Sahana, V., Gowane, G., Nayee, N., Khan, K., Yadav, A., Paul, R., Vohra, V., Alex, R. (2026) Enhancing genomic prediction accuracy for breeds with numerically small reference population using multi-breed evaluation strategies, *Proceedings 13th World Congress on Genetics Applied to Livestock Production* 23964: https://doi.org/10.31274/wcgalp.23964 ©2026 Sahana, et al. CC BY-NC-ND

Enhancing genomic prediction accuracy for breeds with numerically small reference population using multi-breed evaluation strategies

V.N. Sahana[1*], G.R. Gowane[1], N. Nayee[2], K.D. Khan[1], A. Yadav[1], R. Paul[1], V. Vohra[1] and R. Alex[1]

[1]ICAR-National Dairy Research Institute, Karnal 132001, Haryana, India;. [2]National Dairy Development Board, Anand 388001, Gujarat, India;[*]Corresponding author: sahanavn01@gmail.com

Introduction

In developing countries, the implementation of genomic selection is constrained by limited genotyping resources, scarcity of performance records, small herd sizes, and management heterogeneity, which hinder the establishment of sufficiently large reference populations (Mrode *et al.,* 2019). Accurate prediction of genomic breeding values (GBV) is central to the success of genomic selection and is primarily driven by reference population size and availability, posing a major challenge for breeds with small reference population. Additional factors such as heritability, effective population size, linkage disequilibrium (LD), selective genotyping, and marker density further influence predictive accuracy (Goddard, 2009; Gowane *et al.,* 2019). Although across-breed genomic prediction offers a potential strategy to overcome these limitations, difference in LD, allele frequency, and linkage phase across breeds may reduce GBV prediction accuracy. Multi-breed reference populations leverage information from genetically related breeds to improve prediction accuracy in breeds with small reference populations without additional genotyping or phenotyping: however, success depends on genetic relatedness and shared haplotype structure. Given the limited reference population of indigenous Indian cattle and the uncertain yet promising potential of multi-breed approaches, this study evaluates multi-breed genomic evaluation strategies to enhance GBV accuracy in breeds with small reference populations maintained under smallholder production systems.

Materials & Methods

Data: Phenotypic (305-day first lactation milk yield, 305-DMY) and genotypic data (IndusChip Bovine 50K SNP) for 1,298 Gir, 1,291 Sahiwal, and 500 Kankrej animals calved between 2011 to 2022 from data obtained from National Dairy Development Board (NDDB), Anand, India were used for analysis. Multi-breed reference populations were created by combining breeds: Gir–Sahiwal (GS), Sahiwal–Kankrej (SK), Gir–Kankrej (GK), and Gir–Sahiwal–Kankrej (GSK).

Genomic Evaluation: Variance components were estimated using genomic restricted maximum likelihood (GREML). Genomic prediction was performed using Genomic best linear unbiased prediction (GBLUP), which has been shown to perform well, relying solely on the genomic relationship matrix (GRM) even in absence of pedigree, especially in loosely structured breeding systems common in India (Gowane *et al.,* 2025). The model fitted was

$$\mathbf{y} = \mathbf{Xb} + \boldsymbol{Z_1}\mathbf{v} + \boldsymbol{Z_2}\mathbf{g} + \mathbf{e} \quad (1)$$

where $\boldsymbol{y}$ is vector of phenotypes, $\boldsymbol{X}$ and $\boldsymbol{b}$ are incidence matrix and vector of fixed effects (overall mean, season and year of calving), $\boldsymbol{Z_1}$ is incidence matrix for random village effect $v \sim N(0, I\sigma_v^2)$, and $\boldsymbol{Z_2}$ allocates records to GBV $g \sim N(0, G\sigma_g^2)$. The residuals were assumed $\boldsymbol{e} \sim N(0, I\sigma_e^2)$.The GRM was constructed using method 1 of VanRaden (2008) as:

$$\boldsymbol{G} = \frac{\boldsymbol{ZZ'}}{\boldsymbol{2\sum p_i(1-p_i)}}$$

where $\boldsymbol{Z}$ is the centered genotype matrix obtained as $\boldsymbol{Z = M - 2p}$, $\boldsymbol{M}$ is the genotype matrix coded as 0, 1, and 2, and $p_i$is the allele frequency of the i^{th} SNP.

In our previous study (Sahana *et al.,* 2025), we compared three multi-breed genomic evaluation strategies *viz*, shared GRM (assuming common SNP effects and allowing relationships across breed), non-shared GRM (assumed breed-specific SNP effects by treating SNPs from different breeds as non-overlapping, resulting in relationships only within breeds and zero relationships across breeds), and the metafounder (MF) approach for improving GBV prediction accuracy in breeds with limited reference data availability. Building on these findings, the present study further evaluated whether a multi-trait genomic best linear unbiased prediction (MT-GBLUP) approach could provide additional gains beyond the previously identified optimal shared GRM model by considering 305-DMY as a different correlated trait across Gir, Sahiwal, and Kankrej, following van den Berg *et al*. (2020) using Blupf90+ software (Misztal *et al.* 2014). Multi-trait GBLUP models were fitted for both two-breed (GS, GK, SK) and three-breed (GSK) reference populations, wherein the tri-variate model were used to the GSK reference population, and bi-variate models were fitted for the GS, GK, and SK combinations. The model used was:

$$\begin{bmatrix} \boldsymbol{y_G} \\ \boldsymbol{y_S} \\ \boldsymbol{y_K} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X_G} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X_S} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{X_K} \end{bmatrix} \begin{bmatrix} \boldsymbol{b_G} \\ \boldsymbol{b_S} \\ \boldsymbol{b_K} \end{bmatrix} + \begin{bmatrix} \boldsymbol{Z_G} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z_S} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{Z_K} \end{bmatrix} \begin{bmatrix} \boldsymbol{g_G} \\ \boldsymbol{g_S} \\ \boldsymbol{g_K} \end{bmatrix} + \begin{bmatrix} \boldsymbol{e_G} \\ \boldsymbol{e_S} \\ \boldsymbol{e_K} \end{bmatrix}$$

where, $\mathbf{y_G}$**,** $\mathbf{y_S}$**,** $\mathbf{y_K}$ are vectors of phenotype in Gir, Sahiwal and Kankrej breed, $\mathbf{X_G}$**,** $\mathbf{X_S}$**,** $\mathbf{X_K}$ are design matrices allocating phenotypes to vectors with fixed effects $\mathbf{b_G}$**,** $\mathbf{b_S}$, and $\mathbf{b_K}$; $\mathbf{Z_G}$**,** $\mathbf{Z_S}$**,** and $\mathbf{Z_K}$ are design matrices allocating phenotypes to vectors with genomic breeding values $\mathbf{g_G}$**,** $\mathbf{g_S}$, and $\mathbf{g_K}$ distributed as N (0,G ⊗ H), with GRM G, and variance and (co)variance matrix of genomic breeding values ,

$$\mathrm{H} = \begin{bmatrix} \boldsymbol{\sigma^2}_{\boldsymbol{G}} & \boldsymbol{\sigma}_{\boldsymbol{g_{GS}}} & \boldsymbol{\sigma}_{\boldsymbol{g_{GK}}} \\ \boldsymbol{\sigma}_{\boldsymbol{g_{GS}}} & \boldsymbol{\sigma^2}_{\boldsymbol{S}} & \boldsymbol{\sigma}_{\boldsymbol{g_{SK}}} \\ \boldsymbol{\sigma}_{\boldsymbol{g_{GK}}} & \boldsymbol{\sigma}_{\boldsymbol{g_{SK}}} & \boldsymbol{\sigma^2}_{\boldsymbol{K}} \end{bmatrix}$$

$\mathbf{e_G}$**,** $\mathbf{e_S}$, and $\mathbf{e_K}$ are residual error vectors distributed as N (0, E⊗R) with residual variance covariance matrix

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{\sigma^2_{e_G}} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\sigma^2_{e_S}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\sigma^2_{e_K}} \end{bmatrix}$$

Validation: Expected GBV accuracies were calculated from the prediction error variance (PEV) according to Mrode (2014). Validation of GBV predictions was performed separately for each reference population using linear regression (LR) method of Legarra and Reverter (2018), by masking 15% of the most recent phenotypes within the validation breed and comparing GBV from the "whole" (all records) vs. "partial" (masked records) dataset. LR estimators of accuracy, bias, and dispersion for validation animals were computed.

Prediction performance obtained using MT-GBLUP was subsequently compared with previously published results based on shared GRM, non-shared GRM, and metafounder approaches (Sahana *et al.,* 2025) to evaluate the relative benefits of the MT-GBLUP.

Results & Discussion

The average 305-DMY was 2663 ± 16.53 kg in Sahiwal, 2249 ± 25.40 kg in Kankrej, and 2086 ± 20.78 kg in Gir. Heritability estimates 305-DMY varied across breeds, with Gir showing the highest heritability (0.30), genetic variance (133500 ± 24530) and GBV accuracy. Kankrej

showed lowest heritability (0.17) and genetic variance (51918 ± 34101), likely reflecting its smaller data size.

Using GBLUP with single-breed reference, average GBV accuracies for Gir, Sahiwal, and Kankrej were 0.65, 0.60, and 0.49, respectively. The lower accuracy for Kankrej reflects its smaller reference population (n = 500) relative to Gir (n = 1291) and Sahiwal, and aligns with its lower heritability.

The MT-GBLUP provided minimal improvements for Kankrej and was markedly lower than the shared GRM approach (GS: 0% vs. 23.4%; GSK: 2.0% vs. 20.9%) as given in Table 1. The limited amount of genotypic data available and weak genetic connectedness among the three breeds may have been insufficient to effectively link the trait across breeds, thereby reducing the effectiveness of the MT-GBLUP. In addition, MT-GBLUP requires estimation of breed-specific genetic variances and between-breed genetic covariances, increasing the number of parameters, which may reduce prediction accuracy when reference population size is limited. Moreover, MT-GBLUP assumes a single average genetic correlation across the genome, whereas true correlations may vary among loci, particularly when breed-specific QTL effects exist. Under such conditions, the increased parameterization of MT-GBLUP models may outweigh their theoretical advantages, especially in small datasets. Consistent with earlier findings (Calus *et al.,* 2018, Gowane *et al.,* 2025), pooling data across breeds and assuming a genetic correlation of unity, as implemented in the single-trait (Shared GRM) models, provided more accurate predictions in the present study.

Table 1 Expected accuracy of GBV prediction using multi-breed reference population with Shared GRM and MT-GBLUP approach (% change relative GBV prediction accuracy using single breed reference population).

Reference population	Shared GRM			MT-GBLUP		
	Gir	Sahiwal	Kankrej	Gir	Sahiwal	Kankrej
GS	0.63 (-3%)	0.63 (+5%)	-	0.65 (0%)	0.59 (-2%)	-
GK	0.63 (-3%)	-	0.64 (+23%)	0.65 (0%)	-	0.49 (0%)
SK	-	0.57 (-6%)	0.58 (+15%)	-	0.59 (-2%)	0.50 (+2%)
GSK	0.61 (-6%)	0.6 0%	0.62 (+21%)	0.66 (+2%)	0.58 (-3%)	0.50 (+2%)

Substantial differences were observed for between breed genetic correlation estimates for 305-DMY derived from two-breed and three-breed multi-breed genomic evaluation using MT-GBLUP, indicating sensitivity of estimates to reference population composition. In particular, GS (0.18±0.41 vs. 0.08±0.42) and GK (0.035±1.07 vs. 0.008±0.93) correlations differed markedly between two-breed and three-breed combinations, whereas SK correlations remained relatively consistent (0.71±4.61 vs. 0.62±1.3). This suggests limited reliable inference regarding the true underlying between breed genetic correlation likely reflecting small reference population sizes. Only Kankrej benefits from shared GRM and MT-BLUP because of small reference size, while independent within-breed evaluations may be more appropriate for breeds with larger reference population.

Validation: LR validation showed that multi-breed prediction using MT-GBLUP produced the lowest GBV accuracy among all four methods (Table 2 and Table 3), with the lowest accuracy observed for the SK reference population. Regression of EBV from “whole” versus “partial” data was below 1, indicating overestimation of GBV. Bias was comparable with the shared GRM approach but higher for the GK reference population.

Table 2 The linear regression estimators (Accuracy and Bias) for the validation animals.

	Accuracy				Bias			
Reference population	Shared GRM	Non-shared GRM	MF	MT-GBLUP	Shared GRM	Non-shared GRM	MF	MT-GBLUP
GS	0.70	0.50	0.68	0.57	-14	-17.8	-21.8	13.9
GK	0.82	0.55	0.79	0.55	-5.85	-4.9	21.9	-38
SK	0.43	0.39	0.41	0.41	0.83	0.27	52	-1.24
GSK	0.78	0.48	0.74	0.61	-9.7	-12	-11	-1.86

Table 3 The linear regression estimators (Dispersion) for the validation animals

	Dispersion			
Reference population	Shared GRM	Non-shared GRM	MF	MT-GBLUP
GS	1.09	0.92	1.18	0.93
GK	0.97	0.96	1.08	0.81
SK	0.84	0.87	0.87	0.81
GSK	1.05	0.94	1.16	0.88

Conclusion

Our study showed that multi-trait GBLUP resulted in only minimal improvements in GBV accuracy for 305-day milk yield in breeds with small reference populations, such as Kankrej and performed worse than the other three multi-breed genomic evaluation methods, while shared GRM-based multi-breed prediction provided the highest improvement. These findings indicate that, under limited genotypic data and weak across-breed relationships, modelling 305-day milk yield as correlated traits across breeds provides no advantage, and shared GRM approach remains the most practical and effective strategy for increasing GBV prediction accuracy in breeds with small reference populations.

References

Calus, M. P. L., Goddard, M. E., Wientjes, Y. C. J., *et al*. (2018). Journal of dairy science, 101(5), 4279–4294. https://doi.org/10.3168/jds.2017-13366

Goddard, M. (2009). Genetica, 136(2), 245–257. https://doi.org/10.1007/s10709-008-9308-0

Gowane, G. R., Alex, R., Worku, D., *et al*. (2025). Tropical animal health and production, 57(3), 149. https://doi.org/10.1007/s11250-025-04407-6

Gowane, G. R., Lee, S. H., Clark, S., *et al*. (2019). Journal of animal breeding and genetics = Zeitschrift fur Tierzuchtung und Zuchtungsbiologie, 136(5), 390–407. https://doi.org/10.1111/jbg.12420

Legarra, A. and Reverter, A. (2018).Genetic Selection Evolution.50, 53. https://doi.org/10.1186/s12711-018-0426-6

Misztal, I., Tsuruta, S., Lourenco, D., *et al*. (2015). Manual for BLUPF90 family of programs (p. 125)

Mrode, R. (2014) Linear Models for the Prediction of Animal Breeding Values, 3rd ed.; CABI: Oxfordshire, UK, 2014.

Mrode, R., Ojango, J. M. K., Okeyo, A. M., *et al*. (2019). Frontiers in genetics, 9, 694. https://doi.org/10.3389/fgene.2018.00694

Sahana, V. N., Gowane, G. R., Nayee, N., *et al*. (2025). Scientific reports, 15(1), 31421. https://doi.org/10.1038/s41598-025-15733-y

van den Berg, I., MacLeod, I. M., Reich, C. M., *et al*. (2020). Journal of dairy science, 103(7), 6276– 6298. https://doi.org/10.3168/jds.2019-17914

VanRaden, P. M. (2008). Journal of dairy science. 91, 4414–4423. https://doi.org/10.3168/jds.2007-0980