Zhang, C., Jiang, W., Jiang, X., Wu, R. (2026) Evaluation of the genotype imputation accuracy between SNP Chips in different densities and their impacts on (g)EBV prediction in pig breeding, *Proceedings 13th World Congress on Genetics Applied to Livestock Production* 24388: https://doi.org/10.31274/wcgalp.24388 ©2026 *Zhang, et al.* CC BY-NC-ND

Evaluation of the genotype imputation accuracy between SNP Chips in different densities and their impacts on (g)EBV prediction in pig breeding

C. Zhang[1*], W. Jiang[1], X. Jiang[1], and R. Wu[1]

[1] *Sichuan WanJiaHao Animal Breeding Company, Meishan, China;* [*]*Corresponding author:* chunyan.czh@gmail.com

Introduction

Genomic selection has revolutionized pig breeding by enabling more accurate selection using genome-wide markers (Meuwissen *et al*., 2001). While high-density SNP arrays provide reliable genomic estimated breeding values (gEBVs), large-scale genotyping remains costly. Low-density panels combined with genotype imputation offer a cost-effective alternative (Habier *et al*., 2009), but imputation accuracy is critical to their success, which is influenced by many factors, such as reference population, haplotype diversity, relationships, and allele frequency (Badke *et al*., 2013; Ding *et al*., 2023). In this study, we investigated imputation among porcine 5K, 50K, and 75K SNP chips in a Yorkshire population to optimize large-scale genotyping strategies. Further, we evaluated the impact of the imputed genotypes on genomic prediction for age and longissimus muscle depth adjusted to 120kg. These findings provide practical guidance for cost-efficient genomic selection in commercial pig breeding programs.

Materials & Methods

Animals and data collection. A total of 18,097 purebred Yorkshire pigs born between 2023–2025 at WJH nucleus farm were used for this study. A deep pedigree was traced back ~10 generations including 54,718 animals. When the pen's average weight reached ~120 kg (110.5 ± 12.2 kg), individual body weight and longissimus muscle depth (measured using ultrasound probing) were collected.

Sampling and Genotyping. Ear tissue samples were collected for 6,626 animals. All animals were genotyped with the KPS Porcine 50K SNP chip (Beijing Compass Biotechnology Co., China). A subset of 124 animals, representing population diversity and family structure, was genotyped with the KPS Porcine 5K SNP chip (Beijing Compass Biotechnology Co., China) and the CAU 75K SNP chip (LASO Biotech, Suzhou, China). This subset sampling was based on pedigree analysis (Zhang *et al*., 2013), including 44 influential ancestors, 61 offspring with at least one parent in the reference set, and 19 from additional diversity picks. The 5K SNP chip is a subset of the 50K chip. The 75K SNP chip has 11,881 SNPs overlapping with the 50K SNP chip.

Experimental design. For imputation from 5K to 50K, 6,502 animals with observed 50K genotypes were used as the reference population, and an additional 124 animals with 5K genotypes served as the validation set. Imputation from 50K to 75K was implemented using two approaches: 1) the 124 animals with observed 75K genotypes were used as the reference to impute the remaining 6,502 animals from 50K to 75K for downstream genomic prediction; and 2) imputation accuracy was evaluated within the 124 animals using 10-fold cross-validation, where animals were randomly divided into 10 folds evenly. In each iteration, one fold was treated as validation (retained 50K only) and the remaining folds as reference (75K). Imputation accuracy was assessed using concordance rate (CR) and dosage squared

correlation (R^2) between observed and imputed genotypes. For breeding value prediction, 115 animals out of the above 124 sampling with phenotypes and both observed and imputed genotypes were used as the validation set. A total of 6,384 animals with phenotypes, genotypes and pedigree information comprised the reference population, in which phenotypes of the 115 validation animals were masked and their descendants and full-siblings were excluded. Four scenarios were compared based on the information used to construct the relationship matrix: pedigree only (S1); observed 50K for both reference and validation animals (S2); observed 50K for the reference and imputed 50K for the validation set (S3); and imputed 75K for the reference and observed 75K for the validation set (S4). Prediction accuracy was calculated as the correlation between (g)EBVs and adjusted phenotypes divided by the square root of heritability, and dispersion bias was the regression coefficient of adjusted phenotypes on (g)EBVs.

Statistical methods. SNPs were aligned to *Sscrofa 11.1*, coded 0/1 for reference/alternate alleles. Individuals with call rate < 90%, SNPs on non-autosomes or MAF < 1% or call rate < 90% or failing Hardy–Weinberg equilibrium test (p-value < 1×10^{-6}) were removed using PLINK 1.9 (Purcell *et al*., 2007). After filtering, 41,017 SNPs from 50K, 5,440 from 5K, and 50,106 SNPs from 75K remained. Imputation was conducted using Beagle 5.2 (Browning *et al*., 2018). The SNPs with imputation $R^2 < 0.8$ were excluded from relationship matrix construction as discussed by Bolormaa *et al*. (2019). The age (AGE) and LMD at measurement were adjusted to the body weight of 120kg. A bi-variate animal model below (Eq. 1) was used to estimate breeding values:

$$\mathbf{y}=\boldsymbol{Xb}+\mathbf{Z}_a\mathbf{a}+\mathbf{Z}_l\mathbf{l}+\mathbf{e} \quad (1)$$

where $\mathbf{y}$ is the vector of phenotype for the traits, $\boldsymbol{b}$ is a vector of fixed effects, including sex and contemporary group (combination of probe date and pen), $\mathbf{a}$ is a vector of random additive genetic effects [$\mathbf{a} \sim N(\mathbf{0}, \mathbf{A}\sigma_a^2)$], where $\mathbf{A}$ is the relationship matrix derived by pedigree (S1) or by genome (S2-4) and σ_a^2 is the additive genetic variance, $\mathbf{l}$ is a vector of litter effects [$\mathbf{l} \sim N(\mathbf{0}, \mathbf{I}\sigma_l^2)$], where $\mathbf{I}$ is an identity matrix and $\mathbf{e}$ is a vector of random residuals with variance σ_e^2 [$\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$], and $\mathbf{X}$, $\mathbf{Z_a}$, and $\mathbf{Z_l}$, are incidence matrices associating $\boldsymbol{b}$, $\mathbf{a}$, and $\mathbf{l}$ with $\mathbf{y}$. All analyses were performed using the HiBLUP package (Yin *et al*., 2023).

Results

Imputation accuracy is summarized in Table 1. When imputing 5K to 50K, the overall CR per-individual was 98.72%, with the highest accuracy observed for individuals with at least one parent represented in the reference set. The overall R^2 was 0.976, with the highest accuracy for SNPs with minor allele frequency (MAF) ≥ 0.2 and the lowest for SNPs with MAF < 0.05. In contrast, imputing 50K to 75K using 10-fold cross-validation resulted in lower accuracy, with an average CR per-individual of 84.94% and an overall R^2 of 0.475. The accuracy of (g)EBVs is shown in Table 2. Compared with pedigree-based EBVs, incorporating real 50K genotypes (gEBV1) increased accuracy from 0.39 to 0.46 for AGE and from 0.26 to 0.69 for LMD. Using imputed 50K from 5K (gEBV2) or imputed 75K from 50K (gEBV3) resulted in a slight decline in prediction accuracy compared with gEBV1. However, the dispersion bias decreased further from 1 for both traits, indicating the inflation of the (g)EBVs when using imputed genotypes.

Table 1. Average Concordance rate (CR) per-animal and average dosage squared correlation (R^2) of the SNPs grouped by minor allele frequency (MAF).

Imputation type	Scenarios	Average CR per animal (%)	R^2 for SNPs with different MAF				
			0.01-0.05	0.05-0.1	0.1-0.2	≥ 0.2	Average
5K→50K	Overall	98.72	0.922	0.971	0.975	0.984	0.976
	With parents	99.23	0.785	0.942	0.984	0.990	0.970
	Without parents	98.87	0.787	0.941	0.983	0.989	0.969
50K→75K	Overall[1]	84.94 (0.01)	0.198 (0.019)	0.373 (0.009)	0.470 (0.004)	0.541 (0.004)	0.475 (0.004)

[1]: mean and standard deviation from the 10-fold cross-validations

Table 2. Accuracy and bias of (g)EBVs based on the relationship source.

Relationship source	AGE		LMD	
	Accuracy	Bias	Accuracy	Bias
EBV: only pedigree (S1)	0.39	0.77	0.26	0.45
gEBV1: real 50K (S2)	0.46	0.57	0.69	1.00
gEBV2: imputed 50K from 5K (S3)	0.44	0.55	0.65	0.95
gEBV3:imputed 75K from 50K (S4)	0.45	0.55	0.66	0.94

Discussion

Genotype imputation. Imputation from 5K to 50K achieved high accuracy, consistent with previous studies in pigs demonstrating that large, well-connected reference populations enable reliable imputation across SNP densities (Habier *et al*., 2009; Badke *et al*., 2013; Ding *et al*., 2023). In contrast, imputation from 50K to 75K using 10-fold cross-validation within 124 animals resulted in substantially lower accuracy (CR ≈ 85%, R^2 ≈ 0.47). Although these 124 animals were selected to represent population diversity, the limited size of the reference set likely restricts the capture of all variants, particularly rare alleles. Similar limitations were reported when reference sizes are small or lack representative haplotypes (Ding *et al*., 2023). The presence of parents in the reference set markedly improved individual-level accuracy, confirming the importance of close familial relationships for haplotype inference in structured pig populations (Habier *et al*., 2009; Badke *et al*., 2013). At the SNP level, imputation accuracy was high for common variants but substantially reduced for rare variants, likely due to the limited linkage disequilibrium for low-frequency alleles. The improved accuracy for rare SNPs observed when all 124 animals were included reflects the contribution of influential ancestors and broader population diversity, consistent with findings from expanded pig reference panels (Ding *et al*., 2023).

Impacts on (g)EBVs. Compared to only using pedigree, the incorporation of observed 50K genotypes substantially improved prediction accuracy of AGE from 0.39 to 0.46 and LMD from 0.26 to 0.69, which is similar to the reports in Yorkshire by Badke *et al.* (2014) and other results reviewed by Calus *et al.* (2014), further confirming the well-established advantages of genomic selection over pedigree-based evaluation (Meuwissen *et al*., 2001). Replacing the observed 50K genotypes with imputed genotypes for the selection candidates yielded a slight reduction in gEBVs accuracy, accompanied by slightly increased dispersion bias as well. The similar results were reported in Yorkshire by Badke *et al.* (2014), reflecting that the small imputation errors may propagate through the genomic relationship matrix and impact the estimation accuracy. Notably, using the imputed 75K genotypes had a slightly higher prediction accuracy compared with using the imputed 50K, but did not have a higher prediction accuracy relative to using the observed 50K. The relatively low imputation accuracy of the 75K due to the limited reference set might be the main reason. This implies that the benefit of genomic evaluation not only depends on the higher density of markers, more importantly on the imputation accuracy, variants calling reliability, and population structure (Habier *et al*., 2013; Calus *et al*., 2014).

Conclusions

With an existing large, well-connected 50K reference population, using the 5K SNP chip imputed to 50K provides a cost-effective strategy for large-scale genotyping with minimal loss in genomic prediction accuracy. Continued 50K genotyping of a small number of nucleus replacement parents can further enhance imputation performance for the coming descendants. In contrast, upgrading from 50K to 75K by imputation alone remains challenging with the current reference size. Its impact on genomic prediction needs further evaluation once sufficient imputation accuracy is achieved.

Acknowledgements

This work was financially supported by Sichuan Haiju Yinzhi plan (26RCYJ0050).

References

Badke Y.M., Bates R.O., Ernst C.W., Raney N.E., Cantet R.J.C., et al. (2013) BMC Genet. 14:8. doi.org/10.1186/1471-2156-14-8

Badke Y.M., Bates R.O., Ernst C.W., Fix J., and Steibel J.P. (2014) G3. 4(4):623–631. doi.org/10.1534/g3.114.010504

Bolormaa S., Chamberlain A.J., Khansefid M., Stothard P., Swan A.A., et al. (2019) Genet Sel Evol 51. doi.org/10.1186/s12711-018-0443-5

Browning B.L., Zhou Y., and Browning S.R. (2018) Am. J. Hum. Genet. 103(3):338-348. doi.org/10.1016/j.ajhg.2018.07.015

Calcus M.P.L., Bouwman A.C., Hickey J.M., Veerkamp R.F., and Mulder H.A. (2014) Animal 8(11): 1743-1753. doi:10.1017/S1751731114001803

Ding R., Savegnago R., Liu J., Long N., Tan C., et al. (2023) Commun Biol. 6(1):577. doi:10.1038/s42003-023-04933-9

Habier D., Fernando R.L., and Dekkers J.C.M. (2009) Genetics 182(1):343–353. doi: 10.1534/genetics.108.100289

Meuwissen T.H.E., Hayes B.J., and Goddard M.E. (2001) Genetics 157:1819–1829. doi.org/10.1093/genetics/157.4.1819

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A., et al. (2007) Am. J. Hum. Genet. 81(3):559–575. doi.org/10.1086/519795

Yin L., Zhang H., Tang Z., Yin D., Fu Y., et al. (2023) Nucleic Acids Res. 51(8):3501–3514. doi.org/10.1093/nar/gkad074

Zhang P., Zhan X., Rosenberg N.A., and Zöllner S. (2013) Genetics 195(2):319–330. doi.org/10.1534/genetics.113.154591