scientific reports

Check for updates

OPEN

Disentangling clustering configuration intricacies for divergently selected chicken breeds

Anatoly B. Vakhrameev[1], Valeriy G. Narushin[2,3], Tatyana A. Larkina[1], Olga Y. Barkova[1], Grigoriy K. Peglivanyan[1], Artem P. Dysin[1], Natalia V. Dementieva[1]✉, Alexandra V. Makarova[1], Yuri S. Shcherbakov[1], Marina V. Pozovnikova[1], Yuri V. Bondarenko[4], Darren K. Griffin[5]✉ & Michael N. Romanov[5,6]✉

Divergently selected chicken breeds are of great interest not only from an economic point of view, but also in terms of sustaining diversity of the global poultry gene pool. In this regard, it is essential to evaluate the classification (clustering) of varied chicken breeds using methods and models based on phenotypic and genotypic breed differences. It is also important to implement new mathematical indicators and approaches. Accordingly, we set the objectives to test and improve clustering algorithms and models to discriminate between various chicken breeds. A representative portion of the global chicken gene pool including 39 different breeds was examined in terms of an integral performance index, i.e., specific egg mass yield relative to body weight of females. The generated dataset was evaluated within the traditional, phenotypic and genotypic classification/clustering models using the *k*-means method, inflection points clustering, and admixture analysis. The latter embraced SNP genotype datasets including a specific one focused on the performance-associated *NCAPG-LCORL* locus. The *k*-means and inflection points analyses showed certain discrepancies between the tested models/submodels and flaws in the produced cluster configurations. On the other hand, 11 core breeds were identified that were shared between the examined models and demonstrated more adequate clustering and admixture patterns. These findings will lay the foundation for future research to improve methods for clustering as well as genome- and phenome-wide association/mediation analyses.

Abbreviations

BTB	Bantam type breeds
CV	Cross-validation
DPB	Dual purpose breeds
EMB	Egg-meat breeds
ETB	Egg type breeds
EY	Egg mass yield
FB	Fancy breeds
GB	Game breeds
GCM	Genotypic clustering model
IPM	Inflection points model
K	Number of ancestral populations
MEB	Meat-egg breeds
MTB	Meat type breeds

[1]Russian Research Institute of Farm Animal Genetics and Breeding – Branch of the L. K. Ernst Research Science Center for Animal Husbandry, Pushkin, St. Petersburg, Russia. [2]Research Institute for Environment Treatment, Zaporozhye, Ukraine. [3]Vita-Market Ltd, Zaporozhye, Ukraine. [4]Sumy National Agrarian University, Sumy, Ukraine. [5]School of Biosciences, University of Kent, Canterbury, UK. [6]L. K. Ernst Federal Research Centre for Animal Husbandry, Dubrovitsy, Podolsk, Moscow Oblast, Russia. ✉email: dementevan@mail.ru; D.K.Griffin@kent.ac.uk; m.romanov@kent.ac.uk

PCM	Phenotypic clustering model
RRIFAGB	Russian Research Institute of Farm Animal Genetics and Breeding
S	Silhouette score
SNP	Single nucleotide polymorphism
SS	Sum of square distances
SSE	SS within groups
SSG	SS between groups
SST	Total SS
TCM	Traditional classification model
W	Body weight

The global chicken gene pool has been shaped during thousands of years of domestication and demographic history of diverse chicken breeds. These meet versatile human needs for table eggs, poultry meat and aesthetic preferences, culminating in a wide variety of chicken breeds with valuable genomic and phenomic features. They have arisen on different continents and in different countries as a consequence of artificial selection for certain phenotypic (productive) traits and specialized interbreeding (e.g.,[1–5]). Moiseyeva et al.[6] established four major evolutionary lineages of chicken breed formation: (1) egg type (ETB), (2) meat type (MTB), (3) game (GB), and (4) Bantam (BTB; or miniaturized type) breeds. Comparing the phenotypic and genotypic features of a large sample of the world gene pool, Larkina et al.[7] added two more chicken breed formation categories: dual purpose (DPB), including egg-meat (EMB) and meat-egg (MEB) subtypes, and fancy (FB; or ornamental) breeds.

In our previous studies[7,8], we considered three main classification (clustering) models for the evolutionarily determined subdivision of the global chicken breed gene pool. These were: (1) traditional classification model (TCM) generally accepted in poultry breed categorization; (2) phenotypic clustering model (PCM) built according to a suite of phenotypic/performance traits; and (3) genotypic clustering model (GCM; including its two variants, GCM1 and GCM2) based on single nucleotide polymorphism (SNP) genotypes at the well-known *NCAPG-LCORL* locus associated with chicken performance[7,9–11]. This locus has been identified in mammals as a locus associated with body growth and development. Its significant associations with height were shown for the Liangzhou donkey[12], cattle[13], as well as in relation to body weight and skeletal size in sheep[14,15]. Significant SNPs at this locus appear to influence egg weight[16], oviduct size[17] and internal organ mass in chickens[18]. With some preference in favor of PCM, it was, however, very difficult to decide unambiguously which of the classification (clustering) models of breeds was the most suitable.

In cluster analysis, especially when looking for plausible distribution configurations of species, breeds or populations, one often turns to the use of *k*-means clustering[19–22] as well as the elbow method of clustering (e.g.,[23]). A nonhierarchical *k*-means technique is a popular method of multivariate analysis[24], which was also used in cluster analysis to describe the egg-laying patterns of hens (e.g.,[25]). This algorithm seeks to minimize the total square deviation of cluster points from these clusters' centers[21]. The elbow method is a heuristic used in determining the number of clusters in a dataset by plotting the explained variation function and picking the optimal number of clusters at the elbow point of the explained variation curve (e.g.,[26,27]). The elbow method is also applicable for inferring ancestral populations in the admixture analysis based on multi-locus SNP genotypes[28]. For instance, Larkina et al.[7] and Abdelmanova et al.[29] applied it to choosing the optimal number of clusters (ancestral populations) for interpretation of chicken breed clustering.

In this regard, we set ourselves the goal of testing and improving the well-known clustering algorithms based on, or including, the *k*-means and elbow methods. We also, where possible, established novel algorithms to discriminate between various chicken breeds. Using 39 breeds representing a fairly large portion of the world chicken gene pool, i.e., ~ 6% of the FAO estimate of known chicken breeds[30], we analyzed the respective datasets for the above three classification (clustering) models described in the previous study[7]. This enabled generation of new insights into clustering configuration intricacies for divergently selected chicken breeds that can be useful in future genome- and phenome-related research.

Methods

Chicken breeds. A broad sampling of the global chicken gene pool encompassed a total of 759 hens from the 39 breeds (populations) maintained at the Russian Research Institute of Farm Animal Genetics and Breeding (RRIFAGB) bioresource collection farm (Table 1). The 39 populations were purebred, except a meat-type population of three-way hybrids (White Cornish × (Brahma Light × Sussex Light)) bred inter se.

We assessed the following key phenotypic (performance) traits in each breed: mean egg mass yield per hen for 52 weeks of life (*EY*), and mean body weight of sexually mature females at the age of 52 weeks (*W*). *EY*, in turn, was calculated as the product of mean egg weight at 35 weeks of age and egg production for 52 weeks. For subsequent analyses, we introduced an integral performance coefficient for each breed, *EY*/*W*, obtained by dividing *EY* by *W* (Table 1).

For subsequent analysis, we omitted GCM2, since it was the only one of the analyzed models that did not allow us to assess the degree of belonging to ETB among any of the identified performance-based breed types. In GCM1, however, we did not have a breakdown into EMT and MET breeds; therein, we conventionally designated such breeds as DPB[7]. Accordingly, if in the first two models it turned out, for example, that the Pantsirevka Black breed belongs to DPB:EMT breeds, and in GCM1 this was a DPB (Table 1), then we conditionally considered that we had a match of breed types in all three models. For convenience, we further refer to GCM1 as GCM. Venn diagram plotting[31] was used for visualizing the number of different breed types shared between the three classification (clustering) models, i.e., TCM, PCM, and GCM (Table 1).

Breed name	Breed code	*n*[a]	Breed type by model[b]			No. of hens	*EY*/*W*
			TCM	PCM	GCM		
Cochin Bantam	CB	1	FB	BTB	MTB	20	10.06
Red White-tailed Dwarf	RWD	2	MTB	BTB	MTB	18	8.57
Bantam Mille Fleur	BMF	3	FB	BTB	ETB Ia	20	6.60
Russian White	**RWG**	**4**	**ETB**	**ETB**	**ETB Ib**	**30**	**5.83**
Pushkin	Pu	5	DPB:EMB	DPB:MEB	MTB	20	5.29
New Hampshire	NH	6	DPB:EMB	DPB:MEB	MTB	19	5.13
Hamburg Silver Spangled Dwarf	HSSD	7	FB	BTB	ETB Ia	20	5.10
Leningrad Mille Fleur	LMF	8	DPB:EMB	DPB:MEB	DPB IIb	21	5.06
Leghorn Light Brown	**LLB**	**9**	**ETB**	**ETB**	**ETB Ia**	**19**	**4.77**
Leningrad Golden-and-gray	LGG	10	DPB:EMB	DPB:EMB	MTB	20	4.62
Aurora Blue	AB	11	DPB:MEB	ETB	DPB IIa	20	4.54
Amrock	Ar	12	DPB:MEB	ETB	DPB IIa	20	4.46
Rhode Island Red	RIR	13	DPB:EMB	DPB:EMB	ETB Ib	32	4.43
Pavlov Spangled	PS	14	FB	FB	ETB Ia	20	4.36
Poland White-crested Black	PWB	15	FB	FB	ETB Ia	18	4.23
Pantsirevka Black	**PB**	**16**	**DPB:EMB**	**DPB:EMB**	**DPB IIa**	**17**	**4.16**
Russian Crested	RC	17	FB	DPB:EMB	ETB Ib	19	4.12
Frizzle	F	18	FB	ETB	DPB IIa	20	4.07
Plymouth Rock Barred	**PRB**	**19**	**DPB:MEB**	**DPB:MEB**	**DPB IIb**	**18**	**4.06**
Zagorsk Salmon	ZS	20	DPB:EMB	DPB:MEB	MTB	18	3.81
Tsarskoye Selo	**Ts**	**21**	**DPB:MEB**	**DPB:MEB**	**DPB IIb**	**20**	**3.75**
Naked Neck	NN	22	DPB:MEB	ETB	ETB Ib	20	3.74
Sussex Light	SL	23	DPB:MEB	DPB:EMB	DPB IIa	20	3.72
Poltava Clay	**PC**	**24**	**DPB:MEB**	**DPB:MEB**	**DPB IIb**	**17**	**3.72**
Silkie White	SW	25	FB	BTB	MTB	19	3.70
Minorca Black	MB	26	ETB	DPB:EMB	ETB Ia	19	3.56
Brahma Light	BL	27	FB	DPB:EMB	ETB Ib	20	3.50
Pavlov White	PW	28	FB	FB	ETB Ia	15	3.46
Australorp Black Speckled	**ABS**	**29**	**DPB:MEB**	**DPB:MEB**	**DPB IIa**	**20**	**3.34**
Australorp Black	**AoB**	**30**	**DPB:MEB**	**DPB:MEB**	**DPB IIb**	**9**	**3.31**
Pervomai	**Pm**	**31**	**DPB:MEB**	**DPB:MEB**	**DPB IIa**	**20**	**3.27**
Brahma Buff	BB	32	FB	DPB:EMB	MTB	20	3.24
Faverolles Salmon	FS	33	DPB:MEB	DPB:EMB	ETB Ib	20	3.23
Orloff Mille Fleur	OMF	34	GB	DPB:MEB	DPB IIa	20	3.03
Ukrainian Muffed	UM	35	FB	DPB:MEB	DPB IIa	18	2.92
Yurlov Crower	**YC**	**36**	**DPB:MEB**	**DPB:MEB**	**DPB IIa**	**20**	**2.92**
Moscow Game	MG	37	GB	GB	DPB IIb	20	2.73
Uzbek Game	UG	38	GB	GB	MTB	19	2.48
White Cornish × (Brahma Light × Sussex Light)	**WC × (BL × SL)**	**39**	**MTB**	**MTB**	**MTB**	**14**	**1.66**

Table 1. Chicken breeds used in this study respective to breed classification (clustering) models[7] and ranked by *EY*/*W* values. [a]*n*-values are conditional serial numbers assigned to chicken breeds according to the degree of descending *EY*/*W* index. [b]*TCM* traditional classification model, *PCM* phenotypic clustering model, *GCM* genotypic clustering model. Breed types as categorized by Larkina et al.[7]: *ETB* egg, *DPB* dual purpose, *EMB* egg-meat, *MEB* meat-egg, *MTB* meat, *GB* game, *BTB* Bantam, *FB* fancy. Breeds defined as core breeds are shown in bold, and those grouped by the inflection points clustering are highlighted in a different color.

***k*-means clustering.** To analyze and show graphically the distribution of chicken breeds for each of the three models, we employed the *k*-means method implemented as a cluster analysis webtool elsewhere[32]. Herewith, values of the aforementioned *EY*/*W* coefficient were used. Within each model, the 39 breeds were grouped and successively tested in three following ways (Supplementary data S1). Firstly, the original arrangement of breeds characteristic of a model was employed, i.e., using their breakdown by a breed type as defined in Larkina et al.[7]. Secondly, the modified model clustering arrangement was applied based on descending sorting by mean *EY*/*W* values per breed type. Thirdly, the modified clustering was used based on descending sorting by the greatest *EY*/*W* values per breed type. That is, a total of three submodel distribution graphs of 39 breeds were plotted for each model as follows: TCM-1, TCM-2, TCM-3; PCM-1, PCM-2, PCM-3; and GCM-1, GCM-2, GCM-3.

To identify in which of the three models/submodels the distribution of breeds by *EY*/*W* values most adequately conformed to the breakdown of breeds into types, the following *k*-means clustering measures were calculated[32]: *k*, number of optimal required clusters identified by the elbow method and tested within this experiment in the range $1 \le k \le 11$; SSE (within groups), the sum of square (SS) distances from the points to the cluster centers within breed type groups; SSG (between groups), SS from the cluster centers to the average vector; SST (total), SS from the points to the average vector (i.e., SST = SSE + SSG); mean SSE by group; mean *S* (Silhouette[22]) score; and number of outliers. When considering the *S* score range [−1; 1] for a cluster object, its greatest score (i.e., close to 1) conforms to the situation when the object belongs to its specific cluster. A negative *S* value implies that the object is wrongly assigned to this cluster and misclassified. When a cluster contains only one object, Rousseeuw[22] suggested that a value of zero may arbitrarily be assigned, i.e., $S = 0$. However, because the cluster analysis webtool[32] plotted this single object coinciding with its cluster center, we believe that in this case the object completely belongs to its cluster and, therefore, $S = 1$. Consequently, when calculating mean *S* scores for submodels, we used $S = 1$ for single object clusters. This allowed us to overcome a certain bias in submodel *S* scores, if an arbitrary value of zero (as suggested by Rousseeuw[22]) were, otherwise, assigned.

The *k*-means clustering configurations and measures (Supplementary data S1 and S2) were carefully examined to determine for which of the models/submodels the accepted breakdown of breeds into types was most adequately described using the *EY*/*W* coefficient.

Inflection points clustering. We chose the elbow method of clustering (e.g.,[33]) as a basis of further analysis. Despite the elbow method's advantages, one can point out at least one essential drawback: such an elbow cannot always be unambiguously detected when employing this heuristic procedure[23,32].

The elbow method shortcomings can be overcome if (i) a respective clustering index, e.g., an integral coefficient, is chosen and utilized appropriately for producing the breakdown of objects, and (ii) a mathematical algorithm is developed to determine unambiguously the boundaries between clusters. As aforementioned, the ratio of egg production characteristic of layers (i.e., the total mass of eggs produced during a certain laying period) to their weight, *EY*/*W*, served as the integral coefficient. The proposed calculation algorithm resulted in determining the coordinates of the inflection points at which the *EY*/*W*-related function changes the direction of its convexity. To do this, the respective functional dependence had to be approximated by some mathematical function (e.g., a higher-order polynomial), its second derivative being determined and equated to zero (e.g.,[34]). The found values were the desired inflection points of the function, enabling judgement of the boundaries of the corresponding clusters.

SNP genotyping and admixture analysis. As described in detail elsewhere (e.g.,[7]), genome-wide SNP scanning results generated using an Illumina Chicken 60 K SNP iSelect BeadChip (Illumina, San Diego, CA, USA) were processed with the following PLINK 1.9 program[35] filters: –geno 0.2, –hwe 0.0001, and –maf 0.05. Out of 57,636 original SNP markers, 44,200 SNPs remained after the filtering. Linkage disequilibrium (LD) between pairs of SNPs was estimated using the *D*' coefficient proposed by Lewontin[36] and Pearson's r^2 correlation coefficient[37]. Next, we generated admixture models for 11 core breeds using the ADMIXTURE program[28] (with the preset number of 5 iterations) and SNP genotype data for the whole genome and, separately, for the *NCAPG-LCORL* locus. An elbow method-based cross-validation (CV) error plot for determining the number of ancestral populations (K) was produced using Microsoft Excel, with the optimal number being defined using the lowest CV error value from those computed for K = 2 to 6 (using genome-wide SNP dataset) or K = 2 to 6 (at the *NCAPG-LCORL* locus). ADMIXTURE bar plots were visualized in RStudio v. 4.1.0[38] using the POPHELPER library[39]. To provide additional support to the admixture analysis, principal component analysis (PCA) and phylogenetic analysis were performed. PCA analysis was implemented using PLINK 1.9[35], and the results were visualized using the ggplor2 library in R[40]. The phylogenetic tree was built on the basis of pairwise genetic distances (F_{ST}) using the iTOL online service[41].

Ethics approval and consent to participate. All experiments complied with the ARRIVE guidelines and were carried out in accordance with the EU Directive 2010/63/EU for animal experiments. The RRIFAGB—Branch of the L. K. Ernst Federal Research Center for Animal Husbandry provided ethical approval for all research using chickens within the framework of the present study (Protocol No. 2020-4 dated 3 March 2020).

Results

Overall model assessment. For the original 39 breeds, three Larkina et al.[7] models, i.e., TCM, PCM and GCM, were revisited and compared (Table 1). Although there were some discrepancies between the models in classifying (clustering) this breed set, several breeds fall into the same-type or similar classes/clusters in all the three models (as shown in bold in Table 1 and also visualized in the Venn diagram (Fig. 1)).

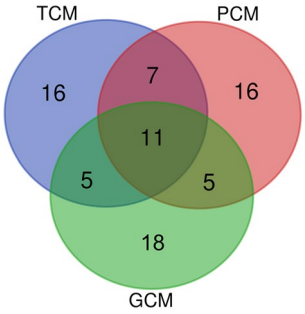

Figure 1. Venn diagram representing distribution of the studied 39 chicken breeds between the three classification (clustering) models: *TCM* traditional classification model, *PCM* phenotypic clustering model, *GCM* genotypic clustering model. Eleven breeds shared between the three models are core breeds.

In other words, particular categories for such breeds were confirmed by all three models, and these were the following 11 chicken populations: Russian White, Leghorn Light Brown (ETB); Pantsirevka Black (EMB); Plymouth Rock Barred, Tsarskoye Selo, Poltava Clay, Australorp Black Speckled, Australorp Black, Pervomai, Yurlov Crower (MEB); and the White Cornish hybrid population (MTB). We conditionally named them *core breeds* (as seen in Fig. 1). For instance, the Russian White breed was included in ETB in each of the three models, the same was true for the Leghorn Light Brown, etc.

k-means clustering. Using the selected appropriate webtool for the *k*-means analysis[32], we were able to analyze mathematically and express graphically the available model/submodel datasets (Supplementary data S1). A summary of full statistics resulting from the *k*-means clustering is presented in Supplementary data S2. Due to similar breed ranking, GCM-2 and GCM-3 statistics turned out to be identical.

Among the nine submodels (Supplementary data S2), numbers (k) of original and optimal required clusters coincided for each of TCM-3, PCM-1 and PCM-3. However, even for them, the plotted cluster configurations differed from the original classification (clustering) models. SSE values ranged between 12.4 (in PCM-3) and 18.4 (in TCM-3). SST values varied from 150.4 (in TCM-2) and 216.7 (in TCM-3). Mean S score was the lowest in GCM-1 (0.37 ± 0.18) and the greatest in TCM-3 (0.54 ± 0.22). Pairwise comparison of mean S scores resulted in significantly greater S values TCM-1 vs GCM-1 ($P < 0.05$), TCM-2 vs GCM-2/GCM-3 ($P < 0.05$), TCM-3 vs GCM-1 and GCM-2/GCM-3 ($P < 0.001$), PCM-1 vs GCM-1 and GCM-2/GCM-3 ($P < 0.01$), PCM-2 vs GCM-1 and GCM-2/GCM-3 ($P < 0.01$), and PCM-3 vs GCM-1 and GCM-2/GCM-3 ($P < 0.01$). Number of cluster outliers per submodel was zero (in PCM-2 and GCM-2/GCM-3) to three (in TCM-1 and GCM-1).

Overall, judging from a total of cluster measures and configurations produced (Supplementary data S1 and S2), GCM submodels seemed to conform to the respective original clustering model to a lesser extent as compared to TCM and PCM submodels. However, none of the submodels looked ideal in this respect.

Additionally, we performed the *k*-means clustering analysis for the 11 core breeds produced (Supplementary data S1 and S2). Their distribution almost ideally conformed to the respective breed types (Fig. 2). Mean S score for the 11 core breeds (0.6291 ± 0.2677) tended to be greater than those for the 39-breed models TCM and PCM and was significantly greater as compared to GCM ($P < 0.001$; Supplementary data S2). Also, there were no outliers in the 11-breed model.

Inflection points model. After calculating EY/W, data for the 39 breeds were ranked from largest to smallest, resulting in chicken breeds arranged in descending order as shown in Table 1 (see also the further details in Supplementary data S3). On the corresponding graph, the pattern of change in EY/W values looked like that shown in Fig. 3A.

The resulting functional dependence was approximated by the following polynomial:

$$\frac{EY}{W} = 11.6480005 - 2.1083905n + 0.23582978n^2 - 0.0124689n^3 + 0.00030623n^4 - 0.00000283n^5 \quad (1)$$

$R = 0.992$.

The degree of correspondence of the approximate dependence to actual data is shown in Fig. 3b.

Next, we defined the first and second derivatives of Eq. (1):

$$\left(\frac{EY}{W}\right)' = -2.1083905 + 0.47165956n - 0.0374067n^2 + 0.00122492n^3 - 0.00001415n^4$$

$$\left(\frac{EY}{W}\right)'' = 0.47165956 - 0.0748134n + 0.00367476n^2 - 0.0000566n^3$$

and equated the second derivative to zero:

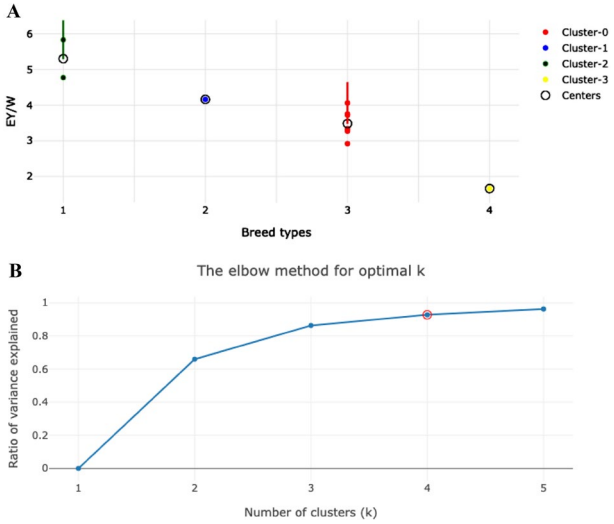

Figure 2. Cluster analysis for distribution of the 11 core breeds using the *k*-means (**A**) and elbow (**B**) methods. Breed (sub)types (**A**): 1, egg type (cluster-2); 2, egg-meat subtype (cluster-1); 3, meat-egg subtype (cluster-0); and 4, meat type (cluster-3). Optimal number of clusters (*k*) was 4 (**B**). *EY*/*W*, integral performance index as a ratio of egg mass yield (*EY*) and female body weight (*W*).

$$0.47165956 - 0.0748134n + 0.00367476n^2 - 0.0000566n^3 = 0$$

$$n^3 - 64.925088n^2 + 1321.79152n - 8333.2078 = 0 \quad (2)$$

By solving the cubic Eq. (2), the following roots, i.e., inflection points, were found: $n_1 = 12.5$, $n_2 = 21.6$, $n_3 = 30.8$. According to the condition of adequate root definition, the third derivative of Eq. (1) should not be equal to zero. That is,

$$\left(\frac{EY}{W}\right)''' = -0.0748134 + 0.00734952n - 0.00016986n^2 \neq 0$$

Thus, the inflection points were correctly defined.

Collectively, we suggest that the analyzed 39 chicken breeds can be conditionally divided into the following four clusters: 1 to 12, ETB; 13 to 21, EMB; 22 to 30, MEB; and 31 to 39, MTB. In Table 1, each cluster is highlighted with a certain color. Since Larkina et al.[7] described the three classification (clustering) models that had their own designations (TCM, PCM, and GCM), we can come up with a name and designation for this model, too, suggesting the inflection points model (IPM).

Subsequently, we tried to focus only on those 11 breeds that were conditionally named core breeds and perform the appropriate IPM clustering for these 11 breeds based on the *EY*/*W* index (Fig. 3C; Supplementary data S4). Accordingly, inflection points were identified at $n_1 = 3.996$, $n_2 = 5.311$, and $n_3 = 8.171$. If considering only 11 core breeds, it seems that a completely clear and plausible dependency graph could be obtained, starting with true ETB (without any "impurities" of non-relevant breeds) on the left side of the graph and ending with one real MTB on the right.

SNP genotyping and admixture analysis. Using genotypes in the 39 breeds for a total of five validated SNPs at the *NCAPG-LCORL* locus, LD analysis between SNP pairs (Supplementary Table S1) showed that some SNPs should have been omitted due to their complete heterozygosity in these breeds. In general, between the five SNP substitutions for all breeds, an average to weak LD level was observed. Full LD ($r^2 = 1$) was found in six breeds such as CB (between GGaluGA265966 and Gga_rs14491028, and Gga_rs15619223 and Gga_rs14491017), Pu (between GGaluGA265969 and Gga_rs14491017), PS (between GGaluGA265966 and Gga_rs14491017), ZS (between rs14491017 and Gga_rs14491028), Ts (between GGaluGA265969 and Gga_rs14491017), and WC × (BL × SL) (between GGaluGA265969 and Gga_rs14491017). Admixture analysis for the 11 core breeds demonstrated the lowest error in the CV procedure at K = 5 (0.12105; Fig. 4A). Furthermore, a certain pattern of genetic differentiation was produced and visualized with the respective ADMIXTURE bar plots (Fig. 4B).

While K = 5 conformed to the most optimal and probable number of clusters (ancestral populations) (Fig. 4B), each breed type already had its own specific genetic structure at K = 3, although showing multiple instances of admixture and introgression from other breeds, except for LLB, a typical ETB, and the three-way crossbred population belonging to MTB. Particularly when K = 3, two ETB were predominantly characterized with a common ancestry of green color, one MTB had mostly a red-colored ancestry, and eight DPB (including both EMB and MEB) had a mixed ancestry of several colors. A similar pattern of clustering was observed at K = 4 and 5.

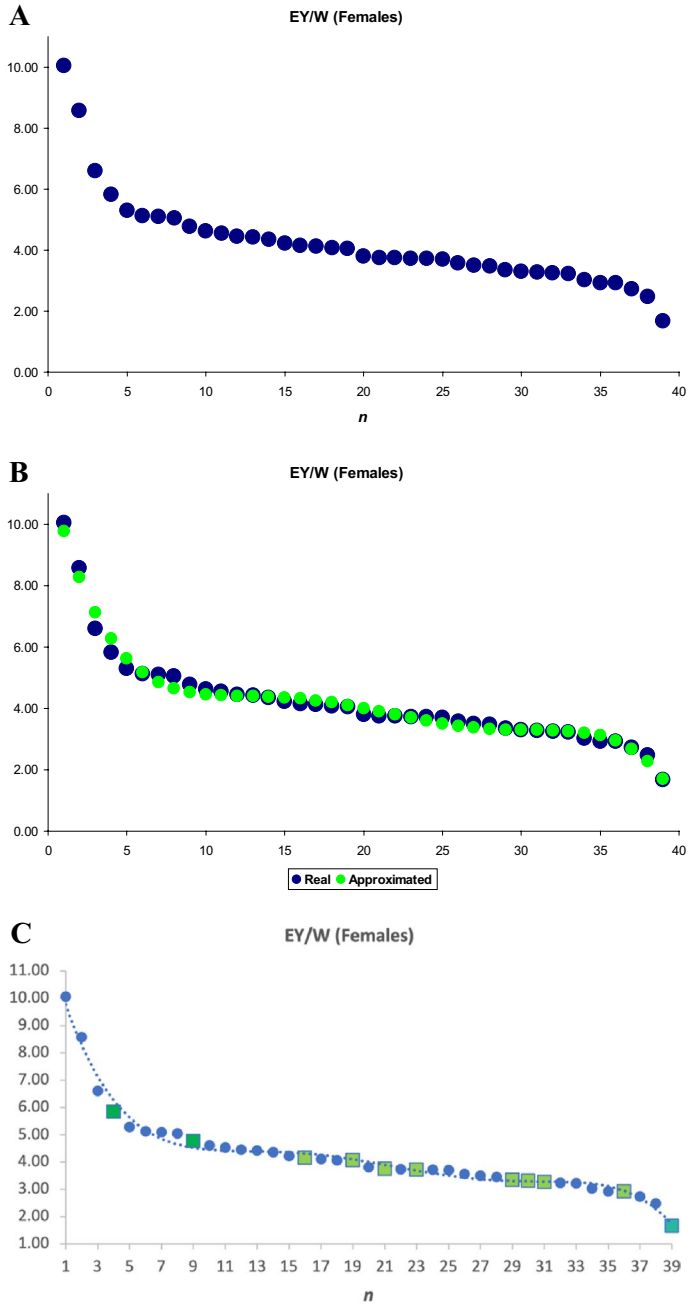

Figure 3. Graph of change in mean *EY*/*W* values in females of the 39 chicken breeds studied. (**A**) original dataset; (**B**) correspondence of the approximated dependence (green trendline) to actual data (blue curve); and **C** original dataset including the respective trendline (blue dotted lines) and 11 core breeds (filled green square). *EY*/*W*, integral performance index as a ratio of egg mass yield (*EY*) and female body weight (*W*); *n* (1 to 39; see Table 1), conditional serial numbers assigned to breeds of chickens according to the degree of descending *EY*/*W* index.

Whole-genome SNP genotypes resulted in even clearer patterns of population structure for each core breed (Fig. 5B). The most optimal number of ancestral populations was achieved at K = 9 (Fig. 5B).

Additionally, we conducted PCA analysis for the 11 core breeds using the same whole-genome genotype dataset (Fig. 6). Two ETB were remotely located from the rest breeds, with LLB being on the right side and RWG at the bottom of the PCA plot. Although all other breeds were located rather crowded on the plot, they were still quite separated from each other, especially WC × (BL × SL) of meat type and two MEB, YC and PRB. The PCA plot for the 39 breeds studied (see Supplementary Fig. S1) had five clusters: three for BTB (HSSD, SW–CB, BMF), one for two related breeds (BL–BB), and a conglomerate of other indistinctly separated breeds. In the phylogenetic tree (Fig. 7), there was an ETB cluster of LLB–RWG. The MTB WC × (BL × SL) formed a separate cluster joining further with the MEB PRB that is a maternal stock of another broiler cross. Other DPB occupied

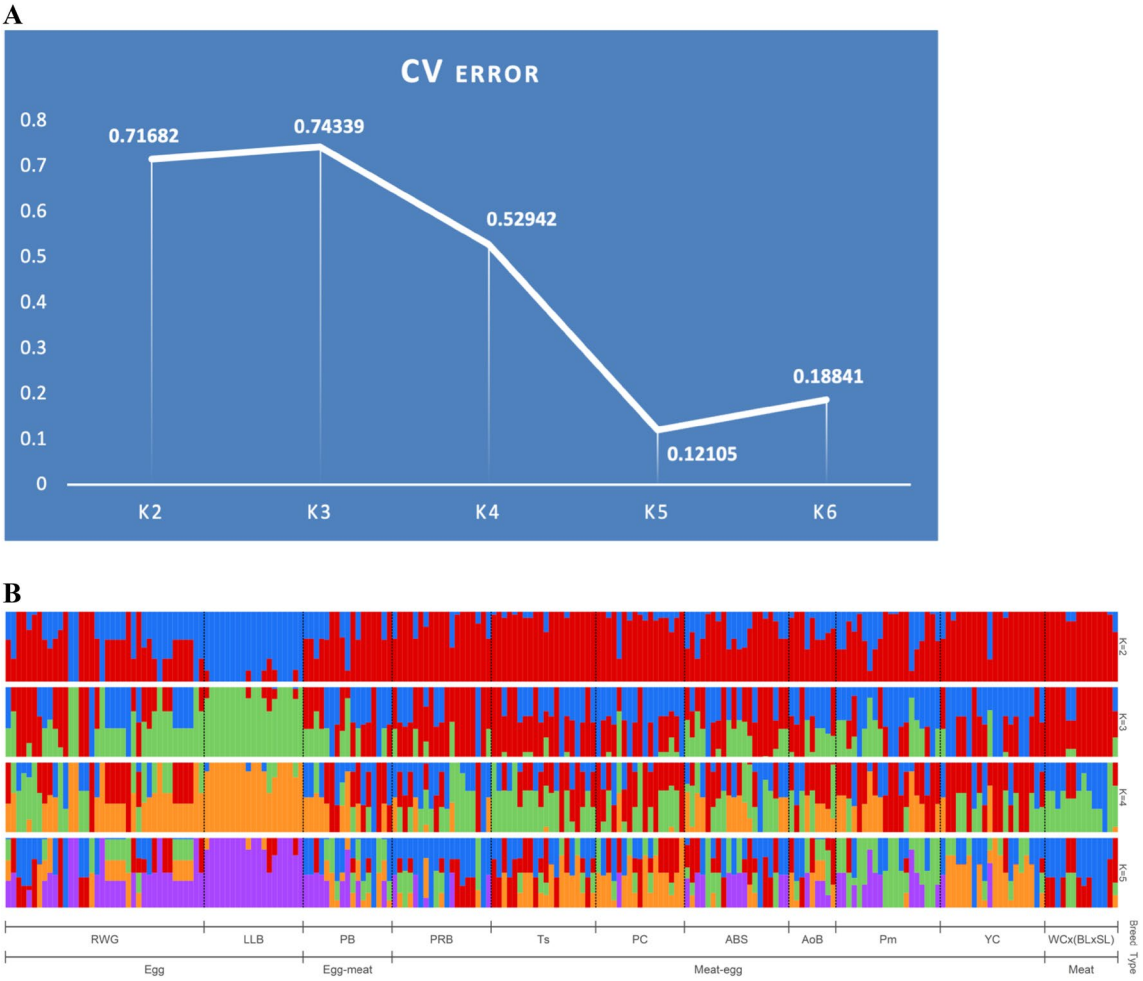

Figure 4. Population structure based on the genetic variation in the 11 core breeds genotyped for five SNP markers at the *NCAPG-LCORL* locus. (**A**) Elbow method-based analysis of cross-validation (CV) error values depending on number of ancestral populations (K). (**B**) Admixture bar plots generated by Bayesian clustering using the ADMIXTURE program. Each admixture plot represents a cluster structure of the studied breeds/breed types depending on number of ancestral populations (K), with the latter being optimal at K = 5. Core breeds: *RWG* Russian White, *LLB* Leghorn Light Brown, *PB* Pantsirevka Black, *PRB* Plymouth Rock Barred, *Ts* Tsarskoye Selo, *PC* Poltava Clay, *ABS* Australorp Black Speckled, *AoB* Australorp Black, *Pm* Pervomai, *YC* Yurlov Crower, *WC* × *(BL* × *SL)* White Cornish × (Brahma Light × Sussex Light).

own clusters and single branches. Overall, the PCA and phylogenetic analyses provided an additional support to, and a proper comparison with, the admixture analysis outcome of the various breeds examined.

Discussion

Assessing the diversity and genetic admixture of chicken breeds from local and world gene pools is an initial and important step to further the process of poultry breeding. It includes a powerful contemporary technique of genomic selection and the emerging field of phenome-wide association/mediation studies[42,43]. In the present study, we explored a number of clustering techniques to establish if the resulting breed groupings make biological sense. In particular, the data analysis, by implementing the integral performance coefficient *EY*/*W* (Fig. 3), allows us to evaluate the genetic potential of poultry performance.

Our *k*-means-based study (Supplementary data S1, Fig. 2) is consistent with the established notion that when there has been any prior group classification of cases, nonhierarchical clustering by the *k*-means method is a useful multivariate exploratory approach (e.g.,[25]). Its strength is that this technique focuses on categorizing groups in order to reduce variance within, and increase variance between, groups[44].

While considering the previous three models, TCM, PCM and GCM[7], we noticed that many breeds showed a different type (category) of clustering in different models. For example, a breed could conditionally be attributed to ETB using the first model, MEB according to the second and MTB according to the third. This may create some noise and bias in the analysis of 39 breeds, including their examination and clustering using the *k*-means (Supplementary data S1, Fig. 2) and new IPM approach (Fig. 3, Supplementary data S3 and S4).

We introduced a novel index such as the ratios of total egg mass yield to body weight of females (*EY*/*W*) and ranked them by descending order (Fig. 3). Logically, the higher the index, the higher egg-type properties

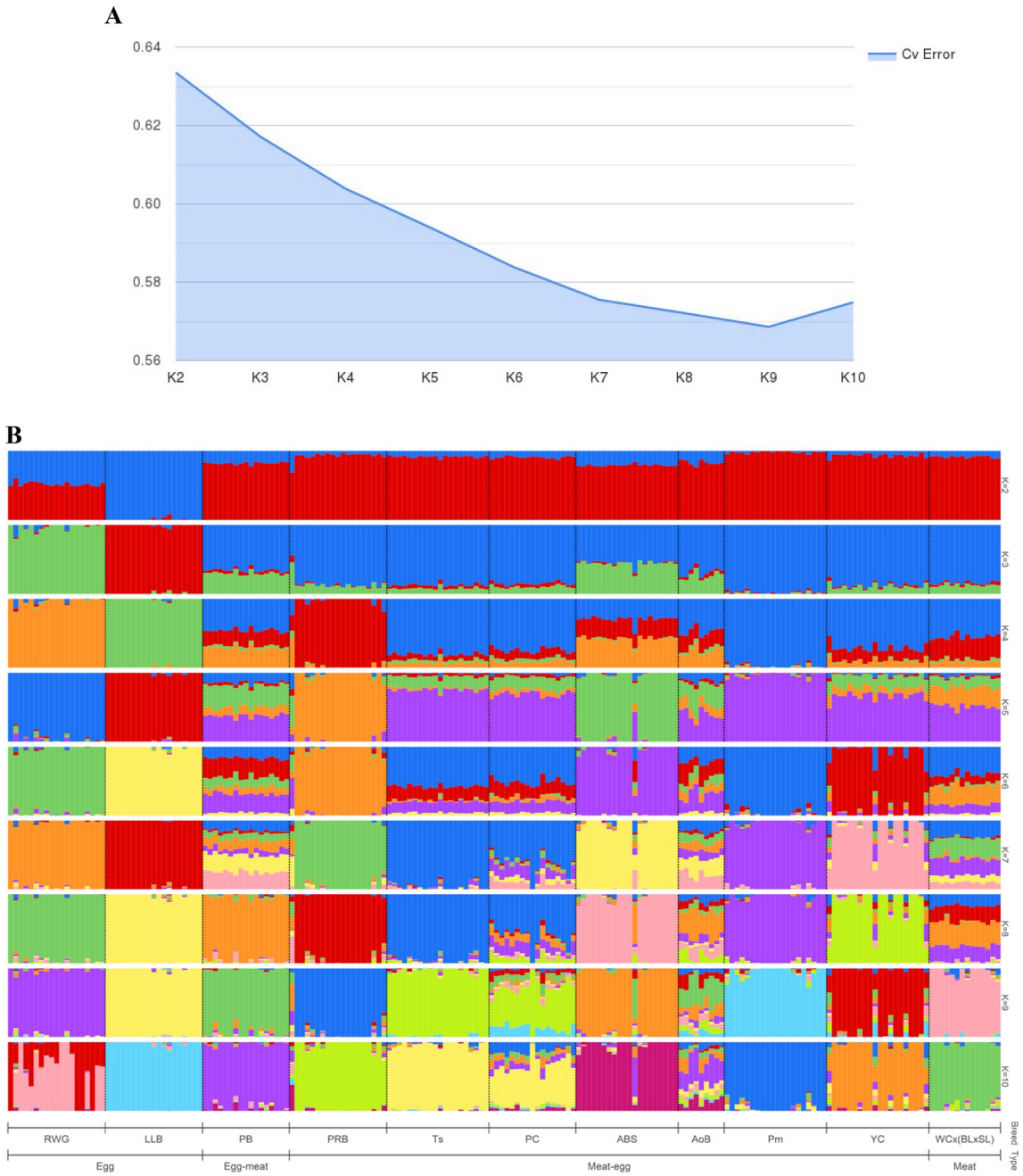

Figure 5. Population structure based on genome-wide genotypes in the 11 core breeds. (**A**) Elbow method-based analysis of cross-validation (CV) error values depending on number of ancestral populations (K). (**B**) Admixture bar plots generated by Bayesian clustering using the ADMIXTURE program. Each admixture plot represents a cluster structure of the studied breeds/breed types depending on number of ancestral populations (K), with the latter being optimal at K = 9. Core breeds: *RWG* Russian White, *LLB* Leghorn Light Brown, *PB* Pantsirevka Black, *PRB* Plymouth Rock Barred, *Ts* Tsarskoye Selo, *PC* Poltava Clay, *ABS* Australorp Black Speckled, *AoB* Australorp Black, *Pm* Pervomai, *YC* Yurlov Crower, *WC* × *(BL* × *SL)* White Cornish × (Brahma Light × Sussex Light).

of the layer/breed will be. There may be a rational kernel in this model; however, it is important to evaluate the adequacy of getting breeds into certain clusters using IPM. Herewith, one should bear in mind the following considerations. In particular, dwarf breeds (Bantams) have never been regarded as ETB. Apparently, DPB (both subtypes) are also not considered purely egg breeds. Therefore, in the first cluster (ETB, 1 to 12), only two breeds, the Russian White and the Leghorn Light Brown, are generally recognized as true egg breeds. The second, third and fourth clusters are also significantly mixed due to the presence of very different breeds in terms of performance (purpose of use) and genetic admixture. Thus, it can be assumed that the proposed model may not be much better than the four models described in Larkina et al.[7]. This, however, is not a flaw of new or old models; the examined breed composition itself is simply heterogenous from a genetic standpoint (see Fig. 4B), being

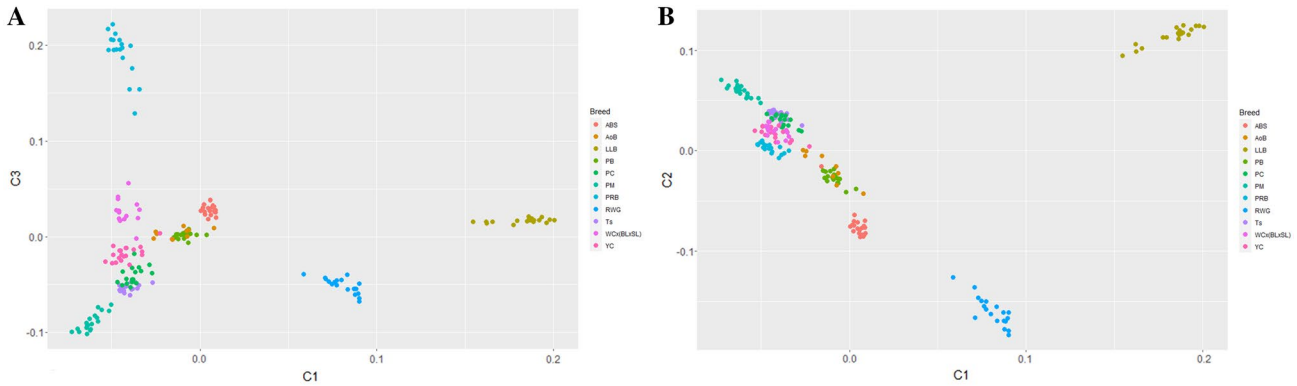

Figure 6. Results of PCA analysis of the 11 core chicken breeds. (**A**) First (C1) and third (C3) components. (**B**) First (C1) and second (C2) components.

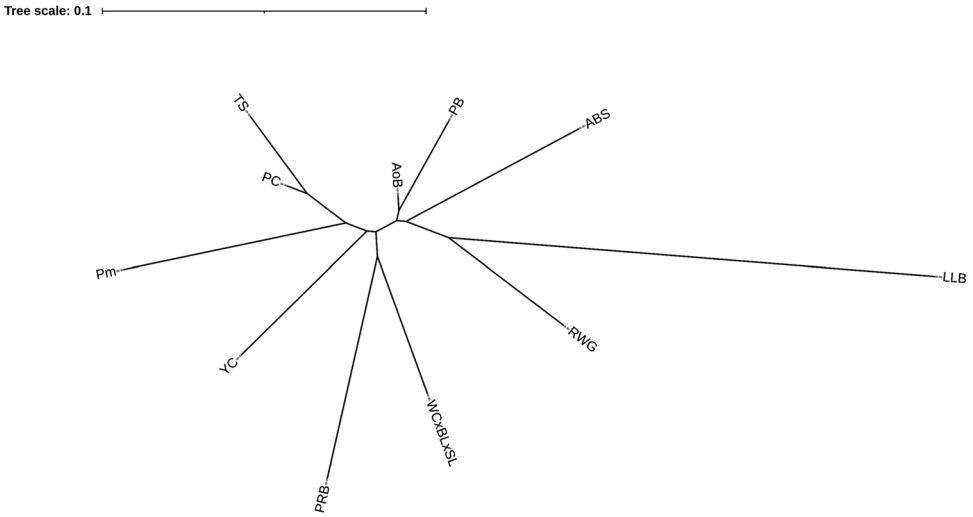

Figure 7. Phylogenetic tree based on pairwise F_{ST} genetic distances and built using the Neighbor-Net method.

often synthetic (composite) by origin and genetic structure. Nonetheless, we posit that the data obtained can be used, for instance, in the search for new mathematical models that allow *for* looking at the chicken breed gene pool from other and very interesting mathematical points of view. Perhaps, as a matter of discussion, it is worth assuming that the proposed model, in principle, is able to suggest (to a varying degree of certainty) the following four groups (Fig. 3): BTB, 1 to 3; a large group of ETB, DPB, FB and BTB (not included in the first group), 4 to 36; GB, 37 to 38; and MTB, 39. Therefore, these results also seem to us to be quite interesting for bearing in mind at developing mathematical models further for the clustering of chicken breeds.

Perhaps, the insufficient resolving power of IPM is explained by the fact that this model assumes the subdivision of breeds according to one specific index, albeit an integral one, i.e., the specific egg productivity relative to the body weight of laying hens. Most probably, BTB have large values of this proposed *EY*/*W* index mainly not due to its numerator, i.e., a supposedly high (or even greater than in true layers) level of egg performance, but due to its denominator, i.e., a significantly lower body weight (after all, they are dwarf, miniature chickens). It is difficult to imagine, of course, that BTB would be the preferred breeds for industrial scale egg production. This would makes no economic sense, and the commercial companies have not switched to dwarf layers. On the other hand, if BTB females, according to this index, lay eggs at the same (or even higher) level as classic ETB breeds, it might make sense to look at them in terms of including them in breeding programs aimed at developing breeds that produce more eggs.

In addition, one could think of some other integral indicator, for example, taking into account any external or other characteristic of chickens. For instance, Vakhrameev and Makarova[45] listed different integral indices that describe exterior features, and one could take a closer look at these indices. In this case, clustering patterns may arise that do not follow generally accepted models. Rather, they are determined according to specific economically important traits and, consequently, to capabilities of a given breed to realize and improve its own egg performance-relevant genetic potential as a result of artificial selection.

It should also be noted whether dwarf breeds lay eggs of proper quality and nutritional value. When compared by egg weight, the differences between ETB and related breeds are fairly small. In general, among all breed groups (clusters) there is a certain uniformity in this trait. It can be assumed that this breakdown of breeds

has occurred not in terms of how breeds of one or another selected performance trait should look like, but in terms of their degree of biological predisposition to egg production. Furthermore, this breed breakdown was obtained strictly in accordance with mathematical rules, which adds a certain attractiveness to it. In principle, in the first approximation, the breeds were, indeed, sorted according to the degree of their egg-type properties (left side of the graph in Fig. 3A) or meat-type properties (graph's right side). Proceeding from the 39 chicken breeds (Fig. 3A) to 11 core breeds (Fig. 3C) resulted in more plausible clustering configuration pursuant to the respective *EY*/*W* function curve and inflection points. However, it is worth noting that a certain problem may arise in this case, which lies in the fact that by cutting the number of points, we can thereby smooth the curve (Supplementary data S4). Therefore, the inflection points do not become obvious, and it is not always possible to determine them. The proposed new method may seem rather controversial, so when aiming at developing a new, more suitable technique, one should plan to verify it in further studies using additional data. In any case, such search for an integral assessment of phenotypic traits can make an important contribution to developing genome- and phenome-related studies[43] and strategies of germ plasm preservation[46,47].

Finally, the admixture analysis results obtained are of special interest, since they were inferred from whole-genome SNP genotypes (Fig. 5) and those at the well-known *NCAPG-LCORL* genomic locus (Fig. 4) associated with productive traits in chickens. At K = 2, population structure of the 11 core breeds conformed to two basic ancestries as postulated by Moiseyeva et al.[6], i.e., ETB (blue-colored in Figs. 4b and 5b) and MTB (red-colored). More mixed ancestries were revealed at K = 3, 4, 5 and so forth, although specific admixture patterns could generally be tracked for an individual breed and each of ETB, DPB and MTB groups. These admixture patterns for the 11 core breeds appeared to be more biologically meaningful than those previously described for the 39 breeds[7] and did not contradict clustering configurations that resulted from using other methods tested here.

Conclusions

The present study examined the importance of different methods for untangling complex clustering configurations among divergently selected chicken breeds representing a wide sampling of the world gene pool. We have demonstrated that different breeds can be classified (clustered) in one way or another depending on the chosen methods and (sub)models as well as on the degree of their genetic admixture. To this end, we have proposed a new integral indicator (i.e., *EY*/*W*), which links the main phenotypic traits in chickens, i.e., egg performance and body weight, as well as a mathematical model based on sorting breeds by inflection points. Future studies will use these findings to improve chicken breed clustering techniques as well as in genome- and phenome-wide association/mediation analyzes (e.g.,[1,42,43]) to elucidate cause-and-effect relationships between economically important characteristics, phenotypes, and SNP genotypes, including those at key associated loci, e.g., *NCAPG-LCORL* as we explored here.

Data availability

All data generated or analyzed during the present study are available from the corresponding author on reasonable request. The datasets supporting the conclusions of this article are included in the main manuscript and supplemental materials.

Received: 1 September 2022; Accepted: 23 January 2023
Published online: 27 February 2023

References

1. Momen, M. *et al.* Including phenotypic causal networks in genome-wide association studies using mixed effects structural equation models. *Front. Genet.* **9**, 455. https://doi.org/10.3389/fgene.2018.00455 (2018).
2. Silva, F. F., Morota, G. & Rosa, G. J. M. Editorial: High-throughput phenotyping in the genomic improvement of livestock. *Front. Genet.* **12**, 707343. https://doi.org/10.3389/fgene.2021.707343 (2021).
3. Bondarenko, Yu. V., Rozhkovsky, A. V., Romanov, M. N. & Bogatyr, V. P. [The use of genetical systems in the development of autosex crosses of egg-laying chickens]. *Ptitsevodstvo (Kiev)* **42**, 11–14 (1989).
4. Romanov, M. N. Using phenetic approaches for studying poultry populations under preservation and breeding. In *Proceedings of the 5th World Congress on Genetics Applied to Livestock Production. Gene Mapping, Polymorphisms, Disease Genetic Markers, Marker Assisted Selection, Gene Expression, Transgenes, Non-Conventional Animal Products, Conservation Genetics, Conservation of Domestic Animal Genetic Resources*, Vol. 21. 556–559 (1994).
5. Khvostyk, V., Tereshchenko, O., Zakharchenko, O. & Bondarenko, Yu. [Influence of "adding blood" of cocks of foreign crosses upon economically beneficial attributes of meat-egg hens of domestic selection]. *Visn. Agrar. Nauki [Bull. Agric. Sci.]* **95**(9), 44–48. https://doi.org/10.31073/agrovisnyk201709-08 (2017).
6. Moiseyeva, I. G., Romanov, M. N., Nikiforov, A. A., Sevastyanova, A. A. & Semyenova, S. K. Evolutionary relationships of Red Jungle Fowl and chicken breeds. *Genet. Sel. Evol.* **35**(4), 403–423. https://doi.org/10.1186/1297-9686-35-5-403 (2003).
7. Larkina, T. A. *et al.* Evolutionary subdivision of domestic chickens: Implications for local breeds as assessed by phenotype and genotype in comparison to commercial and fancy breeds. *Agriculture* **11**(10), 914. https://doi.org/10.3390/agriculture11100914 (2021).
8. Romanov, M. N. *et al.* [Comparative analysis of phenotypic traits in various breeds representing the world poultry gene pool]. In [*Materials of the 3rd International Scientific and Practical Conference on Molecular Genetic Technologies for Analysis of Gene Expression Related to Animal Productivity and Disease Resistance*], 52–63 (Sel'skokhozyaistvennye tekhnologii, 2021). https://doi.org/10.18720/SPBPU/2/z21-43.
9. Moreira, G. C. M. *et al.* Genome-wide association scan for QTL and their positional candidate genes associated with internal organ traits in chickens. *BMC Genomics* **20**(1), 669. https://doi.org/10.1186/s12864-019-6040-3 (2019).
10. Kudinov, A. A. *et al.* Genome-wide association studies targeting the yield of extraembryonic fluid and production traits in Russian White chickens. *BMC Genomics* **20**(1), 270. https://doi.org/10.1186/s12864-019-5605-5 (2019).
11. Liu, J., Zhou, J., Li, J. & Bao, H. Identification of candidate genes associated with slaughter traits in F2 chicken population using genome-wide association study. *Anim. Genet.* **52**(4), 532–535. https://doi.org/10.1111/age.13079 (2021).
12. Wang, G. *et al.* Genome-wide analysis reveals selection signatures for body size and drought adaptation in Liangzhou donkey. *Genomics* **114**(6), 110476. https://doi.org/10.1016/j.ygeno.2022.110476 (2022).

13. Raza, S. H. A. *et al.* Genome-wide association studies reveal novel loci associated with carcass and body measures in beef cattle. *Arch. Biochem. Biophys.* **694**, 108543. https://doi.org/10.1016/j.abb.2020.108543 (2020).
14. He, S. *et al.* Genome-wide scan for runs of homozygosity identifies candidate genes related to economically important traits in Chinese Merino. *Animals* **10**(3), 524. https://doi.org/10.3390/ani10030524 (2020).
15. Posbergh, C. J. & Huson, H. J. All sheeps and sizes: A genetic investigation of mature body size across sheep breeds reveals a polygenic nature. *Anim. Genet.* **52**(1), 99–107. https://doi.org/10.1111/age.13016 (2021).
16. Yi, G. *et al.* Genome-wide association study dissects genetic architecture underlying longitudinal egg weights in chickens. *BMC Genomics* **16**, 746. https://doi.org/10.1186/s12864-015-1945-y (2015).
17. Shen, M. *et al.* A genome-wide study to identify genes responsible for oviduct development in chickens. *PLoS ONE* **12**(12), e0189955. https://doi.org/10.1371/journal.pone.0189955 (2017).
18. Dou, T. *et al.* Genetic architecture and candidate genes detected for chicken internal organ weight with a 600 K single nucleotide polymorphism array. *Asian-Australas. J. Anim. Sci.* **32**(3), 341–349. https://doi.org/10.5713/ajas.18.0274 (2019).
19. Forgy, E. W. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics* **21**(3), 768–769 (1965).
20. MacQueen, J. B. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* Vol. 1, 281–297 (University of California Press, 1967).
21. Lloyd, S. P. Least squares quantization in PCM. *IEEE Trans. Inf. Theory.* **28**(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489 (1982).
22. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
23. Ketchen, D. J. & Shook, C. L. The application of cluster analysis in strategic management research: An analysis and critique. *Strat. Manage. J.* **17**(6), 441–458. https://doi.org/10.1002/(SICI)1097-0266(199606)17:6%3c441::AID-SMJ819%3e3.0.CO;2-G (1996).
24. Rencher, A. C. *Methods of Multivariate Analysis* 2nd edn. (Wiley, 2002). https://doi.org/10.1002/0471271357.
25. Savegnago, R. P. *et al.* Estimates of genetic parameters, and cluster and principal components analyses of breeding values related to egg production traits in a White Leghorn population. *Poult. Sci.* **90**(10), 2174–2188. https://doi.org/10.3382/ps.2011-01474 (2011).
26. Nyambo, D. G., Luhanga, E. T., Yonah, Z. O. & Mujibi, F. D. Application of multiple unsupervised models to validate clusters robustness in characterizing smallholder dairy farmers. *Sci. World J.* **2019**, 1020521. https://doi.org/10.1155/2019/1020521 (2019).
27. Raccagni, W. & Ntalampiras, S. Acoustic classification of cat breed based on time and frequency domain features. in *2021 30th Conference of Open Innovations Association FRUCT*, 184–189 (IEEE, 2021).
28. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**(9), 1655–1664. https://doi.org/10.1101/gr.094052 (2009).
29. Abdelmanova, A. S. *et al.* Unveiling comparative genomic trajectories of selection and key candidate genes in egg-type Russian White and meat-type White Cornish chickens. *Biology* **10**(9), 876. https://doi.org/10.3390/biology10090876 (2021).
30. Weigend, S., Romanov, M. N. & Rath, D. Methodologies to identify, evaluate and conserve poultry genetic resources. in *XXII World's Poultry Congress & Exhibition: Participant List & Full Text CD + Book of Abstracts, Istanbul, Turkey, 8–13 June 2004*, 84 (WPSA Turkish Branch, 2004).
31. Van de Peer Lab. *Draw Venn Diagram. Bioinformatics Evolutionary Genomics* (Ghent University, 2017). http://bioinformatics.psb.ugent.be/webtools/Venn/.
32. Statistics Kingdom. *Cluster Analysis: K-Means Clustering.* (2017). https://www.statskingdom.com/cluster-analysis.html.
33. Zhao, Q., Hautamaki, V. & Fränti, P. Knee point detection in BIC for detecting the number of clusters. In *Lecture Notes in Computer Science* Vol. 5259 (eds Blanc-Talon, J. *et al.*) 664–673 (Springer, 2008). https://doi.org/10.1007/978-3-540-88458-3_60.
34. Math24. *Inflection points. Mathematical analysis. Higher Mathematics* (2016). https://web.archive.org/web/20160428133446/http://math24.ru/точки-перегиба.html.
35. Chang, C. C. *et al.* Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7. https://doi.org/10.1186/s13742-015-0047-8 (2015).
36. Wellek, S. & Ziegler, A. A genotype-based approach to assessing the association between single nucleotide polymorphisms. *Hum. Hered.* **67**(2), 128–139. https://doi.org/10.1159/000179560 (2009).
37. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**(2), 263–265. https://doi.org/10.1093/bioinformatics/bth457 (2005).
38. RStudio Team. *RStudio: Integrated Development for R. Version 4.1.0. RStudio* (PBC, 2021).
39. Francis, R. M. POPHELPER: An R package and web app to analyse and visualize population structure. *Mol. Ecol. Resour.* **17**(1), 27–32. https://doi.org/10.1111/1755-0998.12509 (2017).
40. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2009).
41. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res.* **47**(W1), W256–W259. https://doi.org/10.1093/nar/gkz239 (2019).
42. Jiang, F. Y. *et al.* Research progress in the methodology used in phenome-wide association studies. *Zhonghua Liu Xing Bing Xue Za Zhi* **43**(7), 1154–1161. https://doi.org/10.3760/cma.j.cn112338-20211104-00853 (2022).
43. AG2PI (Agricultural Genome to Phenome Initiative). *Intermediate Omics Data-Enabled Genomic Prediction and Mediation Analysis.* AG2PI Workshop #14 (2022). https://www.ag2pi.org/workshops-and-activities/workshop-2022-07-26/.
44. Hair, J. F. Jr., Black, W. C., Babin, B. J. & Anderson, R. E. *Multivariate Data Analysis: A Global Perspective* 7th edn. (Pearson Education Prentice Hall, 2010).
45. Vakhrameev, A. B. & Makarova, A. V. [*Exterior Assessment of Chickens: Monograph*]. *Electronic Resource (CD-R).* (Russian Research Institute of Farm Animal Genetics and Breeding Branch of the L. K. Ernst Federal Research Center for Animal Husbandry; Publishing House of FSBSI FRC VIZh named after L. K. Ernst, 2021).
46. Tagirov, M. T., Tereshchenko, L. V. & Tereshchenko, A. V. [Substantiation of the possibility of using primary germ cells as material for the preservation of poultry genetic resources]. *Ptakhivnytstvo* **58**, 464–473 (2006).
47. Tereshchenko, O. V., Katerinich, O. O., Pankova, S. M. & Borodai, V. P. [Formation of genetic resources of domestic breeds of poultry in the context of food security of the state]. *Sučasne Ptahivnictvo* **7**(8), 19–21 (2015).

Acknowledgements

We thank the USDA Chicken GWMAS Consortium, Cobb-Vantress Inc., and Hendrix Genetics B.V. for access to the developed 60K chicken SNP chip produced by Illumina Inc. for the GWMAS Consortium.

Author contributions

A.B.V.: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Validation, Visualization, Writing—original draft preparation. V.G.N.: Conceptualization, Formal analysis, Methodology, Software, Visualization, Writing—original draft preparation, Writing—review and editing. T.A.L.: Data curation, Investigation, Project administration, Supervision, Validation. O.Y.B.: Investigation. G.K.P.: Data curation,

Investigation, Resources. A.P.D.: Writing—original draft preparation. N.V.D.: Data curation, Funding acquisition. A.V.M.: Investigation, Resources. Y.S.S.: Formal analysis, Methodology, Software, Visualization. M.V.P.: Data curation. Y.V.B.: Writing—original draft preparation. D.K.G.: Writing—review and editing. M.N.R.: Conceptualization, Formal analysis, Methodology, Software, Visualization, Writing—original draft preparation, Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the Ministry of Science and Higher Education of the Russian Federation (State Assignment Program No. 0445-2021-0010). Generation of admixture models for distinguishing clusters based on SNP genotype data were produced with financial support of the Ministry of Science and Higher Education of the Russian Federation, Grant No. 075-15-2021-1037 (Internal No. 15.BRK.21.0001).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-28651-8.

Correspondence and requests for materials should be addressed to N.V.D., D.K.G. or M.N.R.