

Single nucleotide polymorphism marker combinations for classifying Yeonsan Ogye chicken using a machine learning approach

Eunjin Cho¹, Sunghyun Cho², Minjun Kim³, Thisarani Kalhari Ediriweera¹, Dongwon Seo^{1,4}, Seung-Sook Lee⁵, Jihye Cha⁶, Daehyeok Jin⁷, Young-Kuk Kim¹ and Jun Heon Lee^{1,3*}

¹Department of Bio-AI Convergence, Chungnam National University, Daejeon 34134, Korea

²Research and Development Center, Insilicogen Inc., Yongin 19654, Korea

³Division of Animal and Dairy Science, Chungnam National University, Daejeon 34134, Korea

⁴Research Institute TNT Research Company, Jeonju 54810, Korea

⁵Yeonsan Ogye Foundation, Nonsan 32910, Korea

⁶Animal Genome & Bioinformatics, National Institute of Animal Science, Rural Development Administration, Wanju 55365, Korea

⁷Animal Genetic Resources Research Center, National Institute of Animal Science, Rural Development Administration, Hamyang 50000, Korea



Received: Apr 13, 2022

Revised: Jul 15, 2022

Accepted: Aug 1, 2022

*Corresponding author

Jun Heon Lee

Department of Bio-AI Convergence,
Chungnam National University,
Daejeon 34134, Korea.

Tel: +82-42-821-5779

E-mail: junheon@cnu.ac.kr

Copyright © 2022 Korean Society of
Animal Sciences and Technology.

This is an Open Access article
distributed under the terms of the
Creative Commons Attribution
Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted
non-commercial use, distribution, and
reproduction in any medium, provided
the original work is properly cited.

ORCID

Eunjin Cho

<https://orcid.org/0000-0003-4800-1603>

Sunghyun Cho

<https://orcid.org/0000-0002-2657-9159>

Minjun Kim

<https://orcid.org/0000-0002-8173-8431>

Thisarani Kalhari Ediriweera

<https://orcid.org/0000-0001-9649-6925>

Abstract

Genetic analysis has great potential as a tool to differentiate between different species and breeds of livestock. In this study, the optimal combinations of single nucleotide polymorphism (SNP) markers for discriminating the Yeonsan Ogye chicken (*Gallus gallus domesticus*) breed were identified using high-density 600K SNP array data. In 3,904 individuals from 198 chicken breeds, SNP markers specific to the target population were discovered through a case-control genome-wide association study (GWAS) and filtered out based on the linkage disequilibrium blocks. Significant SNP markers were selected by feature selection applying two machine learning algorithms: Random Forest (RF) and AdaBoost (AB). Using a machine learning approach, the 38 (RF) and 43 (AB) optimal SNP marker combinations for the Yeonsan Ogye chicken population demonstrated 100% accuracy. Hence, the GWAS and machine learning models used in this study can be efficiently utilized to identify the optimal combination of markers for discriminating target populations using multiple SNP markers.

Keywords: Breed identification, Yeonsan Ogye, Single nucleotide polymorphism, Machine learning

INTRODUCTION

Economically, it is important that different breeds of livestock can be easily identified. Consumers often encounter processed products, including meats, at markets and it is necessary to identify the origin, breed, and species of the animals used in products. Several Korean studies have described tools

Dongwon Seo
<https://orcid.org/0000-0003-0548-7068>
 Seung-Sook Lee
<https://orcid.org/0000-0002-1898-5448>
 Jihye Cha
<https://orcid.org/0000-0002-9705-2979>
 Daehyeok Jin
<https://orcid.org/0000-0001-5091-4271>
 Young-Kuk Kim
<https://orcid.org/0000-0002-8329-9348>
 Jun Heon Lee
<https://orcid.org/0000-0003-3996-9209>

Competing interests

No potential conflict of interest relevant to this article was reported.

Funding sources

This study was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-01441, Artificial Intelligence Convergence Research Center [Chungnam National University]) and the “Cooperative Research Program for Agriculture Science and Technology Development (No. PJ015785)” of the Rural Development Administration, Korea.

Acknowledgements

The SYNBREED data were kindly provided by Dr. Steffen Weigend from the Institute of Farm Animal Genetics (ING), Friedrich-Loeffler-Institut, Neustadt, Germany, and the Korean native chicken data were provided from the Animal Genetic Resources Research Center, National Institute of Animal Science, and Yeonsan Ogye Foundation, Korea.

Availability of data and material

Upon reasonable request, the datasets of this study can be available from the corresponding author.

Authors' contributions

Conceptualization: Cho S, Seo D, Lee SS, Lee JH.

Data curation: Cho E, Cho S, Seo D, Lee SS, Cha J, Jin D.

Formal analysis: Cho E, Cho S.

Methodology: Cho E, Cho S, Kim M, Ediriweera TK, Seo D.

Software: Cho E, Cho S, Seo D.

Validation: Cho E, Cho S, Kim M, Seo D, Kim YK, Lee JH.

Investigation: Cho E, Cho S, Seo D.

Writing - original draft: Cho E.

Writing - review & editing: Cho E, Cho S, Kim M, Ediriweera TK, Seo D, Lee SS, Cha J, Jin D, Kim YK, Lee JH.

Ethics approval and consent to participate

This research has been approved by the Institutional Animal Care and Use Committee (IACUC) of Chungnam National University (202103A-CNU-061).

for determining the breed of Korean native chicken (KNC; *Gallus gallus domesticus*) used in various products [1,2]. However, the current traceability system in Korea only considers chicken meat and egg quality. The ability to discriminate between different chicken breeds using a genetic approach could improve consumer confidence while also safeguarding unique genetic resources.

Yeonsan Ogye, one of the KNC breed, is characterized by black feathers, skin, and bones, and considered an important element of Korean heritage. Globally, only a few chicken breeds display similar black plumage to Yeonsan Ogye, including Ayam Cemani from Indonesia, H'mong from Vietnam, and Svarthöna from Sweden [3,4]. In general, the techniques used for identifying specific chicken breeds are based on morphological characteristics, but it is sometimes challenging to morphologically distinguish breeds with similar phenotypes.

Genetic information could be applied for precise breed identification. Various genetic markers have been developed and used to obtain genetic information. Typically, microsatellite (MS) markers are utilized for the identification of various livestock breeds [5–7]. However, as MS markers have unique characteristics, they are not always reflective of the entire genome, and some also have high mutation rates [8]. In addition, research using MS markers requires significant human input, and the interpretation of the results is subjective.

Single nucleotide polymorphism (SNP) markers could overcome the limitations of MS markers [9]. Recently, genotyping methods using SNP arrays have been developed over several generations, and the cost of genotyping continues to fall. Hence, a large amount of SNP data is available for application as genotype biomarkers and can rapidly provide accurate information for breed identification. However, identifying optimal SNP markers for specific populations using high-density SNP chips is still quite complex.

Machine learning using classification models is possible to deal with the large genotype data effectively. The classification model is a process of distinguishing the class of new input data based on learned data with labels through various algorithms. In particular, the Random Forest (RF) and AdaBoost (AB) algorithms are effectively used to reduce overfitting, handle large data, and select the important variables.

The objective of this study was to determine optimal SNP marker combinations to discriminate a target chicken population (Yeonsan Ogye) from other breeds using two machine learning algorithms (RF and AB).

MATERIALS AND METHODS

This research has been approved by the Institutional Animal Care and Use Committee (IACUC) of Chungnam National University (202103A-CNU-061).

An overview of the procedure used for identifying SNP markers to discriminate the Yeonsan Ogye breed is provided in Fig. 1.

Samples and genotypes

Three data sets were used in this study: Sets 1 and 2 for selecting SNP markers, and Set 3 for validation (Table 1). Sets 1 and 2 consisted of 3,904 individuals from 198 chicken breeds, genotyped with a 600K SNP array (Affymetrix, Santa Clara, CA, USA) [10]. Set 1 constituted populations of KNC from the Korean National Institute of Animal Science (NIAS), including Yeonsan Ogye (189 birds), and other indigenous (208 birds from five lines) and adapted KNC (218 birds) breeds. Set 2 consisted of commercial chickens (CC; 34 broilers and 20 layers) and various other global chicken breeds from the SYNBREED project in Germany [11]. The SYNBREED dataset included 3,235 individuals and 174 breeds from 32 countries, including Africa, South America, Asia, and Europe.

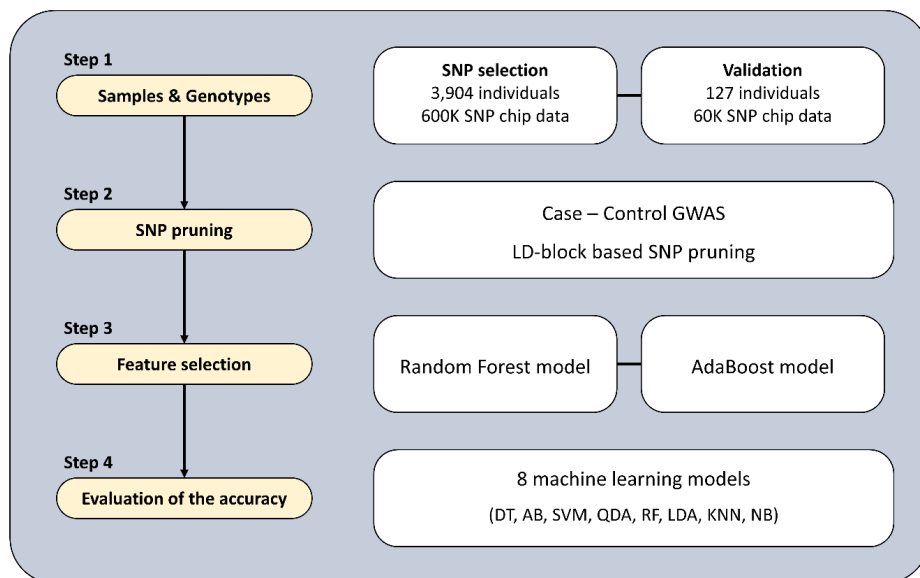


Fig. 1. Marker combination selection process for classification of the Yeonsan Ogye chicken breed. SNP, single nucleotide polymorphism; GWAS, genome-wide association study, LD, linkage disequilibrium; DT, decision tree; AB, AdaBoost; SVM, support vector machine; QDA, quadratic discriminant analysis; RF, Random Forest; LDA, linear discriminant analysis; KNN, K-Nearest Neighbor; NB, Naïve Bayes.

Table 1. Summary of the samples used in this study

Objective	Class	Population	Number of animals	Total
SNP selection (600K SNP array)	Set 1	KNC		3,904
		Indigenous	208	
		Adapted	218	
	Set 2	Yeonsan Ogye	189	
		Commercial		
		Broiler	34	
	Layer	20		
	SYNBREED	3,235		
Validation test (60K SNP array)	Set 3	Yeonsan Ogye	67	127
		KNC	30	
		Yeonsan Ogye and White Leghorn crossbreed	30	

SNP, single nucleotide polymorphism; KNC, Korean native chicken.

Set 3 consisted of Yeonsan Ogye (67 birds) and KNC (30 birds from two lines), genotyped using a custom 60K SNP array made by our research team, and an F2 generation crossbreed population of Yeonsan Ogye and White Leghorn (30 birds) genotyped with an Illumina 60K SNP array (Illumina, San Diego, CA, USA) [12].

Data pre-processing and single nucleotide polymorphism pruning

A total of 542,717 common SNPs was derived from Sets 1 and 2, and there were two major quality control (QC) cut-offs: genotyping rate $\geq 90\%$ and minor allele frequency ≥ 0.05 . For determining Yeonsan Ogye-specific SNPs, the derived SNPs were subjected to a case-control genome-wide association study (GWAS) performed using PLINK 1.9 software [13]. In that analysis, the case

group was the Yeonsan Ogye population, and the control group comprised all other populations. The significant SNPs were figured out based on the Bonferroni-corrected p -value ($\alpha = 0.01$). The linkage disequilibrium (LD) was calculated, and LD block-based SNP pruning was conducted to select one SNP per 50 LD blocks.

Feature selection

Machine learning was applied for the feature selection of pruned SNP markers to reduce the number of SNP markers and identify optimal markers. Feature importance values were calculated through two machine learning models: RF using the “randomForest” package in R software [14] and AB using the “adabag” R package [15]. SNPs with importance values higher than the point at which feature importance rapidly decreased were classified as optimum markers. Principal component analysis (PCA) was conducted to verify the SNP marker selections.

Evaluation of accuracy

To resolve data imbalances before analysis, only one individual was randomly selected from each of the 197 populations in the control group. To confirm the accuracy of discrimination for the Yeonsan Ogye chicken population, 70% of the total data were used as the training set, and the remaining 30% as the test set, based on five repeated 10-fold cross-validation. Eight different machine learning models were employed to evaluate the accuracy: Decision Tree (DT), AB, Support Vector Machine (SVM), Quadratic Discriminant Analysis (QDA), RF, Linear Discriminant Analysis (LDA), K-Nearest Neighbor (KNN), and Naïve Bayes (NB) [16–18]. Principle components 1 (RF, 47.4%; AB, 45.3%) and 2 (RF, 5.9%; AB, 5.4%) values, derived from the PCA for marker selection, were used to build these eight classification models with the “caret” R package [19].

$$\text{Class} \sim \text{PC1} + \text{PC2}$$

For performance verification, each machine learning model was assessed based on confusion matrix values: accuracy, specificity, sensitivity (recall), precision, and F1-score.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{Sensitivity (Recall)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where TP is true-positive (number of correct predictions for the case group), TN is true-negative (number of correct predictions for the control group), FP is false-positive (number of incorrect predictions for the case group) and FN is false-negative (number of incorrect predictions for the control group).

Validation tests

Validation tests were conducted on independent populations to validate the discriminatory performance of the selected marker combinations. Set 3 was used for validation analysis; the data were genotyped using 60K SNP arrays. Minimac3 and Minimac4 software were used for data imputation prior to the analysis [20].

RESULTS

Genetic clusters

PCA of the 600K SNP genotype data for the entire population was performed. Fig. 2 shows the genetic clustering for each population. The indigenous KNC populations were clustered separately from the other groups, while the adapted KNC populations tended to cluster with CC such as broilers and layers. Contrary to this, the Yeosan Ogye population was well differentiated from both the SYNBREED and Korean populations.

Single nucleotide polymorphism pruning and feature selection

A case-control GWAS was performed to determine significant SNP markers. The target breed, Yeosan Ogye, was the case group, and the other populations comprised the control group. The GWAS revealed 285,227 significant SNPs based on a Bonferroni corrected p -value of < 0.01 . As well as LD blocks, 100,799 haplotype blocks were distinguished. Ultimately, 120 SNPs were

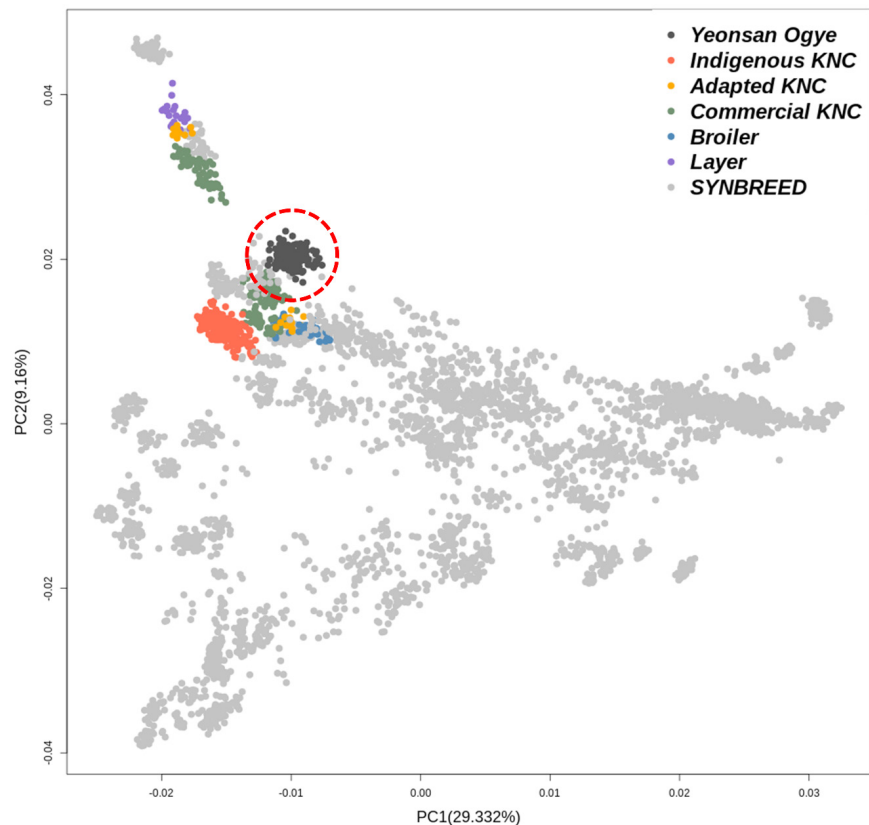


Fig. 2. Results of principal component analysis of 600K single nucleotide polymorphism genotype data. Note that Yeosan Ogye (within the red circle) is distinct from the other Korean breeds, and the foreign SYNBREED populations [11] with CC-BY. KNC, Korean native chicken.

extracted through LD-based SNP pruning of 151,062 markers common to both the GWAS results and LD blocks. In a final step, 38 (RF) and 43 (AB) SNPs were identified as the optimal marker combinations. According to the PCA of these SNP combinations, the Yeonsan Ogye population was accurately distinguished from the control group species (Fig. 3).

Evaluation of classification accuracy

Using the 38 and 43 optimal SNP combinations described above, all eight machine learning algorithms discriminated the Yeonsan Ogye population perfectly (Fig. 4 and 5) according to the confusion matrix values (i.e., accuracy = 1.00) (Table 2).

In total, 30 markers from the imputation results overlapped with the previously selected marker combinations, and distinguished the Yeonsan Ogye and control group populations accurately; the confusion matrix values were all 1.00 (Fig. 6 and 7), except for that of QDA (0.97) based on AB feature selection.

DISCUSSION

Optimal strategies for breed identification are essential for protecting livestock pedigree, and for industrial research. Native chickens are a particularly important target for biodiversity conservation; chickens are able to adapt well to new environments [21]. Park et al. [22] reported that the provision of breed information for native chickens promoted consumption.

Genotyping methods have been developed over several generations, and the cost of genotyping continues to decline. Hence, extensive genotype data are available for use as biomarkers. SNP markers have been used for genetic classification based on PCA, F-statistics, and genotype frequencies [23–25]. However, identifying optimal SNP markers to identify specific breeds using high-density SNP chips is still quite challenging.

In this study, several markers were identified based on GWAS and LD pruning results and using high-density 600K SNP chip data. Johnson et al. [26] and Wallace et al. [27] explained that it is challenging to determine whether genetic markers identified through GWAS are causative genes in response to LD. Bakshi et al. [28] stated that more informative results can be obtained by removing

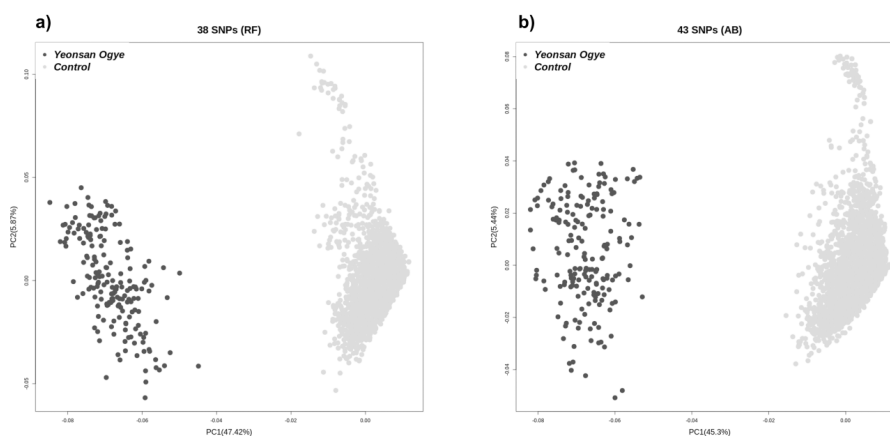


Fig. 3. Results of principal component analysis using optimal marker combinations selected by two machine learning models. Two marker combinations could discriminate Yeonsan Ogye (black) from other control populations (gray). a) The result of Random Forest feature selection process and b) The result of AdaBoost feature selection process. SNP, single nucleotide polymorphism; RF, Random Forest; AB, AdaBoost.

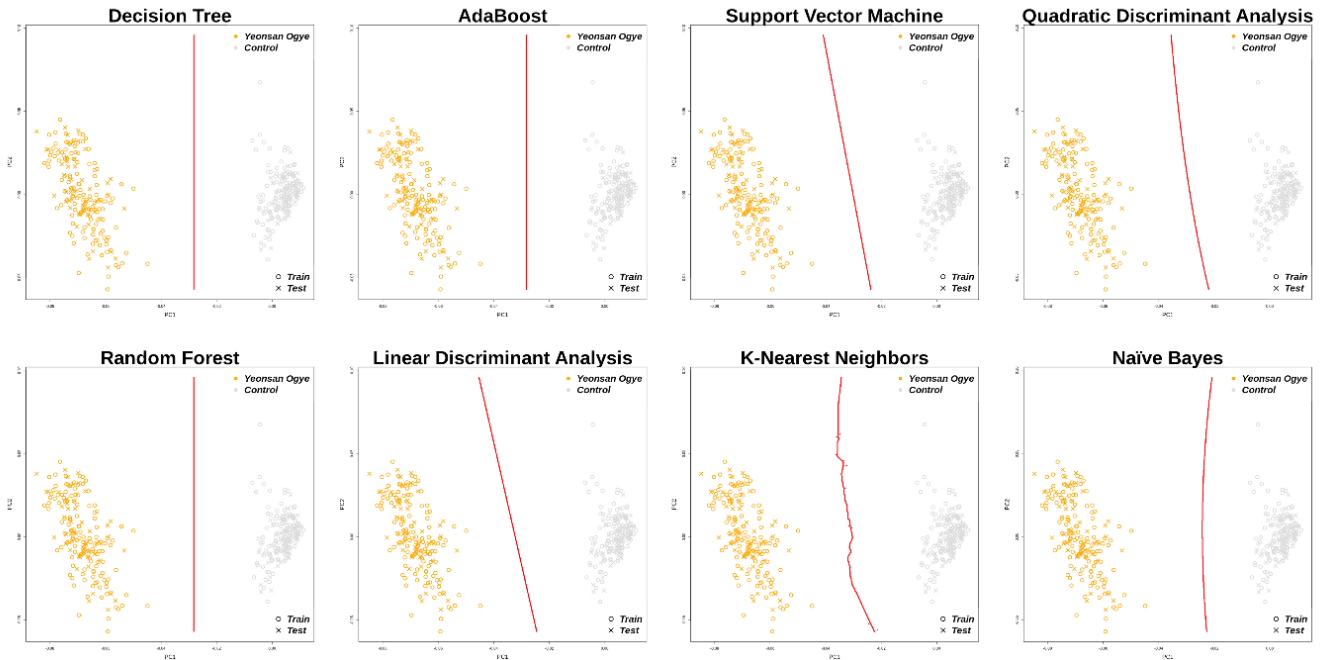


Fig. 4. Classification results for eight machine learning models using 38 markers identified via a Random Forest feature selection process. All machine learning models could discriminate Yeonsan Ogye (case, yellow) from both the other Korean and global chicken populations (control, gray). The red lines are the classification trend lines for the machine learning models.

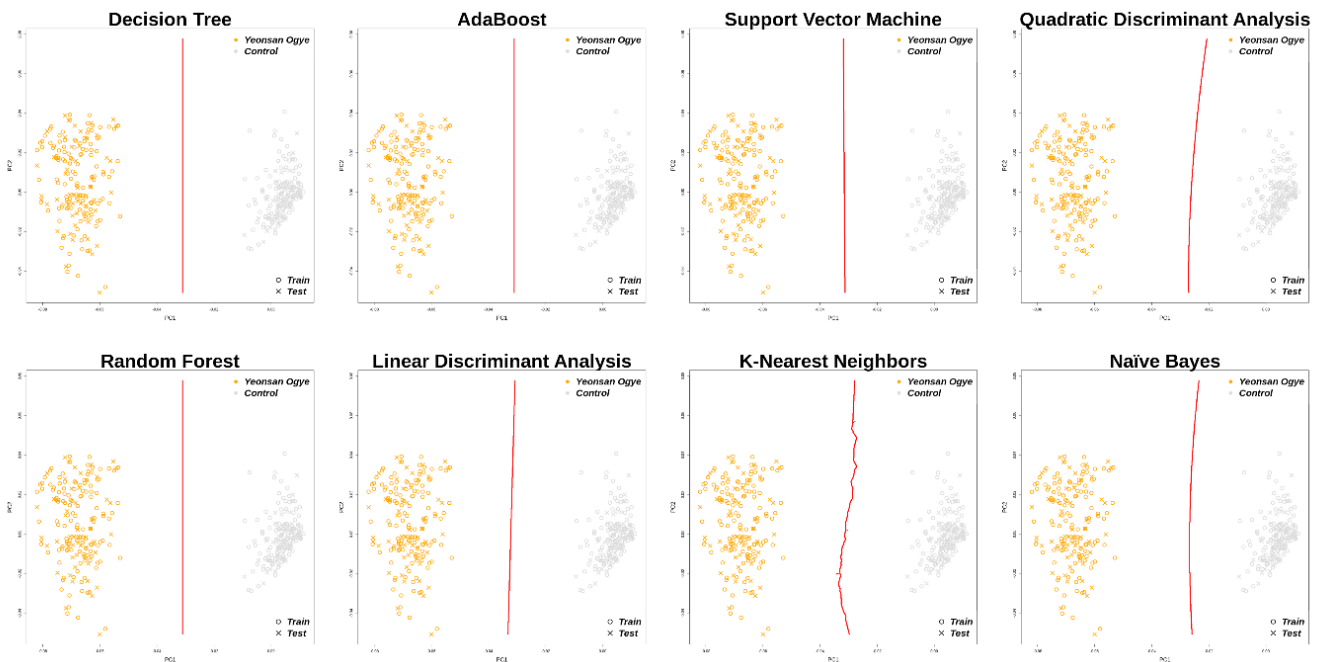


Fig. 5. Classification results for eight machine learning models using 43 markers identified via an AdaBoost feature selection process. All machine learning models could discriminate Yeonsan Ogye (case, yellow) from both the other Korean and global chicken populations (control, gray). The red lines are the classification trend lines for the machine learning models.

Table 2. Classification accuracies for the different machine learning models using optimal marker combinations

Method	Model	Accuracy	Specificity	Sensitivity	Precision	F1-score
RF	DT	1.00	1.00	1.00	1.00	1.00
	AB	1.00	1.00	1.00	1.00	1.00
	SVM	1.00	1.00	1.00	1.00	1.00
	QDA	1.00	1.00	1.00	1.00	1.00
	RF	1.00	1.00	1.00	1.00	1.00
	LDA	1.00	1.00	1.00	1.00	1.00
	KNN	1.00	1.00	1.00	1.00	1.00
	NB	1.00	1.00	1.00	1.00	1.00
AB	DT	1.00	1.00	1.00	1.00	1.00
	AB	1.00	1.00	1.00	1.00	1.00
	SVM	1.00	1.00	1.00	1.00	1.00
	QDA	1.00	1.00	1.00	1.00	1.00
	RF	1.00	1.00	1.00	1.00	1.00
	LDA	1.00	1.00	1.00	1.00	1.00
	KNN	1.00	1.00	1.00	1.00	1.00
	NB	1.00	1.00	1.00	1.00	1.00

DT, decision tree; AB, AdaBoost; SVM, support vector machine; QDA, quadratic discriminant analysis; RF, Random Forest; LDA, linear discriminant analysis; KNN, K-Nearest Neighbor; NB, Naive Bayes.

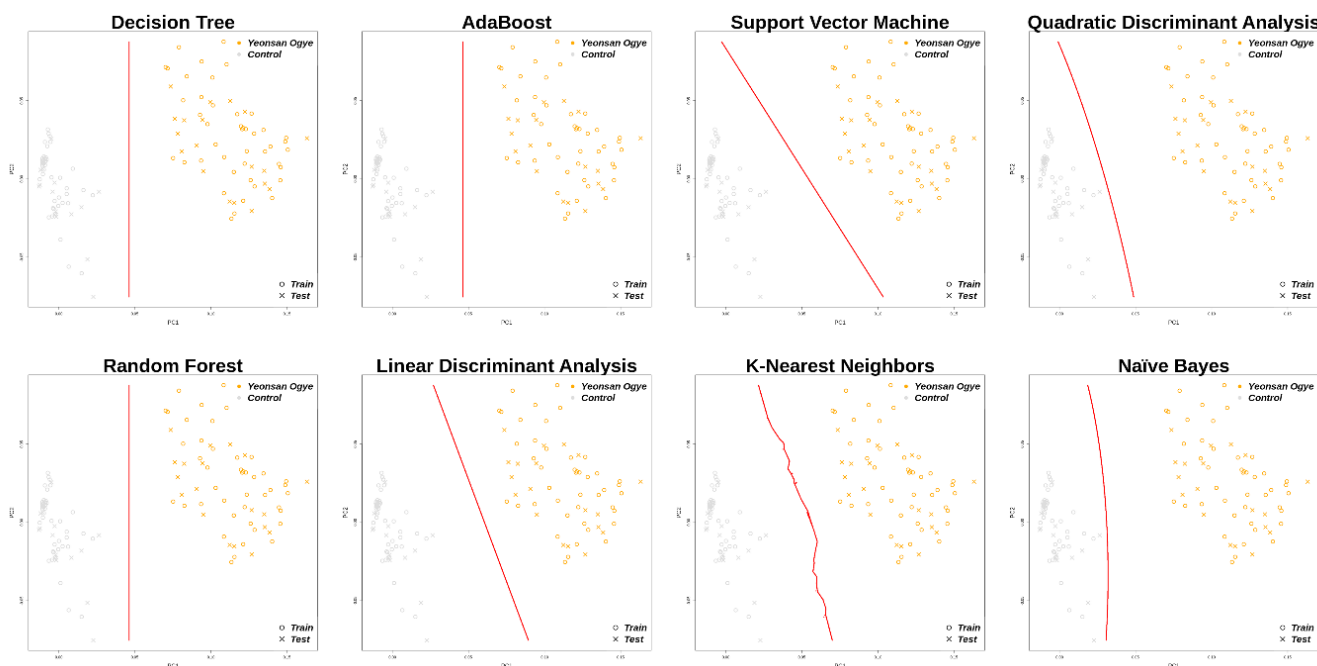


Fig. 6. Validation test results for eight machine learning models using 30 markers identified via a Random Forest feature selection process. All machine learning models could discriminate Yeonsan Ogye (case, yellow) from the other Korean chicken breeds, and the Yeonsan Ogye and White Leghorn crossbred (control, gray). The red lines are the classification trend lines for the machine learning models.

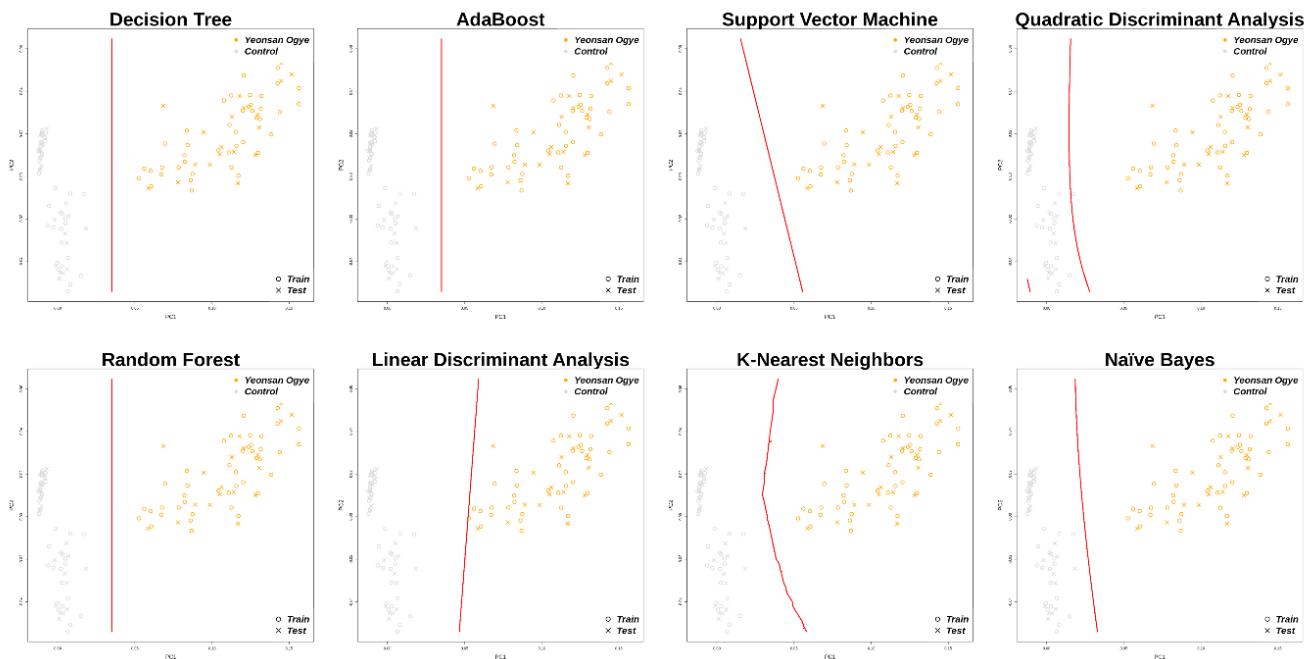


Fig. 7. Validation test results for eight machine learning models using 30 markers identified via an AdaBoost feature selection process. All machine learning models could discriminate Yeonsan Ogye (case, yellow) from the other Korean chicken breeds, and the Yeonsan Ogye and White Leghorn crossbreed (control, gray). The red lines are the classification trend lines for the machine learning models.

SNPs with strong LD relationships from the analysis. In our analysis, the target breed, Yeonsan Ogye, was effectively discriminated using SNP markers selected with consideration of LD.

Machine learning is an artificial intelligence technology for classifying data and making predictions. We applied machine learning algorithms to identify SNP marker combinations for Yeonsan Ogye classification through GWAS and LD pruning. Machine learning has been used to select SNP markers for various livestock species [29–32]. Moreover, applying feature selection to GWAS results can reduce dimensionality and overfitting errors when identifying markers, resulting in more accurate predictions [33].

In this study, RF and AB models were used to determine optimal SNP marker combinations; 38 and 43 significant SNP markers were identified, respectively, and both sets showed remarkable classification power. Notably, 14 SNPs were shared between the two marker sets, and it was possible to differentiate the target population with sufficient accuracy (more than 98%) using those markers. In addition to accuracy, other confusion matrix evaluation indices, such as sensitivity (recall) and precision, also demonstrated the high classification power of the marker combinations.

The precise results obtained herein could be explained by the fact that the Yeonsan Ogye chicken is a genetically unique breed. The PCA plot of the 600K genotype data showed that the Yeonsan Ogye population was clustered separately from the other breeds. Further, Yeonsan Ogye chicken had a gene pool independently from the entirely black chickens in the SYNBREED group, such as Cemani and Sumatran from Indonesia, and Silkies from China.

The marker combinations identified for the Yeonsan Ogye pure line (PL) showed impressive results in the validation test. Two of five KNC lines and the Yeonsan Ogye-White Leghorn crossbreed were included in the control group for the validation test. The 30 SNPs were common to both SNP marker sets and correctly differentiated KNC and Yeonsan Ogye, as also seen during the SNP marker selection process. The Yeonsan Ogye and White Leghorn crossbreeds were also clearly

distinguished; the phenotypes of the individuals comprising this F2 generation were very diverse. The marker combinations showed the ability to perfectly discriminate pure Yeosan Ogye birds, even from other chicken breeds with a similar phenotype.

Generally, the chickens available on the market are CC produced by using PLs through three- or four-way crossbreeding. Since breed-specific markers are identified using PLs, the applicability to breeds that have not been verified via the marker selection process is limited. Although verification analysis was performed on the crossbreeds in this study, it would be complicated to apply it to crossbreeds other than White Leghorn. Ultimately, the discriminatory power of the optimal SNP marker combinations identified herein must be verified through application to other populations.

CONCLUSION

We identified two optimal SNP combinations for accurately classifying the Yeosan Ogye chicken breed through a machine learning approach. The results indicated that, through GWAS, LD, and feature selection, machine learning models could be applied for identifying other breeds.

REFERENCES

1. Park MH, Oh JD, Jeon GJ, Kong HS, Yeon SH, Sang BD, et al. Method discrimination for product traceability and identification of Korean native chicken using microsatellite DNA. *Korean J Org Agric*. 2004;12:451-61.
2. Suh S, Cho CY, Kim JH, Choi SB, Kim YS, Kim H, et al. Analysis of genetic characteristics and probability of individual discrimination in Korean indigenous chicken brands by microsatellite marker. *J Anim Sci Technol*. 2013;55:185-94. <https://doi.org/10.5187/JAST.2013.55.3.185>
3. Dharmayanthi AB, Terai Y, Sulandari S, Zein MSA, Akiyama T, Satta Y. The origin and evolution of fibromelanosis in domesticated chickens: genomic comparison of Indonesian Cemani and Chinese Silkie breeds. *PLOS ONE*. 2017;12:e0173147. <https://doi.org/10.1371/journal.pone.0173147>
4. 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:e1002412. <https://doi.org/10.1371/journal.pgen.1002412>
5. Choi NR, Seo DW, Jemaa SB, Sultana H, Heo KN, Jo C, et al. Discrimination of the commercial Korean native chicken population using microsatellite markers. *J Anim Sci Technol*. 2015;57:5. <https://doi.org/10.1186/s40781-015-0044-6>
6. Oh JD, Song KD, Seo JH, Kim DK, Kim SH, Seo KS, et al. Genetic traceability of black pig meats using microsatellite markers. *Asian-Australas J Anim Sci*. 2014;27:926-31. <https://doi.org/10.5713/ajas.2013.13829>
7. Serrano M, Calvo JH, Martínez M, Marcos-Carcavilla A, Cuevas J, González C, et al. Microsatellite based genetic diversity and population structure of the endangered Spanish Guadarrama goat breed. *BMC Genet*. 2009;10:61. <https://doi.org/10.1186/1471-2156-10-61>
8. Fischer MC, Rellstab C, Leuzinger M, Roumet M, Gugerli F, Shimizu KK, et al. Estimating genomic diversity and population differentiation: an empirical comparison of microsatellite and SNP variation in *Arabidopsis halleri*. *BMC Genomics*. 2017;18:69. <https://doi.org/10.1186/s12864-016-3459-7>
9. Karniol B, Shirak A, Baruch E, Singrün C, Tal A, Cahana A, et al. Development of a 25-plex

- SNP assay for traceability in cattle. *Anim Genet.* 2009;40:353-6. <https://doi.org/10.1111/j.1365-2052.2008.01846.x>
10. Kranis A, Gheyas AA, Boschiero C, Turner F, Yu L, Smith S, et al. Development of a high density 600K SNP genotyping array for chicken. *BMC Genomics.* 2013;14:59. <https://doi.org/10.1186/1471-2164-14-59>
 11. Malomane DK, Simianer H, Weigend A, Reimer C, Schmitt AO, Weigend S. The SYNBREED chicken diversity panel: a global resource to assess chicken diversity at high genomic resolution. *BMC Genomics.* 2019;20:345. <https://doi.org/10.1186/s12864-019-5727-9>
 12. Groenen MAM, Megens HJ, Zare Y, Warren WC, Hillier LW, Crooijmans RPMA, et al. The development and characterization of a 60K SNP chip for chicken. *BMC Genomics.* 2011;12:274. <https://doi.org/10.1186/1471-2164-12-274>
 13. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81:559-75. <https://doi.org/10.1086/519795>
 14. Breiman L. Random forests. *Mach Learn.* 2001;45:5-32. <https://doi.org/10.1023/A:1010933404324>
 15. Alfaro E, Gamez M, Garcia N. adabag: an R package for classification with boosting and bagging. *J Stat Softw.* 2013;54:1-35. <https://doi.org/10.18637/jss.v054.i02>
 16. Crisci C, Ghattas B, Perera G. A review of supervised machine learning algorithms and their applications to ecological data. *Ecol Model.* 2012;240:113-22. <https://doi.org/10.1016/j.ecolmodel.2012.03.001>
 17. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak.* 2019;19.1:281. <https://doi.org/10.1186/s12911-019-1004-8>
 18. Zhang MQ. Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc Natl Acad Sci USA.* 1997;94.2:565-8. <https://doi.org/10.1073/pnas.94.2.565>
 19. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw.* 2008;28:1-26. <https://doi.org/10.18637/jss.v028.i05>
 20. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. *Nature Genet.* 2016;48:1284-7. <https://doi.org/10.1038/ng.3656>
 21. Hoffmann I. Climate change and the characterization, breeding and conservation of animal genetic resources. *Anim Genet.* 2010;41:32-46. <https://doi.org/10.1111/j.1365-2052.2010.02043.x>
 22. Park S, Kim N, Kim W, Moon J. The Effect of Korean native chicken breed information on consumer sensory evaluation and purchase behavior. *Food Sci Anim Resour.* 2022;42:111-27. <https://doi.org/10.5851/kosfa.2021.e67>
 23. Heaton MP, Harhay GP, Bennett GL, Stone RT, Grosse WM, Casas E, et al. Selection and use of SNP markers for animal identification and paternity analysis in U.S. beef cattle. *Mamm Genome.* 2002;13:272-81. <https://doi.org/10.1007/s00335-001-2146-3>
 24. Suekawa Y, Aihara H, Araki M, Hosokawa D, Mannen H, Sasazaki S. Development of breed identification markers based on a bovine 50K SNP array. *Meat Sci.* 2010;85:285-8. <https://doi.org/10.1016/j.meatsci.2010.01.015>
 25. Wilkinson S, Wiener P, Archibald AL, Law A, Schnabel RD, McKay SD, et al. Evaluation of approaches for identifying population informative markers from high density SNP chips.

- BMC Genet. 2011;12:45. <https://doi.org/10.1186/1471-2156-12-45>
26. Johnson RC, Nelson GW, Troyer JL, Lautenberger JA, Kessing BD, Winkler CA, et al. Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics*. 2010;11:724. <https://doi.org/10.1186/1471-2164-11-724>
 27. Wallace JG, Bradbury PJ, Zhang N, Gibon Y, Stitt M, Buckler ES. Association mapping across numerous traits reveals patterns of functional variation in maize. *PLOS Genet*. 2014;10:e1004845. <https://doi.org/10.1371/journal.pgen.1004845>
 28. Bakshi A, Zhu Z, Vinkhuyzen AA, Hill WD, McRae AF, Visscher PM, et al. Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. *Sci Rep*. 2016;6:32894. <https://doi.org/10.1038/srep32894>
 29. Seo D, Cho S, Manjula P, Choi N, Kim YK, Koh YJ, et al. Identification of target chicken populations by machine learning models using the minimum number of SNPs. *Animals*. 2021;11:241. <https://doi.org/10.3390/ani11010241>
 30. Matukumalli LK, Grefenstette JJ, Hyten DL, Choi IY, Cregan PB, Van Tassell CP. Application of machine learning in SNP discovery. *BMC Bioinformatics*. 2006;7:4. <https://doi.org/10.1186/1471-2105-7-4>
 31. Schiavo G, Bertolini F, Galimberti G, Bovo S, Dall'Olio S, Nanni Costa L, et al. A machine learning approach for the identification of population-informative markers from high-throughput genotyping data: application to several pig breeds. *Animal*. 2020;14:223-32. <https://doi.org/10.1017/S1751731119002167>
 32. Xu Z, Diao S, Teng J, Chen Z, Feng X, Cai X, et al. Breed identification of meat using machine learning and breed tag SNPs. *Food Control*. 2021;125:107971. <https://doi.org/10.1016/j.foodcont.2021.107971>
 33. Bermingham ML, Pong-Wong R, Spiliopoulou A, Hayward C, Rudan I, Campbell H, et al. Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Sci Rep*. 2015;5:10312. <https://doi.org/10.1038/srep10312>