Genetics Selection Evolution

RESEARCH ARTICLE

Open Access

Predictive ability of multi-population genomic prediction methods of phenotypes for reproduction traits in Chinese and Austrian pigs

Xue Wang[1], Zipeng Zhang[1], Hehe Du[1], Christina Pfeiffer[2], Gábor Mészáros[2] and Xiangdong Ding[1*]

Abstract

Background Multi-population genomic prediction can rapidly expand the size of the reference population and improve genomic prediction ability. Machine learning (ML) algorithms have shown advantages in single-population genomic prediction of phenotypes. However, few studies have explored the effectiveness of ML methods for multi-population genomic prediction.

Results In this study, 3720 Yorkshire pigs from Austria and four breeding farms in China were used, and single-trait genomic best linear unbiased prediction (ST-GBLUP), multitrait GBLUP (MT-GBLUP), Bayesian Horseshoe (BayesHE), and three ML methods (support vector regression (SVR), kernel ridge regression (KRR) and AdaBoost.R2) were compared to explore the optimal method for joint genomic prediction of phenotypes of Chinese and Austrian pigs through 10 replicates of fivefold cross-validation. In this study, we tested the performance of different methods in two scenarios: (i) including only one Austrian population and one Chinese pig population that were genetically linked based on principal component analysis (PCA) (designated as the "two-population scenario") and (ii) adding reference populations that are unrelated based on PCA to the above two populations (designated as the "multi-population scenario"). Our results show that, the use of MT-GBLUP in the two-population scenario resulted in an improvement of 7.1% in predictive ability compared to ST-GBLUP, while the use of SVR and KKR yielded improvements in predictive ability of 4.5 and 5.3%, respectively, compared to MT-GBLUP. SVR and KRR also yielded lower mean square errors (MSE) in most population and trait combinations. In the multi-population scenario, improvements in predictive ability of 29.7, 24.4 and 11.1% were obtained compared to ST-GBLUP when using, respectively, SVR, KRR, and AdaBoost.R2. However, compared to MT-GBLUP, the potential of ML methods to improve predictive ability was not demonstrated.

Conclusions Our study demonstrates that ML algorithms can achieve better prediction performance than multitrait GBLUP models in multi-population genomic prediction of phenotypes when the populations have similar genetic backgrounds; however, when reference populations that are unrelated based on PCA are added, the ML methods did not show a benefit. When the number of populations increased, only MT-GBLUP improved predictive ability in both validation populations, while the other methods showed improvement in only one population.

*Correspondence:
Xiangdong Ding
xding@cau.edu.cn
Full list of author information is available at the end of the article

Background

Today, genomic prediction [1] is widely accepted and has been successfully implemented in animal and plant breeding schemes [2–4]. However, a large reference population size is key to accurate genomic prediction [5, 6]. For small reference populations, e.g., breeds or strains of livestock with small populations, it is rather difficult to obtain a sufficiently large reference population, which limits the predictive ability of genomic prediction. A potential option is to combine multiple populations to construct a large reference population, i.e., multi-population genomic prediction, such as for Holstein populations in the EuroGenomics [7] and North American consortia [6]. The advantages of multi-population genomic prediction have been widely verified and results in dairy cattle suggest that this cost-effective strategy can significantly improve the predictive ability of genomic prediction for numerically small breeds if the reference population is made up of individuals from closely-related breeds [7, 8]. Similarly, studies in pigs have also indicated that, compared with a single reference population, joint reference populations with different genetic backgrounds can further improve the predictive ability of genomic prediction and reduce prediction bias, especially for the prediction of reproductive traits with a low heritability [9]. Likewise, in beef cattle, Bonifazi et al. [10] combined age-adjusted weaning weight phenotypes and genomic data from five Limousin populations using a single-step approach and demonstrated the advantage of using a combined reference population. Cardoso et al. [11] evaluated the potential of improving tick resistance in beef cattle breeds from seven countries through multitrait genomic selection and also demonstrated the benefits of combining data from different breeds.

Currently, the prevailing methods for computing genomic estimated breeding values (GEBV) are the genomic best linear unbiased prediction (GBLUP) method, which is implemented by estimating the variance components and solving the mixed model equations of Henderson [12], and Bayesian methods with different priors using a Markov chain Monte Carlo (MCMC) methods to estimate the required (genetic) parameters [13–15]. Multitrait models are often used for multi-population genomic prediction, which treat the same trait in different populations as different traits and can be used to capture genotype-by-environment (G × E) interactions between populations. However, these models have limitations when there are many environments, as more genetic parameters need to be estimated and model convergence may be difficult to achieve. Thus, it remains difficult to find a method that can perform well when multiple populations with different genetic backgrounds are combined for genomic prediction.

Recently, machine learning (ML) algorithms have been widely and successfully used in gene screening, genotype imputation, genomic prediction, and protein structure and function prediction [16–20]. For genomic prediction, ML algorithms differ from conventional methods in that, as nonparametric methods, they are able to flexibly capture hidden relationships between genotype and phenotype in an adaptive manner, while making few or no specific distributional assumptions for predictors [21]. Accordingly, ML is potentially attractive for handling higher-order nonlinear relationships in high-dimensional genomic data (*e.g.,* epistasis, dominance, or G × E interactions), which are more likely to exist in multi-population genomic prediction. Several studies have shown that ML algorithms, such as support vector machine regression (SVR), kernel ridge regression (KRR), and the AdaBoost ensemble algorithm, have advantages over GBLUP and Bayes B in predicting genomic-enabled prediction values [18, 22, 23].

In spite of growing interest, little research has explored the effectiveness of ML methods for multi-population genomic prediction. Faville et al. [24] compared the prediction performance of GBLUP, kinship using genotyping-by-sequencing (GBS) with depth adjustment (KGD), random forest, and ridge regression models for genomic prediction in five perennial ryegrass populations and found that the predictive ability of KGD and GBLUP were marginally superior or equal to that of ridge regression (RR) and random forest (RF) computational approaches. Moreover, the use of multiple populations without constructing a joint genomic relationship matrix is also challenging for ML methods. Therefore, it is necessary to further explore the feasibility of ML methods for multi-population genomic prediction of phenotypes. In this study, a joint reference population comprised of Chinese and Austrian pig populations was established and the performances of ML methods and of single- and multi-trait GBLUP and Bayesian methods were evaluated in these two populations with similar genetic backgrounds to determine the optimal methods to improve the predictive ability for phenotypes of reproduction traits in Chinese and Austrian pigs.

Methods

Ethics statement

The entire procedure for blood sample collection was carried out in strict accordance with the protocol approved by the Animal Care and Use Committee of China Agricultural University (Permit Number: DK996) and Austrian Pig Breeders Association.

Population and phenotypes

In this study, five Yorkshire pig populations were used. One was provided by the Austria Pig Breeding Association, referred to as 'Austria', and the others were sourced from four breeding farms in China, identified as A, B, C, and D (Table 1). Two reproduction traits, i.e. total number of piglets born (TNB) and number of piglets born alive (NBA), were studied. In the "two-population scenario", we used the Austrian population and Chinese pig population A, which were shown to be genetically linked based on principal component analysis (PCA) of genotypes, for joint genomic prediction. In the "multi-population scenario", we added data from the other three Chinese populations, which had more dispersed genetic backgrounds, to the Austrian and A populations to assess the impact of expanding the reference population size on the predictive ability for the Austrian and the A population. This scenario was termed the "multi-population scenario". Because the main aim of this study was to explore whether adding unrelated populations could improve the prediction ability for two genetically connected populations, we did not make predictions for the other populations in the multi-population scenario.

Derivation of corrected phenotypes

To avoid double-counting parental information, corrected phenotypic values ($\mathbf{y_c}$) derived from pedigree-based estimated breeding values (EBV) were used as response variables in all genomic prediction analyses [25]. For this purpose, single-trait repeatability models were used to estimate EBV for TNB and NBA and genetic parameters, separately for each population. In the model, the fixed effect was herd-year-season, and the random effects were additive genetics ($\mathbf{a}$), permanent environment ($\mathbf{p_e}$), and residuals ($\mathbf{e}$). The random effects were assumed to have the following distributions: $\mathbf{a} \sim N(\mathbf{0}, \mathbf{A}\sigma_a^2)$, $\mathbf{p_e} \sim N(\mathbf{0}, \mathbf{I}\sigma_{pe}^2)$, and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$, where $\mathbf{A}$ is the pedigree-based relationship matrix, $\mathbf{I}$ is the identity matrix, and σ_a^2, σ_{pe}^2, and σ_e^2 are the variances of additive genetic effects, permanent environment effects of sows, and residuals, respectively. The $\mathbf{A}$ matrix for each population was constructed separately using the pedigree of each population, and a total of 14,118 animals were traced across the five populations. The number of generations and full-sib and half-sib families for each population are listed in Additional file 1: Table S1. The estimated genetic parameters are in Table 1. EBV were calculated using the DMUAI procedure in the DMU software [26]. The $\mathbf{y_c}$ values were calculated as the EBV plus the average estimated residuals for multiple parities of a sow following Song et al. [27].

Genotype data and imputation

In this study, the Austrian pigs were genotyped using a customized 50 K SNP panel (59,319 SNPs, Austria 50 K), while all Chinese populations were genotyped using the PorcineSNP50 BeadChip (Illumina, CA, USA) (50,697 SNPs, PorcineSNP50) and population A was also genotyped using the GenoBaits Porcine SNP 50 K panel (Molbreeding, China) (52,000 SNPs, GBTS50K). As shown in Fig. 1a, 31,062 SNPs were shared between the Austria 50 K and PorcineSNP50 panels, 33,192 SNPs between the Austria 50 K and GBTS50K panels, and 30,998 SNPs between the PorcineSNP50 and GBTS50K panels. In total, 3720 individuals were genotyped, and the number of genotyped individuals in each population is in Table 1. The individuals genotyped with the Austria 50 K and PorcineSNP50 panels were imputed to GBTS50K using the Beagle 5.0 software [28], with the reference population for imputation comprising 4839 Yorkshire pigs from multiple farms in China (including the four farms in this study). The theoretical imputation accuracy was assessed

Table 1 Summary of the five Yorkshire populations, number of genotyped individuals and heritability estimates (h^2)

SNP panels	Population	Trait[a]	σ^2_a	σ^2_e	h^2(SE)	Number of records	Birth year	Genotyped animals
Austria 50K	Austria	TNB	0.71	6.69	0.09 (0.03)	3713	2006–2017	591
		NBA	0.49	5.71	0.08 (0.03)			
PorcineSNP50 BeadChip; GenoBaits Porcine SNP 50 K	A	TNB	1.26	8.95	0.12 (0.03)	2841	2016–2020	742
		NBA	0.99	7.49	0.11 (0.03)			
PorcineSNP50 BeadChip	C	TNB	1.62	11.46	0.12 (0.02)	4144	2015–2018	1153
		NBA	1.55	11.45	0.12 (0.02)			
	B	TNB	0.59	7.77	0.07 (0.01)	2209	2015–2018	550
		NBA	0.37	6.73	0.05 (0.01)			
	D	TNB	1.05	11.53	0.08 (0.03)	1209	2018–2019	684
		NBA	0.58	10.96	0.05 (0.02)			

[a] *TNB* total number of piglets born, *NBA* number of piglets born alive, *SE* standard error

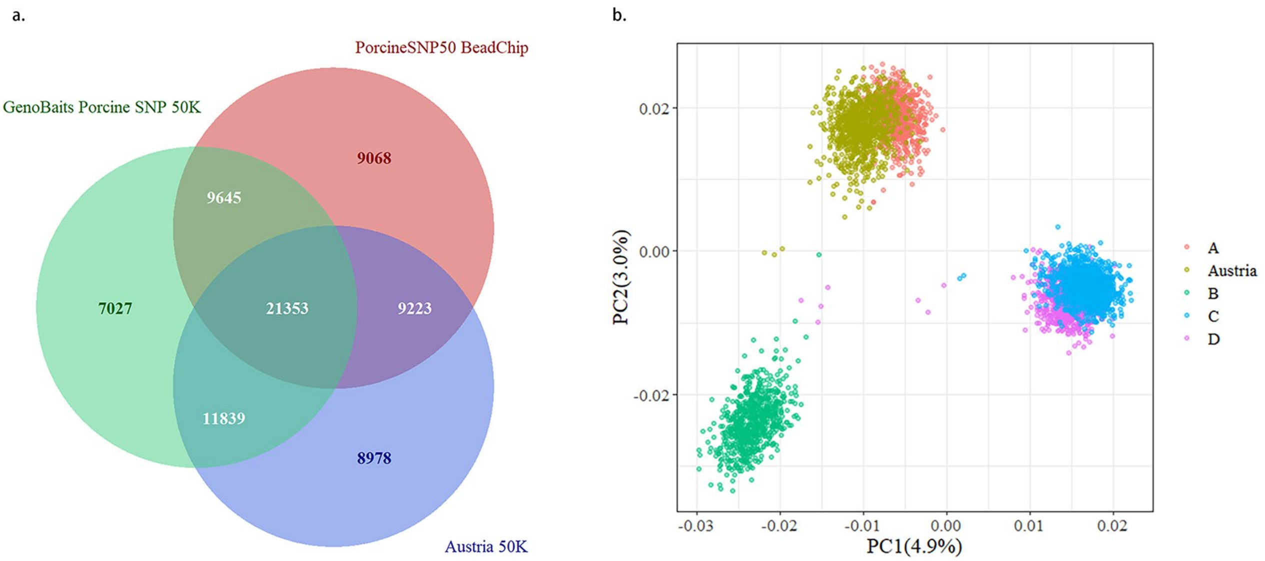

Fig. 1 Venn diagram of the number of shared SNPs between panels (**a**) and principal component analysis of the Austrian and Chinese Yorkshire populations (**b**). *PC1* first principal component, *PC2* second principal component

by the dosage R-squared (DR^2), which is the estimated squared correlation between the estimated allele dose and the true allele dose that is calculated in Beagle 5.0 [28].

No common individuals were genotyped by both the Austria 50 K and the other two SNP panels. Thus, to further evaluate the practical effect of genotype imputation, first, all the non-missing SNPs (1112 SNPs) among the SNPs shared between the Austria 50 K and GBTS50K panels were set as missing. In addition, 2000 SNPs were randomly selected from the 29,488 shared non-missing SNPs of the PorcineSNP50 and GBTS50K panels and set as missing; then, Beagle 5.0 was used for re-imputation and the genotype concordance rate (CR) was calculated to evaluate imputation accuracy. After imputation, the PLINK software [29] was used to remove SNPs with a minor allele frequency (MAF) lower than 0.05 and a call rate less than 0.90 and animals with a call rate less than 0.90. After genotype quality control, all remaining individuals (3720) and 47,734 SNPs on autosomes were retained for further analysis.

Principal component analysis and linkage disequilibrium

To analyse the population structure of the five populations, PCA was performed on the SNP genotypes using the GCTA software [30] and a matrix of eigenvectors, in descending order, representing the principal components (PC), with PC1 having the largest eigenvalue, was created. The LD between each pair of SNPs was measured as r_{LD} and r^2_{LD} (i.e. the square of r_{LD}) [31], using

$$r_{LD} = \frac{f(AB) - f(A)f(B)}{\sqrt{f(A)f(a)f(B)f(b)}},$$

where f(AB), f(A), f(B), f(a), and f(b) are the observed frequencies of haplotype AB and alleles A, B, a, and b, respectively. The average r^2_{LD} across all chromosomes was calculated for each population, and the consistency of LD between the Austria and A populations and with other populations was measured using the correlation of r_{LD} values of pairs of adjacent SNPs on each autosome.

Statistical models

Single-trait GBLUP (ST-GBLUP), multitrait GBLUP (MT-GBLUP), BayesHE, and three ML methods (SVR, KRR and Adaboost.R2) were evaluated. For all methods, the response variables were the corrected phenotypes $\mathbf{y_c}$ and the independent variables for the three ML methods were the vectors of SNP genotypes, encoded as 0, 1, and 2.

Single-trait GBLUP (ST-GBLUP)

The model used for ST-GBLUP was:

$$\mathbf{y_c} = 1\mu + \mathbf{Za} + \mathbf{e},$$

where $\mathbf{y_c}$ is the vector of corrected phenotypes of genotyped individuals; μ is the overall mean, and **1** is a vector of ones; **a** is the vector of additive genetic effects, assumed distributed $N(\mathbf{0}, \mathbf{G}\sigma^2_a)$, where σ^2_a is the additive genetic variance and **G** is the genomic relationship matrix; **Z** is the incidence matrix allocating records to **a**; **e** is the vector of random errors, assumed distributed

$N(\mathbf{0}, \mathbf{I}\sigma_e^2)$, where $\mathbf{I}$ is the identity matrix and σ_e^2 is the residual variance. The $\mathbf{G}$ matrix was constructed following the first method proposed by VanRaden [32]. The ST-GBLUP model was fitted using the DMU software [26] and the variance components were estimated using the average information restricted maximum likelihood (AI-REML) algorithm implemented of the DMUAI procedure.

Multitrait GBLUP (MT-GBLUP)

For the two-population scenario, the MT-GBLUP model was:

$$\begin{bmatrix} \mathbf{y}_{c1} \\ \mathbf{y}_{c2} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix} \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix},$$

where $\begin{bmatrix} \mathbf{y}_{c1} \\ \mathbf{y}_{c2} \end{bmatrix}$ are the vectors of corrected phenotypes for trait 1 and trait 2 (the same trait in the A and Austrian populations); $\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ is the vector of overall means for trait 1 and trait 2; $\begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix}$ are the vectors of additive genetic effects of the two traits, assumed distributed $N(\mathbf{0}, \mathbf{M}\otimes\mathbf{G})$, where $\mathbf{M} = \begin{bmatrix} \sigma_{a1}^2 & \sigma_{a12} \\ \sigma_{a12} & \sigma_{a2}^2 \end{bmatrix}$ represents the genetic variance and covariance matrix of the two traits; $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are the incidence matrices allocating records to $\mathbf{a}_1$ and$\mathbf{a}_2$, respectively ; $\begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix}$ are the vectors of random residual errors, assumed to be normally distributed $N(0, \mathbf{R}\otimes\mathbf{I})$, where $\mathbf{I}$ is the identity matrix and $\mathbf{R} = \begin{bmatrix} \sigma_{e1}^2 & \sigma_{e12} \\ \sigma_{e12} & \sigma_{e2}^2 \end{bmatrix}$ is the residual variance and covariance matrix.

For the multi-population scenario, the MT-GBLUP model can be expressed as:

$$\begin{bmatrix} \mathbf{y}_{c1} \\ \mathbf{y}_{c2} \\ \mathbf{y}_{c3} \\ \mathbf{y}_{c4} \\ \mathbf{y}_{c5} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Z}_3 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{Z}_4 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{Z}_5 \end{bmatrix} \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \mathbf{a}_3 \\ \mathbf{a}_4 \\ \mathbf{a}_5 \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \\ \mathbf{e}_4 \\ \mathbf{e}_5 \end{bmatrix},$$

where $[\mathbf{y}_{c1}, \mathbf{y}_{c2}, \mathbf{y}_{c3}, \mathbf{y}_{c4}, \mathbf{y}_{c5}]^{\mathrm{T}}$ are the vectors of corrected phenotypes for traits 1 to 5 (the same trait in the five populations); $[\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3, \boldsymbol{\mu}_4, \boldsymbol{\mu}_5]^{\mathrm{T}}$ are the vectors of overall means for traits 1 to 5; $[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4, \mathbf{a}_5]^{\mathrm{T}}$ are the vectors of additive genetic effects of the five traits, assumed distributed $N(\mathbf{0}, \mathbf{M}\otimes\mathbf{G})$, where

$$\mathbf{M} = \begin{bmatrix} \sigma_{a1}^2 & \sigma_{a12} & \sigma_{a13} & \sigma_{a14} & \sigma_{a15} \\ \sigma_{a12} & \sigma_{a2}^2 & \sigma_{a23} & \sigma_{a24} & \sigma_{a25} \\ \sigma_{a13} & \sigma_{a23} & \sigma_{a3}^2 & \sigma_{a34} & \sigma_{a35} \\ \sigma_{a14} & \sigma_{a24} & \sigma_{a34} & \sigma_{a4}^2 & \sigma_{a45} \\ \sigma_{a15} & \sigma_{a25} & \sigma_{a35} & \sigma_{a45} & \sigma_{a5}^2 \end{bmatrix}$$

is the genetic variance and covariance matrix of the five traits; $\mathbf{Z}_1$ to $\mathbf{Z}_5$ are the incidence matrices allocating records to $\mathbf{a}_1$ to $\mathbf{a}_5$, respectively; $[\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_5]^{\mathrm{T}}$ are the vectors of random residual errors, assumed distributed $N(\mathbf{0}, \mathbf{R}\otimes\mathbf{I})$, where

$$\mathbf{R} = \begin{bmatrix} \sigma_{e1}^2 & \sigma_{e12} & \sigma_{e13} & \sigma_{e14} & \sigma_{e15} \\ \sigma_{e12} & \sigma_{e2}^2 & \sigma_{e23} & \sigma_{e24} & \sigma_{e25} \\ \sigma_{e13} & \sigma_{e23} & \sigma_{e3}^2 & \sigma_{e34} & \sigma_{e35} \\ \sigma_{e14} & \sigma_{e24} & \sigma_{e34} & \sigma_{e4}^2 & \sigma_{e45} \\ \sigma_{e15} & \sigma_{e25} & \sigma_{e35} & \sigma_{e45} & \sigma_{e5}^2 \end{bmatrix}$$

represents the residual variance and covariance matrix.

For the MT-GBLUP methods, the $\mathbf{G}$ matrix was also constructed according to the first method proposed by VanRaden [32], and the variance components were calculated using the DMU software [26]. For the multi-population scenario, (co)variance component parameters were estimated using bivariate analyses for each pair of populations with the AI-REML algorithm of the DMU software [26], and the (co)variance matrix was converted into a positive definite matrix using a bending procedure [33]. The R package mbend was used to implement unweighted bending of variance components [34], with the other parameters set to default values (i.e. max.iter, small.positive and method were 10000, 0.0001 and "hj", respectively). Given that the variance components in both scenarios were estimated using the complete dataset, we calculated the genetic correlation for each trait among the five populations based on the variance component results of MT-GBLUP in the multi-population scenario. The estimated values of genetic (co)variance and genetic correlation between pairs of the five populations are in Additional file 1: Table S2. Genetic correlations were calculated as $\frac{\sigma_{a12}}{\sqrt{\sigma_{a1}^2\sigma_{a2}^2}}$.

Bayes Horseshoe

Bayes horseshoe (Bayes HE), developed by Shi et al. [35], is a Bayesian model based on global–local priors (i.e., the global parameter shrinks the marker effects to zero, whereas the local parameter allows a marker to escape from shrinkage) to increase the flexibility and adaptability of hyperparameter estimation. BayesHE exhibited the highest or second highest predictive ability compared with traditional Bayesian methods such as BayesA and BayesB [35]. The model fitted for Bayes HE was:

$$\mathbf{y_c} = \mathbf{1}\mu + \sum\nolimits_{k=1}^{m} \mathbf{x}_k \beta_k + \mathbf{e},$$

where $\mathbf{y_c}$ is the vector of corrected phenotypes of genotyped animals, μ is the overall mean, $\mathbf{x}_k$ is the genotype vector of the k-th marker, and m is the number of markers; β_k is the allele substitution effect of the k-th marker, assumed to be distributed as $\beta_k \sim N(0, \lambda_k^2 \tau^2)$, where the local parameter λ_k follows a half-t distribution of $\lambda_k \sim \mathrm{half} - \mathrm{t}^{+}(\upsilon, 1)$ $(\upsilon \sim \mathrm{G}(\mathrm{a}, \mathrm{c}))$ and the global parameter τ follows a positive half-Cauchy distribution of $\tau \sim \mathrm{C}^{+}(0, \mathrm{N}^{-1})$; $\mathbf{e}$ is the vector of random residuals, assumed distributed as $N(\mathbf{0}, \mathbf{I}\sigma_e^2)$. In this study, the first form of BayesHE (BayesHE1) was used, in which a is equal to 4 and c is equal to 1 [35]. Using in-house scripts written in Fortran 95 [35], the Markov chain Monte Carlo (MCMC) chain was run for 50,000 cycles, with the first 20,000 cycles discarded as burn-in, and every 50th sample of the remaining 30,000 iterations was saved for inferring posterior statistics.

Support vector regression

SVR is a classic algorithm for dealing with regression problems in machine learning. It can use the nonlinear kernel function (e.g. radial basis function (RBF) kernel) to map the input data of the original space into the high-dimensional kernel space and model and predict in the high-dimensional kernel space [36]. Therefore, we can construct linear models in the feature space to address regression problems. SVR was fitted using the following model:

$$\mathrm{f}(\mathbf{x}) = \mathrm{b} + \mathbf{h}(\mathbf{x})^{\mathrm{T}} \mathbf{w},$$

where $\mathrm{f}(\mathbf{x})$ is the predicted value, $\mathbf{h}(\mathbf{x})^{\mathrm{T}}$ denotes the kernel function, representing a nonlinear transformation of the predictor variables in $\mathbf{x}$ (i.e., genotype vector), $\mathbf{w}$ is the vector of weights, and b is the intercept. In the context of 'ε-insensitive' SVM regression, the loss was calculated only when the absolute value of the discrepancy between $\mathrm{f}(\mathbf{x}_{\mathrm{i}})$ and $\mathrm{y_i}$ exceeded some constant (ε). The SVR problem can be formalized as [36]:

$$\min_{\mathrm{w,b}} \frac{1}{2} \|\mathbf{w}\|^2 + \mathrm{C} \sum\nolimits_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{V}_{\varepsilon}\left(\mathrm{y_{ci}} - \mathrm{f}(\mathbf{x}_{\mathrm{i}})\right),$$

where $\mathrm{V}_{\varepsilon}(\mathrm{z}) = \begin{cases} 0, & \mathit{if}\, |\mathrm{z}| < \varepsilon; \\ |\mathrm{z}| - \varepsilon, \mathit{otherwise.}' \end{cases}$

with C being the regularization constant, $\mathbf{y_c}$ the vector of corrected phenotypes, $||\cdot||$ is the norm in Hilbert space, n is the sample size, z is the error (i.e., $\mathrm{y_{ci}} - \mathrm{f}(\mathbf{x}_{\mathrm{i}})$), and V_{ε} is the 'ε-insensitive' loss. After optimization, SVR can be expressed as:

$$\mathrm{f}(\mathbf{x}) = \sum\nolimits_{\mathrm{i}=1}^{\mathrm{n}} \left(\widehat{\mathrm{a}}_{\mathrm{i}} - \mathrm{a_i}\right) \mathrm{K}(\mathbf{x}, \mathbf{x}_{\mathrm{i}}) + \mathrm{b},$$

where $\widehat{\mathrm{a}}_{\mathrm{i}}$ and $\mathrm{a_i}$ denote positive weights assigned to each observation and estimated from the data, and $\mathrm{K}(\mathbf{x}, \mathbf{x}_{\mathrm{i}})$ represents the inner product of $\mathbf{x}$ (a new input data point) and $\mathbf{x}_{\mathrm{i}}$ (the i-th data point in the training dataset) after being mapped to a high-dimensional space through a kernel function (i.e. $\mathrm{K}(\mathbf{x}, \mathbf{x}_{\mathrm{i}}) = \mathbf{h}(\mathbf{x}) \cdot \mathbf{h}(\mathbf{x}_{\mathrm{i}})^{\mathrm{T}}$). A grid search was used to identify the optimal kernel function and the hyperparameters for C and gamma (the parameter controlling kernel width in the RBF kernel).

Kernel ridge regression

Kernel ridge regression (KRR), as a nonlinear regression method, can be used to effectively mine the nonlinear structure of data [37]. Like SVR, KRR uses a nonlinear kernel function ($\boldsymbol{\phi}(\mathbf{x}_{\mathrm{i}})$) to map the original data into a high-dimensional feature space and then builds a ridge regression model in the feature space for prediction. The linear regression model is expressed as $\mathrm{y_i} = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_{\mathrm{i}})$, where $\boldsymbol{\beta}$ denotes the weight vector. KRR employs regularized least squares to ascertain the weight vector $\boldsymbol{\beta}$ by minimizing the following objective function [37]:

$$\min \mathrm{L_{KRR}} = \frac{1}{2} \|\boldsymbol{\beta}\|^2 + \frac{1}{2\lambda} \sum\nolimits_{\mathrm{i}=1}^{\mathrm{n}} \left(\mathrm{y_{ci}} - \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_{\mathrm{i}})\right)^2,$$

where λ is the regularization constant. By computing the derivative of $\mathrm{L_{KRR}}$ with respect to $\boldsymbol{\beta}$ and setting the resulting equations to zero, the resultant weight vector $\boldsymbol{\beta}$ is determined as follows:

$$\boldsymbol{\beta} = \left(\boldsymbol{\phi}^{\mathrm{T}} \boldsymbol{\phi} + \lambda \mathbf{I}\right)^{-1} \boldsymbol{\phi}^{\mathrm{T}} \mathbf{y_c},$$

where $\boldsymbol{\phi}$ contains the mapped samples $\boldsymbol{\phi}(\mathbf{x}_{\mathrm{i}})$ in its rows. In other words, $\boldsymbol{\phi}(\mathbf{x}_{\mathrm{i}})$ represents the vector obtained by applying the feature mapping to a single data point $\mathbf{x}_{\mathrm{i}}$, while $\boldsymbol{\phi}$ represents the feature matrix of the entire dataset. $\mathbf{I}$ is the identity matrix. Using the representer's theorem, $\boldsymbol{\beta}$ can be expressed in relation to the dual weights α as:

$$\boldsymbol{\beta} = \sum\nolimits_{i=1}^{n} \alpha_{i}\boldsymbol{\phi}(\mathbf{x}_{i}) = \boldsymbol{\phi}^{T}\alpha.$$

Hence, the closed-form solution for the dual weight α is obtained as follows:

$$\alpha = \left(\boldsymbol{\phi}^{T}\boldsymbol{\phi} + \lambda\mathbf{I}\right)^{-1}\mathbf{y_c} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y_c},$$

where $\mathbf{K}$ is the kernel matrix (i.e., Gram matrix), and $K_{ij} = K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \boldsymbol{\phi}(\mathbf{x}_i) \cdot \boldsymbol{\phi}(\mathbf{x}_j)^{T}$. If the number of training samples is n, the kernel matrix can be expressed as:

$$\mathbf{K} = \begin{bmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & K(\mathbf{x}_1, \mathbf{x}_2) & \cdots & K(\mathbf{x}_1, \mathbf{x}_n) \\ K(\mathbf{x}_2, \mathbf{x}_1) & K(\mathbf{x}_2, \mathbf{x}_2) & \cdots & K(\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & \vdots & \vdots \\ K(\mathbf{x}_n, \mathbf{x}_1) & K(\mathbf{x}_n, \mathbf{x}_2) & \cdots & K(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}_{n\times n}$$

Ultimately, given a new test sample $\mathbf{x}_i$ (i.e. the genotype vector of individual i), the predicted output is derived using dual weights, and the similarity between the test sample $\mathbf{x}_i$ and all training samples is employed for prediction. Thus, the expression of KRR is:

$$y(\mathbf{x}_i) = \mathbf{k}'(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y_c},$$

where $y(\mathbf{x}_i)$ is the predicted value of sample $\mathbf{x}_i$, $\mathbf{k}' = \mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)$ (j = 1,2,3,…,n), and the expanded form of $\mathbf{k}'$ is:

$$\mathbf{k}' = \begin{bmatrix} K(\mathbf{x}_i, \mathbf{x}_1) \\ K(\mathbf{x}_i, \mathbf{x}_2) \\ \vdots \\ K(\mathbf{x}_i, \mathbf{x}_n) \end{bmatrix}^{T}$$

Like SVR, a grid search was used to find the optimal kernel function, λ, and the RBF kernel parameter gamma.

AdaBoost.R2

The expression for AdaBoost.R2 can be written as [38]:

$$f(\mathbf{x}) = \sum\nolimits_{t=1}^{M}\left(\log\frac{1}{\varepsilon_t}\right)f_t(\mathbf{x}) \Big/ \sum\nolimits_{t=1}^{M}\left(\log\frac{1}{\varepsilon_t}\right);$$

where $f(\mathbf{x})$ is the final predicted value and $f_t(\mathbf{x})$ is the predicted value of the *t*-th weak learner; ε_t is the error rate ($\varepsilon_t = \overline{L}_t / \left(1 - \overline{L}_t\right)$); $\overline{L}_t$ is the average loss, and $\overline{L}_t = \sum_{i=1}^{m} L_t(i)D_t(i)$, where $L_t(i)$ is the error between the predicted value and the true value (i.e. corrected phenotype) of the i-th individual, and $D_t(i)$ is the distribution of weights and $D_{t+1}(i) = \frac{D_t(i)\beta_t^{(1-L_t(i))}}{Z_t}$, where Z_t is the normalization factor such that the sum of $D_{t+1}(i)$ is 1. In this study, to reduce the hyperparameter tuning time and to reuse the experience of some of the used learners, KRR was used as the base learner for AdaBoost.R2 (abbreviated as AdaBoost.R2_KRR). When optimizing the hyperparameters of AdaBoost.R2, the optimal number of base learners was unstable and difficult to determine. Therefore, we used the default number of base learners (i.e., 50).

For the above three ML methods, genomic prediction was performed with the help of the sklearn package for Python (V0.22). In addition, considering that the optimal hyperparameters of ML methods might differ between populations, it is not reasonable to train the model for prediction by directly combining populations, especially for populations with different genetic backgrounds. Therefore, a population-specific hyperparameter optimization strategy based on cross-validation was adopted for all three ML methods. For the two-population scenario, first, the reference populations for genomic selection of each population were randomly divided into five groups, four of which were treated as the training set, and the remaining set was treated as the test set. Then, the training sets of the two populations were combined to construct a joint training set, the Pearson correlation between corrected phenotypes $\mathbf{y_c}$ and predicted genetic effects of the test set under different hyperparameter combinations was calculated separately for each population by grid search, and the hyperparameter combination with the highest average Pearson correlation was used as the optimal hyperparameter when predicting this population. The same strategy was used for the multi-population scenario to divide the Austrian and A training but the other three Chinese populations were not divided but were directly added to the training sets. The optimal hyperparameters for multi-population genomic prediction in the two-population and multi-population scenarios are shown in Tables 2 and 3, respectively.

Cross-validation and genomic predictive ability

Because of the inconsistency in birth years between the Austria and A populations (as shown in Table 1) and the small size of both populations, we did not adopt the strategy of using young individuals as the validation set. Instead, 10 replicates of fivefold cross-validation (CV) were performed to estimate the predictive ability, mean square error (MSE), and unbiasedness. For the two-population scenario, all individuals from each population were randomly divided into five groups, with four groups serving as the reference population and the remaining group as the validation population; the reference populations of the two populations were then combined to create a joint reference population. For the multi-population scenario, the same partitioning strategy was used for the Austrian and A reference sets but the other three

Table 2 Optimal hyperparameters for machine learning methods for two-population genomic prediction obtained through a grid search

Trait[a]	Method	Population	
		Austria	A
TNB	SVR	kernel=‘rbf’, C=4, gamma=0.0001	kernel=‘rbf’, C=5, gamma=0.0001
	KRR	kernel=‘rbf’, λ = 0.0001, gamma=0.0001	kernel=‘rbf’, λ = 5, gamma=0.0001
	Adaboost.R2	KRR_kernel=‘rbf’, KRR_ λ= 0.001, KRR_gamma=0.0001	KRR_kernel=’rbf’, KRR_ λ = 2.5, KRR_gamma=0.0001
NBA	SVR	kernel=‘rbf’, C=4, gamma=0.0001	kernel=‘rbf’, C=5, gamma=0.0001
	KRR	kernel=‘rbf’, λ = 0.0001, gamma=0.0001	kernel=‘rbf’, λ = 3, gamma=0.0001
	Adaboost.R2	KRR_kernel=’rbf’, KRR_ λ = 0.01, KRR_gamma=0.0001	kernel=’rbf’, KRR_ λ = 2, KRR_gamma=0.0001

[a] *TNB* total number of piglets born, *NBA* number of piglets born alive, *SVR* support vector regression, *KRR* kernel ridge regression

Table 3 The optimal hyperparameters for machine learning methods for multi-population genomic prediction obtained through a grid search

Trait[a]	Method	Population	
		Austria	A
TNB	SVR	kernel=‘rbf’, C=3, gamma=0.0001	kernel=‘rbf’, C=1, gamma=0.0001
	KRR	kernel=‘rbf’, λ = 5, gamma=0.0002	kernel=‘rbf’, λ = 2, gamma=0.0001
	Adaboost.R2	KRR_kernel=‘rbf’, KRR_ λ = 0.1, KRR_gamma=0.0001	KRR_kernel=’rbf’, KRR_ λ = 0.1, KRR_gamma=0.0001
NBA	SVR	kernel=‘rbf’, C=3, gamma=0.0001	kernel=‘rbf’, C=5, gamma=0.0001
	KRR	kernel=‘rbf’, λ=4, gamma=0.0002	kernel=‘rbf’, λ = 2, gamma=0.00005
	Adaboost.R2	KRR_kernel=’rbf’, KRR_ λ = 0.05, KRR_gamma=0.0001	kernel=’rbf’, KRR_ λ =0.05, KRR_gamma=0.0001

[a] *TNB* total number of piglets born, *NBA* number of piglets born alive, *SVR* support vector regression, *KRR* kernel ridge regression

Chinese populations were not divided; instead, they were directly added to the reference population. Notably, for both scenarios, the reference and validation populations for genomic prediction were the same for all methods for each replicate of the fivefold CV.

For all methods, predictive ability was assessed as the Pearson correlation between corrected phenotypes $\mathbf{y_c}$ and predicted values (i.e., breeding values in GBLUP and BayesHE, and genetic effects including additive and non-additive effects in ML methods) of the validation population, prediction bias was calculated as the regression of $\mathbf{y_c}$ on predicted values of the validation population, and MSE was computed as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left(f_i' - y_{ci}'\right)^2,$$

where n represents the number of animals in each validation data set of the fivefold CV, and the vectors $\mathbf{f}'$ and $\mathbf{y}_\mathbf{c}'$ represent the centralized predicted values (i.e., predicted value minus the mean of the validation population) and centralized observed values (i.e., corrected phenotype minus the mean of the validation population), respectively. The overall predictive ability, MSE, and unbiasedness were the averages of 10 replicates of fivefold CV. In addition, for all evaluation metrics (predictive ability, MSE and unbiasedness), the standard error of the fivefold CV results for each replicate was calculated, and the final standard error reported was the average of 10 replicates. To compare the prediction accuracy of different methods, multiple t-tests (with P-values adjusted using the Bonferroni method) were conducted based on the outcomes of 10 replicates.

Results

Population structure and genetic parameters

Due to the lack of genetic exchange between the five pig populations, no pedigree connections between the different populations were observed. However, the PCA SNP genotypes of the five populations (Fig. 1b) showed that the genetic backgrounds of the Austrian and A pigs are relatively closer than those of the other three Chinese populations.

Table 4 illustrates the LD between adjacent SNPs across populations. The average r^2_{LD} of adjacent SNPs on each chromosome in the Austrian, A, B, C, and D populations ranged from 0.33 to 0.42, 0.34 to 0.43, 0.34 to 0.44, 0.34

Table 4 Linkage disequilibrium between adjacent SNPs by chromosome and populations

Chr	Length (Mb)	Number of SNPs	Mean distance (kb)	Mean r^2_{LD}					Cor[a]						
				Austria[c]	A[c]	B[c]	C[c]	D[c]	Aus-A	Aus-B	Aus-C	Aus-D	A–B	A–C	A–D
1	261.26	5177	48.14	0.41	0.41	0.44	0.44	0.44	0.49	0.29	0.42	0.42	0.29	0.39	0.41
2	144.32	3133	44.59	0.37	0.39	0.39	0.39	0.39	0.30	0.31	0.37	0.36	0.26	0.29	0.33
3	126.4	2849	42.51	0.36	0.36	0.37	0.39	0.39	0.51	0.34	0.36	0.28	0.21	0.36	0.35
4	124.7	2851	38.9	0.41	0.41	0.40	0.44	0.45	0.41	0.29	0.34	0.34	0.29	0.32	0.31
5	99.43	2222	43	0.35	0.36	0.36	0.37	0.38	0.40	0.38	0.35	0.34	0.29	0.45	0.39
6	162.85	2839	53.6	0.40	0.40	0.42	0.41	0.42	0.45	0.36	0.40	0.44	0.36	0.41	0.42
7	115.97	2914	38.11	0.34	0.35	0.36	0.36	0.37	0.41	0.31	0.29	0.28	0.30	0.32	0.35
8	132.37	2871	43.73	0.34	0.36	0.36	0.37	0.38	0.46	0.32	0.33	0.33	0.35	0.33	0.29
9	132.86	3026	41.98	0.37	0.39	0.37	0.41	0.41	0.51	0.32	0.46	0.36	0.27	0.41	0.43
10	65.89	1832	34.64	0.33	0.34	0.36	0.34	0.34	0.40	0.31	0.37	0.38	0.31	0.38	0.34
11	75.35	1773	40.78	0.33	0.34	0.35	0.36	0.36	0.36	0.22	0.36	0.22	0.21	0.37	0.37
12	58.11	1719	33.2	0.39	0.40	0.38	0.41	0.41	0.45	0.25	0.32	0.28	0.25	0.27	0.26
13	198.36	3795	49.51	0.42	0.43	0.43	0.44	0.44	0.51	0.49	0.50	0.48	0.33	0.47	0.47
14	135.07	3359	38.37	0.41	0.41	0.41	0.44	0.44	0.37	0.30	0.26	0.31	0.30	0.33	0.29
15	133.57	2739	44.63	0.40	0.40	0.40	0.43	0.43	0.32	0.33	0.35	0.41	0.32	0.32	0.38
16	75.58	1776	39.79	0.34	0.35	0.34	0.36	0.37	0.29	0.27	0.31	0.30	0.19	0.35	0.25
17	59.77	1605	36.19	0.34	0.34	0.34	0.37	0.38	0.35	0.26	0.25	0.25	0.40	0.28	0.35
18	53.08	1254	39.22	0.34	0.36	0.37	0.36	0.37	0.36	0.28	0.34	0.37	0.25	0.40	0.38
Mean	–	–	41.72	0.37	0.35	0.38	0.39	0.40	0.41	0.32	0.35	0.34	0.29	0.36	0.35
Total	2154.94[b]	47.734[b]	–	–	–	–	–	–	–						

[a] Austria-A/B/C/D: the correlation of r_{LD} of pairs of adjacent SNPs between the Austrian and A/B/C/D populations; A-B/C/D: the correlation of r_{LD} of pairs of adjacent SNPs between the A and B/C/D populations

[b] Across the 18 porcine autosomes

[c] Yorkshire pig populations from Austria, A, B, C and D

to 0.44, and 0.34 to 0.45, respectively, and the average r^2_{LD} across all chromosomes in the Austria, A, B, C, and D populations were equal to 0.37, 0.35, 0.38, 0.39, and 0.40, respectively. The average correlation of r_{LD} (i.e., LD consistency) between adjacent SNPs across all chromosomes was 0.41, 0.32, 0.35, and 0.34 for Austria with populations A, B, C, and D, respectively; for population A, the average correlation of r_{LD} with populations B, C, and D was 0.29, 0.36, and 0.35, respectively. The correlations of r_{LD} also suggested the genetic relationship between the Austria and A populations is closer than that between the Austria and other populations.

Pedigree-based estimates of heritability for each population using the single-trait repeatability model during the derivation of corrected phenotypes are in Table 1. The heritability estimates for TNB and NBA ranged from 0.05 to 0.12 in the five populations.

Genotype imputation accuracy

The imputation accuracy at different MAF intervals and for each chromosome is illustrated in Fig. 2. DR^2 was sensitive to MAF when the Austria 50 K genotypes were imputed to GBTS50K and was relatively low for markers with a MAF lower than 0.05 (Fig. 2a). No significant differences in DR^2 and CR were observed between chromosomes (Fig. 2b, d). Overall, the average DR^2 values of the Austria 50 K and PorcineSNP50 genotypes imputed to GBTS50K were 0.89 and 0.97, respectively, while the average CR of the Austria 50 K and PorcineSNP50 genotypes imputed to GBTS50K was 0.96 and 0.99, respectively. Accordingly, both imputations were sufficiently accurate to analyse the three SNP panels together.

Genomic prediction in the two-population scenario

Comparison of ML methods with other methods

Figure 3 presents the average predictive ability, MSE and unbiasedness values for the TNB and NBA traits in the two-population scenario (the raw values of Fig. 3 are in Additional file 1: Table S3). In all population and trait combinations, the highest predictive abilities were achieved using SVR and KRR. However, the predictive ability of AdaBoost.R2 was not significantly higher than that of MT-GBLUP for all traits and populations. When SVR, KRR, and Adaboost.R2 were used, improvements in average predictive ability were 4.5, 5.3, and −0.8% (averaged for two traits and two populations), respectively,

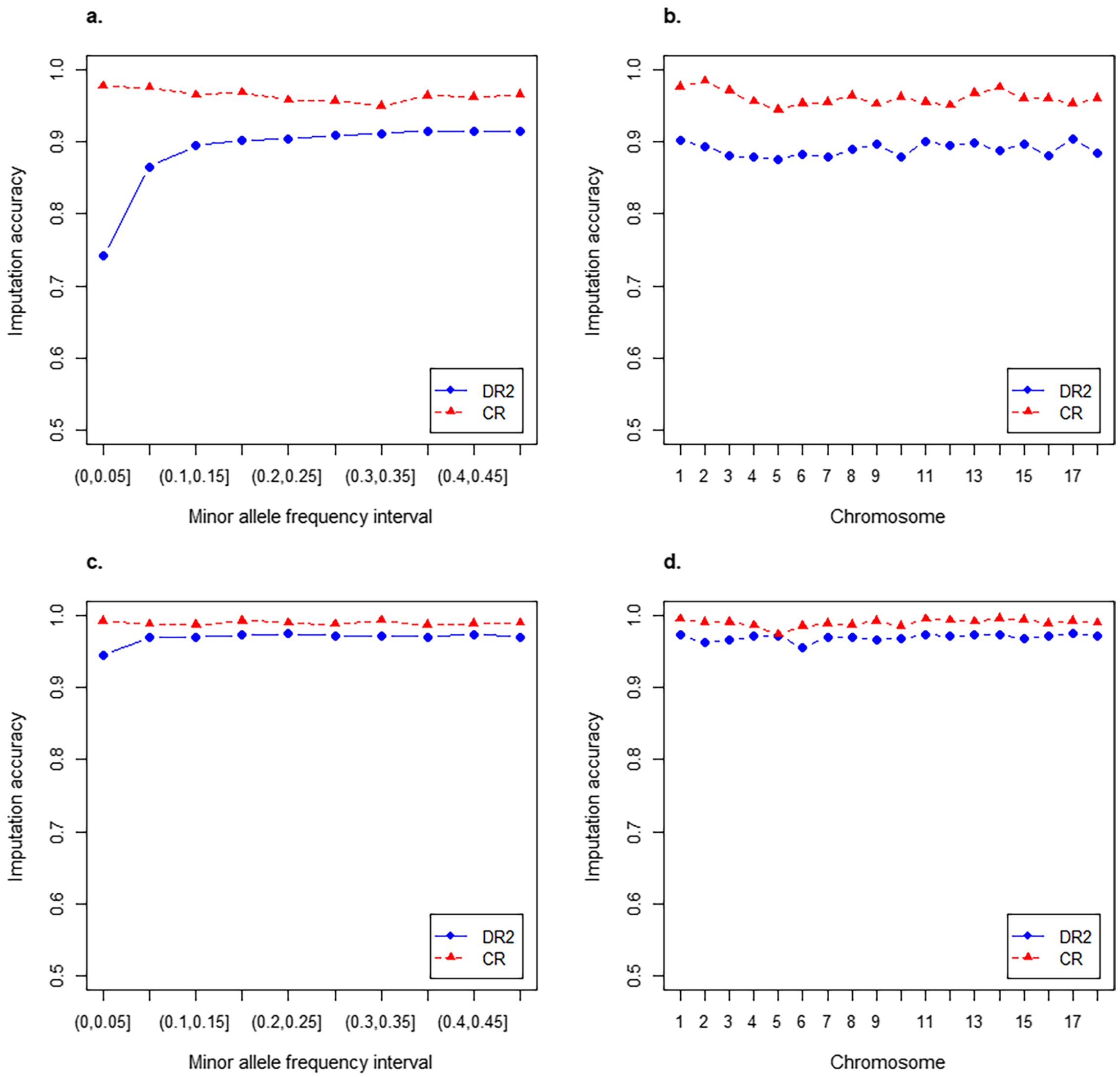

Fig. 2 Average imputation accuracy for SNPs with different minor allele frequency (MAF) intervals and chromosomes. **a**, **b** Average imputation accuracy of Austria 50 K genotypes to GenoBaits Porcine SNP 50 K. **c**, **d** Average imputation accuracy of the PorcineSNP50 BeadChip genotypes to GenoBaits Porcine SNP 50 K. DR2: average estimated squared correlation between the imputed and the true allele dose; genotype concordance rate (CR): average ratio of the correctly imputed genotypes

compared to MT-GBLUP and 11.9, 12.6, and 6.1% (averages of both traits and both populations), respectively, compared to ST-GBLUP. The average improvement in predictive ability of the ML methods compared to ST-GBLUP, MT-GBLUP, and BayesHE were 10.2, 3.0, and 10.8%, respectively (averages of both traits and both populations). Although the predictive abilities of SVR and KRR were higher than those of MT-GBLUP in all population and trait combinations, this superiority was significant only for the A population and not for the Austrian population. Compared to MT-GBLUP, the improvement obtained with ML was greater for population A than for the Austrian population, while compared to ST-GBLUP, it was smaller for population A than for the Austrian population. Compared to MT-GBLUP, SVR, and KRR produced lower MSE in most scenarios, while Adaboost. R2 resulted in lower MSA than MT-GBLUP only for A population. In terms of unbiasedness, in the A population, AdaBoost.R2 generated the largest bias for both TNB and NBA traits, while the other methods yielded small biases (i.e., the regression coefficients were close to 1); in the Austria population, the biases of all three ML methods were much larger than those of ST-GBLUP, MT-GBLUP, and BayesHE.

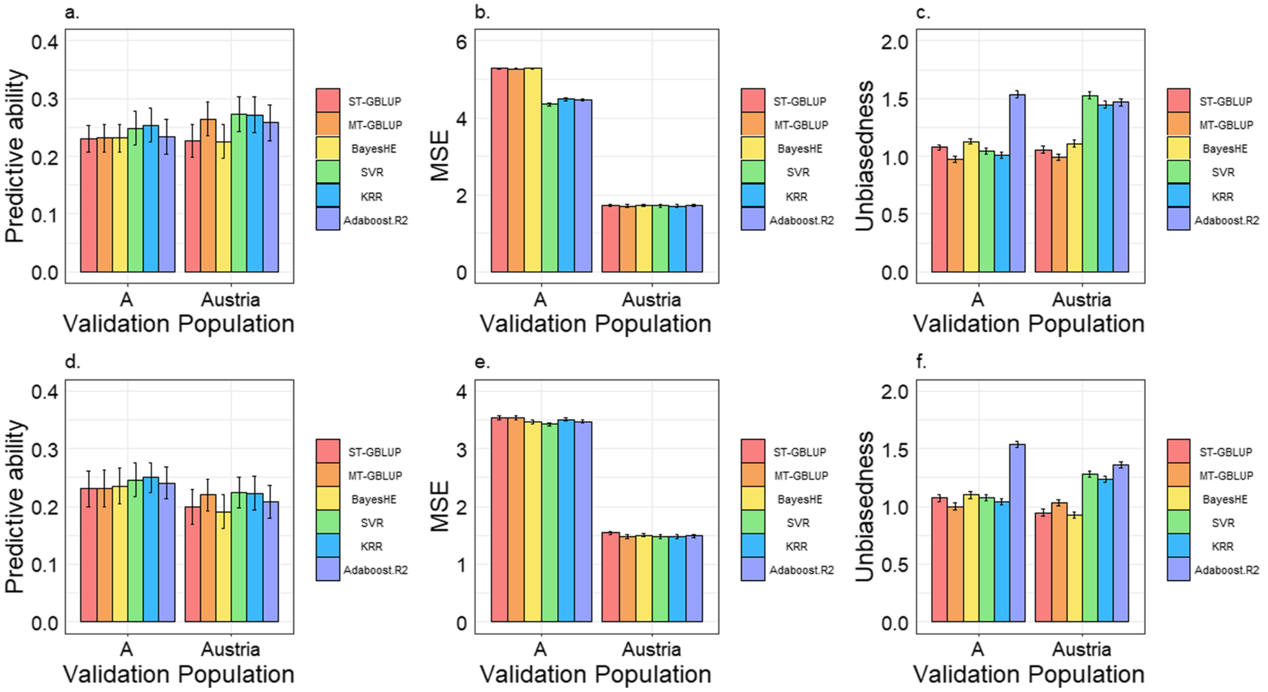

Fig. 3 Predictive ability, mean squared error (MSE), and bias for genomic predictions for two populations with genetically linked backgrounds. **a**–**c** Predictive ability, MSE, and bias for total number of piglets born (TNB). **d**–**f** Predictive ability, MSE, and bias for number of piglets born alive (NBA). ST-GBLUP: single-trait genomic best linear unbiased prediction model; MT-GBLUP: multitrait GBLUP model. The error bar represents the standard error

Comparison of ST-GBLUP, MT-GBLUP and BayesHE

We found that the predictive ability was higher when MT-GBLUP was used compared to ST-GBLUP, with average improvements of 8.8 and 5.5% for TNB and NBA, respectively. Bias of predictions from MT-GBLUP were also closer to 1 (i.e., smaller bias) than for ST-GBLUP and MT-GBLUP yielded smaller MSE.

As shown in Fig. 3, the predictive ability of ST-GBLUP was similar to that of BayesHE for any population and none of these differences were statistically significant. However, MSE for TNB prediction were lower for ST-GBLUP than for BayesHE, but the opposite was observed for NBA (see Additional file 1: Table S3). For TNB, BayesHE resulted in a larger bias than ST-GBLUP (average regression values were 1.12 and 1.07, respectively) but for NBA, the biases were similar for these two methods.

In addition to the comparison of prediction performance, estimates of the genetic correlation estimated using MT-GBLUP between the Austrian and A populations were 0.62 and 0.40 for TNB and NBA, respectively, with standard errors of 0.125 and 0.152, respectively (see Additional file 1: Table S2).

Genomic prediction in the multi-population scenario

Comparison of different methods

The genomic predictive abilities when the reference population size was expanded to multiple populations are shown in Fig. 4 (the raw values of Fig. 4 are also in Additional file 1: Table S3). Our results indicate that the ML methods did not demonstrate an overall advantage over MT-GBLUP in terms of predictive ability. Compared to MT-GBLUP, the predictive abilities of SVR and KRR improved by on average 1.0 and 1.6%, respectively, in A population, while they decreased by on average 4.4 and 11.5%, respectively, in the Austrian population (averages of both traits); Adaboost.R2 did not show any improvement in predictive ability compared to MT-GBLUP for either trait and either population. However, compared to ST-GBLUP, ML methods generally showed higher predictive abilities; 29.7, 24.4, and 11.1% higher for SVR, KRR, and AdaBoost.R2, respectively (averages of both traits and both populations). Moreover, for both traits, the ML methods yielded greater improvements over ST-GBLUP for the prediction of the Austrian population than for that of the A population. This might be due to the fact that when the reference population is enlarged from two populations to multiple populations, the predictive ability of ST-GBLUP for the Austrian population

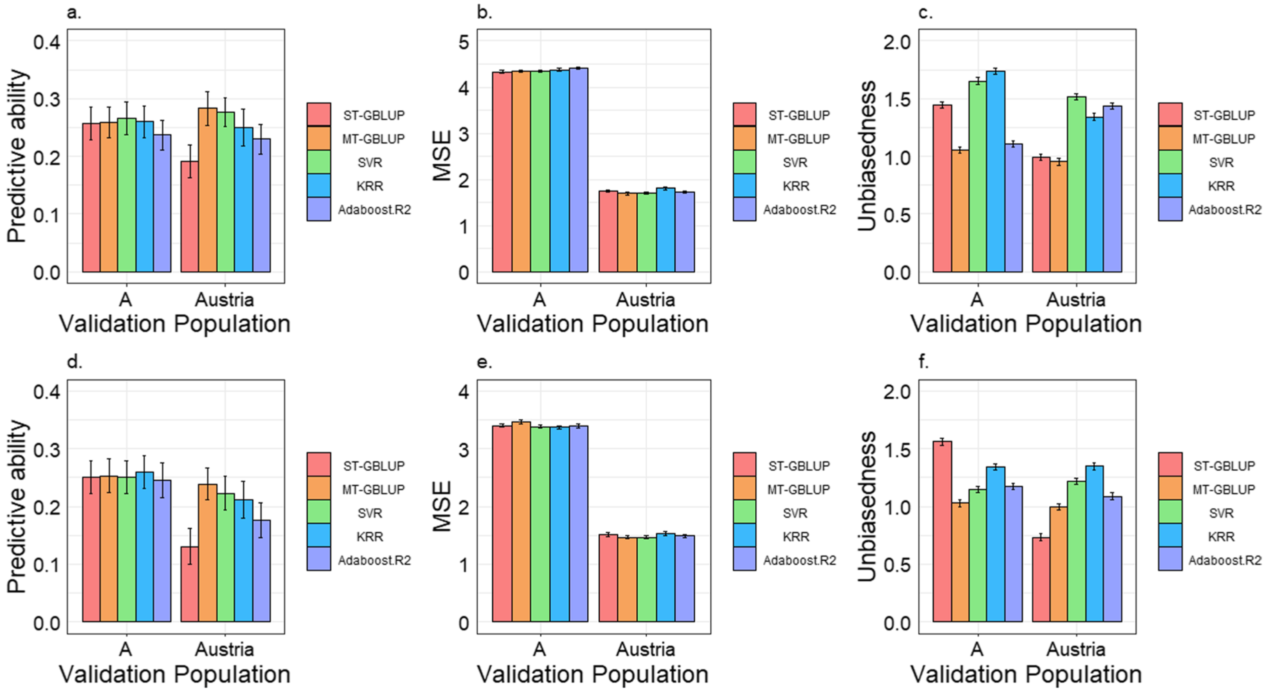

Fig. 4 Predictive ability, mean squared error (MSE), and bias for genomic predictions for multiple populations with different genetic backgrounds. **a**–**c** Predictive ability, MSE, and bias for total number of piglets born (TNB). **d**–**f** Predictive ability, MSE, and bias for number of piglets born alive (NBA). ST-GBLUP: single-trait genomic best linear unbiased prediction model; MT-GBLUP: multitrait GBLUP model. The error bar represents the standard error

decreased, while it slightly increased for the A population. When predicting the Austrian population, average improvements of 57.6, 46.4 and 27.4% (averages of both traits) over ST-GBLUP were observed for SVR, KRR and AdaBoost.R2, respectively. However, when predicting the A population, the improvements were small, at 1.8 and 2.4% (averages of both traits) (averages of both traits) for SVR and KRR, respectively, while a decrease of 5.1% (averages of both traits) was observed for Adaboost.R2 compared to ST-GBLUP. It should also be noted that the predictive abilities for the A population were higher than those for the Austrian population for almost all traits and methods, as shown in Fig. 4.

In terms of MSE, MT-GBLUP produced the lowest MSE for most traits and populations; SVR and AdaBoost.R2 performed similar to or better than ST-GBLUP, while KRR produced higher MSE than ST-GBLUP for most traits and populations. Concerning bias, MT-GBLUP was unbiased, with values close to 1, for all population and trait combinations. In contrast, the other methods exhibited significant deviations from 1 for almost all scenarios. The ML methods resulted in larger biases than ST-GBLUP for all trait and population combinations, except for NBA in the A population, which resulted in regression coefficients closer to 1, although still relatively large (> 1.14).

Comparison of the two-population and multi-population scenarios

Figures 3 and 4 show the predictive abilities and MSE in the two-population and multi-population scenario, respectively, with the underlying values reported in Additional file 1: Table S3. When enlarging the reference population of genomic prediction from two to multiple populations, ST-GBLUP, KRR, and Adaboost.R2 all improved predictive ability in population A (with average improvements of, respectively, 10.2, 3.2, and 1.5% across the two traits), while their predictive ability decreased in the Austrian population (with an average decrease of, respectively, 25.0, 6.5, and 13.1% across the two traits). When enlarging the reference population, MT-GBLUP showed an improvement in predictive ability for all population and trait combinations, and this improvement was statistically significant for most population and trait combinations. Although improvements in predictive ability when moving from two-populations to multi-populations were generally also observed for SRV (average improvement of 2.5% across all populations and traits), statistical significance was only found for the TNB in population A (improvement of 6.83%). In addition, lower MSE were observed when using MT-GBLUP and SVR for multiple populations than for two populations for all trait and population combinations but similar or lower MSE were obtained when using ST-GBLUP, KRR, and AdaBoost.R2 for multiple populations than for two populations. Moreover, when moving from two populations to multiple populations, prediction bias generally increased for ST-GBLUP and the ML methods, but remained approximately the same for MT-GBLUP.

Discussion

The predictive ability of multi-population genomic prediction is affected by a variety of factors, such as differences in the LD between markers and quantitative trait loci (QTL) across populations, QTL segregating in one population only and not the other, differences in MAF between populations, and differences in allele substitution effects due to the different backgrounds of the populations [39, 40]. In this study, we explored the effectiveness of joint genomic prediction in Chinese and Austrian pigs in a two-population scenario and a multi-population scenario.

As Robertson [41] suggested, 0.80 is the biologically important threshold for G×E interactions. The genetic correlation for the TNB and NBA traits between the Austria and A populations were 0.62 and 0.40, respectively (see Additional file 1: Table S2). Therefore, G×E interactions are likely to exist between the two populations. In the two-population scenario, multitrait models can be used to improve the predictive ability of genomic prediction due to their ability to capture the G×E interactions between populations, but higher predictive ability can be further gained using ML methods. In the multi-population scenario, apart from the Austria and A populations, populations C and D also had similar genetic backgrounds to each other (Fig. 1b) but the genetic correlations between them were lower than those between the Austria and A populations (see Additional file 1: Table S2), suggesting the possibility of G×E interactions between the C and D populations as well. Considering that large G×E interactions were not accounted for in single-trait models, ST-GBLUP also achieved an overall lower predictive ability compared to MT-GBLUP and ML methods, and MT-GBLUP resulted in a smaller bias than other methods.

In the two-population scenario, compared to ST-GBLUP, the predictive ability of MT-GBLUP improved by on average 8.8 and 5.5% for TNB and NBA, respectively. In addition, multitrait models can also be used to account for different phenotype scales between populations, particularly in the context of international joint evaluations, in which traits are not defined in the same way and country-specific scale effects may apply [42]. Nevertheless, the predictive abilities of SVR and KRR were higher than that of MT-GBLUP in almost all cases (Fig. 3). In the multi-population scenario, MT-GBLUP is also an ideal “benchmark” for comparison, but it failed to converge when the covariances between populations were estimated using a 5-trait model. Therefore, a series of bivariate analyses were performed and the estimated genetic parameters were combined into the final (co)variance matrix. This approach of using a series of bivariate analyses is commonly practised in international dairy and beef evaluations to deal with multiple populations [43]. Alternative approaches could be used to estimate (co)variance components, such as a Bayesian sampling approach (e.g., Gibbs sampling) instead of the REML method [44] that was used here. When using multitrait models to handle multiple populations, in spite of genetic differences populations, those with similar genetic backgrounds can be collectively modelled as a single population, which reduces the number of traits in the model, significantly reducing computational demands. The same approach is applicable in scenarios involving different breeds or G×E interactions, since farms from similar areas tend to exhibit fewer G×E interactions and can thus be regarded as a joint population. Therefore, multitrait models are still feasible when dealing with data from numerous populations.

When comparing MT-GBLUP with ML methods, in the two-population scenario, ML methods demonstrated superior predictive abilities with similar or smaller biases compared to MT-GBLUP in the A population, while they showed comparable predictive abilities to MT-GBLUP for the Austrian population but with increased biases; in the multi-population scenario, ML methods did not show higher predictive abilities than MT-GBLUP but greater prediction bias. Finally, the ML methods exhibited greater computational efficiency, particularly for the multi-population scenario, as shown in Additional file 1: Table S4. In addition, when the number of populations increased, only MT-GBLUP yielded an improvement in predictive ability in both populations, while the other methods yielded an improvement only in population A and not in the Austrian population. This could be because the added populations had significantly lower genetic correlations with the Austrian population compared to their correlations with population A (as shown in Additional file 1: Table S2); thus, they may have greater G×E interactions with the Austrian population. If G×E interactions are not adequately accounted for, adding such populations in the multi-population scenario might decrease the predictive ability. However, the only model that was proficient in handling this multi-population scenario is MT-GBLUP, as it can account for such G×E interactions. Consequently, only MT-GBLUP improved the predictive ability for the Austrian population in the multi-population scenario.

In this study, we did not find consistency between PCA and the genetic correlation of specific traits. In spite of the distinct genetic backgrounds of populations A, B, and C based on PCA (Fig. 1b), the estimates of genetic correlations for TNB and NBA between them were high (see Additional file 1: Table S2). The estimates of these genetic correlations may, however, be biased because the genetic markers may not accurately represent differences in allele frequencies at causal loci between populations [45]. Wientjes et al. [46] demonstrated that unbiased estimates of genetic correlation can be obtained from genomic relationships based on causal loci. However, when the non-causal SNPs that are used to estimate genetic correlation between populations do not have similar properties as the causal loci (e.g., similar pattern of allele frequencies), estimates of the genetic correlation can be biased [45]. Estimates of genetic correlations between populations can also be affected when the genetic effects captured by the genotyped markers have higher or lower genetic correlations than the portion that is not captured by markers [47]. Finally, genetic correlations differ across traits [48–50] and are also affected by potential differences in the environments that the populations are exposed to [51].

Among the three ML methods, SVR and KRR had higher predictive ability than ST-GBLUP. These findings were consistent with other studies, showing the superiority of SVR and KRR in terms of predictive ability for genomic prediction of phenotypes [16, 18, 52, 53]. ML methods have demonstrated an ability to use nonadditive effects and improve the predictive ability [16]. Moreover, the ML methods employed in this study adopted a population-specific hyperparameter optimization strategy, ensuring that the most suitable hyperparameters could be obtained for each population. In addition, the optimal hyperparameters were determined using a grid search based on the Pearson correlation coefficient between corrected phenotypes $\mathbf{y_c}$ and predicted genetic effects, thereby ensuring that the identified optimal hyperparameters were closer to the global optimum. These factors contribute to the advantages of ML algorithms over ST(MT)-GBLUP and Bayesian methods in the two-population scenario. However, when adding unrelated reference populations, the data were not evenly distributed across PCA (Fig. 1b). The bandwidth parameter of the RBF kernel (i.e., the inverse of gamma) is affected by the distribution and local density of data points [54]. Therefore, using an RBF kernel with a fixed gamma value for all individuals may not adequately adapt to local features within the data, thereby struggling to establish connections between populations. In contrast, the MT-GBLUP method captures genetic connections between populations through the $\mathbf{G}$ matrix. Consequently, MT-GBLUP still achieved higher predictive abilities in that scenario. For ML methods, developing an RBF kernel that dynamically adjusts the bandwidth for each data point based on local density (i.e., reducing bandwidth within clusters to capture more subtle local features, while increasing bandwidth in sparser areas between clusters to establish connections across different clusters) could enable the model to better adapt to local features and distributions of data, which may be beneficial for multi-population datasets.

ML methods also have limitations. (1) Traditional ML methods often fail to decompose variance components, making the calculation of heritabilities and genetic correlations a challenge. (2) ML methods often exhibit sensitivity to minor perturbations (a change or disturbance to the original data) and noise in the data, which could lead to instability in the model output (as in the case of Adaboost.R2 in our study). (3) Hyperparameter optimization is required during the model training process. The time required for hyperparameter optimization with grid search is influenced by experience (i.e. the selection of the hyperparameter ranges and values in grid search) and novices may need some time to explore the optimal hyperparameters. However, some automated hyperparameter search strategies, such as random search, Bayesian optimization, and gradient-based hyperparameter optimization, can significantly improve optimization efficiency without being affected by experience. It should be noted that in practical breeding, hyperparameter optimization is not an issue because automatic hyperparameter optimization can be integrated into the analysis workflow. In addition, as long as there are no significant changes in the reference population, there is no need to re-optimize hyperparameters and previously trained models can be directly employed for prediction. And (4), traditional ML methods often struggle to facilitate practical selection to improve breeding values; because ML methods can capture nonlinear relationships in genomic data, the genetic effects that they predict also include non-additive effects (e.g., dominance and epistasis). Therefore, traditional ML methods are difficult to apply directly in practical applications for selection to improve breeding values, and more research is required to explore the dissection of additive effects.

Because the primary aim of this study was to compare the performance of phenotype prediction methods and because of limitations in sample size and birth year, a random CV strategy was adopted to achieve more stable results. However, in practical breeding, selection often requires using older animals in the training set to predict younger individuals. The predictive ability and bias obtained from CV may not represent those of practical predictive ability and bias because the relationship between the training and validation sets is different.

As an ensemble learning method, AdaBoost.R2 did not show advantages compared to SVR and KRR, for which there are three possible explanations. (1) The number of AdaBoost.R2 iterations (i.e., the number of base learners) has a strong impact on the predictive ability of the model [55]. However, the small size of each population and the differences between populations in the training set increased the differences between replicate training sets, thereby rendering the determination of the optimal number of base learners unstable and challenging (results not shown). As a result, we did not identify the optimal number of base learners but, instead, we used the default number of base learners, which may have somewhat compromised the performance of AdaBoost.R2. (2) Since AdaBoost.R2 mainly focuses on reducing bias through a stepwise boosting approach, it often overlooks variance control, leading to an increase in model overfitting. To increase the diversity of base learners and reduce the risk of overfitting, a strong ensemble must be constructed based on learners with fairly weak generalization performance (such as the classification and regression tree (CART) decision tree or the multi-layer perceptron (MLP) neural network) [55]. However, our choice to use the more robust KRR as the base learner for AdaBoost was primarily motivated by our desire to use fewer base learners and to capitalize on the tuning experience (i.e. the approximate range of hyperparameters) gained from the learners employed in this study (e.g., SVR and KRR), thereby significantly reducing hyperparameter optimization time. This may have resulted in a decrease in the predictive ability of AdaBoost.R2. And (3) The performance of AdaBoost.R2 is sensitive to abnormal samples, which refer to data points that significantly deviate from the norm or general distribution of the dataset, often due to measurement error or anomalies [55]. For multi-population genomic prediction, the individuals in the validation population and some of the individuals in the reference population come from different populations, and abnormal samples from different populations may be assigned greater weights in iterations, while samples from the same population may be assigned smaller weights, thus affecting the final predictive ability, which may also be the reason for the large bias observed for Adaboost.R2.

ML methods provide new options for multi-population genomic prediction. However, as the sample size increases, the computing time of traditional ML methods that model each population separately increases. In contrast, in transfer learning, previously trained models can be reused, and the knowledge learned in the source domain (e.g. large population) can be used to help learning tasks in the target domain (e.g. small population) according to the similarity of data, tasks (e.g. genomic prediction of phenotypes), and models; therefore, the model trained on big data can be transferred to small datasets successfully [56, 57].

Conclusions

In this study, we used ST-GBLUP, MT-GBLUP, BayesHE, and three ML methods for multi-population genomic prediction of phenotypes reproductive traits in Chinese and Austrian pigs. Our results demonstrated that the MT-GBLUP method showed advantages over ST-GBLUP in both two-population and multi-population scenarios. When enlarging the reference population from two populations to multiple populations, ST-GBLUP and ML methods produced overall larger bias, while MT-GBLUP generally achieved similar bias. Compared to MT-GBLUP, ML methods possess potential to improve the genomic prediction ability of both populations in the two-population scenario, while in the multi-population scenario, the advantages of ML methods were not demonstrated.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12711-024-00915-5.

Supplementary material 1: Table S1. Number of generations, full-sib and half-sib families for each population when constructing the A matrix. Table S2. Estimates of genetic variance (block-diagonals), covariance (upper triangular blocks; italic) and genetic correlation (lower triangular blocks; in bold; standard errors in parentheses) of multi-trait genomic best linear unbiased prediction (MT-GBLUP). Table S3. Predictive ability, mean squared error (MSE), and unbiasedness of different methods in the two and multi-population scenarios. Table S4. Average computation time to complete each fold of fivefold cross-validation (CV) for all genomic prediction methods.

Acknowledgements

The authors gratefully acknowledge the constructive comments from the reviewers.

Author contributions

XDD designed the experiments. XW performed the statistical analysis and wrote the manuscript. XDD, GM and CP revised the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by grants from the National Key Research and Development Project (2023YFD1300200,2019YFE0106800), China Agriculture Research System of MOF and MARA, and Key R&D Program of Shandong Province (2022LZGC003), Scientific & Technological Cooperation project No. CN 08/2020 supported by the Austrian Agency for International Cooperation in Education and Research (OeAD-GmbH).

Availability of data and materials

The datasets used or analysed during the present study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Animal samples used in this study were approved by the Animal Care and Use Committee of China Agricultural University and Austrian Pig Breeders Association. There was no use of human participants, data or tissues.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]State Key Laboratory of Animal Biotech Breeding, Key Laboratory of Animal Genetics and Breeding of Ministry of Agriculture and Rural Affairs, National Engineering Laboratory of Animal Breeding, College of Animal Science and Technology, China Agricultural University, Beijing, China. [2]University of Natural Resources and Life Sciences, Vienna, Austria.

Received: 22 December 2022 Accepted: 30 May 2024
Published online: 26 June 2024

References

1. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29.
2. de Roos AP, Schrooten C, Veerkamp RF, van Arendonk JA. Effects of genomic selection on genetic improvement, inbreeding, and merit of young versus proven bulls. J Dairy Sci. 2011;94:1559–67.
3. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: genomic selection in dairy cattle: progress and challenges. J Dairy Sci. 2009;92:433–43.
4. Heffner EL, Jannink JL, Sorrells ME. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Genome. 2011;4:65–75.
5. Goddard M. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 2009;136:245–57.
6. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, et al. Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci. 2009;92:16–24.
7. Lund MS, de Roos APW, de Vries AG, Druet T, Ducrocq V, Fritz S, et al. A common reference population from four European Holstein populations increases reliability of genomic predictions. Genet Sel Evol. 2011;43:43.
8. Zhou L, Heringstad B, Su G, Guldbrandtsen B, Meuwissen THE, Svendsen M, et al. Genomic predictions based on a joint reference population for the Nordic Red cattle breeds. J Dairy Sci. 2014;97:4485–96.
9. Song H, Zhang J, Jiang Y, Gao H, Tang S, Mi S, et al. Genomic prediction for growth and reproduction traits in pig using an admixed reference population. J Anim Sci. 2017;95:3415–24.
10. Bonifazi R, Calus MPL, Ten Napel J, Veerkamp RF, Michenet A, Savoia S, et al. International single-step SNPBLUP beef cattle evaluations for Limousin weaning weight. Genet Sel Evol. 2022;54:57.
11. Cardoso FF, Matika O, Djikeng A, Mapholi N, Burrow HM, Yokoo MJI, et al. Multiple country and breed genomic prediction of tick resistance in beef cattle. Front Immunol. 2021;12:620847.
12. Henderson CR. Sire evaluation and genetic trends. J Anim Sci. 1973;1973:10–41.
13. Gianola D, de Campos G, Hill WG, Manfredi E, Fernando R. Additive genetic variability and the Bayesian alphabet. Genetics. 2009;183:347–63.
14. Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics. 2011;12:186.
15. Gianola D. Priors in whole-genome regression: the bayesian alphabet returns. Genetics. 2013;194:573–96.
16. Long N, Gianola D, Rosa GJ, Weigel KA. Application of support vector regression to genome-assisted prediction of quantitative traits. Theor Appl Genet. 2011;123:1065–74.
17. Kojima K, Tadaka S, Katsuoka F, Tamiya G, Yamamoto M, Kinoshita K. A genotype imputation method for de-identified haplotype reference information by using recurrent neural network. PLoS Comput Biol. 2020;16:e1008207.
18. González-Camacho JM, Ornella L, Pérez-Rodríguez P, Gianola D, Dreisigacker S, Crossa J. Applications of machine learning methods to genomic selection in breeding wheat for rust resistance. Plant Genome. 2018;11:170104.
19. Noé F, De Fabritiis G, Clementi C. Machine learning for protein folding and dynamics. Curr Opin Struct Biol. 2020;60:77–84.
20. Fa R, Cozzetto D, Wan C, Jones DT. Predicting human protein function with multi-task deep neural networks. PLoS ONE. 2018;13:e0198216.
21. González-Recio O, Rosa GJM, Gianola D. Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. Livest Sci. 2014;166:217–31.
22. Liang M, Chang T, An B, Duan X, Du L, Wang X, et al. A stacking ensemble learning framework for genomic prediction. Front Genet. 2021;12:600040.
23. Liang M, Miao J, Wang X, Chang T, An B, Duan X, et al. Application of ensemble learning to genomic selection in chinese simmental beef cattle. J Anim Breed Genet. 2021;138:291–9.
24. Faville MJ, Ganesh S, Cao M, Jahufer MZZ, Bilton TP, Easton HS, et al. Predictive ability of genomic selection models in a multi-population perennial ryegrass training set using genotyping-by-sequencing. Theor Appl Genet. 2018;131:703–20.
25. Guo X, Christensen OF, Ostersen T, Wang Y, Lund MS, Su G. Improving genetic evaluation of litter size and piglet mortality for both genotyped and nongenotyped individuals using a single-step method. J Anim Sci. 2015;93:503–12.
26. Madsen P, Jensen J, Labouriau R, Christensen O, Sahana G, editors. DMU - a package for analyzing multivariate mixed models in quantitative genetics and genomics. In: Proceedings of the 10th World Congress of Genetics Applied to Livestock Production: 17–22 August 2014; Canada. 2014.
27. Song H, Ye S, Jiang Y, Zhang Z, Zhang Q, Ding X. Using imputation-based whole-genome sequencing data to improve the accuracy of genomic prediction for combined populations in pigs. Genet Sel Evol. 2019;51:58.
28. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet. 2009;84:210–23.
29. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
30. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88:76–82.
31. Hill WG, Robertson A. Linkage disequilibrium in finite populations. Theor Appl Genet. 1968;38:226–31.
32. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
33. Jorjani H, Klei L, Emanuelson U. A simple method for weighted bending of genetic (co)variance matrices. J Dairy Sci. 2003;86:677–9.
34. Nilforooshan MA. mbend: an R package for bending non-positive-definite symmetric matrices to positive-definite. BMC Genet. 2020;21:97.
35. Shi S, Li X, Fang L, Liu A, Su G, Zhang Y, et al. Genomic prediction using Bayesian regression models with global-local prior. Front Genet. 2021;12:628205.
36. Müller AC, Guido S. Introduction to machine learning with python: a guide for data scientists. Sebastopol: O'Reilly Media, Inc; 2017.
37. Exterkate P, Groenen PJF, Heij C, van Dijk D. Nonlinear forecasting with many predictors using kernel ridge regression. Int J Forecast. 2016;32:736–53.
38. Shrestha DL, Solomatine DP. Experiments with AdaBoost.RT, an improved boosting scheme for regression. Neural Comput. 2006;18:1678–710.
39. Zhong S, Dekkers JC, Fernando RL, Jannink JL. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a Barley case study. Genetics. 2009;182:355–64.
40. Raymond B, Bouwman AC, Wientjes YCJ, Schrooten C, Houwing-Duistermaat J, Veerkamp RF. Genomic prediction for numerically small breeds, using models with pre-selected and differentially weighted markers. Genet Sel Evol. 2018;50:49.
41. Robertson A. The sampling variance of the genetic correlation coefficient. Biometrics. 1959;15:469–85.

42. Nilforooshan MA, Jorjani H. Invited review: a quarter of a century-International genetic evaluation of dairy sires using MACE methodology. J Dairy Sci. 2022;105:3–21.
43. Jorjani H, Emanuelson U, Fikse WF. Data subsetting strategies for estimation of across-country genetic correlations. J Dairy Sci. 2005;88:1214–24.
44. Bonifazi R, Vandenplas J, ten Napel J, Matilainen K, Veerkamp RF, Calus MPL. Impact of sub-setting the data of the main Limousin beef cattle population on the estimates of across-country genetic correlations. Genet Sel Evol. 2020;52:32.
45. Wientjes YCJ, Calus MPL, Duenk P, Bijma P. Required properties for markers used to calculate unbiased estimates of the genetic correlation between populations. Genet Sel Evol. 2018;50:65.
46. Wientjes YCJ, Bijma P, Vandenplas J, Calus MPL. Multi-population genomic relationships for estimating current genetic variances within and genetic correlations between populations. Genetics. 2017;207:503–15.
47. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh P-R, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015;47:1236–41.
48. Karoui S, Carabaño MJ, Díaz C, Legarra A. Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. Genet Sel Evol. 2012;44:39.
49. Zhou L, Lund MS, Wang Y, Su G. Genomic predictions across Nordic Holstein and Nordic Red using the genomic best linear unbiased prediction model with different genomic relationship matrices. J Anim Breed Genet. 2014;131:249–57.
50. Brown CB, Ye CJ, Price AL, Zaitlen N, Asian Genetic Epidemiology Network Type 2 Diabetes Consortium. Transethnic genetic-correlation estimates from summary statistics. Am J Hum Genet. 2016;99:76–88.
51. Falconer DS. The problem of environment and selection. Am Nat. 1952;86:293–8.
52. Zhao T, Fernando R, Cheng H. Interpretable artificial neural networks incorporating Bayesian alphabet models for genome-wide prediction and association studies. G3. 2021;11:228.
53. Ghafouri-Kesbi F, Rahimi-Mianji G, Honarvar M, Nejati-Javaremi A. Predictive ability of random forests, boosting, support vector machines and genomic best linear unbiased prediction in different scenarios of genomic evaluation. Anim Prod Sci. 2016;57:229–36.
54. Kerpicci M, Ozkan H, Kozat SS. Online anomaly detection with bandwidth optimized hierarchical kernel density estimators. IEEE Trans Neural Netw Learn Syst. 2021;32:4253–66.
55. Zhou Z. Machine learning. Beijing: Tsinghua University Press; 2016.
56. Pathak A, Mandana K, Saha G. Ensembled transfer learning and multiple kernel learning for phonocardiogram based atherosclerotic coronary artery disease detection. IEEE J Biomed Health Inform. 2022;26:2804–13.
57. Rogers AW, Vega-Ramon F, Yan J, Del Rio-Chanona EA, Jing K, Zhang D. A transfer learning approach for predictive modeling of bioprocesses using small data. Biotechnol Bioeng. 2022;119:411–22.

Publisher's Note