JAST (Journal of Animal Science and Technology) TITLE PAGE

Upload this completed form to website with submission

ARTICLE INFORMATION	Fill in information in each box below
Article Type	Research Article
Article Title (within 20 words without abbreviations)	Effect of Breed Composition in Genomic Prediction Using Crossbred Pig Reference Population
Running Title (within 10 words)	Effect of Breed Composition in Genomic Prediction
Author	Euiseo Hong [first_author] 1, Yoonji Chung [first_author] 2, Phuong Thanh N. Dinh3, Yoonsik Kim2, Suyeon Maeng4, Young jae Choi3, Jaeho Lee3, Woonyoung Jeong1, Hyunji Choi5, Seung Hwan Lee4
Affiliation	1 Department of Bio-Big Data and Precision Agriculture, Chungnam National University, Daejeon 34134, Korea, Republic of 2 Institute of Agricultural Science, Chungnam National University, Daejeon 34134, Korea, Republic of 3 Department of Bio-AI Convergence, Chungnam National University, Daejeon 34134, Korea, Republic of 4 Division of Animal & Dairy Science, Chungnam National University, Daejeon 34134, Korea, Republic of 5 Division of Animal Genomics and Bioinformatics, National Institute of Animal Science, Jeonbuk-do 55365, Korea, Republic of
ORCID (for more information, please visit https://orcid.org)	Euiseo Hong (https://orcid.org/0000-0003-3078-2560) Yoonji Chung (https://orcid.org/0000-0002-6906-6468) Phuong Thanh N. Dinh (https://orcid.org/0000-0002-3057-0210) Yoonsik Kim (https://orcid.org/0000-0002-5318-7521) Suyeon Maeng (https://orcid.org/0000-0001-9903-3803) Young jae Choi (https://orcid.org/0000-0003-1540-6970) Jaeho Lee (https://orcid.org/0009-0008-7721-8135) Woonyoung Jeong (https://orcid.org/0009-0002-7572-1382) Hyunji Choi (https://orcid.org/0000-0001-9782-6586) 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 State funding sources (grants, funding sources, equipment, and supplies). Include name and number of grant if available.	*This work was supported by Chungnam National University.*
Acknowledgements	
Availability of data and material	
Authors' contributions Please specify the authors' role using this form.	Conceptualization: Hong E, Lee SH Data curation: Hong E, Chung Y, Choi H, Lee SH Formal analysis: Hong E, Chung Y, Lee SH Methodology: Hong E, Chung Y, Dinh PTN, Kim Y Software: Hong E, Maeng S, Choi YJ, Lee J, Jeong W Validation: Hong E Investigation: Hong E Writing - original draft: Hong E Writing - review & editing: Hong E, Chung Y, Dinh PTN, Kim Y, Maeng S, Choi YJ, Lee J, Jeong W, Choi H, Lee SH

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

CORRESPONDING AUTHOR CONTACT INFORMATION

For the corresponding author (responsible for correspondence, proofreading, and reprints)	**Fill in information in each box below**
First name, middle initial, last name	Seung Hwan Lee
Email address – this is where your proofs will be sent	genomicselection46@gmail.com
Secondary Email address	
Address	
Cell phone number	+82-10-5337-1617
Office phone number	+82-42-821-5878
Fax number	

Effect of Breed Composition in Genomic Prediction Using Crossbred Pig Reference Population

Abstract

In contrast to conventional genomic prediction, which typically targets a single breed and circumvents the necessity for population structure adjustments, multi-breed genomic prediction necessitates accounting for population structure to mitigate potential bias. The presence of this structure in multi-breed datasets can influence prediction accuracy, rendering proper modeling crucial for achieving unbiased results. This study aimed to address the effect of population structure on multi-breed genomic prediction, particularly focusing on crossbred reference populations. The predictive accuracy of genomic models was assessed by incorporating genomic breed composition (GBC) or principal component analysis (PCA) into the genomic best linear unbiased prediction (GBLUP) model. The accuracy of five different genomic prediction models was evaluated using data from 354 Duroc × Korean native pig crossbreds, 1,105 Landrace × Korean native pig crossbreds, and 1,107 Landrace × Yorkshire × Duroc crossbreds. The models tested were GBLUP without population structure adjustment, GBLUP with PCA as a fixed effect, GBLUP with GBC as a fixed effect, GBLUP with PCA as a random effect, and GBLUP with GBC as a random effect. The highest predictive accuracies for backfat thickness (0.59) and carcass weight (0.50) were observed in Models 1, 4, and 5. In contrast, Models 2 and 3, which included population structure as a fixed effect, exhibited lower accuracies, with backfat thickness accuracies of 0.40 and 0.53 and carcass weight accuracies of 0.34 and 0.38, respectively. These findings suggest that in multi-breed genomic prediction, the most efficient and accurate approach is either to forgo adjusting for population structure or, if adjustments are necessary, to model it as a random effect. This study provides a robust framework for multi-breed genomic prediction, highlighting the critical role of appropriately accounting for population structure. Moreover, our findings have important implications for improving genomic selection efficiency, ultimately enhancing commercial production by optimizing prediction accuracy in crossbred populations.

Keywords: genomic breed composition, genomic prediction, multi-breed genomic prediction, population structure

Introduction

Accurate prediction of genomic breeding values is a critical component of successful genomic selection, which requires a sufficiently large reference population to reliably estimate marker effects [1]. However, small populations, such as Jersey cattle, often pose challenges owing to the limited reference populations of progeny-tested bulls, leading to less reliable genomic breeding values [2]. Consequently, genetic progress is restricted in breeds without a large reference population. One approach to addressing this limitation is across-breed prediction, which involves the use of a large reference dataset from another breed [3]. Another approach is multi-breed prediction, which combines data from multiple breeds to create a larger, more comprehensive dataset [3]. Both approaches can enhance prediction accuracy for smaller breeds, helping them become more competitive while minimizing the additional costs associated with genotyping and phenotyping.

Empirical studies have demonstrated that the accuracy of across-breed genomic prediction is often near zero and that combining multiple breeds has not yielded significant improvements in accuracy [3, 4]. However, these methods remain promising, particularly when combined with strategies that account for population structure and other sources of variation [5, 6]. Addressing population structure, also referred to as population stratification, is critical for genomic prediction across different breeds. Population structure arises from differences in allele frequencies between subpopulations, which may result from geographic separation, or natural or artificial selection [7]. These differences can lead to spurious marker-trait associations [8, 9], potentially inflating estimates of genomic heritability [10] and introducing bias into genomic prediction accuracy [6].

To mitigate the effects of population structure, it is important to model it appropriately within genomic prediction models, particularly when combining data from multiple breeds. A common method involves incorporating principal components (PCs) derived from genomic data as a fixed effect in the prediction model [7]. However, incorporating PCs as a fixed effect can result in over-correction, as these components are derived from the genomic relationship matrix used in genomic prediction [11]. To address this limitation, in this study, PCs were modeled as a random effect to capture population structure without confounding the genomic relationship matrix. The predictive accuracy of these models was compared with those of models in which PCs were excluded. Additionally, breed composition, another explanatory factor for population structure, was modeled as either a fixed or random effect to adjust for population structure.

In this study, we evaluated the accuracy of genomic predictions using models that incorporated breed composition and PCs as fixed and random effects and compared the results with those of a baseline model. This study aimed to determine whether accounting for population structure using breed composition or PCs can improve genomic prediction accuracy. The findings of this study may provide valuable insights into optimizing genomic prediction models for populations with complex or diverse structures.

Materials and Methods

Animals, genotypes, and phenotypes

The genotype dataset comprised data from 354 Duroc × Korean native pigs (DK), 1,105 Landrace × Korean native pigs (LK), 1,017 Landrace × Yorkshire × Duroc (LYD) crossbreds, along with purebred animals. Crossbred individuals were genotyped using the Illumina PorcineSNP60 Genotyping BeadChip, whereas genotype data for purebred animals were provided by the Centre for Research in Agricultural Genomics [12]. Genotype data for the Korean native pigs (KNPs) among the purebreds were provided by the National Institute of Animal Science in Korea. Details regarding the number of animals, single nucleotide polymorphisms (SNPs), and average observed heterozygosity rate for each breed are presented in **Supplementary Table 1**. The quality control process involved the exclusion of SNPs located on sex chromosomes, with a genotype call rate below 90%, and with a minor allele frequency below 1%. After merging datasets and applying the quality control process, a common set of 24,118 SNPs were retained for analysis.

Phenotypic data revealed differences in backfat thickness and carcass weight among the breeds. The LYD breed exhibited the lowest backfat thickness, whereas the DK breed had the highest backfat thickness. Conversely, the DK breed exhibited the lowest carcass weight, whereas LYD had the highest carcass weight. The carcass performance of the breeds crossed with the KNP was lower than that of LYD. This finding aligns with the known characteristics of the KNP breed, which is known for its good meat quality but poor growth rate [13]. Statistical details for the phenotypes are provided in **Supplementary Table 2**.

Principal Component Analysis

Principal component analysis (PCA) was employed to investigate genetic differences between populations and to correct for population structure. PCA simplifies data complexity while maintaining the underlying relationships among the data points. When applied to biallelic genotype data, PCA identifies the eigenvalues and eigenvectors of the covariance matrix of allele frequencies, thereby reducing the data to a limited number of dimensions known as PCs. Each PC represents a proportion of the total genomic variation. Subsequently, the data are mapped onto the space defined by these PC axes, facilitating the visualization of samples and their distances from each other in a scatter plot. In this visualization, sample overlap indicates shared genetic identity, reflecting common ancestry or origin [14].

Genomic Breed Composition

Genomic breed composition (GBC) was estimated from genomic data using a maximum likelihood model implemented in ADMIXTURE v1.3.0 [15]. ADMIXTURE uses genotype data to cluster individuals into subgroups based on a predetermined number of groups. The projection extension of the ADMIXTURE program allows for estimating ancestry using predefined ancestral population allele frequencies. This extension enables efficient ancestry inference across large genomic datasets, leveraging allele frequencies from reference panels, such as the 1000 Genomes Project. Additionally, the projection approach is particularly advantageous for datasets with significant population distribution imbalances, as such imbalances can adversely affect the accuracy of ancestry inference [16].

The projection extension of the ADMIXTURE program was used to analyze the dataset owing to the imbalance between purebred and crossbred samples. Ancestral population allele frequencies were estimated using the purebred samples, whereas the GBC values of the crossbreds were estimated using the allele frequencies of the purebreds.

Statistical Models

First, PCs and GBCs were calculated for each individual, which were subsequently used in five models to predict genomic estimated breeding values (GEBV). Although additional fixed effects such as age and farm were considered, age information was unavailable, and farm data showed high multicollinearity with the PC and GBC values, which precluded their inclusion.

- Model 1 (NULL) is defined as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{e},$$

where **y** represents the vector of trait records (backfat thickness or carcass weight); **b** indicates the vector of fixed effects, including sex; **X** denotes the design matrix linking fixed effects to the records; **g** represents the vector of random genetic effects, modeled as $\sim N(0, \mathbf{G}\sigma_g^2)$, with **G** being the genomic relationship matrix and σ_g^2 being the genetic variance captured by the SNPs; **Z** indicates the design matrix linking records to animals; and **e** denotes the vector of random deviations, modeled as $\sim N(0, \mathbf{I}\sigma_e^2)$, with **I** as an animal-by-animal identity matrix and σ_e^2 representing the error variance. The GEBV for this model was predicted as $\mathbf{GEBV} = \hat{\mathbf{g}}$. The genomic relationship matrix was constructed using GCTA v1.94.1 software according to the following equation [17]:

$$\mathbf{G}_{jk} = \frac{1}{N}\sum_{i=1}^{N} \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)},$$

where x_{ij} and x_{ik} represent the genotypes (coded as 0, 1, or 2) of individuals *j* and *k* at SNP *i*. p_i indicates the allele frequency of SNP *i*, and *N* denotes the total number of SNPs. The distribution of the diagonal and off-diagonal elements of the genomic relationship matrix is shown in **Supplementary Figure 1**. The mean of the diagonal elements is 1.03, indicating low inbreeding within the population. The mean of the off-diagonal elements is 0, showing that individuals are genetically independent of each other.

- Model 2 (PC_F) is defined as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{e},$$

where **y** represents the vector of trait records; **b** denotes the vector of fixed effects, which includes PC values (20 PCs) and sex; **X** indicates the design matrix linking fixed effects to records; **g** represents the vector of random genetic effects; **Z** denotes the design matrix linking records to animals; and **e** indicates the vector of random deviations. For this model, $\mathbf{GEBV} = \hat{\mathbf{g}}$.

- Model 3 (GBC_F) is defined as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{e},$$

where **y** represents the vector of trait records; **b** denotes the vector of fixed effects, which includes GBC values and sex (here, breed composition values represent the proportion of each individual's genome derived from the four breeds: Duroc, KNP, Landrace, and Yorkshire); **X** indicates the design matrix linking fixed effects to records; **g** represents the vector of random genetic effects; **Z** denotes the design matrix linking records to animals; and **e** indicates the vector of random deviations. For this model, $\mathbf{GEBV} = \hat{\mathbf{g}}$.

- Model 4 (PC_R) is defined as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{Zpc} + \mathbf{e},$$

where **y** indicates the vector of trait records; **b** represents the vector of fixed effects, including sex; **X** denotes the design matrix linking fixed effects to records; **g** indicates the vector of random genetic effects; **pc** denotes the vector of random variables representing groups of PC values, which were clustered using the Gaussian Mixture Model implemented in the ‘mclust’ R package [18]; **Z** indicates the design matrix linking records to animals; and **e** denotes the vector of random deviations. For this model, $\mathbf{GEBV} = \hat{\mathbf{g}} + \widehat{\mathbf{pc}}$.

- Model 5 (GBC_R) is defined as follows:

$$\mathbf{y} = \mathbf{Xf} + \mathbf{Zg} + \mathbf{Zgbc} + \mathbf{e},$$

where **y** represents the vector of trait records; **b** denotes the vector of fixed effects, including sex; **X** indicates the design matrix linking fixed effects to records; **g** represents the vector of random genetic effects; **gbc** denotes the vector of random variables representing groups of GBC values, which were clustered using the Gaussian Mixture Model implemented in the ‘mclust’ R package [18]; **Z** indicates the design matrix linking records to animals; and **e** represents the vector of random deviations. For this model, $\mathbf{GEBV} = \hat{\mathbf{g}} + \widehat{\mathbf{gbc}}$.

Variance components were estimated using the restricted maximum likelihood (REML) method, as implemented in MTG2 [19], for each model. Heritability for the traits was estimated using the formula $h^2 = \widehat{\sigma_g^2}/(\widehat{\sigma_g^2} + \widehat{\sigma_e^2})$. The accuracy of GEBVs for each of the five models was calculated as $r(\mathbf{GEBV}, \mathbf{y})$, where **y** represents the phenotypes corrected for fixed effects [20]. A 5-fold cross-validation approach was used to validate the models. In this method, animals were randomly divided into five groups, with each group treated as the validation set while the remaining groups constituted the reference set.

Results

Principal Components Analysis

PCA was performed to explore genetic structure across populations. The analysis revealed that the first PC (PC1) accounted for 43.9% of the total genetic variance, whereas the second PC (PC2) constituted 13.6% of the variance (**Figure 1**). The PCA plot revealed a clear separation among the crossbred populations, indicating distinct genetic backgrounds. However, the LYD population exhibited greater dispersion along the first two PCs, suggesting more considerable genetic variation within this group. This observed variation is likely attributed to the presence of F1 hybrids in the dataset, which primarily combined Landrace and Yorkshire genetics, thereby increasing the overall diversity observed in this population.

Genomic Breed Composition

The breed composition of the crossbred populations was evaluated using ADMIXTURE analysis; the results are depicted in **Figure 2**. The analysis was conducted in unsupervised mode using genomic data from purebred samples, and the estimated breed allele frequencies were subsequently used to infer breed membership coefficients for the crossbred individuals.

In the LYD population, the estimated breed composition revealed an average contribution of 31%, 33%, and 36% from Landrace, Yorkshire, and Duroc, respectively (**Table 1**). The presence of F1 animals, as indicated by the PCA, was corroborated by the breed composition analysis, where the contribution of the Landrace and Yorkshire breeds showed that the F1 crossbreds were indeed hybrids of these two pure breeds. The variation in breed composition within the LYD population was not substantial, with standard deviations of 0.13, 0.12, and 0.19 for Landrace, Yorkshire, and Duroc, respectively. Similarly, the DK and LK populations exhibited balanced breed compositions. In the DK population, the average breed composition was 63% Duroc and 37% KNP, with minimal variation between individuals (SD = 0.05 for both breeds). The LK population had an average composition of 61% Landrace and 39% KNP, and low variation was also observed across individuals (SD = 0.06 for both breeds). These results suggest that the parental breeds had relatively balanced genetic contributions, as evidenced by the minimal variation in breed composition between individuals within the DK and LK populations.

Genetic Parameter Estimates

Heritability estimates for backfat thickness and carcass weight were derived from five different models; the associated variance components are detailed in **Table 2**. The estimates of genetic additive variance (V_g) and error variance (V_e) were used to calculate heritability for each trait.

Model 1 (NULL), which did not account for population structure, yielded the highest heritability estimates, with a heritability value of 0.44 ± 0.03 for backfat thickness and 0.31 ± 0.03 for carcass weight. The elevated heritability estimates for this model may be attributed to its lack of adjustments for potential confounding factors related to breed differences. Models 2 (PCA_F) and 3 (GBC_F), which incorporated population structure as a fixed effect, yielded lower heritability estimates; Model 2 estimated heritability for backfat thickness at 0.41 ± 0.03 and carcass weight at 0.26 ± 0.03, whereas Model 3 estimated these factors at 0.44 ± 0.03 and 0.27 ± 0.03, respectively. These reductions in heritability suggest that accounting for population structure as a fixed effect can decrease the perceived genetic influence on the traits. Models 4 (PCA_R) and 5 (GBC_R) included additional genetic variance components (V_{pc} and V_{gbc}) to account for population structure as a random effect. In Model 4, the genetic variance (V_g) was estimated at 13.2 ± 1.3 and V_{pc} at 1.6 ± 1.7 for backfat thickness, contributing an additional heritability of 0.05 ± 0.05 to the base estimate of 0.41 ± 0.04. For carcass weight, V_g was estimated at 28.1 ± 3.6 and V_{pc} at 23.7 ± 15.1, contributing an additional heritability of 0.19 ± 0.1 to the base estimate of 0.23 ± 0.04. Model 5 demonstrated similar patterns, although V_{gbc} for backfat thickness was close to zero. These models typically yielded heritability estimates similar to those of Model 1 for backfat thickness; however, for carcass weight, they provided a more nuanced understanding of genetic effects by accounting for population structure as a separate effect.

Accuracy of Genomic Estimated Breeding Values

The accuracy of GEBVs was evaluated using five models; the results are summarized in **Table 3** and depicted in **Figure 3**. Model 1 (NULL), Model 4 (PCA_R), and Model 5 (GBC_R) exhibited the highest accuracy for predicting both backfat thickness and carcass weight. These models achieved an average accuracy of 0.59 for backfat thickness and 0.50 for carcass weight, with minimal variation across replicates (SD = 0.01 for backfat thickness and between 0.03 to 0.04 for carcass weight).

Models that incorporated population structure as a fixed effect (Models 2 and 3) demonstrated lower accuracies for GEBVs. For backfat thickness, Model 2 (PCA_F) achieved a mean accuracy of 0.40 ± 0.03, whereas Model 3 (GBC_F) yielded a mean accuracy of 0.53 ± 0.04. The accuracy for carcass weight in these models was reduced similarly, with Model 2 achieving an accuracy of 0.34 ± 0.03 and Model 3 yielding an accuracy of 0.38 ± 0.02. These results suggest that modeling population structure as a fixed effect captures population differences but compromises GEBV accuracy. In contrast, modeling population structure as a random effect captures genetic variation due to breed differences without adversely affecting GEBV accuracy.

The Spearman rank correlation coefficient of GEBV between all models showed that all models were highly correlated with each other (except Model 2 in backfat thickness), ranging from 0.59 to 0.60. In carcass weight, Models 1, 4, and 5 had high Spearman correlation coefficients with each other, but models 2 and 3 had low correlation coefficients with the other models, ranging from 0.39 to 0.70 (**Figure 4**). Models that did not correct for population structure and models that corrected for population structure as a random effect had similar genomic prediction patterns.

Discussion

In multi-breed genomic predictions, using a reference population that encompasses multiple breeds inevitably introduces differences in population structure across these breeds. Therefore, this study aimed to assess prediction accuracy while adjusting population structure as either a fixed or random effect in multi-breed genomic predictions. The findings revealed that adjusting for population structure as a fixed effect resulted in decreased accuracy, whereas treating it as a random effect did not yield any improvements in accuracy. These results suggest that in multi-breed genomic predictions, the genomic relationship matrix sufficiently accounts for population structure, indicating that a model without adjustments for population structure is the most efficient.

Genotypic versus pedigree-based breed composition

GBC highlights the superior accuracy of genotypic data over that of pedigree information in determining breed composition. Pedigree records often contain inaccuracies or are incomplete, which can result in erroneous breed composition estimates [21, 22]. In contrast, using genomic data with tools such as ADMIXTURE provides a more precise assessment [23]. The findings of this study revealed that the breed compositions calculated using ADMIXTURE closely aligned with those expected from complete pedigree records, thereby corroborating previous research that emphasizes the reliability of genomic data for estimating breed composition in admixed populations [23].

Effect of population structure on genomic estimated breeding values

The effect of population structure on the estimation of genetic parameters is a well-established concern in genomic studies. Population structure can lead to false-positive associations [24], which may result in inflated heritability estimates [10] and biased accuracies in genomic predictions [6]. To address this issue, this study incorporated PCs and GBCs into GBLUP models as fixed or random effects.

Notably, the inclusion of PCs or GBCs as fixed effects resulted in decreased accuracy of GEBVs compared to those of models that excluded these factors. This reduction in accuracy may stem from the redundancy between the information provided by these variables and that captured by the genomic relationship matrix. Essentially, the genomic relationship matrix already encompasses much of the population structure information; therefore, adding PCs or breed composition as fixed effects could result in double-counting, leading to overcorrection and reduced model accuracy [11, 25]. In contrast, treating PCs and GBC as random effects did not yield any improvement in prediction accuracy. This result suggests that the additional genetic variance components captured by these random effects did not provide significant new information beyond what was already accounted for by the genomic relationship matrix. Similarly, previous studies have demonstrated that incorporating population structure as a random effect does not enhance the accuracy of genomic predictions [25]. However, the advantage of including breed as a random effect within the model, as GEBVs are divided into two components. Specifically, a model with a random effect splits the genetic variance into within-breed and across-breed GEBVs, thereby facilitating the understanding of how predictions differ within and across breeds [25].

These findings hold significant implications for the optimal design of genomic prediction models. Although accounting for population structure is crucial to avoid biases, these results indicate that the genomic relationship matrix within the GBLUP framework sufficiently captures the required information. Consequently, additional adjustments for population structure, whether as fixed or random effects, may be unnecessary and could even negatively affect prediction accuracy. These findings support the growing consensus that simpler models that rely on the genomic relationship matrix without further correction for population structure are often the most effective [25].

This study focused on carcass traits and therefore did not explicitly include heterozygosity, even though crossbred animals were used. However, recent findings suggest that including heterozygosity in genomic predictions for maternal traits can improve prediction accuracy [26]. Therefore, future research on maternal traits in genomic prediction models may benefit from considering heterozygosity as a factor to further enhance prediction accuracy.

Implications for multi-breed genomic prediction

Our findings have significant implications in the field of multi-breed genomic prediction. This study demonstrated that the genomic relationship matrix alone could effectively capture breed differences within multi-breed populations, thereby eliminating the necessity for additional corrections for population structure. This circumvention is particularly advantageous in multi-breed contexts, where genetic relationships among breeds can vary widely, facilitating accurate predictions of breeding values for selection decisions.

Given the observed decrease in accuracy when population structure was included as a fixed effect, future studies and practical applications of genomic prediction should prioritize models that incorporate the genomic relationship matrix as the primary tool for capturing genetic variance. This approach is more straightforward and ensures higher accuracy in predicting breeding values, which is crucial for effectively managing and improving crossbred populations.

In conclusion, this study underscores the robustness of the genomic relationship matrix in accounting for population structure within multi-breed genomic prediction. The findings suggest that, although population structure is an important consideration, the genomic relationship matrix is sufficient for capturing the relevant genetic variance, modeling additional corrections unnecessary. This insight is valuable for optimizing genomic prediction models in crossbred populations and enhancing the accuracy of GEBV predictions.

References

1. Goddard M. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 2009;136(2):245-57.

2. Thomasen JR, Guldbrandtsen B, Su G, Brøndum RF, Lund MS. Reliabilities of genomic estimated breeding values in Danish Jersey. Animal. 2012;6(5):789-96.

3. Olson KM, VanRaden PM, Tooker ME. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. J Dairy Sci. 2012;95(9):5378-83.

4. Erbe M, Hayes B, Matukumalli L, Goswami S, Bowman P, Reich C, et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 2012;95(7):4114-29.

5. Hoze C, Fritz S, Phocas F, Boichard D, Ducrocq V, Croiseau P. Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population. J Dairy Sci. 2014;97(6):3918-29.

6. Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, Allison DB, et al. Beyond Missing Heritability: Prediction of Complex Traits. Plos Genet. 2011;7(4).

7. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics. 2006;38(8):904-9.

8. Pritchard JK, Donnelly P. Case-control studies of association in structured or admixed populations. Theor Popul Biol. 2001;60(3):227-37.

9. Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. Nature Genetics. 2004;36(5):512-7.

10. Visscher PM, Yang JA, Goddard ME. A Commentary on 'Common SNPs Explain a Large Proportion of the Heritability for Human Height' by Yang et al. (2010). Twin Res Hum Genet. 2010;13(6):517-24.

11. Janss L, de Los Campos G, Sheehan N, Sorensen D. Inferences from genomic models in stratified populations. Genetics. 2012;192(2):693-704.

12. Burgos-Paz W, Souza CA, Megens HJ, Ramayo-Caldas Y, Melo M, Lemús-Flores C, et al. Porcine colonization of the Americas: a 60k SNP story. Heredity. 2013;110(4):321-30.

13. Park J, Kim Y, Jung H-J, Park B, Lee J, Moon H. Comparison of meat quality and physicochemical characteristics of pork between Korean native black pigs (KNBP) and Landrace by market weight. J Anim Sci Technol. 2005;47(1):91-8.

14. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. Plos Genet. 2006;2(12):2074-93.

15. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655-64.

16. Shringarpure S, Xing EP. Effects of Sample Selection Bias on the Accuracy of Population Structure and Ancestry Inference. G3-Genes Genom Genet. 2014;4(5):901-11.

17. Yang JA, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. Am J Hum Genet. 2011;88(1):76-82.

18. Scrucca L, Fraley C, Murphy TB, Raftery AE. Model-based clustering, classification, and density estimation using mclust in R: Chapman and Hall/CRC; 2023.

19. Lee SH, Van der Werf JH. MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. Bioinformatics. 2016;32(9):1420-2.

20. Lourenco DAL, 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. Genetics Selection Evolution. 2015;47.

21. Kuehn LA, Keele JW, Bennett GL, McDaneld TG, Smith TPL, Snelling WM, et al. Predicting breed composition using breed frequencies of 50,000 markers from the US Meat Animal Research Center 2,000 Bull Project. J Anim Sci. 2011;89(6):1742-50.

22. Funkhouser SA, Bates RO, Ernst CW, Newcom D, Steibel JP. Estimation of genome-wide and locus-specific breed composition in pigs. Transl Anim Sci. 2017;1(1):36-44.

23. Gobena M, Elzo MA, Mateescu RG. Population Structure and Genomic Breed Composition in an Angus-Brahman Crossbred Cattle Population. Front Genet. 2018;9.

24. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010;11(7):459-63.

25. Hayes BJ, Copley J, Dodd E, Ross EM, Speight S, Fordyce G. Multi-breed genomic evaluation for tropical beef cattle when no pedigree information is available. Genetics Selection Evolution. 2023;55(1):71.

26. Iversen MW, Nordbø Ø, Gjerlaug-Enger E, Grindflek E, Lopes MS, Meuwissen T. Effects of heterozygosity on performance of purebred and crossbred pigs. Genetics Selection Evolution. 2019;51:1-13.

Tables

Table 1. Genomic breed composition by breeds.

Population	Breed	Min	Median	Max	Mean	SD
Landrace × Yorkshire × Duroc	**Landrace**	0.06	0.27	0.90	0.31	0.13
	Yorkshire	0.05	0.30	0.89	0.33	0.12
	Duroc	0	0.43	0.75	0.36	0.19
Duroc × KNP	**Duroc**	0.49	0.63	0.75	0.63	0.05
	KNP	0.25	0.37	0.51	0.37	0.05
Landrace × KNP	**Landrace**	0.43	0.62	0.75	0.61	0.06
	KNP	0.25	0.38	0.57	0.39	0.06

Table 2. Variance components and heritability estimates from five models for backfat thickness and carcass weight traits. Variance components are the genetic additive variance (V_g) and the error variance (V_e). In addition, the Model 4 (PC_R) and the Model 5 (GBC_R) estimates additional genetic variance components (V_{pc} and V_{gbc}).

Model		**Variance components**		**Heritabilities**	
		Backfat thickness (mm)	**Carcass weight (kg)**	**Backfat thickness (mm)**	**Carcass weight (kg)**
1 (NULL)	V_g	13.5 ± 1.3	31.4 ± 3.6	0.44 ± 0.03	0.31 ± 0.03
	V_e	17.1 ± 0.8	69.2 ± 2.7		
2 (PC_F)	V_g	12.1 ± 1.3	24.9 ± 3.7	0.41 ± 0.03	0.26 ± 0.03
	V_e	17.5 ± 0.8	71.3 ± 2.8		
3 (GBC_F)	V_g	13.7 ± 1.3	26.0 ± 3.4	0.44 ± 0.03	0.27 ± 0.03
	V_e	17.1 ± 0.8	70.6 ± 2.7		
4 (PC_R)	V_g	13.2 ± 1.3	28.1 ± 3.6	0.41 ± 0.04	0.23 ± 0.04
	V_{pc}	1.6 ± 1.7	23.7 ± 15.1	0.05 ± 0.05	0.19 ± 0.1
	V_e	17.2 ± 0.8	69.9 ± 2.7		
5 (GBC_R)	V_g	13.5 ± 1.3	27.1 ± 3.5	0.44 ± 0.03	0.23 ± 0.04
	V_{gbc}	0.2 ± 0.3	22.3 ± 14.3	0	0.19 ± 0.1
	V_e	17.1 ± 0.8	70.2 ± 2.8		

Table 3. Mean and standard deviation of GEBV accuracy for five prediction methods.

Model	Backfat thickness (mm)		Carcass weight (kg)	
	Mean	SD	Mean	SD
1 (NULL)	0.59	0.01	0.50	0.04
2 (PCA_F)	0.40	0.03	0.34	0.03
3 (GBC_F)	0.53	0.04	0.38	0.02
4 (PCA_R)	0.59	0.01	0.50	0.03
5 (GBC_R)	0.59	0.01	0.50	0.03

Figures

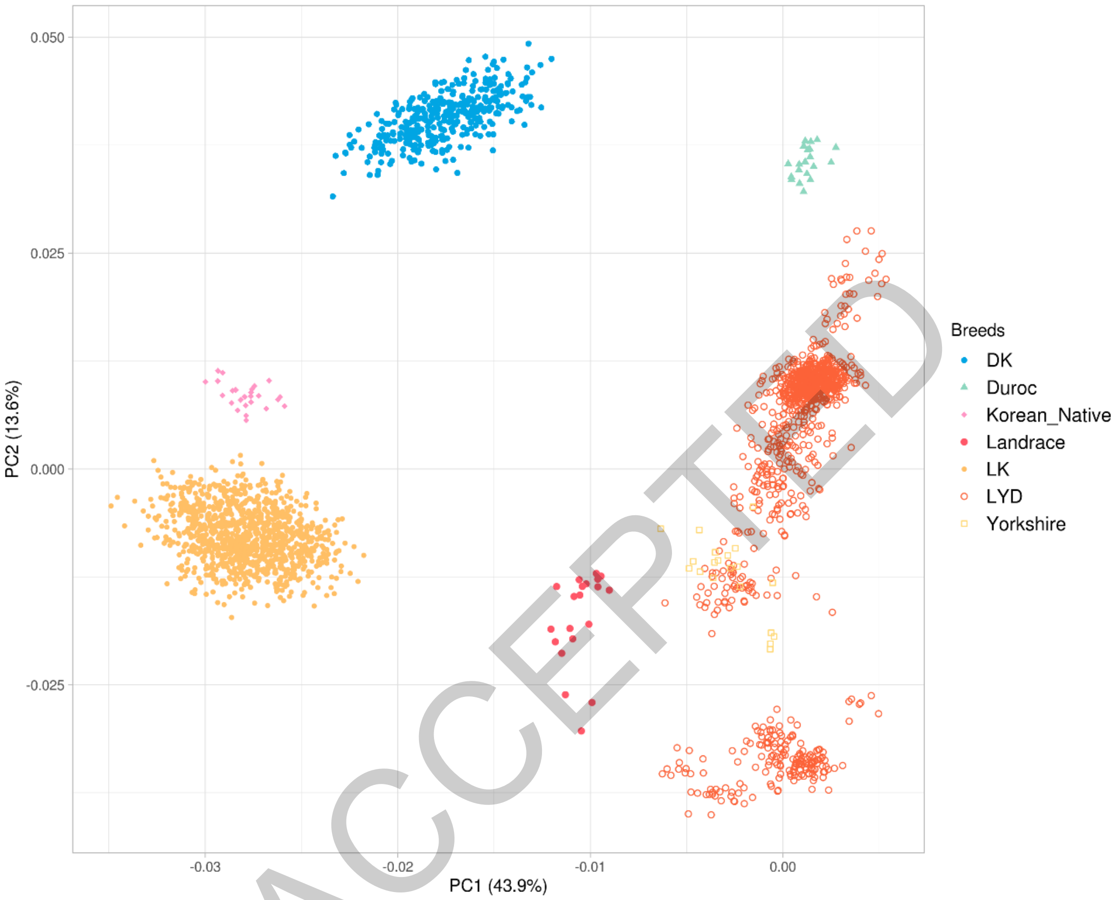

Figure 1. Population distribution across the first and second principal components.

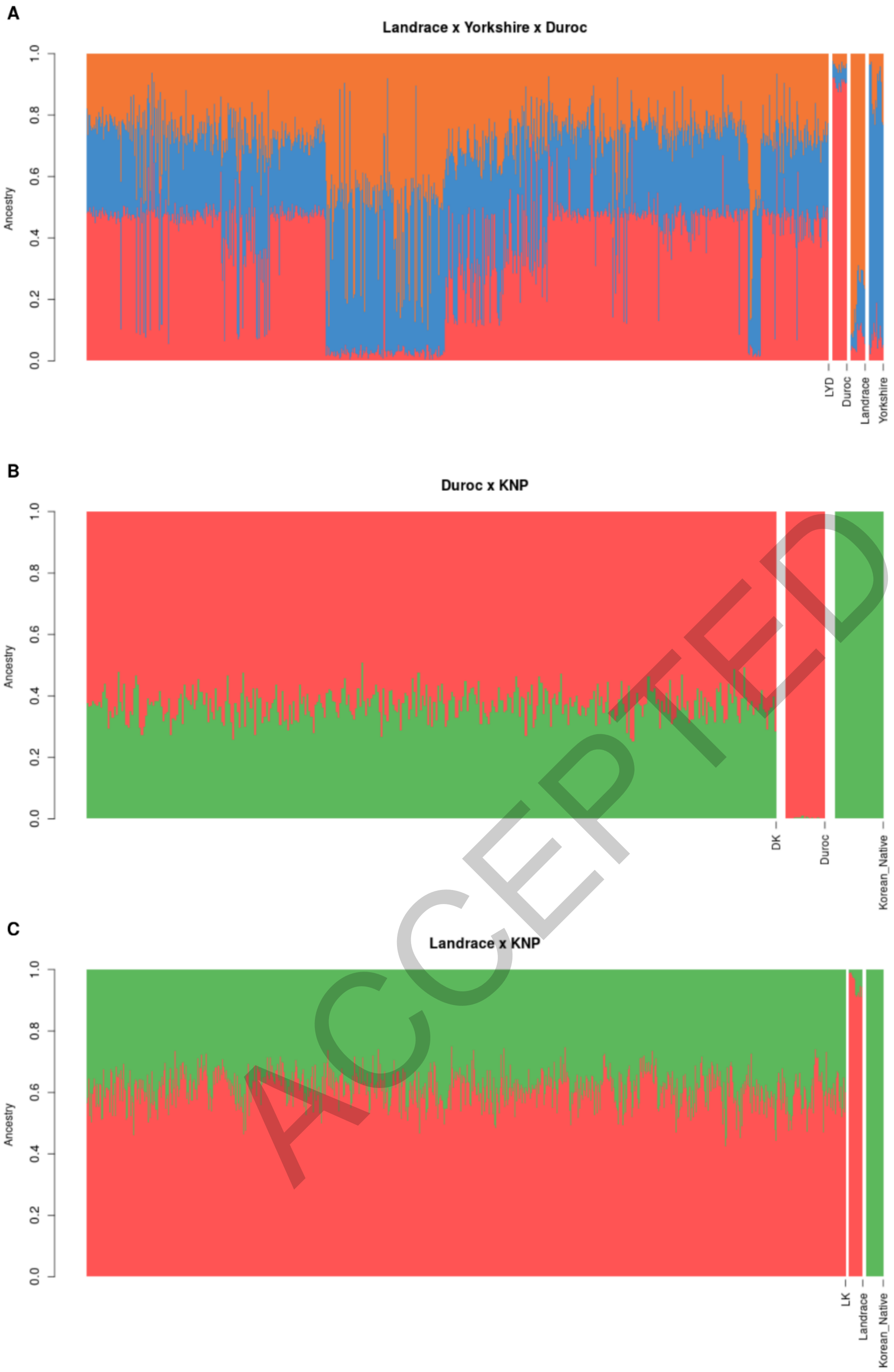

Figure 2. Bar plot of the *Q* matrix from an ADMIXTURE run, showing the proportion of the genome contributed by each breed. A shows the LYD population, B shows the DK population, and C shows the LK population. Each vertical bar represents an individual.

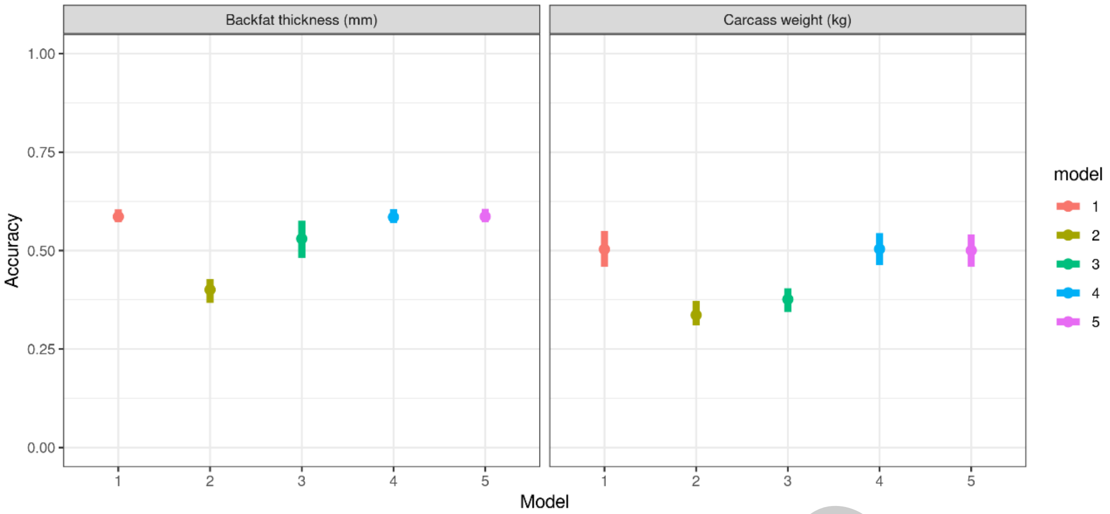

Figure 3. GEBV accuracy of five prediction models. From left to right, the models are Model 1 (NULL), Model 2 (PCA_F), Model 3 (GBC_F), Model 4 (PCA_R), and Model 5 (GBC_R). The dots represent the average accuracy, and the lines indicate the standard deviation.

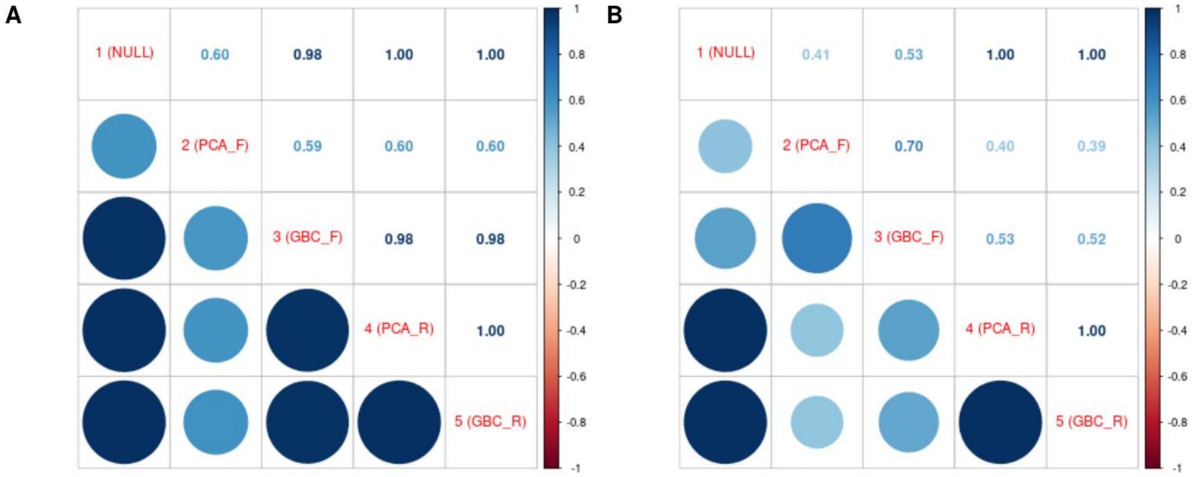

Figure 4. Spearman correlation between models. A represents backfat thickness and B represents carcass weight.