27(15):2156–8. <https://doi.org/10.1093/bioinformatics/btr330>.
29. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81(3):559–75. <https://doi.org/10.1086/519795>.
30. McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Peric M, Barac-Lauc L, et al. Runs of homozygosity in European populations. *Am J Hum Genet.* 2008;83(3):359–72. <https://doi.org/10.1016/j.ajhg.2008.08.007>.
31. Xu NY, Si W, Li M, Gong M, Larivière JM, Nanaei HA, et al. Genome-wide scan for selective footprints and genes related to cold tolerance in Chantecler chickens. *Zool Res.* 2021;42(6):710–20. <https://doi.org/10.24272/j.issn.2095-8137.2021.189>.
32. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29(1):15–21. <https://doi.org/10.1093/bioinformatics/bts635>.
33. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods.* 2015;12(4):357–60. <https://doi.org/10.1038/nmeth.3317>.
34. Fu W, Wang R, Nanaei HA, Wang J, Hu D, Jiang Y. GD v2.0: a major update of the ruminant functional and evolutionary genomics database. *Nucleic Acids Res.* 2022;50(D1):D1091–9. <https://doi.org/10.1093/nar/gkab887>.
35. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc.* 2016;11(9):1650–67. <https://doi.org/10.1038/nprot.2016.095>.
36. Kryuchkova-Mostacci N, Robinson-Rechavi M. A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform.* 2017;18(2):205–14. <https://doi.org/10.1093/bib/bbw008>.
37. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15(12):550. <https://doi.org/10.1186/s13059-014-0550-8>.
38. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012;9(4):357–9. <https://doi.org/10.1038/nmeth.1923>.
39. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. *Gigascience.* 2021;10(2). <https://doi.org/10.1093/gigascience/giab008>.
40. Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* 2014;42(Web Server issue):W187–191. <https://doi.org/10.1093/nar/gku365>.
41. Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. *Nat Protoc.* 2012;7(9):1728–40. <https://doi.org/10.1038/nprot.2012.101>.
42. Navarro Gonzalez J, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, et al. The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res.* 2021;49(D1):D1046–57. <https://doi.org/10.1093/nar/gkaa1070>.
43. Li D, Mei H, Shen Y, Su S, Zhang W, Wang J, et al. ECharts: a declarative framework for rapid construction of web-based visualization. *Visual Informatics.* 2018;2(2):136–46. <https://doi.org/10.1016/j.visinf.2018.04.011>.
44. Li D, Hsu S, Purushotham D, Sears RL, Wang T. WashU Epigenome Browser update 2019. *Nucleic Acids Res.* 2019;47(W1):W158–65. <https://doi.org/10.1093/nar/gkz348>.
45. Deng W, Nickle DC, Learn GH, Maust B, Mullins JI. ViroBLAST: a stand-alone BLAST web server for flexible queries of multiple databases and user's datasets. *Bioinformatics.* 2007;23(17):2334–6. <https://doi.org/10.1093/bioinformatics/btm331>.
46. Cendron F, Mastrangelo S, Tolone M, Perini F, Lasagna E, Cassandro M. Genome-wide analysis reveals the patterns of genetic diversity and population structure of 8 Italian local chicken breeds. *Poult Sci.* 2021;100(2):441–51. <https://doi.org/10.1016/j.psj.2020.10.023>.
47. Loog L, Thomas MG, Barnett R, Allen R, Sykes N, Paxinos PD, et al. Inferring Allele Frequency Trajectories from Ancient DNA Indicates That Selection on a Chicken Gene Coincided with Changes in Medieval Husbandry Practices. *Mol Biol Evol.* 2017;34(8):1981–90. <https://doi.org/10.1093/molbev/msx142>.
48. Dunn L, Jull MA. On the inheritance of some characters of the silky fowl. *J Genet.* 1927;19(1):27–63. <https://doi.org/10.1007/BF02983116>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

