JABG. 2023 December, 7(4)157-166
https://doi.org/10.12972/jabng.20230017

Print ISSN 1226-5543
Online ISSN 2586-4297

Research Article

Effect of population structure on the accuracy of genomic prediction for carcass traits in Hanwoo cattle: a simulation study

Bomin Kim[1,†], Waruni Ekanayake[1,†], Jeongwoen Shin[2], Yoonsik Kim[1], Dooho Lee[1], Yeongkuk Kim[3], Seung Hwan Lee[1*]

[1]Division of Animal and Dairy Sciences, College of Agriculture and Life Science, Chungnam National University

[2]TNT Research Co, Jeonju, 54810, South Korea

[3]Quantomic research & solution, Daejeon, 34134, South Korea

***Corresponding author: Seung Hwan Lee,** Division of Animal & Dairy Science, Chungnam National University, Daejeon, Korea, E-mail: slee46@cnu.ac.kr

[†] These authors contributed equally to this work.

ABSTRACT

Improvement of quantitative traits in livestock is an essential goal of animal breeding. For the achievement of this goal, an accurate breeding value estimation is needed. The genetic relationship between reference and test groups is a key factor in determining the accuracy of genomic estimated breeding value (GEBV) thus, the structure of the reference population is crucial for efficient genomic selection. By the number of sharing parents, the population structure can be divided into half-sibling and full-sibling families. Also, the population structure of Hanwoo cattle primarily consists of half-sibling families because of the production system. Therefore, comparing half-sibling and full-sibling families is challenging in the Hanwoo cattle population, yet an important issue in the direction of breeding strategy in Korea. The objective of this study was to compare the accuracy of GEBV between different family structures and investigate efficient family size in the reference population using simulated data. 6 different populations were simulated using QMSim software, and the individuals in the last generations were separated into reference and test groups. The GEBV was calculated using BLUPF90 software. Practical accuracy was between 0.36-0.52 in half-sibling families and 0.55-0.77 in full-sibling families. The increase rate of accuracy was highest at the sibling size of 20, with practical accuracy of 0.52 in half-sibling families and 0.77 in full-sibling families. As a result, the most efficient population structure for genomic prediction was a sibling size of 20 in a full-sibling family.

Keywords: Genomic estimated breeding value, Genomic selection, Hanwoo, Population structure, Reference population

INTRODUCTION

The main goal of animal breeding is to select individuals with good genetic merit for economic traits as parents and to produce offspring with desired traits (C.M. Dekkers, 2012). Traits of interest are expressed in terms of values such as phenotype and breeding value. Breeding value explains the expected offspring performance of an individual by variable information and is therefore an efficient source for selection in animal breeding. However, knowing the animals' true breeding value or real genetic potential is impractical (Oldenbroek & Van der Waaij, 2020). Thus, the researcher's objective is to predict the real genetic potential with Estimated Breeding Value (EBV). Genomic best linear unbiased prediction (GBLUP) is one of the methods used to estimate genomic breeding value (Henderson, 1984). GBLUP accepts genomic information to construct relationship information and estimates breeding value by applying the correlation and path coefficient methods to Mendelian inheritance (Zhang et al., 2021).

Received December 06, 2023 **Revised** December 20, 2023 **Accepted** December 20, 2023

Selection response is a value that indicates the genetic difference between generations. In other words, selection response represents the power of selection. The selection response should be of considerable value to successfully achieve the goal of animal breeding in a short period. Selection response is influenced by selection intensity, genetic variance, generation interval, and the accuracy of estimated breeding value (EBV) (S. H. Lee et al., 2015). Using the appropriate reference population, for instance, animals who have close genetic relationships with the test population, is vital for the higher accuracy of the EBV (Clark et al., 2012). The importance of population structure in reference data has been demonstrated in previous studies (Clark & Van Der Werf, 2013). The EBVs calculated using closely related individuals as a reference showed a higher correlation with phenotypic information than those using individuals distantly related to the animals in the test population. The population structure consists mainly of half-siblings and full-siblings. Half-siblings and full-siblings explain the relationship between individuals sharing one parent and both parents respectively. In Hanwoo cattle, the half-sibling family structure is relatively more prominent than the full-sibling family structure because the Hanwoo production system consists of only a few males, known as Korean proven bulls (KPN), that are used for mating (S.-H. Lee et al., 2014). Although this unique population structure of Hanwoo cattle makes comparing half-sibling and full-sibling families challenging, it provides a distinctive opportunity to study the effect of population structure on the accuracy of EBV. It is worthwhile to investigate how the population structure would affect the EBV accuracy. Moreover, examining the impact of the efficient number of offspring in a family in reference data may play an important role in the selection of effective breeding strategies in the Hanwoo population. The objective of this study was to compare the accuracy of GEBV between different population structures and to investigate the efficient family size in reference populations using simulated data.

MATERIALS AND METHODS

Simulation

Phenotype and genotype data were simulated using QMSim software (Sargolzaei & Schenkel, 2009) to mimic the population structure and genotypes of carcass traits in Hanwoo cattle (Table 1). A 50K marker density panel was simulated to generate bi-allelic markers evenly distributed across 29 autosomes with a length of 2,349 Mbp. A total of 6 different Hanwoo cattle populations were simulated to generate half-sibling or full-sibling data with varying numbers of siblings in the family.

Table 1. Population structure and simulation parameters(continued).

Parameter	Half-sibling	Full-sibling
Step 1: Historical Generations		
Number of generations (size) – phase 1	1,000 (1,000)	
Number of generations (size) – phase 2	20 (3,150, 6,150)	1,020 (3,150)
Number of males in the last historical generation	150	
Step 2: Recent Generations		
Number of founder males from the HG	150	75
Number of founder females from the HG	1,500, 3,000, 6,000	3,000
Number of offspring per dam		10, 20, 40
Number of generations	3	
Ratio of male progenies	50%	
Mating design	EBV, positive assertive	
Replacement ratio for males	30%	

Table 1. Population structure and simulation parameters.

Parameter	Half-sibling	Full-sibling
Step 2: Recent Generations		
Replacement ratio for females	30%	
Selection design	High EBV	
Culling design	Low EBV	
EBV estimation method	BLUP	
Trait heritability	0.45	
QTL heritability	0.2	
Phenotypic variance	1.0	
Genome		
Number of chromosomes	28	
Total length	2349	
Number of markers	1724	
Marker distribution	Random	
Number of QTLs	100	
QTL distribution	Random	
MAF for markers	Equal	
MAF for QTL	Equal	
Additive allelic effects for markers		
Additive allelic effects for QTL	Gamma distribution (shape = 0.4)	
Complete LD in the first HG	Mq	
Rate of missing marker genotypes	0.01	
Rate of marker genotyping error	0.005	

The historical population (HP) phase 1 was created by randomly mating 1,000 individuals over 1,000 generations to create linkage disequilibrium (LD) in the population. The size of HP phase 2 was then increased over 20 generations to create founders for more recent generations. The proportion of sires was maintained at less than 10% across all populations in recent generations to mimic the sex ratio of the Hanwoo cattle population used for mating. The number of founders and litter size were altered between populations to create populations with different family structures. In the half-sibling population, the number of offspring per dam was fixed at one, and the number of half-siblings in each family was controlled by the number of dams in the recent population. However, in the full-sibling population, the number of offspring per dam was set at 10, 20, and 40 to create a population corresponding to the full-sibling family size (Figure 1).

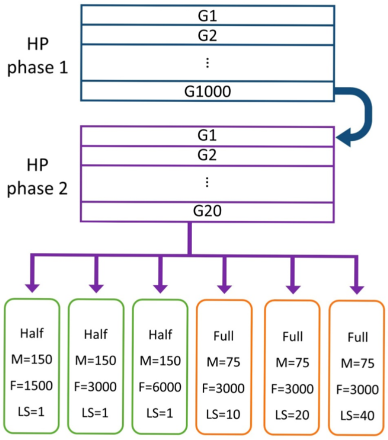

Figure 1. Diagram of simulated population structure; HP: historical population, Half: half-sib, Full: full-sib, M: number of male progenies, F: number of female progenies, LS: Litter size.

Structure of test and reference data

Individuals in the last generation of the recent population were used to create test and reference data. The size of the population and the number of siblings within the family are described in Table 2. In each population, 100 individuals were randomly selected for the test data, and the rest were used as reference data.

Table 2. Population size in the last generation of the recent population.

Number of siblings	Half-sibling	Full-sibling
10	1,500	3,000
20	3,000	3,000
40	6,000	3,000

Genotype relationship matrix

Genotype quality control (Hardy-Weinberg test: 0.0001, genotyping rate: 0.1, minor allele frequency: 0.01) was performed using the PLINK 1.9 software (Purcell et al., 2007) for the genomic data used for the analysis, and the GCTA software (Yang et al., 2011) was used to calculate genomic relationship matrix (GRM) which incorporates genomic level relationships between individuals. GRM was calculated based on genomic information using the equation:

$$G = \frac{(M-P)(M-P)'}{2\sum_{j=1}^{m} p_j(1-\ p_j)}$$

Where M is the marker allele matrix for each individual and P is a matrix comprising the frequency of the second allele, p_j. By dividing $(M-P)(M-P)'$ by $\sum_{j=1}^{m} p_j(1-\ p_j)$, G matrix becomes similar to the pedigree-based relationship matrix and can be used as relationship matrix (Lourenco et al., 2020). The relationships between reference and test groups were considered to calculate the variance of relatives. Figure 2 shows the considered elements in the GRM.

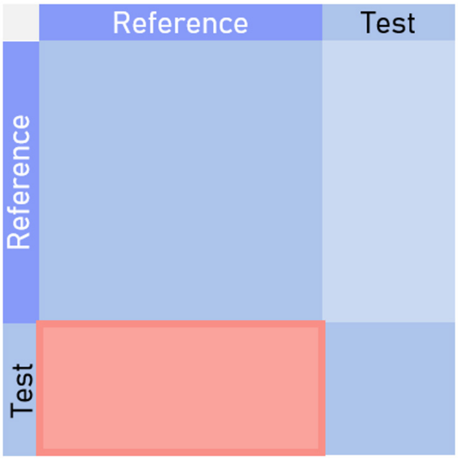

Figure 2. Structure of GRM. The elements in the red square were used to calculate the variance of relatives.

Genomic prediction

The genomic prediction model considered sex as a fixed effect and animal effects as a random effect. Therefore, the GEBV was calculated based on the following model (Henderson, 1984):

$$y = Xb + Wu + e$$

where y, b, u, and e are the vectors of the phenotypes, fixed effects, animal effects, and residual errors, respectively. X is the incidence matrix for fixed effects, and W is the incidence matrix for animal effects. This model can also be expressed as:

$$\begin{bmatrix} X'X & X'W \\ W'X & W'W + G^{-1}\frac{\sigma^2{}_e}{\sigma^2{}_u} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ W'y \end{bmatrix}$$

where $\sigma^2{}_{\mathrm{e}}$, $\sigma^2{}_{\mathrm{u}}$ are the residual variance, the additive genetic variance respectively and G is the genomic relationship matrix.

Genetic variance and residual variance were calculated by AIREMLF90 based on reference data and GEBV was calculated by BLUPF90 (Aguilar et al., 2018).

The accuracy of genomic prediction (practical accuracy) was calculated for each half-sibling and full-sibling population. The accuracy of GEBV in each population was calculated by the Pearson correlation coefficient (r^2) between phenotypes and GEBVs of individuals in test data set by equation (Benesty et al., 2009):

$$r_{a,b} = \frac{E(ab)}{\sigma_a \sigma_b}$$

where a and b are random variables, $E(ab)$ is the cross-correlation between a and b, and $\sigma_{\mathrm{a}} = E(a^2)$ and $\sigma_{\mathrm{b}} = E(b^2)$ are the variances of the random variables a and b.

The expected accuracy of genomic prediction was calculated in two ways. First expected accuracy (theoretical accuracy) was calculated by using parameters from mixed-model equations (Henderson, 1975):

$$r_{EBV} = \sqrt{1 - \frac{PEV}{var(g)}}$$

where prediction error variance (PEV) was calculated using the true breeding value of test individuals. In addition, the second expected accuracy (correlation accuracy) was calculated considering the family structure of populations (Oldenbroek & Van der Waaij, 2020):

$$r_{half} = \sqrt{\frac{\frac{1}{16}nh^2}{1 + (n-1)(\frac{1}{4}h^2 + c^2{}_{HS})}}, \quad r_{full} = \sqrt{\frac{\frac{1}{4}nh^2}{1 + (n-1)(\frac{1}{2}h^2 + c^2{}_{FS})}}$$

where n, h^2 and c^2 are the number of siblings, heritability, and common environmental effect respectively. The common environmental effect was assumed to be zero.

RESULTS

Table 3 depicts the variance of relatives of half-sibling and full-sibling families with different levels of EBV accuracy. Individuals in the test group were divided into three groups with a high, middle, and low theoretical accuracy of EBV by the first and third quartile numbers of the theoretical accuracy. The variance of relatives displayed higher values with higher EBV accuracy. Moreover, full-sibling families showed a greater overall variance of relatives compared to half-sibling families.

Table 3. Variance of relatives of half-sibling and full-sibling families with different levels of EBV accuracy.

EBV Accuracy	Variance of Relatives	
	Half-sibling	Full-sibling
High accuracy	0.0013	0.0073
Medium accuracy	0.0009	0.0043
Low accuracy	0.0005	0.0023

In test data, full-sibling families exhibited higher levels of the overall mean of practical accuracy compared to half-sibling families, as shown in Figure 3. Full-sibling families with both 20 and 40 siblings equally recorded the highest value of practical accuracy of 0.77 and the lowest accuracy (0.36) was observed corresponding to the half-sibling families with 10 siblings.

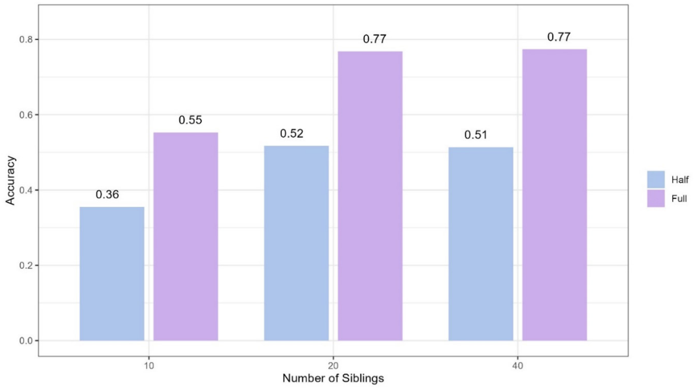

Figure 3. Practical genomic prediction accuracy in half-sibling and full-sibling family.

Figure 4 presents the theoretical accuracy, correlation accuracy, and practical accuracy concerning the half-sibling population in the test data. For test data, the mean theoretical accuracy was 0.42, 0.49, and 0.55 for families with 10, 20, and 40 siblings respectively. Correlation accuracies were observed as 0.37, 0.42, and 0.46 and practical accuracies were 0.36, 0.52, and 0.51 for 10, 20, and 40 sibling families respectively. In 10 and 40 sibling families, the practical accuracy was observed to be lower than the theoretical accuracy. However, in 20 sibling families, the mean of practical accuracy exhibited a slightly higher value compared to the theoretical accuracy. The theoretical accuracy displayed the highest figure with 40 sibling families (0.55), in contrast, 20 sibling families were observed to have the highest practical accuracy (0.52).

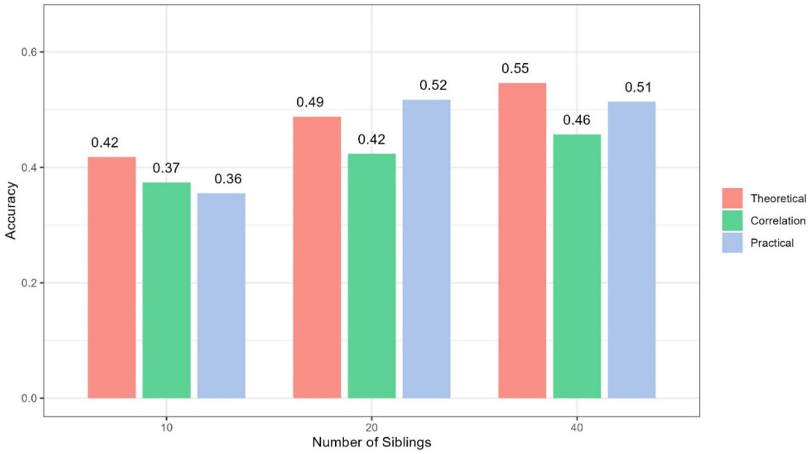

Figure 4. Expected and practical genomic prediction accuracy in half-sibling family.

The theoretical, correlation, and practical accuracies of full-sibling families in the test population (Figure 5) showed the same pattern as half-sibling families with moderate changes in the accuracy levels. Families with 10, 20, and 40 siblings demonstrated similar levels of 0.63, 0.67, and 0.69, theoretical accuracies respectively. Nevertheless, the practical accuracy showed an increasing trend from 0.55 for 10 sibling families to 0.77 for 20 sibling families and then stayed constant from 20 to 40 sibling families. The mean practical accuracy displayed a lower-than-expected value in 10 sibling families. Correlation accuracy slightly increased from 0.61 to 0.68, from 10 to 40 sibling families.

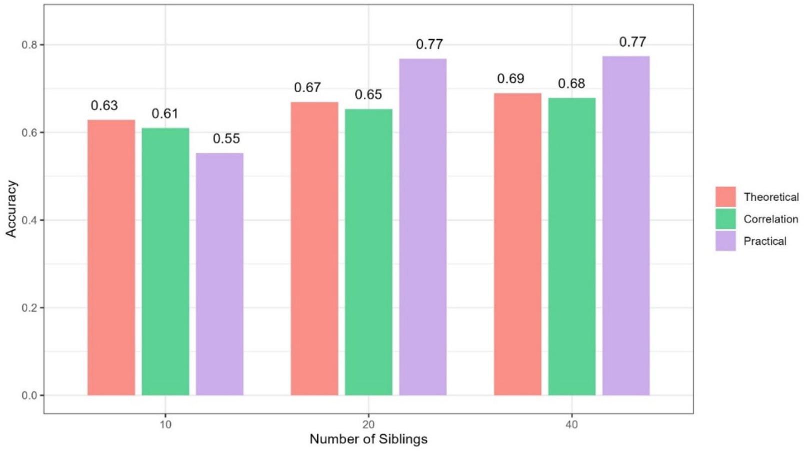

Figure 5. Expected and practical genomic prediction accuracy in full-sibling family.

DISCUSSION

The success of genomic selection in cattle breeding programs significantly relies on understanding the impact of the structure of the reference population on prediction accuracy. The objective of this study was to evaluate the GEBV accuracy of animals from different population structures and family sizes in reference data. Half-sibling and full-sibling populations and families with 10, 20, and 40 siblings were used and estimated breeding values from each population were compared. The practical accuracy of EBV showed higher accuracy in full-sibling families than in half-sibling families. This result is expected to be the effect of different family structures in the reference population. Recent studies(Clark et al., 2012; S. H. Lee et al., 2017) have shown the effect of population structure on genomic prediction accuracy. When animals in the main population are in a close relationship with reference animals, the genomic prediction accuracy tends to be higher. Close relationships imply that more SNP-phenotype associations estimated are shared by both reference and test populations thus genomic information of reference data better explains the genomic information of the individuals in the test data (Oldenbroek & Van der Waaij, 2020). Assuming that parents of the last generation are not related, the relationship of individuals in a half-sibling family is either 0 or 0.25, and the relationship of individuals in a full-sibling family is 0 or 0.5. As the individuals in simulation data came from a single historical population, this value may be larger. However, the gap in the relationship between siblings and unrelated individuals is still different between half-sibling and full-sibling families. As a result, the variance of relatives in half-sibling families is smaller than in full-sibling families. Hence, genomic information of full-sibling-based reference is comparatively more effective in explaining genomic information of the test data resulting in higher prediction accuracy.

For a better understanding of the effect of family size on prediction accuracy, the expected and practical accuracies were compared between different family sizes of 10, 20, and 40 siblings. The expected accuracy of GEBV in half-sibling families exhibited an increasing tendency as the number of siblings in the family increased. With a large number of siblings in the family, there is more information available to predict the breeding values leading to higher prediction accuracy (Chu et al., 2019). However, unlike the expectation, practical accuracy did not show a significant difference when the number of siblings was increased from 20 to 40. While the main reason for this observation is expected to be the proportion of explainable information, it can also be a result of the difference in data. Assuming that every individual in the test data came from a different family, the proportion of siblings in the test data in reference was about 0.33 in all populations of half-sibling families. In this case, the size of reference data was assumed to have a low impact on the accuracy of GEBV (Habier et al., 2010). Nevertheless, the low value of practical accuracy in a half-sibling family with 10 siblings may be the effect of small reference data.

For full-sibling families, unlike in half-sibling families, the proportion of siblings in test data in reference data increased with the number of siblings in the population. However, in the full-sibling population, the practical accuracy did not show a significant difference between populations with 20 and 40 siblings. This may indicate that the reference data constructed with the number of siblings above 20 in the family has the same ability to explain the genomic information of the test data.

Based on the results of this study, is it suggested that using reference data with a full-sibling family structure had a higher advantage in genomic prediction than the reference data with a half-sibling structure. Moreover, in both half and full-sibling families, the accuracy of genomic prediction exhibited the best efficiency at a sibling size of 20 in a family.

This study employed simulation data to examine the effect of population structure and family size of reference data on prediction accuracy. For a deeper understanding of the impact of the reference population structure on GEBV accuracy, it is worthwhile to investigate the concept using real data.

CONCLUSION

The results from simulation data demonstrated that in genomic prediction, reference data with a full-sibling family structure is more advantageous than reference data with a half-sibling structure. Moreover, sibling size 20 in a family provided the highest accuracy of genomic prediction.

CONFLICT OF INTERESTS

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

ACKNOWLEDGEMENTS

This research was financially supported by the Ministry of Small and Medium-sized Enterprises (SMEs) and Startups (MSS), Korea, under the “Regional Specialized Industry Development Program+(R&D, S3370836)” supervised by the Korea Technology and Information Promotion Agency for SMEs (TIPA).

Bomin belongs to the Artificial Intelligence Convergence Research Center as a Master's student at Chungnam National University. Their research support was supported by the Institute of Information & communications Technology Planning & evaluation (IITP) grant funded by the Korean government (MSIT) (No.RS-2022-00155857, Artificial Intelligence Convergence Research Center (Chungnam National University)).

REFERENCES

Aguilar, I., Tsuruta, S., Masuda, Y., Lourenco, D. A. L., Legarra, A., & Misztal, I. (2018). BLUPF90 suite of programs for animal breeding with focus on genomics. In *10th World Congress on Genetics Applied to Livestock Production*.

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson Correlation Coefficient. In: Noise Reduction in Speech Processing. In *Springer Topics in Signal Processing* (Vol. 2).

Chu, T. T., Bastiaansen, J. W. M., Berg, P., & Komen, H. (2019). Optimized grouping to increase accuracy of prediction of breeding values based on group records in genomic selection breeding programs. *Genetics Selection Evolution*, *51*(1). https://doi.org/10.1186/s12711-019-0509-z

Clark, S. A., Hickey, J. M., Daetwyler, H. D., & van der Werf, J. H. J. (2012). The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genetics, Selection, Evolution : GSE*, *44*. https://doi.org/10.1186/1297-9686-44-4

Clark, S. A., & Van Der Werf, J. (2013). Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values. *Methods in Molecular Biology*, *1019*. https://doi.org/10.1007/978-1-62703-447-0_13

C.M. Dekkers, J. (2012). Application of Genomics Tools to Animal Breeding. *Current Genomics*, *13*(3). https://doi.org/10.2174/138920212800543057

Habier, D., Tetens, J., Seefried, F. R., Lichtner, P., & Thaller, G. (2010). The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genetics Selection Evolution*, *42*(1). https://doi.org/10.1186/1297-9686-42-5

Henderson, C. R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics*, *31*(2). https://doi.org/10.2307/2529430

Henderson, C. R. (1984). Applications of Linear Models in Animal Breeding. In *Univ. of Guelph, Guelph, Canada.*

Lee, S. H., Cho, Y. M., Lee, J. H., & Oh, S. J. (2015). Implementation of genomic selection in Hanwoo breeding program. *Korean Journal of Agricultural Science*, *42*(4). https://doi.org/10.7744/cnujas.2015.42.3.397

Lee, S. H., Clark, S., & Van Der Werf, J. H. J. (2017). Estimation of genomic prediction accuracy from reference populations with varying degrees of relationship. *PLoS ONE*, *12*(12). https://doi.org/10.1371/journal.pone.0189775

Lee, S.-H., Park, B.-H., Sharma, A., Dang, C.-G., Lee, S.-S., Choi, T.-J., Choy, Y.-H., Kim, H.-C., Jeon, K.-J., Kim, S.-D., Yeon, S.-H., Park, S.-B., & Kang, H.-S. (2014). Hanwoo cattle: origin, domestication, breeding strategies and genomic selection. *Journal of Animal Science and Technology*, *56*(1). https://doi.org/10.1186/2055-0391-56-2

Lourenco, D., Legarra, A., Tsuruta, S., Masuda, Y., Aguilar, I., & Misztal, I. (2020). Single-step genomic evaluations from theory to practice: using SNP chips and sequence data in blupf90. *Genes*, *11*(7). https://doi.org/10.3390/genes11070790

Oldenbroek, K., & Van der Waaij, L. (2014). Animal Breeding and Genetic Improvement. *UPA. Twenty-Five Years of Progressive Agrarian Unionism.*

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3). https://doi.org/10.1086/519795

Sargolzaei, M., & Schenkel, F. S. (2009). QMSim: A large-scale genome simulator for livestock. *Bioinformatics*, *25*(5). https://doi.org/10.1093/bioinformatics/btp045

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, *88*(1). https://doi.org/10.1016/j.ajhg.2010.11.011

Zhang, J., Liu, F., Reif, J. C., & Jiang, Y. (2021). On the use of GBLUP and its extension for

GWAS with additive and epistatic effects. *G3: Genes, Genomes, Genetics*, *11*(7). https://doi.org/10.1093/g3journal/jkab122

AUTHORS INFORMATION

Seung Hwan Lee: https://orcid.org/0000-0003-1508-4887

Bomin Kim: https://orcid.org/0000-0003-4289-0796

Waruni Ekanayake: https://orcid.org/0000-0003-2988-0315

Jeongwoen Shin: https://orcid.org/0000-0001-7131-4080

Yoonsik Kim: https://orcid.org/0000-0002-5318-7521

Dooho Lee: https://orcid.org/0000-0002-2174-7897

Yeongkuk Kim: https://orcid.org/0000-0002-6530-2304