

[Click here to view linked References](#)

The genomics data for 78 chickens from 15 populations

1
2
3 Diyan Li^{1†}, Tiandong Che^{1†}, Binlong Chen^{1†}, Shilin Tian^{1,2†}, Xuming Zhou^{3†}, Guolong Zhang^{4†},
4 Miao Li¹, Xiaoling Zhao¹, Huadong Yin¹, Yan Wang¹, Ruiqiang Li², Qing Zhu¹ and Mingzhou
5 Li¹
6

7
8 ¹ Institute of Animal Genetics and Breeding, College of Animal Science and Technology,
9 Sichuan Agricultural University, Chengdu, China

10 ² Beijing Novogene bioinformatics Technology Co., Ltd, Beijing, China

11 ³ Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard
12 Medical School, Boston, USA

13 ⁴ Department of Animal Science, Oklahoma State University, Stillwater, Oklahoma, USA
14

15
16
17
18 † These authors contributed equally to this work.

19 **Correspondence:** zhuqingsicau@163.com; mingzhou.li@sicau.edu.cn.
20
21

22 Abstract

23 **Background:** Since the domestication of the red jungle fowl (*Gallus gallus*) (dating back to
24 ~10,000 B.P.) in Asia, domestic chickens (*Gallus gallus domesticus*) have been subjected to the
25 combined effects of natural selection and human-driven artificial selection; this has resulted in
26 marked phenotypic diversity in a number of traits, including behavior, body composition, egg
27 production and skin color. Population genome variations through diversifying selection have
28 not been fully reported.

29
30
31 **Findings:** The whole genomes of 78 domestic chickens were sequenced to an average of
32 approximately 18-fold coverage depth for each individual with a total of 1.69T reads. Together
33 with the available genomes of 5 wild red jungle fowl and 8 Xishuangbanna game fowls, we
34 conducted a comparative genomics analysis of 91 individuals from 16 populations. After
35 aligning approximately 21.9-Gb high-quality reads of each individual to the reference genome,
36 we identified 5.27-6.65 million SNPs for each population. These SNPs included 1.1 million
37 novel SNPs in 15 populations that were absent from the current chicken dbSNP database. These
38 data are abundant resources for poultry breeding.

39
40
41 **Conclusions:** The data will serve as a valuable resource for investigating diversifying selection
42 and candidate genes for selective breeding in chicken.
43

44
45 **Keywords:** Chicken, Genome diversity, Population genomics, Next generation sequencing
46
47

48 Data description

49 Genome sequencing

50
51 The 78 blood samples were collected from the wing vein, this experimentation was approved
52 by the Institutional Animal Care and Use Committee of Sichuan Agricultural University under
53 permit number YCS-B20100804. Genomic DNA was extracted from these samples using
54 phenol following standard procedures. The whole genomes of 36 Tibetan fowls from the
55 Qinghai-Tibet Plateau and 42 domestic fowls from Szechwan Basin (**Figure 1 and Additional**
56 **file 1: Table S1**) were sequenced on the Illumina HiSeq 2000 platform. The raw reads were
57 cleaned by removing low-quality paired reads, which mainly resulted from base-calling
58
59
60

1 duplicates and adapter contamination. Reads with $\geq 10\%$ of unidentified nucleotides, >10 nt
2 aligned to the adapter, allowing $\leq 10\%$ mismatches, $>50\%$ of bases with Phred quality <33 , and
3 putative PCR duplicates generated in the library construction process were removed. In addition,
4 previously published genome sequence data from two other populations (red jungle fowls and
5 Xishuangbanna game fowls) (GenBank accession number PRJNA241474) were downloaded
6 and analyzed.
7

8 **Data generation and analysis**

9 ***Read mapping***

10 The high-quality paired-end reads were mapped to the reference chicken genome [1]
11 (Galgal4.78) using BWA software [2] with the command ‘mem -t 10 -k 32’. Then, BAM
12 alignment files were generated using SAMtools [3]. Next, we improved the alignment results
13 by filtering the alignment reads with mismatches ≤ 5 and mapping quality=0. The alignment
14 results were then corrected using Picard (<http://sourceforge.net/projects/picard/>) with two core
15 commands. The ‘AddOrReplaceReadGroups’ command was used to replace all read groups in
16 the INPUT file with a new read group and assign all reads to this read group in the OUTPUT
17 BAM. The ‘FixMateInformation’ command was used to ensure that all mate-pair information
18 was in sync between each read and its mate pair. If multiple read pairs had identical external
19 coordinates, only the pair with the highest mapping quality was retained. Reads were realigned
20 around the indels by first identifying the regions for realignment where at least one read
21 contains an indel with a cluster of mismatching bases around it.
22
23
24
25
26
27
28
29
30

31 ***SNP calling***

32 To identify high-credibility variation in the 91 chickens, the highest-accuracy alignment was
33 first processed using the ‘mpileup’ program in SAMtools with the parameters ‘-C -D -S -m 2 -
34 F 0.002 -d 1000’ [3]. The variants were filtered for downstream analysis by requiring a
35 minimum coverage of 4 and a maximum coverage of 200, a minimum RMS mapping quality
36 of 20 and no gaps present within a 3-bp window. We then detected genomic variants for each
37 chicken breed using the Genome Analysis Toolkit (GATK, version v3.1) [4] with the
38 HaplotypeCaller-based method. Before calling variants, the base quality scores were
39 recalibrated using GATK, which provides empirically accurate base quality scores for each base
40 in every read. After SNP calling, we applied variant quality recalibration to exclude potential
41 false-positive variant calls. We used the command ‘VariantFiltration’ with the parameter ‘--
42 filterExpression "QD < 10.0 || FS > 60.0 || MQ < 40.0 || ReadPosRankSum < -8.0" -G_filter
43 "GQ<20"’. The package ANNOVAR[5] was used to annotate SNPs causing nonsense and
44 missense mutations.
45
46
47
48
49
50
51
52
53
54

55 ***Analysis of the population structure and evolutionary history***

56 The phylogenetic relationships of the individual genomes were estimated using principle
57 component analysis (PCA) with the population-scale SNPs using the EIGENSOFT4.2 [6]
58
59
60
61
62
63
64
65

software package, and the eigenvectors were obtained from the covariance matrix using the R function reigen.

Findings

The whole genomes of 78 domestic chickens were sequenced to an average of approximately 18-fold coverage depth for each individual with a total of 1.69T reads (**Additional file 1: Table S1**). The general phenotypic differences between RJF, Tibetan and Sichuan local chicken breeds were presented in **Additional file 1: Table S2**. Together with the available genomes of 5 wild red jungle fowl and 8 Xishuangbanna game fowls (approximately 16.6-fold coverage for each individual), we conducted a comparative genomics analysis of 91 individuals from 15 domestic populations and an ancestry population. After aligning approximately 21.9-Gb high-quality reads of each individual to the reference genome assembly (Galgal4.78) using the Burrows-Wheeler Aligner (BWA) tool [2], we performed population single nucleotide polymorphism (SNP) calling using both SAMtools and the Genome Analysis Toolkit (GATK). We identified 5.27-6.65 million SNPs for each population that were confirmed by both softwares. These SNPs included 1.1 million novel SNPs in 15 populations that were absent from the current chicken dbSNP database (Build 145) (**Figure 1**).

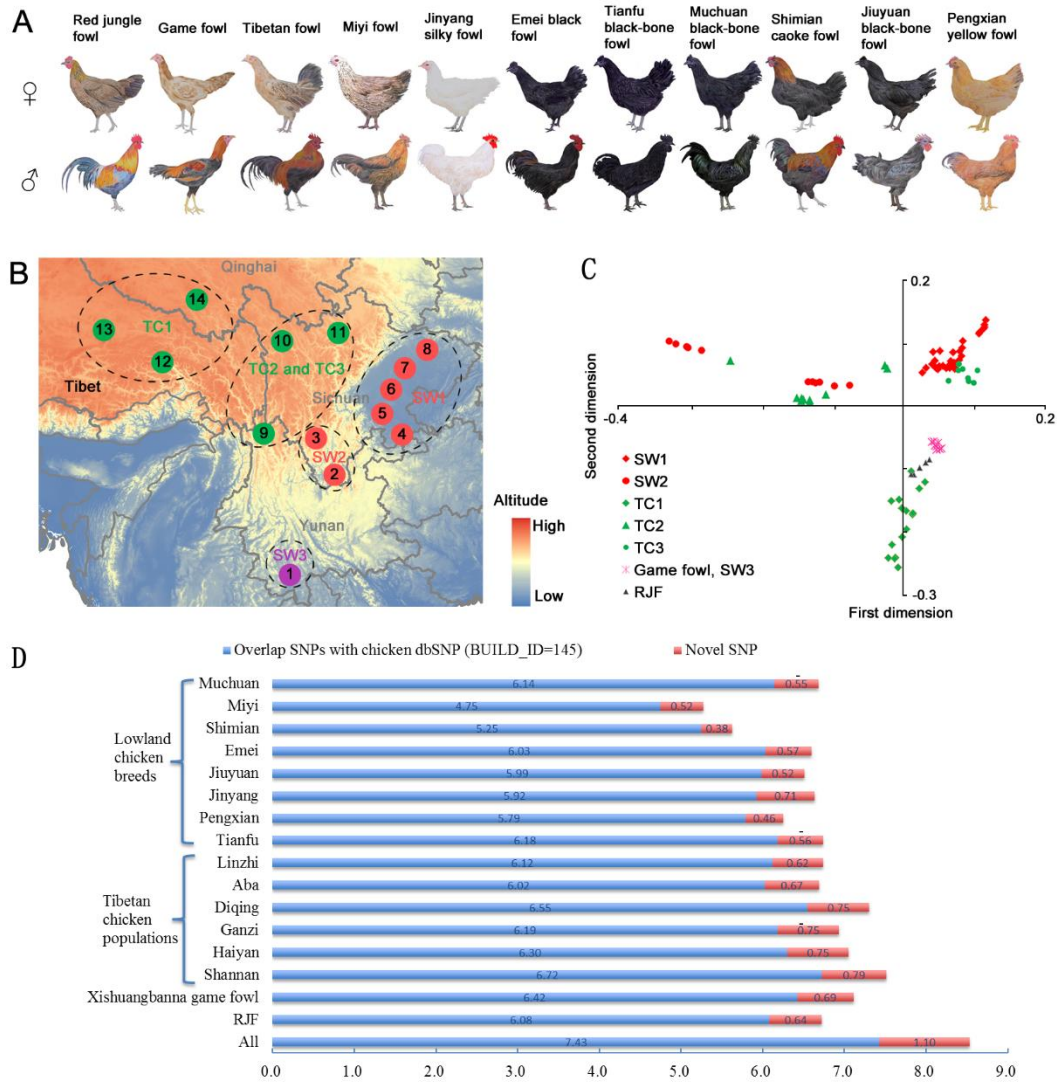


Figure 1 Phenotypic differences, map of sampling localities, population structure and SNPs of the 91 chickens. (A) The striking phenotypic differences among the nine domestic chicken breeds and red jungle fowls (RJFs) analyzed in this study. (B) Geographic distribution of the chicken populations. Red and green localities represent eight lowland and six highland chicken populations used in this study, respectively (1-14 represent Xishuangbanna, Miyi, Jinyang, Muchuan, Emei, Jiuyuan, Chengdu, Pengxian, Diqing, Ganzi, Aba, Shannan, Linzhi and Haiyan, respectively). (C) Principal component plots. The first dimension and second dimension are shown. The fraction of the variance explained was 8.91% for eigenvector 1 ($P < 0.05$, Tracy-Widom test) and 7.43% for eigenvector 2 ($P < 0.05$, Tracy-Widom test). (D) Comparison of identified SNPs in the 15 chicken populations and red jungle fowl with the public database of chicken variants (dbSNP, build 145).

Nucleotide polymorphism and diversity in each population were analyzed using the method of sequence diversity statistics (θ_π and θ_ω) [7]. We observed a relatively higher genetic diversity in all 6 TC sampling locations (pairwise nucleotide diversity, $\theta_\pi = 2.63 \times 10^{-3}$, and average nucleotide polymorphism $\theta_\omega = 2.02 \times 10^{-3}$) compared to the other domestic populations ($\theta_\pi = 2.43 \times 10^{-3}$ and $\theta_\omega = 2.03 \times 10^{-3}$). These values were comparable to those of the RJFs ($\theta_\pi = 2.58 \times 10^{-3}$ and $\theta_\omega = 2.38 \times 10^{-3}$) (**Additional file 1: Table S3**). Particularly, more than half of the SNPs were detected in intergenic regions in each population, suggesting that changes at regulatory sites may have played a prominent role in diversifying selection of various chicken breeds.

1 Phylogenetic analysis based on genome-wide SNPs using the neighbor-joining (NJ)
2 method revealed the segregation of 15 domestic populations and wild RJFs into seven distinct
3 lineages (SW3, SW2, SW1, TC1, TC2, TC3 and a RJF group (**Figure 1**). A similar pattern of
4 clustering was also observed based on principal component analysis (PCA) using
5 EIGENSOFT4.2 [6]. These distinct distribution patterns and expansion signatures suggest that
6 the divergent Tibetan clades may have originated from different regions, such as Yunnan,
7 southwest China and/or surrounding areas [8].
8
9

10 **Availability of supporting data**

11 The sequencing data for this project have been deposited in the NCBI sequence read archive
12 (SRA) under accession number SRP067615. All supplementary figures and tables are provided
13 in Additional file 1.
14
15

16 **Additional file**

17 Additional file 1: Table S1, Table S2 and Table S3. (xlsx 24KB)
18
19

20 **Funding**

21 This work was supported by China Agricultural Research System (CARS-41), the 12th Five
22 Year Plan for breeding program in Sichuan-Selective breeding of new breeds and the synthetic
23 strains in laying hens (2011NZ0099-7), National Natural Science Foundation of China
24 (31160432) and Sichuan Provincial Department of Science and Technology Program
25 (2015JQ0023).
26
27

28 **Authors' contributions**

29 Q.Z., and M.L. designed and supervised the project. B.C., M.L., H.Y., Y.W., X.Z., G.Z. and X.Z.
30 collected and generated the data, and performed the preliminary bioinformatic analyses. S.T.,
31 filtered the data and performed the majority of the population genetic analysis. D.L. and T.C.
32 wrote the manuscript.
33
34

35 **Competing financial interests**

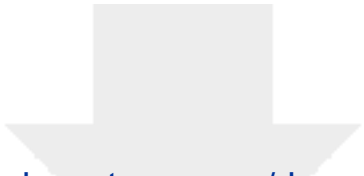
36 The authors declare no competing financial interests.
37
38

39 **References**

- 40 1. Hillier LW, Miller W, Birney E, Groenen MAM, Crooijmans RPMA, Aerts J, et al. Sequence and
41 comparative analysis of the chicken genome provide unique perspectives on vertebrate
42 evolution. *Nature*. 2004; 432: 695-716.
 - 43 2. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform.
44 *Bioinformatics*. 2010; 26: 589-595.
 - 45 3. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map
46 format and SAMtools. *Bioinformatics*. 2009; 25: 2078-2079.
 - 47 4. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome
48 Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.
49 *Genome Res*. 2010; 20: 1297-1303.
- 50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

5. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M,Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*. 2005; 21: 3674-3676.
6. Nick P, Price AL,David R. Population structure and eigenanalysis. *Plos Genetics*. 2006; 2: 2074-2093.
7. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A*. 1979; 76: 5269-5273.
8. Ming-Shan W, Yan L, Min-Sheng P, Li Z, Zong-Ji W, Qi-Ye L, et al. Genomic Analyses Reveal Potential Independent Adaptation to High Altitude in Tibetan Chickens. *Molecular Biology & Evolution*. 2015; 32: 1880-9.



Click here to access/download
Supplementary Material
Additional file 1.xlsx

