RESEARCH ARTICLE

J Anim Sci Technol 2020;62(4):438-448
https://doi.org/10.5187/jast.2020.62.4.438

JAST
Journal of Animal Science and Technology
pISSN 2672-0191 eISSN 2055-0391

Comparison of prediction accuracy for genomic estimated breeding value using the reference pig population of single-breed and admixed-breed

Soo Hyun Lee[1#], Dongwon Seo[1#], Doo Ho Lee[1], Ji Min Kang[1], Yeong Kuk Kim[1], Kyung Tai Lee[2], Tae Hun Kim[2], Bong Hwan Choi[2]* and Seung Hwan Lee[1]*

[1]*Division of Animal and Dairy Science, Chungnam National University, Daejeon 34134, Korea*
[2]*Animal Genomics and Bioinformatics Division, National Institute of Animal Science, RDA, Wanju 55365, Korea*

Received: May 20, 2020
Revised: Jun 18, 2020
Accepted: Jun 18, 2020

#These authors contributed equally to this work.

***Corresponding author**
Seung Hwan Lee
Division of Animal and Dairy Science,
Chungnam National University,
Daejeon 34134, Korea.
Tel: +82-42-821-5772
E-mail: slee46@cnu.ac.kr

Bong Hwan Choi
Animal Genomics & Bioinformatics
Division, National Institute of Animal
Science, RDA, Wanju 55365, Korea.
Tel: +82-63-238-7304
E-mail: bhchoi@korea.kr

Copyright © 2020 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
Soo Hyun Lee
https://orcid.org/0000-0001-5257-2068
Dongwon Seo
https://orcid.org/0000-0003-0548-7068

Abstract

This study was performed to increase the accuracy of genomic estimated breeding value (GEBV) predictions for domestic pigs using single-breed and admixed reference populations (single-breed of Berkshire pigs [BS] with cross breed of Korean native pigs and Landrace pigs [CB]). The principal component analysis (PCA), linkage disequilibrium (LD), and genome-wide association study (GWAS) were performed to analyze the population structure prior to genomic prediction. Reference and test population data sets were randomly sampled 10 times each and precision accuracy was analyzed according to the size of the reference population (100, 200, 300, or 400 animals). For the BS population, prediction accuracy was higher for all economically important traits with larger reference population size. Prediction accuracy was ranged from −0.05 to 0.003, for all traits except carcass weight (CWT), when CB was used as the reference population and BS as the test. The accuracy of CB for backfat thickness (BF) and shear force (SF) using admixed population as reference increased with reference population size, while the results for CWT and muscle pH at 24 hours after slaughter (pH) were equivocal with respect to the relationship between accuracy and reference population size, although overall accuracy was similar to that using the BS as the reference.

Keywords: Korean native pig, Genomic prediction, Admixed reference, Genome-wide association study

INTRODUCTION

Genomic selection is a useful way to enhance economically important traits in domestic animals. Previous studies showed that using reference populations with abundant markers and a large size increases the prediction accuracy of estimated breeding value (EBV) [1]. However, in small size of reference population, obtaining an appropriate reference population comprising individuals of the same breed is difficult, leading to low accuracy of predictions. As an alternative approach, use of an admixed popu-

Doo Ho Lee
https://orcid.org/0000-0002-2174-7897
Ji Min Kang
https://orcid.org/0000-0002-4907-2203
Yeong Kuk Kim
https://orcid.org/0000-0002-6530-2304
Kyung Tai Lee
https:// orcid.org/0000-0003-0990-4818
Tae Hun Kim
https:// orcid.org/0000-0003-1621-3281
Bong Hwan Choi
https:// orcid.org/0000-0002-4795-3285
Seung Hwan Lee
https://orcid.org/0000-0003-1508-4887

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

Funding sources
This work was carried out with the support of "Cooperative Research Program for Agriculture Science and Technology Development (Project No. PJ011191012017)" Rural Development Administration, Korea.

Acknowledgements
This research was partly supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. NRF-2019R1F1A1057605).

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

Authors' contributions
Conceptualization: Choi BH, Lee Seung Hwan.
Data curation: Lee Soo Hyun, Lee DH, Kim YK.
Formal analysis: Lee Soo Hyun, Seo D.
Methodology: Lee Soo Hyun, Lee DH, Kim YK.
Software: Kang JM, Kim YK.
Validation: Seo D, Lee KT, Kim TH, Lee Seung Hwan.
Investigation: Lee KT, Kim TH, Choi BH.
Writing - original draft: Lee Soo Hyun, Seo D.
Writing - review & editing: Lee KT, Choi BH, Lee Seung Hwan.

Ethics approval and consent to participate
This article does not require IRB/IACUC approval because there are no human and animal participants.

lation including the target population as a reference has been recommended [2,3]. Such admixed populations can be used as a reference when breeds are defined as their link by genotypes. When the reference population comprises a breed that is distinct from the test population, they must be genetically related, rather than related by pedigree. Genetic markers can explain the relationships among all individuals in a genomic relationship matrix. In addition, with greater linkage disequilibrium (LD), the prediction accuracy of EBV should increase [4-6]. In this point of view, this study was performed to determine the prediction accuracy of EBV using an admixed reference population consisting of crossbred Korean native pig and Landrace pigs (CB).

MATERIALS AND METHODS

Genotypes and phenotypes of collected samples

In accordance with the ethical guidelines, a total of 1,289 pigs (695 Berkshire [BS] and 594 cross breed (CB) blood samples were collected by veterinarians and were genotyped using a Porcine 60K SNP chip (Illumina, San Diego, CA, USA) (Table 1). These samples were provided by the National Institute of Animal Science (Jeonju, Korea); 25 KNP and 20 Landrace purebred samples were also provided to confirm the genetic relationships. Quality control (QC) was performed on each population; 41,594 and 39,002 BS and CB single nucleotide polymorphisms (SNPs) remained after QC (missing chromosomes with 11,166 and 4,214 markers, minor allele frequency less than 1% with 359 and 5,606 markers, missing genotypes over than 10% with 10,030 and 433 markers for BS and CB, respectively) and were merged to yield a single admixed population (Table 1). After merging, with in common and overwrapped markers, 45,875 SNPs remained. The phenotypes of the 1,289 animals were measured (backfat thickness [BF], carcass weight [CWT], muscle pH at 24 hours after slaughter [pH], and shear force [SF]). The sex and slaughter age of all animals were recorded in the phenotype measurement processes.

Analyses prior to genomic estimated breeding value (GEBV) prediction: population structure and genome-wide association study (GWAS)

The population structure was evaluated, and association studies were conducted to enable further analyses. Visualization of the population structure is useful to determine genetic relationships among breeds. Using each 20 samples genotype information from each BS, CB, Landrace, and KNP populations, principal component analysis (PCA) was performed to generate clusters, determine any shared principal components, and detect any incorrectly classified individuals. Furthermore, plots of LD by distance, within populations and among breeds, were generated. A GWAS of the traits of interest was performed for genetic comparison between the CB and BS, and to determine any significant loci or LD relationships. The GWAS was performed based on a mixed

Table 1. Genotype information for the studied population

	Number of animal	Original genotype	SNPs removed by QC			Number of SNPs after QC
			Not located or located on sex chromosome	Minor allele frequency (< 0.01[1)])	Missing genotype (> 0.1[2)])	
Berkshire (BS)	695	63149	11,166	359	10,030	41,594
Korean native × Landrace crossbreed (CB)	594	49255	4,214	5,606	433	39,002

[1)] Alleles removed when minor allele frequency < 1%.

[2)] Alleles removed when genotype is missing from > 10% of the entire population.

SNPs, single nucleotide polymorphisms; QC, quality control.

linear model generated using GCTA software (ver. 1.25.3 [7]). Bayesian mixture model was created using the BayesR program (default option with 0, 0.0001, 0.001, 0.01 effect sizes of mixture; 50000 MCMC chain; 20,000 burnin; 10th thin interval). Proportion of variance for specific SNP was calculated as follow:

$$\text{Variance explained}\,(\%) = \frac{\left(2 \times p \times q \times \beta^2\right)}{Va}$$

Information on the genetic contributions to traits was obtained from a previous study [8]. The PCA, LD analysis, and data processing were performed using PLINK 1.9 [9] and R software (R Development Core Team, Vienna, Austria) [10]. Data were visualized in the R environment.

Procedure for predicting breeding value

To compare the prediction accuracy of breeding value between single-breed and admixed reference populations, both the reference and test animal data sets were randomly sampled 10 times each. There is no intersect animals among test and reference population. The GEBV predictions were performed using all test and reference set combinations, and mean accuracy was assessed according to the size of the reference population. Prediction accuracy using a single-breed reference population was determined for each breed (250 test animals each) by reference population size (100, 200, 300, or 400 animals) (Fig. 1). For the analysis involving the admixed reference population, the reference population size was the same as in the previous scenario. Admixed reference included each breed with an equal ratio. 125 individuals were randomly selected from each of the two breeds as test animals.

A genetic relationship matrix was built using GCTA (ver. 1.25.3 [7]) and ASReml 4.1 [11] was used for genomic prediction. The model used in this study was as follows:

$$y \sim \mu + Xb + Zu + e$$

where y indicates the measured phenotype, μ is the overall mean, X and Z are design matrices related to fixed effects and effects, respectively, b and u are vectors of fixed and genetic effects, respec-

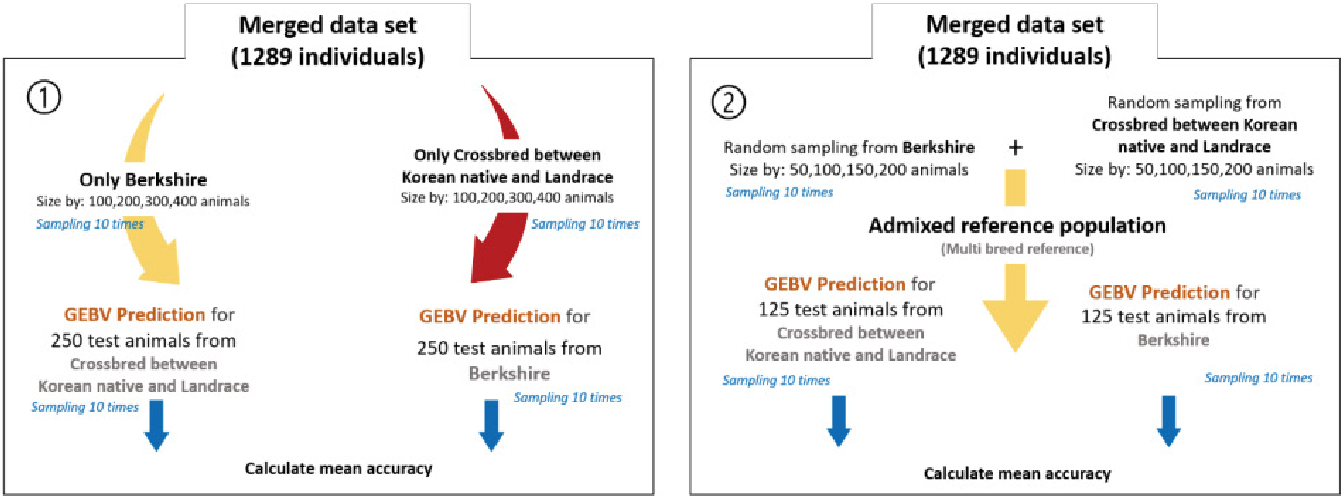


Fig. 1. Schematic of the breeding value predictions with and without use of the admixed reference population (1 and 2, respectively).

tively, and *e* indicates error variance. The prediction accuracy was given by the correlation between GEBV and own phenotype using the following equation [12]:

$$\text{Accuracy}_{\text{corr}} = \text{cor}\,(\text{GEBV}, \text{Y})$$

RESULTS

Population structure and genome-wide association study (GWAS)

An overview of the population genetic structure was obtained by the PCA and GWAS prior to genomic prediction (Figs. 2–6). First, in order to compare the populations with the same sample size, 20 samples SNP genotype information such as KPN and Landrace purebred populations were randomly extracted from BS and CB, respectively. As shown in Fig. 2, each population forms a distinct cluster; the first and second principal components explain 12.89% and 9.38% of the variance in the population genetic structure, respectively. On the axis of the first component, the Landrace and BS populations are located close to each other, with the KNP population being more distant. On the axis of the second component only, the KNP population was located towards the middle.

LD was examined in each population by distance. (Figs. 3 and 4). KNP has clearly stronger LD pattern than those for BS, CB, and Landrace, while BS showed the weakest correlations, and the

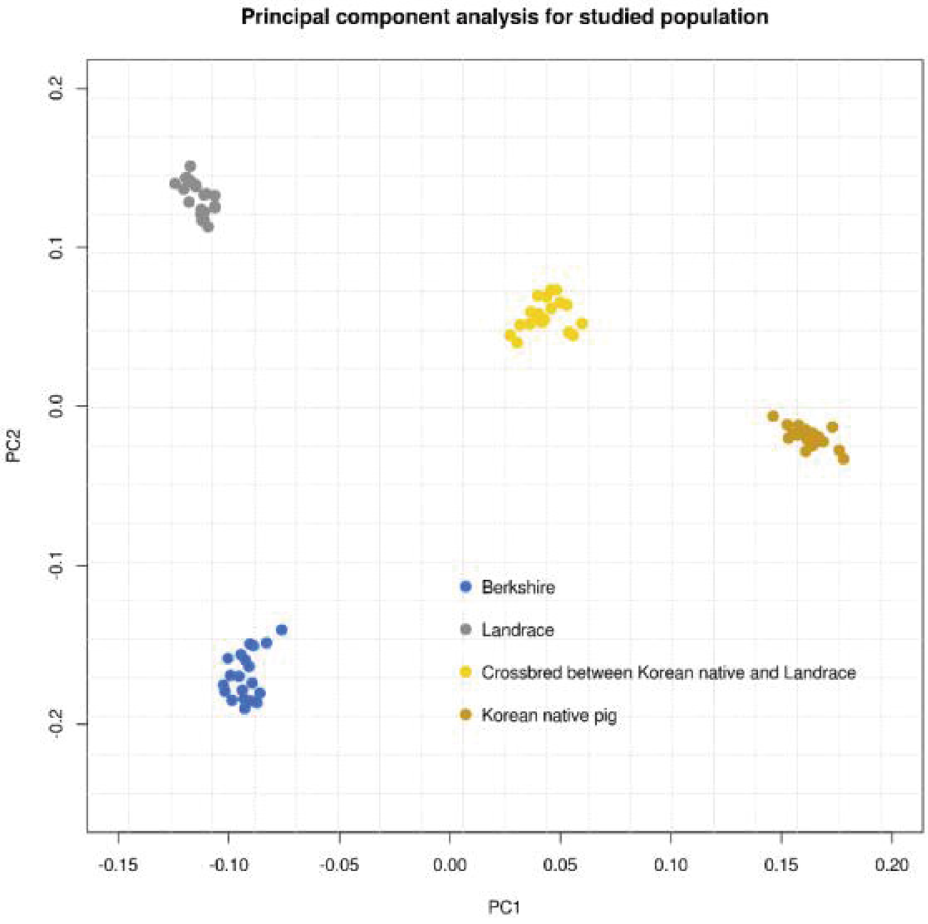


Fig. 2. Principal component analysis among the studied population. 20 samples per each population were used to confirm the genetic relationship.

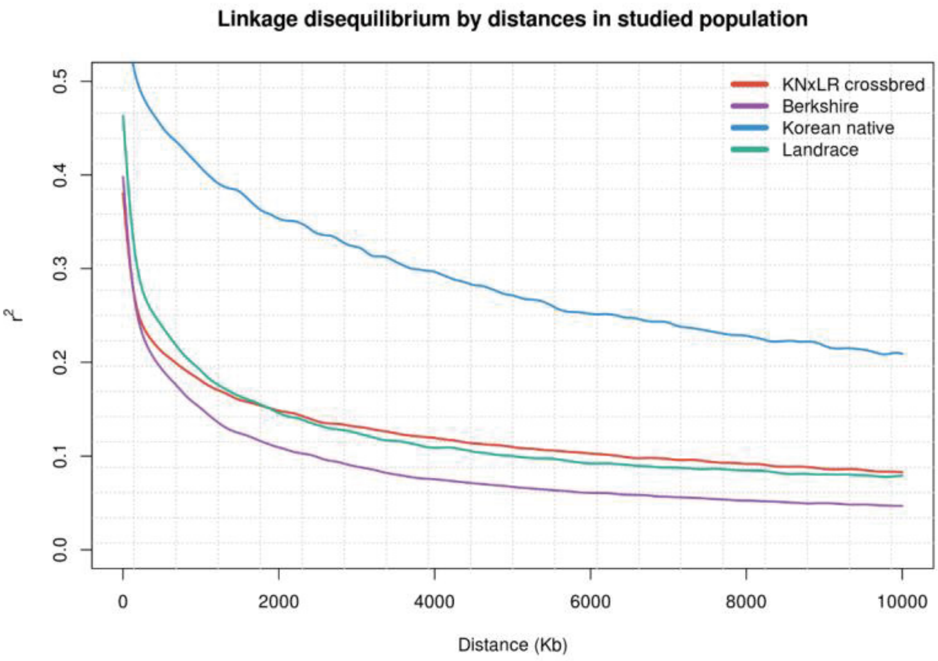


Fig. 3. Linkage disequilibrium (LD) by genetic distance for the different breeds. KN, Korean native pig; LR, Landrace.

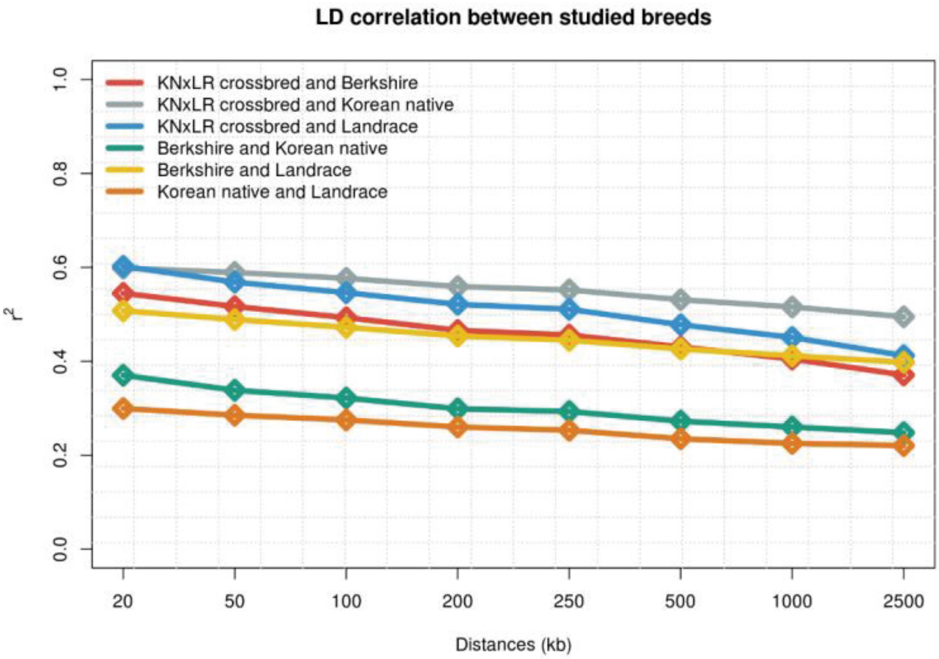
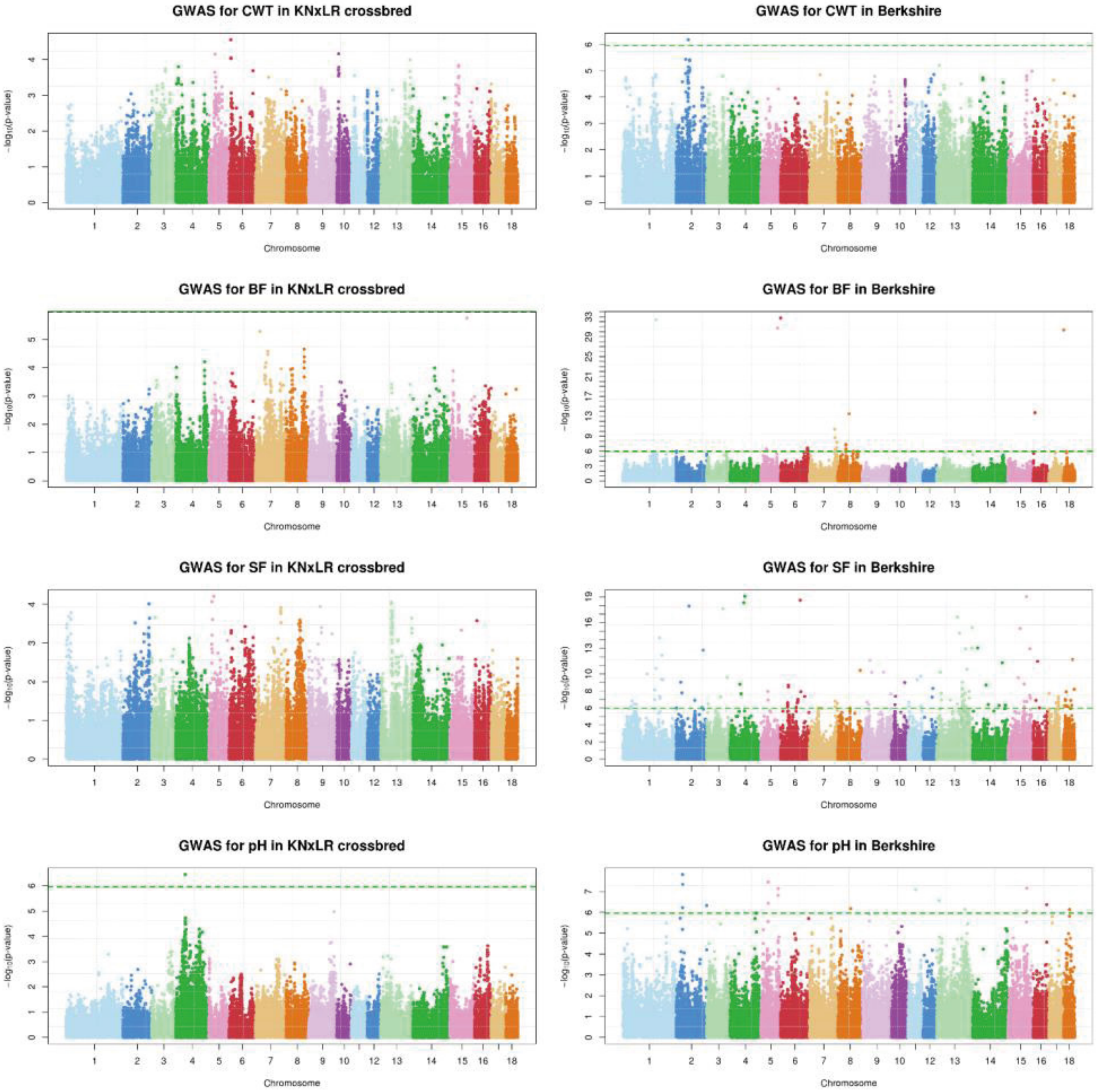


Fig. 4. Linkage disequilibrium (LD) by genetic distance: correlations between breeds. KN, Korean native pig; LR, Landrace.

differences between those of CB and Landrace were small. In terms of the correlations between breed pairs, those of KNP and Landrace, and KNP with BS, were weakest, and that of CB with KNP was strongest, followed by CB with Landrace (Fig. 4).

https://doi.org/10.5187/jast.2020.62.4.438

The GWAS, which used a mixed linear model (Fig. 5), showed that there were no significant SNPs for any trait in common, based on a significance threshold of 1.08 × 10^{-6}, between CB and BS (with Bonferroni correction applied). BS had significant SNPs for all traits, while CB had significant SNPs only for pH. In a Bayesian mixture model, the genetic contribution of CB to all markers was ~0%, while BS made a contribution of > 2.5% contribution to BF, and > 1% to pH and SF (Fig. 6).


GWAS for CWT in KNxLR crossbred

GWAS for CWT in Berkshire

GWAS for BF in KNxLR crossbred

GWAS for BF in Berkshire

GWAS for SF in KNxLR crossbred

GWAS for SF in Berkshire

GWAS for pH in KNxLR crossbred

GWAS for pH in Berkshire

Fig. 5. GWAS based on a mixed linear regression model of all traits in the Berkshire and crossbreed (CB) populations. CWT, carcass weight; BF, backfat thickness; SF, shear force; pH, muscle pH at 24 hours after slaughter; KN, Korean native pig; LR, Landrace.

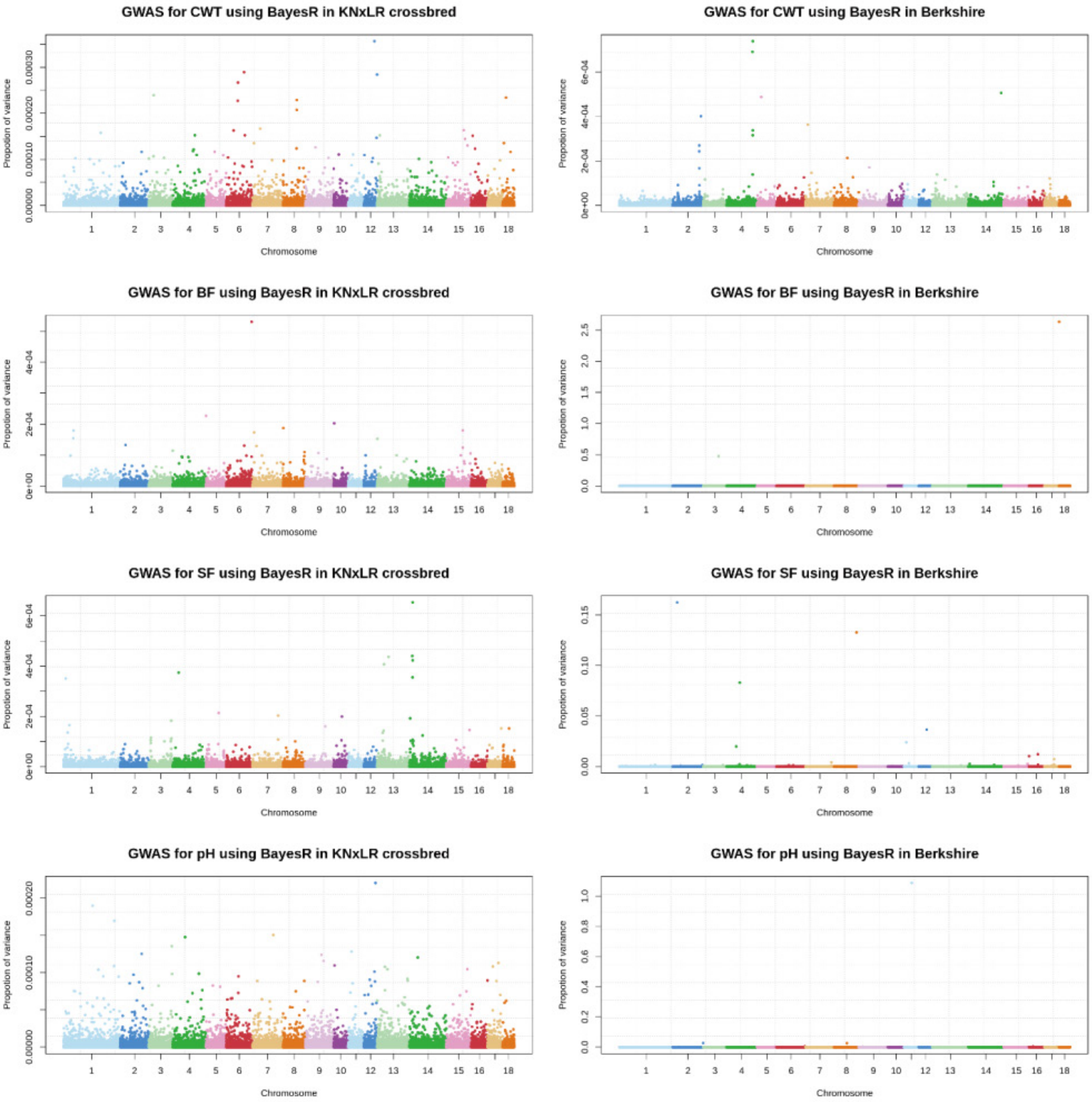


Fig. 6. GWAS based on a Bayesian mixture model of all traits in the Berkshire and crossbreed (CB) populations. CWT, carcass weight; BF, backfat thickness; SF, shear force; pH, muscle pH at 24 hours after slaughter; KN, Korean native pig; LR, Landrace.

Comparison for prediction accuracies of genomic estimated breeding value between admixed and single-breed reference populations

The prediction accuracy was zero or negative when using CB and BS as the reference and test populations, respectively. Increasing the size of the reference population did not affect the accuracy of the predictions for any trait except CWT, which increased by 6.26% between reference population sizes of 100 and 400. Use of the admixed population as the same pattern of reference increased the accuracy of the predictions for the BS population by 0.004, 0.013, 0.024, and 0.035 for CWT, BF, SF, and pH, respectively (Table 2; Fig. 7).

Using CB and BS as the test and reference populations, respectively, the prediction accuracy was zero or negative for all traits except CWT. The accuracy of the predictions for the CB population, when using the admixed population as the reference, increased marginally with increasing size of the reference population, but was not markedly higher compared to when the BS was the reference population.

DISCUSSION

Our GWAS results showed that the prediction accuracy of breeding value varied according to the degree to which a trait is favored. The prediction accuracy of single-breed and admixed reference population-based was shown to depend on the quantitative trait locus (QTL) and relationship among population [13]; the current study did not deal with QTLs, but carefully suggested that GWAS can also be associated with predict breeding value. Prediction accuracy with respect to genomic selection varies by both the LD between markers and QTLs, and genomic relationships (obtained by population structure analysis) [14]. In this study, the prediction accuracy for highly associated traits was higher when the admixed reference population was used, for example for BF, SF, and pH (but not CWT) in the BS population. In contrast, the CB population had no traits that were highly associated with those in the BS population, except pH. For BS, using both the single-breed and admixed reference populations, prediction accuracy for CWT was low compared to the other traits. In CB, the accuracy rates for CWT and pH were not markedly different when using the single-breed or admixed reference population; furthermore, these two traits were less strongly associated in the mixed linear model for the BS population. For CB, prediction accuracy for BF and SF was higher with a larger admixed reference population. When we use the admixed reference population that contains both test population breed in this study, relationship among them possibly

Table 2. Prediction accuracy for each scenario

Variables	Reference size	Accuracy of CB when using admixed reference				Accuracy of BS when using admixed reference				Accuracy of BS when using CB reference				Accuracy of CB when using BS reference			
		Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
BF	100	–0.022	0.135	–0.203	0.241	0.102	0.101	–0.104	0.240	0.010	0.051	–0.092	0.087	–0.042	0.057	–0.119	0.024
	200	0.036	0.091	–0.046	0.229	0.127	0.101	–0.032	0.238	0.016	0.062	–0.077	0.125	–0.033	0.060	–0.124	0.044
	300	0.027	0.056	–0.034	0.129	0.140	0.092	–0.003	0.259	0.003	0.067	–0.085	0.101	–0.041	0.047	–0.113	0.040
	400	0.057	0.079	–0.030	0.203	0.143	0.084	0.004	0.304	0.004	0.054	–0.087	0.079	–0.025	0.035	–0.074	0.038
CWT	100	0.012	0.095	–0.152	0.305	0.055	0.101	–0.073	0.222	0.001	0.063	–0.098	0.121	0.049	0.054	–0.070	0.123
	200	0.056	0.100	–0.067	0.310	0.064	0.078	–0.042	0.166	0.044	0.065	–0.050	0.158	0.043	0.038	–0.017	0.099
	300	0.035	0.096	–0.072	0.316	0.072	0.099	–0.073	0.253	0.056	0.076	–0.097	0.155	0.032	0.056	–0.031	0.153
	400	0.062	0.094	–0.058	0.300	0.066	0.087	–0.052	0.233	0.063	0.075	–0.074	0.202	0.043	0.055	–0.018	0.161
pH	100	–0.013	0.066	–0.089	0.113	0.131	0.160	–0.146	0.426	0.010	0.053	–0.059	0.096	–0.012	0.041	–0.075	0.051
	200	–0.013	0.092	–0.174	0.099	0.188	0.095	0.064	0.342	0.008	0.053	–0.073	0.091	–0.019	0.050	–0.088	0.069
	300	0.005	0.108	–0.181	0.193	0.160	0.108	–0.007	0.293	0.011	0.058	–0.071	0.098	–0.001	0.041	–0.067	0.075
	400	0.000	0.112	–0.179	0.198	0.258	0.102	0.084	0.438	0.006	0.065	–0.107	0.114	–0.001	0.042	–0.086	0.064
SF	100	0.038	0.072	–0.088	0.175	0.258	0.103	0.069	0.420	–0.027	0.066	–0.124	0.083	0.001	0.057	–0.076	0.085
	200	0.073	0.093	–0.044	0.181	0.293	0.072	0.206	0.394	–0.020	0.045	–0.116	0.030	–0.050	0.052	–0.118	0.015
	300	0.070	0.098	–0.058	0.240	0.299	0.053	0.228	0.391	–0.028	0.044	–0.086	0.046	–0.019	0.062	–0.135	0.058
	400	0.094	0.117	–0.071	0.249	0.339	0.068	0.242	0.443	–0.032	0.049	–0.115	0.045	–0.018	0.045	–0.110	0.048

CB, cross breed of Korean native pig and Lanrace; BS, Berkshire; SD, standard deviation; BF, backfat thickness; CWT, carcass weight; pH, muscle pH at 24 hours after slaughter; SF, shear force.

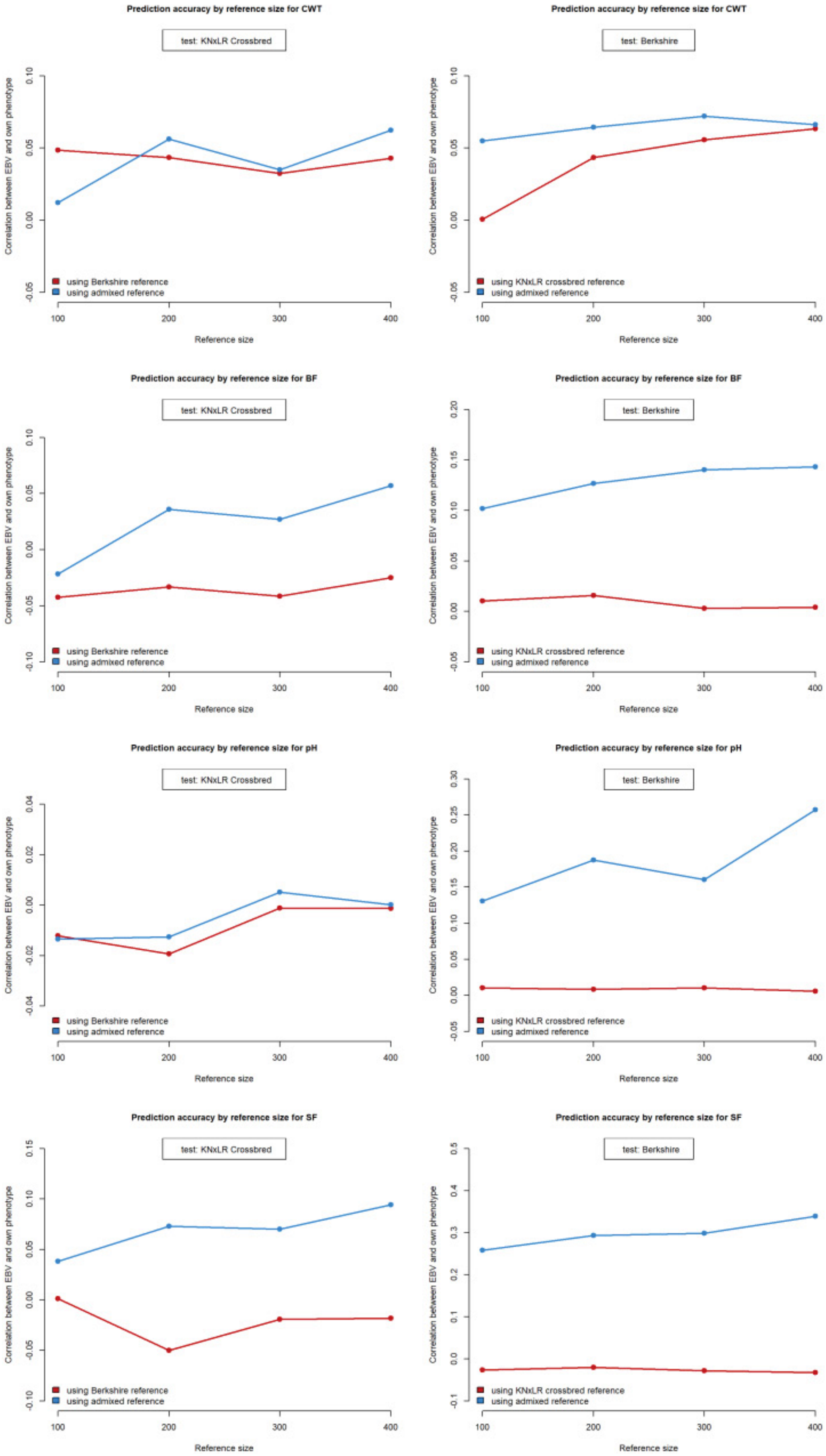


Fig. 7. Genomic estimated breeding value prediction accuracy for the Berkshire and crossbred population by reference population size and breed. CWT; carcass weight, BF; backfat thickness, SF; shear force, pH; muscle pH at 24 hours after slaughter, KN; Korean native pig, LR; Landrace.

be dense. As we mentioned above, following the Wientjes et al. [13], accuracy can be improved how they are related. The haplotypes for specific trait in BS also have a chance to affect accuracy on CB when predicting GEBV. Thus, use of an admixed reference population with traits associated with those in the reference population possibly improved the prediction accuracy of breeding value for test population.

A Bayesian approach is recommended for genomic predictions involving multi-breed populations [15]. A study in dairy cattle indicated that LD does not persist across breeds, except over short genetic distance (< 10 kb) [16]. Some of the putative markers have possibly linked with QTL in LD, while in across breed or multi-breed, low LD relatedness among breeds that already depicted in LD correlation has a small impact on prediction accuracy. Using the Bayesian method also allows us to focus on the QTL rather than LD [17]. As shown in Fig. 6 of this study, the BS population has an advantage with regard to markers with the high genetic contribution in BF, SF, and pH.

This study aimed to provide data that could facilitate improvement and conservation of the KNP. Due to the small size of the KNP population as the reference population, CB (included KNP genotype information) data was also used as the additional reference population. Nevertheless, this approach can be to improve prediction accuracy of breeding value and may facilitate phenotype development by following suggestions. Firstly, LD phases may have been broken down when breeds are crossed, which could be advantageous in some circumstances, for example by increasing the chance of uncovering causal variants for the target trait. Second, the crossing of genetically different populations results in genetic and phenotypic variance, which can lead to high performance animals than those of the previous generation. Though we couldn't find out putative markers or clear prediction accuracy patterns based on the CB reference, aspect of accuracy with CB using admixed population as a reference can provide valuable information when composing reference population. Furthermore, it is presumed that using the admixed population as a reference population contributes to EBV accuracy by sharing the phenotype associated Berkshire haplotype information while utilizing the relatedness of reference population with the test population.

The current pig improvement system of the Korean pig industry is relying on abroad seed stocks mainly on private farms and pig unions. For this reason, breeding plans and improvement goals are kept confidential and are not disclosed. To address these challenges, the National Institute of Animal Science has been running a Swine Genetic Improvement Network Program since 2008 (https://www.pignet.or.kr). This program aims to select Korean breeding pigs by establishing a system for genetic evaluation at the national level through exchanges and network connection of high-performance pigs among domestic pigs. Therefore, in order to establish a system for selecting and interacting with excellent pigs, it is considered that it is necessary to build an efficient reference population for estimating more accurate EBV as well as understanding the phenotype of the pigs on each farm. The result of this study is expected that the phenotype EBV estimation using the admixed reference population requires verification using various populations and additional samples, but it can provide useful information for the genetic improvement of KNP along with a Swine Genetic Improvement Network Program.

REFERENCES

1. Zhong S, Dekkers JCM, Fernando RL, Jannink JL. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics. 2009;182:355-64.
2. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: genomic selection in dairy cattle: progress and challenges. J Dairy Sci. 2009;92:433-43.

3. Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K, Goddard ME. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol. 2009;41:51.
4. Goddard M. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 2009;136:245-57.
5. Moghaddar N, Swan AA, van Der Werf JHJ. Comparing genomic prediction accuracy from purebred, crossbred and combined purebred and crossbred reference populations in sheep. Genet Sel Evol. 2014;46:58.
6. Hidalgo AM, Bastiaansen JWM, Lopes MS, Harlizius B, Groenen MAM, de Koning DJ. Accuracy of predicted genomic breeding values in purebred and crossbred pigs. G3 (Bethesda). 2015;5:1575-83.
7. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Human Genet. 2011;88:76-82.
8. Bhuiyan MSA, Lim D, Park M, Lee SH, Kim YK, Gondro C, et al. Functional partitioning of genomic variance and genome-wide association study for carcass traits in Korean Hanwoo cattle using imputed sequence level SNP data. Front Genet. 2018;9:217.
9. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
10. R Core Team. R: a language and environment for statistical computing [Internet]. 2013 [cited 2020 May 2]. http://finzi.psych.upenn.edu/R/library/dplR/doc/intro-dplR.pdf
11. Gilmour AR, Gogel BJ, Cullis BR, Welham SJ, Thompson R. ASReml user guide release 4.1 structural specification [Internet]. 2015 [cited 2020 May 2]. https://www.hpc.iastate.edu/sites/default/files/uploads/ASREML/UserGuideStructural.pdf.
12. Lourenco DA, Fragomeni BO, Tsuruta S, Aguilar I, Zumbach B, Hawken RJ, et al. Accuracy of estimated breeding values with genomic information on males, females, or both: an example on broiler chicken. Genet Sel Evol. 2015;47:56.
13. Wientjes YCJ, Calus MPL, Goddard ME, Hayes BJ. Impact of QTL properties on the accuracy of multi-breed genomic prediction. Genet Sel Evol. 2015;4s7:42.
14. Daetwyler HD, Kemper KE, van der Werf JHJ, Hayes BJ. Components of the accuracy of genomic prediction in a multi-breed sheep population. J Anim Sci. 2012;90:3375-84.
15. Kemper KE, Reich CM, Bowman PJ, Vander Jagt CJ, Chamberlain AJ, Mason BA, et al. Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. Genet Sel Evol. 2015;47:29.
16. de Roos APW, Hayes BJ, Goddard ME. Reliability of genomic predictions across multiple populations. Genetics. 2009;183:1545-53.
17. Meuwissen T, Hayes B, Goddard M. Genomic selection: a paradigm shift in animal breeding. Anim Front. 2016;6:6-14.

https://doi.org/10.5187/jast.2020.62.4.438