Supplementary information

Large-scale genome-wide SNP analysis reveals the rugged (and ragged) landscape of global ancestry, phylogeny, and demographic history in chicken breeds

Natalia V. DEMENTIEVA[1], Yuri S. SHCHERBAKOV[1], Olga I. STANISHEVSKAYA[1], Anatoly B. VAKHRAMEEV[1], Tatiana A. LARKINA[1], Artem P. DYSIN[1], Olga A. NIKOLAEVA[1], Anna E. RYABOVA[1], Anastasiia I. AZOVTSEVA[1], Olga V. MITROFANOVA[1], Grigoriy K. PEGLIVANYAN[1], Natalia R. REINBACH[1], Darren K. GRIFFIN[2], Michael N. ROMANOV[2,3]

[1]*Russian Research Institute of Farm Animal Genetics and Breeding – Branch of the L. K. Ernst Federal Research Centre for Animal Husbandry, Pushkin, St. Petersburg, 196601, Russia*
[2]*School of Biosciences, University of Kent, Canterbury, CT2 7NJ, UK*
[3]*L. K. Ernst Federal Research Center for Animal Husbandry, Dubrovitsy, Podolsk, Moscow Oblast, 142132, Russia*

Data S1: Genomic SNP-scanning of the Russian White gene pool population for studying its structure and developing the genomic selection methodology

Genomic SNP scan

Fig. S1-1. A Russian White laying hen.

The population of Russian White chickens (RW; Fig. S1-1), selected in the RRIFAGB gene pool for 25 generations using individual selection, was examined using a genome-wide SNP scan to study the genetic features of the structure of its population when comparing the modern subpopulation of the RW, RWG, with the 2001 RWS ancestral subpopulation (Dementeva et al. 2017). Despite the high computerization of the process, the final quality of the decoding of animal genotypes was different, but within the acceptable norm. The quality of sample genotyping exceeded 90%, given the importance of preparing genotypic data as a necessary technical step for subsequent bioinformatic analysis. In view of the different sexes of the animals in the sample, the sex chromosomes were removed in order to prevent distortion of the analysis.

Estimating the genetic distances between subpopulations within the same breed is an important part of controlling variation within a breed. In particular, three subpopulations of the RW breed were studied. Two subpopulations from 2001 (RWP, of the ARPRTI, 10 animals; and RWS, of the RRIFAGB, 6 animals) and a modern RRIFAGB subpopulation (RWG, 170 animals) were analysed. The greatest genetic distance was found between RWP and RWS. The results are presented in Table S1-1 and Fig. S1-2.

Table S1-1. Genetic divergence of pairwise compared subpopulations of the Russian White chicken breed obtained by whole genome SNP scanning.

Subpopulation 1	Subpopulation 2	F_{ST}	SD	Chi squared	*p*-value
RWG (*n*=170)	RWS (*n*=6)	0.112	0.003	34.566	0.000148055
RWP (*n*=10)	RWS (*n*=6)	0.209	0.005	216.678	5.29722e-41
RWP (*n*=10)	RWG (*n*=170)	0.101	0.001	233.848	1.33943e-44

Based on the principal component analysis (PCA; Fig. S1-2), the current RWG subpopulation had a higher variability as compared to both the ancestral RWS and the RWP subpopulation and showed an adequate pattern of the distribution of the studied animals. Significant differences were found between the RRIFAGB subpopulations (RWG and RWS) and the RWP subpopulation ($p<0.05$).

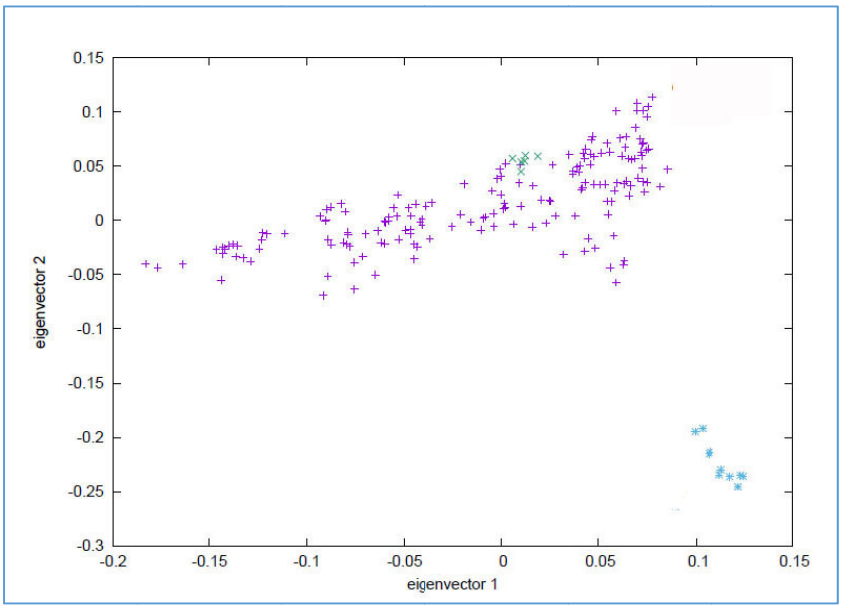

Fig. S1-2. Results of PCA analysis for 33,834 SNP markers in the area of principal components of vector 1 (eigenvector 1) and vector 2 (eigenvector 2). Russian White (RW) subpopulations:

+ , RWG (current RW subpopulation);

× , RWS (RRIFAGB, 2001);

✳ , RWP (ARPRTI, 2001).

Determination of the genetic structure of the current RWG subpopulation and its comparison with the ancestral RWS subpopulation were carried out by the method of multidimensional scaling (MDS) (Fig. S1-3). Linkage disequilibrium (LD) and the frequencies of allele occurrence of by groups were calculated using the PLINK 1.9 program (as implemented by Dementeva et al. 2017). In particular, MDS, which reflects the spatial distribution of animals depending on genetic similarity, was used to analyse the difference between subpopulations of RW chickens. In other words, the farther in space the animals were located on the graph, the less similar they were.

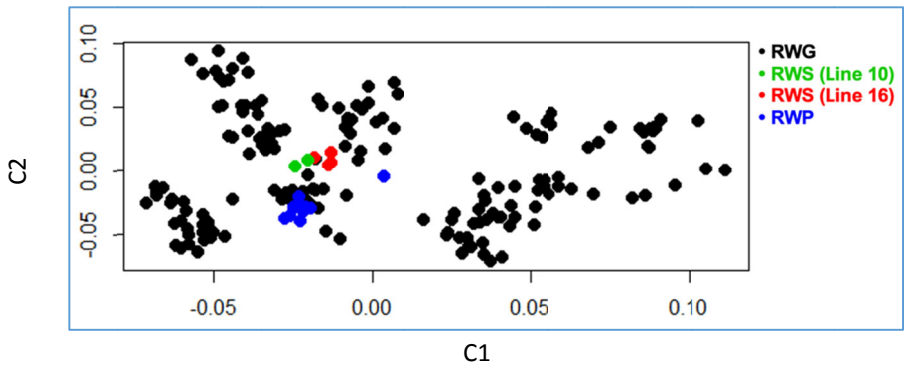

Fig. S1-3. MDS analysis for Russian White subpopulations. Amrock (Ar), Aurora Blue (AB), Australorp Black (AoB), Australorp Black Speckled (ABS), Bantam Mille Fleur (BMF), Brahma Buff (BB), Brahma Light (BL), Cochin Bantam (including three subpopulations CBm1, CBm2 and CBm3), Cochin Blue (CBl), Czech Golden (CG), Faverolles Salmon (FS), Frizzle (F), Hamburg Silver Spangled Dwarf (HSSD), Leghorn Light Brown (or Italian Partridge; LLB), Leningrad Golden-and-grey (LGG), Leningrad Mille Fleur (LMF), Minorca Black (MB), Moscow Game (MG), Naked Neck (NN), New Hampshire (NH), Orloff Mille Fleur (OMF), Pantsirevka Black (PB), Pavlov Spangled (PS), Pavlov White (PW), Pervomai (two subpopulations Pm1 and Pm2), Plymouth Rock Barred (PRB), Poland White-crested Black (PWB), Poltava Clay (PC), Pushkin (Pu), Red White-tailed Dwarf (RWD), Rhode Island Red (RIR), Russian Crested (RC), Russian White (RW, including three subpopulations RWS, RWP and RWG), Silkie White (SW), Sussex Light (two subpopulations SL1 and SL2), Tsarskoye Selo (Ts), Ukrainian Muffed (UM), Uzbek Game (or Kulangi; UG), White Cornish (three subpopulations WC1, WC2 and WC3), Yurlov Crower (YC), Zagorsk Salmon (ZS).

Thus, the population structure was studied and the analysis of the genomic architecture of the RW breed was carried out. The practical value of analysis at the level of various subpopulations was to use its results to assess and control variability in the breed. As a result, it was found that the greatest difference was observed between the historical RWP and RWS bird groups, while the current RWG subpopulation showed higher variability as compared to the ancestral RWS subpopulation. The data obtained suggest the intensive development of the current RWG subpopulation and the successful breeding process. Variation, being the source of breeding progress, was studied more deeply in the current RWG subpopulation. The analysis revealed a clear distribution of animals into four subgroups, and information about the belonging of individuals to subgroups will be used in future breeding work. In particular, it was found that a key factor in the observed distribution was the use of four main roosters in breeding the animals of the sample. A more accurate genomic characterization of the identified groups was carried out by studying the characteristics of the groups in terms of the presence and length of regions where LD of SNP markers was observed. The group of maximum LD was found to be the RWS historical subpopulation. Despite the common position of the third group of the current RWG subpopulation and the ancestral RWS subpopulation on the MDS plot, they differed in the number of monomorphic SNPs. Using a linear model, the significance of the effect of group, chromosome, interaction, and SNP interval on LD was also assessed.

Methodology of genomic selection

One of the bioinformatic criteria for assessing the genetic characteristics of small populations for planning the breeding process, including genomic selection, is the assessment of dynamic changes in the molecular architecture in a population over time. The need for this is due to the preservation of the characteristics of the breed.

The second of the criteria is the structure of the breed, which is based in a small population on the presence of related similarity and is determined by the presence of SNPs that are in LD. The structural features of the groups are studied in terms of the presence and length of regions where LD of SNP markers was observed. The average value of LD, and the number of monomorphic and minor alleles in each group is determined.

The calculation of the frequencies of LD occurrence equal to 1 at various distances, especially large ones, between SNP markers provides information about the saturation of the population with haploblocks. Based on this, a conclusion is made about the accumulation or decay of long LD regions in the population over time.

The third important criterion is the assessment of the level of inbreeding. Using an example of comparing breeds with different levels of genetic diversity, one shows the effect of breeding methods on indicators of genetic variability, such as F_{IS} and ROH (presence and length of homozygous regions).

The fourth important criterion is the formulation of the tasks of the selection process and the determination of solutions. By implementing genomic association analysis, the search for SNPs that are in LD at the localization sites of a putative QTL, as well as the study of the localization of homozygous regions, it is possible to determine markers for conducting a directed selection process.

The development of the genomic selection methodology (Fig. S1-4), stated as a research goal, was carried out on the example of finding solutions when creating a biotechnological line of poultry based on the gene pool RW breed for the production of viral vaccines (Kudinov et al. 2019; Dementieva et al. 2020a). Evaluation of the breeding value of the RW chicken breed on the basis of productivity is a necessary element for building a breeding program and implementing genetic progress in the population. Among modern methods for assessing breeding value, the most promising and proven method is the use of genome-wide information to predict breeding value. This method is widely known as a genomic prediction method, whereas genomic breeding of chickens is little described in the literature due to the limited dissemination of relevant information by large commercial breeding companies.

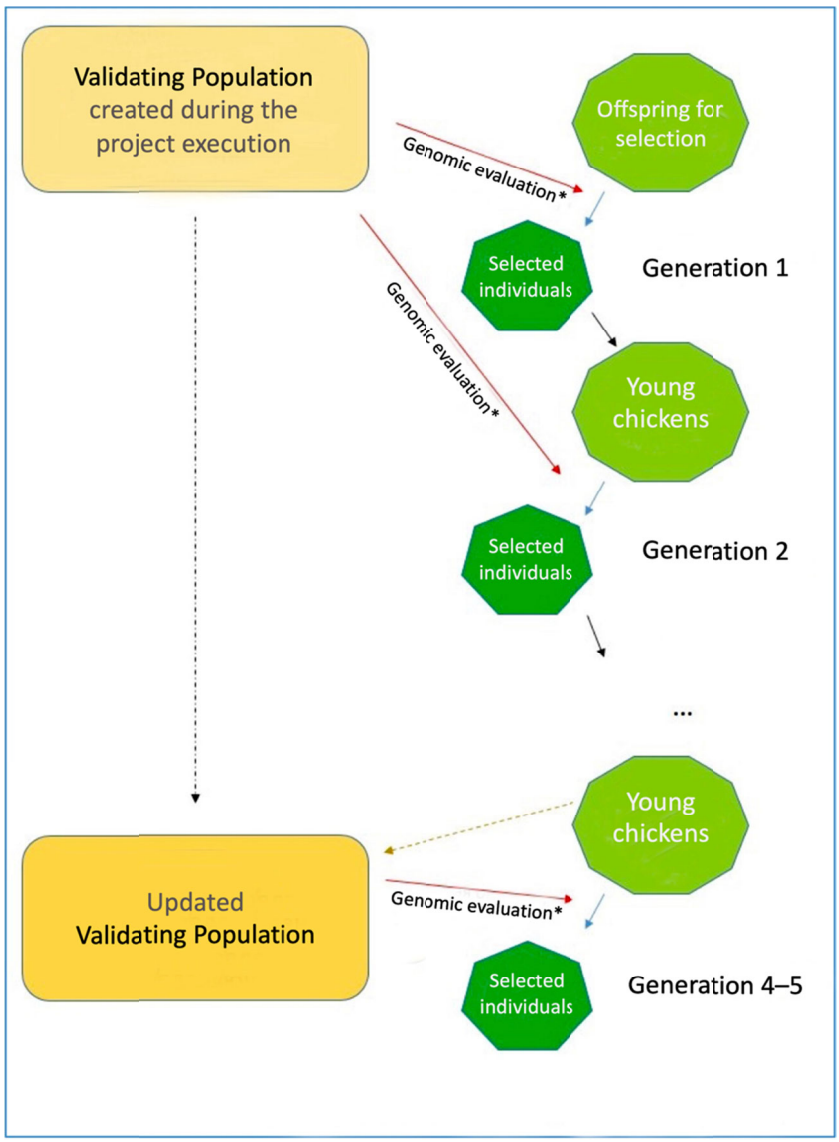

Fig. S1-4. Scheme of genomic selection of the gene pool Russian White chicken breed.

Phenotypic and genealogical information from 428 chickens was collected to develop a methodology for genomic selection of the gene pool RW chickens. Due to the small number of gene pool breeds and flocks, the most adequate of the existing ones is the single step GBLUP genomic assessment method (Christensen and Lund 2010), which simultaneously uses information about the molecular and phenotypic data of the studied population, which does not require prior launch of the traditional BLUP model. The potential of the method lies in the unified use of data on productivity, pedigree and genetic markers in a single calculation space, without loss of information at the stages of a multi-phase process.

To carry out a genomic assessment based on the characteristics of egg production, egg weight and volume of embryonic fluid, a mixed model equation of the following type was developed:

$$Y=H+CG+a+e,$$

where Y is a performance trait, H is a year or generation, CG is place of bird keeping (cage), a is the breeding value of an animal, and e is the unknown residual.

The proposed model was used to calculate the additive and residual variances of the studied traits. Genetic variance is a necessary component of the breeding value model. One of the particular variants of REML can be used for evaluation. In this study, the AI-REML method (Jensen et al. 1997) was used. The defining element of the single-step (ssGBLUP) model is the presence of an inverted H matrix, which is a combination of the relatedness matrix for non-genotyped A22 and genotyped G animals. The creation of the H matrix and its inversion is possible using one of the

methods described in the literature. In particular, the VanRaden method 1 (VanRaden, 2008) was used to create a genomic kinship matrix. It is a combined matrix of kinship and genotypes that makes it possible to successfully use the method on small and closed populations. The absence of the need to divide the population into validating and validated (or estimating and estimated) leads to a balanced use of all data available for analysis. This eliminates the loss of information when a historical animal has phenotypic productivity indicators, but does not have biological material for genotyping, and a current young animal is available for genotyping, but does not have productivity indicators taken into account. This scheme makes it possible to saturate the young population with data from ancestors, while the ancestral part receives potential information about alleles and their frequencies due to the current (genotyped) population. The studied population of the RW chicken breed can be divided into the current RWG subpopulation and the ancestral RWS subpopulation, genotyped in proportions of 65% and 35%, while the availability of data on productivity in genotyped chickens for individual traits will make it possible to assess the transfer of the trait more accurately from ancestor to descendant.

Thus, genomic selection of animals and birds is based on the algorithm for making a selection decision using the results of genomic prediction. The classical breeding model of a population implies the appearance of individuals of average productivity, below and above average, which corresponds to a normal distribution. The classical model of population breeding involves the selection of the best individuals (above average) as parents of future generations. A distinctive feature of poultry farming is a shortened reproductive (genetic) interval, which allows the use of several generations of live chickens with an assessment of breeding value. The greatest potential of genomic selection is the selection of individuals for traits that have lower heritability or require additional costs for their measurement.

References

Dementeva NV, Romanov MN, Kudinov AA, et al., 2017. Studying the structure of a gene pool population of the Russian White chicken breed by genome-wide SNP scan. *Sel'skokhozyaistvennaya Biol*, 52(6):1166-1174. https://doi.org/10.15389/agrobiology.2017.6.1166eng

Kudinov AA, Dementieva NV, Mitrofanova OV, et al., 2019. Genome-wide association studies targeting the yield of extraembryonic fluid and production traits in Russian White chickens. *BMC Genomics*, 20:270. https://doi.org/10.1186/s12864-019-5605-5

Dementieva NV, Fedorova ES, Krutikova AA, et al., 2020a. Genetic variability of indels in the prolactin and dopamine receptor D2 genes and their association with the yield of allanto-amniotic fluid in Russian White laying hens. *Tarım Bilim Derg – J Agric Sci*, 26(3):373-379. https://doi.org/10.15832/ankutbd.483561

Christensen OF, Lund MS, 2010. Genomic prediction when some animals are not genotyped. *Genet Sel Evol*, 42:2. https://doi.org/10.1186/1297-9686-42-2

VanRaden PM, 2008. Efficient methods to compute genomic predictions. *J Dairy Sci*, 91(11):4414-4423. https://doi.org/10.3168/jds.2007-0980