Olabosoye, B., Dekkers, J., Steibel, J. (2026) Comparison of strategies for incorporating functional annotation in genomic prediction, *Proceedings 13th World Congress on Genetics Applied to Livestock Production* 23620: https://doi.org/10.31274/wcgalp.23620
©2026 Olabosoye, et al. CC BY-NC-ND

Comparison of strategies for incorporating functional annotation in genomic prediction

B. Olabosoye[1*], J. Dekkers[1], and J. Steibel[1]

[1]*Iowa State University, Ames, USA;* [*]*Corresponding author: blessing@iastate.edu*

Introduction

With the availability of functional annotation data, there is a need for analysis protocols to incorporate biological knowledge into genomic prediction models (Ma *et al*., 2025). Therefore, identifying and prioritizing functional and trait-specific variants holds promise for further improving genomic prediction accuracy. Several studies have reported modest improvements in prediction accuracy when incorporating functional annotation data, such as gene ontology terms, into genomic best linear unbiased prediction (GBLUP) models (Morgante *et al*., 2020). QTL databases that store results from previous GWAS or linkage analyses provide another source of gene-trait annotation data. Here, we present approaches for prioritizing SNPs for genomic prediction using the following data sources and analytic strategies: i) SNPs prioritized from genomic regions associated with related phenotypic traits in other populations, and ii) SNPs prioritized within genes' start and end positions (genic regions). Overall, the objective of this study was to better understand how leveraging different annotation strategies for prioritizing SNPs affects genetic variance explained and prediction accuracy.

Materials & Methods

Data. The pig dataset was obtained from a previously reported natural disease challenge study (Cheng *et al*., 2020). The dataset includes 4,095 wean-to-finish Landrace x Yorkshire pigs from seven breeding companies that were members of the PigGen Canada research consortium. Of the 2,882 pigs that reached market weight, data on lean yield (LYD) and backfat (BF) were used. Each company contributed pigs for one batch per cycle, for a total of 7 cycles, with each cycle comprising 7 batches. Pigs were genotyped using the 650K Affymetrix Axiom Porcine Genotyping Array. After quality control, 2,551 pigs (426,861 SNPs) and 2,554 pigs (426,873 SNPs) were available for LYD and BF, respectively.

GBLUP model. The classical model for genomic prediction (GBLUP) followed Cheng *et al*. (2020):

$$\mathbf{y}=\mathbf{X\beta}+\mathbf{S\tau}+\mathbf{W\alpha}+\mathbf{u}+\mathbf{e} \quad (1)$$

where $\mathbf{y}$ is the vector of the observed phenotypic records; $\boldsymbol{\beta}$ represents the vector of fixed batch and slaughter date effects and the covariate of slaughter age and weight; $\boldsymbol{\tau}$ represents the vector of random pen within batch effects $\sim N(\mathbf{0}, \mathbf{I}\sigma^2_\tau)$; and $\boldsymbol{\alpha}$ represents the vector of random litter effects $\sim N(\mathbf{0}, \mathbf{I}\sigma^2_\alpha)$. The vector of additive genetic values is denoted by $\mathbf{u}\sim N(\mathbf{0}, \mathbf{G}\sigma^2_u)$. where $\mathbf{G}=\mathbf{ZZ}^\top$, where $\mathbf{Z}$ is the standardized SNP genotype dosage matrix; and $\mathbf{e}$ is the vector of residual errors $\sim N(\mathbf{0}, \mathbf{I}\sigma^2_e)$.

QTL and gene annotation from public databases. Two types of publicly available annotations were used: a) genome-trait associations from related traits, and b) gene annotations. Curated trait-specific associated genes (AG) data were downloaded from PigQTLdb (Release 58, SS11.1; Hu *et al*., 2022). Specifically, all available AG data for LYD-related traits (lean meat+bone in ham, lean cuts weight, estimated carcass lean content, and lean meat %) and BF-related traits (subcutaneous fat thickness and backfat weight) were used.

Also, a null set was created for comparison with true trait-associated genes. The null sets for AG consisted of SNPs mapped to randomly selected genes of equal size as those in AG. In addition to association annotation, gene annotation was retrieved from the Ensembl database, and SNPs were mapped to each gene's start and end positions (genic regions), referred to as genic SNPs.

***f*GBLUP model.** In this *f*GBLUP model, prior biological knowledge from different functional annotation strategies was incorporated into GBLUP, including based on associated genes and genic regions. Like GBLUP and similar to GFBLUP by Edwards *et al*. (2016), *f*GBLUP includes a single random genetic effect (assuming homogeneous variances) and uses a genomic relationship matrix (GRM) for the given annotation strategy. Thus, the model for the functional annotation strategy *f* is the same as Eq. 1, but with $\mathbf{u}$ replaced by $\mathbf{u}_f$, where $\mathbf{u}_f \sim N(0, \mathbf{G}_f \sigma^2_{u_f})$ is the additive genetic values, and $\mathbf{G}_f = \mathbf{Z}_f \mathbf{Z}_f^\top$ is the genomic relationship matrix based on SNPs in the *f* annotation strategy. An extension is *f*GBLUP+R, which includes $\mathbf{u}_r$ as a second additive genetic value with a GRM based on the remaining SNPs.

Genomic prediction. Prediction metrics were evaluated using random 5-fold cross-validation (CV). The predicted value is the genomic estimated breeding value (GEBV), and the observed value is the phenotype adjusted for fixed effects (batch and slaughter age, weight & date) and random effects (pen and litter) based on the full data. In the validation population, prediction metrics were Pearson's correlation coefficient, a measure of prediction accuracy between the GEBV and the adjusted phenotype, and Root Mean Squared Error (RMSE), the square root of the average squared difference between the GEBV and the adjusted phenotype.

Results

Based on *f*GBLUP, estimates of the proportion of variance (h^2) captured by SNPs mapped to associated genes (AG) were 0.16 and 0.07 for LYD and BF, respectively (Table 1), compared to 0.48 and 0.47 for GBLUP, while genic SNPs estimates were similar to all SNPs with GBLUP. With extended *f*GBLUP, the estimate of h^2 was marginally higher for AG+R than for GBLUP for LYD, but similar for BF. For genic SNPs with *f*GBLUP, phenotypic prediction accuracy (0.26) was similar to using all SNPs with GBLUP (Table 2). However, using only SNPs mapped to associated genes resulted in lower accuracy for LYD (0.16) and BF (0.11). Extended *f*GBLUP resulted in -0.32 and 4.34% gain in accuracy over GBLUP for Genic+R and AG+R, respectively, for LYD, and 0.02 and -0.95% accuracy over GBLUP for BF. Compared to the null set, prediction accuracy was lower for SNPs mapped to randomly selected genes than for those mapped to associated genes for both traits.

Discussion

In this study, we demonstrated that prioritizing SNPs using strategies that incorporate prior biological knowledge can influence the variance explained and the prediction accuracy of phenotypes. Prioritizing SNPs by genic region explained over 60% of the total genetic variance for both LYD and BF traits in this study. This is consistent with results from other livestock species (Bhuiyan et al., 2018). Although variants within genic regions explained a significant portion of the total genetic variance, those within associated genes accounted for only a fraction of it. However, the marginal increase in total genetic variance explained by partitioning the functional class (AG) translated into a percentage gain in accuracy for LYD, demonstrating that not all variants contribute equally to trait variability. Our results support the alternative hypothesis that SNPs within genes associated with traits exhibit higher phenotypic prediction accuracy than those within randomly selected genes. These results further support prior

evidence linking these genes to the evaluated traits. As a result, this biological knowledge can be leveraged when selecting candidates in breeding schemes, thereby improving genetic progress. With increased total genetic variance and prediction accuracy, *f*GBLUP has the potential to enhance the effectiveness of candidate selection as more trait-specific biological knowledge, including genes associated with traits, becomes available. Future work could provide a better understanding of these annotation strategies by using breeding value prediction accuracy to assess the influence of prioritizing SNPs guided by prior biological knowledge on associated genes.

Table 1. Estimates of variance components from the full data for LYD and BF based on GBLUP and *f*GBLUP with different annotation strategies.

Annotation Strategy[1]	LYD						BF					
	$\sigma^2_{g_1}$	$\sigma^2_{g_2}$	σ^2_s	σ^2_p	σ^2_e	h^2	$\sigma^2_{g_1}$	$\sigma^2_{g_2}$	σ^2_s	σ^2_p	σ^2_e	h^2
GBLUP	1.12		0.16	0.01	1.06	0.48	5.02		0.83	0.00	4.88	0.47
Genic	1.04		0.17	0.01	1.10	0.45	4.73		0.87	0.00	5.02	0.45
Genic+R	0.65	0.45	0.16	0.01	1.06	0.47	3.18	1.80	0.82	0.00	4.89	0.47
AG	0.37		0.39	0.01	1.54	0.16	0.74		1.89	0.08	7.22	0.07
AG+R	0.15	1.02	0.15	0.00	1.06	0.49	0.04	4.97	0.82	0.00	4.88	0.47

[1]Annotation strategy: GBLUP used all SNPs with no annotation, Genic used genic SNPs, and AG used SNPs mapped to trait-associated genes. The inclusion of SNPs unrelated to the specific annotation strategy is denoted with +R. Variance components: $\sigma^2_{g_1}$ = genetic variance of the specific annotation strategy, $\sigma^2_{g_2}$ = genetic variance not associated with the specific annotation strategy, σ^2_s = litter variance, σ^2_p = pen variance, σ^2_e = residual variance, and h^2 = heritability.

Table 2. Prediction accuracy (ρ) and root mean squared error (*d*) from 5-fold cross-validation for LYD and BF for GBLUP and *f*GBLUP with different annotation strategies.

Annotation Strategy[1]	LYD			BF		
	m^2	ρ	*d*	m^2	ρ	*d*
GBLUP	426,861	0.26	1.13	426,873	0.26	2.41
Genic	211,568	0.26	1.13	211,576	0.26	2.41
Genic+R	211,568+215,293	0.26	1.14	211,576+215,293	0.26	2.41
AG	1042	0.16	1.13	1546	0.11	2.42
AG+R	1,042+425,819	0.27	1.13	1,546+425,327	0.26	2.41
			Null sets[3]			
AG*	602	0.10	1.14	736	0.07	2.42
AG+R*	602+426,259	0.26	1.13	736+ 426,137	0.26	2.41

[1]Annotation strategy: GBLUP used all SNPs with no annotation, Genic used genic SNPs, and AG used SNPs mapped to associated genes. The inclusion of SNPs unrelated to the specific annotation strategy is denoted with +R. [2]*m*: Number of SNP markers in each annotation strategy.

[3]Null sets: For AG*, SNPs were mapped to randomly selected genes of equal size to those in AG.

Conclusions

This study has demonstrated that SNP prioritizing using different strategies to preselect SNPs based on biological knowledge can influence genetic variance explanation and the prediction accuracy of phenotypes. Notwithstanding, the extent of improvement depends on the trait evaluated and the annotation strategy used to prioritize SNPs.

References

Bhuiyan M.S.A., Lim D., Park M., *et al*. (2018) Front. Genet. 9:217. doi.org/10.3389/fgene.2018.00217

Cheng J., Putz A. M., Harding J. C. S., *et al*. (2020) J. Anim Sci. 98(8). doi.org/10.1093/jas/skaa244

Edwards S. M., Sørensen I. F., Sarup P., *et al*. (2016) Genetics 203:1871–1883. doi.org/10.1534/genetics.116.187161

Hu Z.-L., Park C. A., and Reecy J. M. (2022) Nucl. Acids Res. 50(D1):D956–D961. doi.org/10.1093/nar/gkab1116

Ma R., Kuang R., Zhang J., *et al*. (2025) J. Genet. Genom. 52:1588–1599. doi.org/10.1016/j.jgg.2025.03.013

Morgante F., Huang W., Sørensen P., *et al*. (2020) G3 10(12):4599–4613. doi.org/10.1534/g3.120.401847