OXFORD

AMERICAN SOCIETY OF **ANIMAL SCIENCE**

Leveraging low-density crossbred genotypes to offset crossbred phenotypes and their impact on purebred predictions

Natália Galoro Leite,[†,1] **Ching-Yi Chen,**[‡] **William O. Herring,**[‡] **Justin Holl,**[‡] **Shogo Tsuruta,**[†] **and Daniela Lourenco**[†]

[†]Department of Animal and Dairy Science, University of Georgia, Athens, GA, USA, 30602
[‡]Genus PIC, Hendersonville, TN, USA, 37075
[1]Corresponding author: nataliagaloro@uga.edu

Abstract

The objectives of this study were to 1) investigate the predictability and bias of genomic breeding values (**GEBV**) of purebred (**PB**) sires for CB performance when CB genotypes imputed from a low-density panel are available, 2) assess if the availability of those CB genotypes can be used to partially offset CB phenotypic recording, and 3) investigate the impact of including imputed CB genotypes in genomic analyses when using the algorithm for proven and young (**APY**). Two pig populations with up to 207,375 PB and 32,893 CB phenotypic records per trait and 138,026 PB and 32,893 CB genotypes were evaluated. PB sires were genotyped for a 50K panel, whereas CB animals were genotyped for a low-density panel of 600 SNP and imputed to 50K. The predictability and bias of GEBV of PB sires for backfat thickness (**BFX**) and average daily gain recorded (**ADGX**) recorded on CB animals were assessed when CB genotypes were available or not in the analyses. In the first set of analyses, direct inverses of the genomic relationship matrix (**G**) were used with phenotypic datasets truncated at different time points. In the next step, we evaluated the APY algorithm with core compositions differing in the CB genotype contributions. After that, the performance of core compositions was compared with an analysis using a random PB core from a purely PB genomic set. The number of rounds to convergence was recorded for all APY analyses. With the direct inverse of **G** in the first set of analyses, adding CB genotypes imputed from a low-density panel (600 SNP) did not improve predictability or reduce the bias of PB sires' GEBV for CB performance, even for sires with fewer CB progeny phenotypes in the analysis. That indicates that the inclusion of CB genotypes primarily used for inferring pedigree in commercial farms is of no benefit to offset CB phenotyping. When CB genotypes were incorporated into APY, a random core composition or a core with no CB genotypes reduced bias and the number of rounds to convergence but did not affect predictability. Still, a PB random core composition from a genomic set with only PB genotypes resulted in the highest predictability and the smallest number of rounds to convergence, although bias increased. Genotyping CB individuals for low-density panels is a valuable identification tool for linking CB phenotypes to pedigree; however, the inclusion of those CB genotypes imputed from a low-density panel (600 SNP) might not benefit genomic predictions for PB individuals or offset CB phenotyping for the evaluated CB performance traits. Further studies will help understand the usefulness of those imputed CB genotypes for traits with lower PB–CB genetic correlations and traits not recorded in the PB environment, such as mortality and disease traits.

Lay Summary

Crossbred (**CB**) genotypes primarily used for inferring pedigree in commercial farms can be potentially used for genomic prediction and partially offset CB phenotyping. We investigated the predictability and bias of genomic breeding values (**GEBV**) of purebred (**PB**) sires for CB performance when CB genotypes are available, assessed if the availability of those CB genotypes can be used to partially offset CB phenotypic recording, and investigated the impact of including CB genotypes in genomic analyses when using the algorithm for proven and young (**APY**). The predictability and bias of GEBV of PB sires for two CB traits were assessed when CB genotypes were available or not in the analyses. Later, the performance of different APY core compositions accounting for CB genotypes was compared with a random core from a purely PB genomic set. Adding CB genotypes did not improve predictability or reduce the bias of PB sires' GEBV for CB performance, indicating that the inclusion of CB genotypes imputed from a low-density (600 SNP) panel is of no benefit to offset CB phenotyping. With APY, a random core composition from a genomic set with only PB genotypes resulted in the highest predictability and the smallest number of rounds to convergence, although bias increased.

Introduction

The benefits of breed complementarity and heterosis are effectively exploited in the pig industry. Purebred (**PB**) pigs are selected to compose specialized lines and later mated to produce crossbred (**CB**) progeny, commonly known as commercial animals. Genetic selection is performed on PB animals, whereas the phenotypic improvements are expected to occur

at the CB level, primarily in much more challenging conditions than in the nucleus environments (Knol et al., 2016; Garrick, 2017). The difference between the genetic background and the environment experienced by PB and CB animals results in low accuracy of PB breeding values to predict CB performance. However, more accurate predictions can be obtained in a process called combined CB and PB selection (**CCPS**), where performance records of CB progeny are included in the genetic evaluation (Wei and van der Werf, 1994).

Including CB performance information in the genetic evaluations requires pedigree recording to link phenotypes from CB animals to PB parents. However, tracking pedigrees on commercial farms is challenging due to several factors, such as the use of pooled semen, lack or loss of individual identification tags, and the difficult accommodation of pedigree recording along with daily procedures at a commercial farm (Maiorano et al., 2019; Hollifield et al., 2021; See et al., 2021). Those factors might limit the use of CCPS-based systems. A possible way to mitigate such a challenge is by genotyping CB animals using less expensive low-density SNP panels and inferring kinship through genomic information.

Beyond its primary purpose of inferring kinship, the low-density genotypic data on CB animals could be used for genomic prediction and potentially partially offset CB progeny phenotyping while maintaining a constant prediction accuracy for selection candidates. It has been shown through simulation (Dekkers, 2007; See et al., 2020) and empirical studies (Hidalgo et al., 2015; Lourenco et al., 2016; Iversen et al., 2017) that when the breeding goal is to improve CB performance, adding CB genomic information into genomic evaluations could increase response to selection at the commercial level. However, the magnitude of such an increase depends on the trait, the genetic correlation between PB and CB populations, and the relationship between validation and training sets.

Another aspect of using the genotypes from CB animals is the rapid increase in the size of the genotyped population. With a large number of genotyped animals, directly inverting the genomic relationship matrix for GBLUP-based methods may not be feasible. To overcome that limitation, Misztal et al. (2014a) proposed the algorithm for proven and young (**APY**) for obtaining a sparse representation of the inverse of the genomic relationship matrix (**G**). With APY, the direct inversion of **G** is only required for a small set of animals (core animals), whereas the remaining components for noncore animals are obtained based on recursive equations with a linear computational cost. This enables genomic evaluations with millions of genotyped animals in a reasonable time and computing cost (Tsuruta et al., 2021; Cesarani et al., 2022).

The objectives of this study were to 1) investigate the predictability and bias of GEBV of PB sires for CB performance when CB genotypes imputed from a low-density panel are available, 2) assess if the availability of those CB genotypes can be used to partially offset CB phenotypic recording, and 3) investigate the impact of including imputed CB genotypes in genomic analyses when using APY for the inversion of the genomic relationship matrix.

Materials and methods

Animal Care and Use Committee approval was not needed because the information was obtained from pre-existing databases.

Data set

Research data sets were provided by PIC (a Genus company, Hendersonville, TN). Phenotypic information was recorded from 2000 to 2020 and was available for sires from two PB terminal sire lines (**SL1** and **SL2**) and their three-way CB progeny resulting from the cross with F1 dams. For simplicity, SL1 sires and their CB progeny will be referred to as population 1 (**PP1**), and SL2 sires and their CB progeny as population 2 (**PP2**). Pedigree information was available for PB sires and CB animals and traced back three generations from phenotyped animals, resulting in 151,625 pedigree records for PP1 and 246,699 records for PP2. Phenotypes were available for four traits: backfat thickness recorded on PB sires (**BFP**) and CB animals (**BFX**), and average daily gain recorded on PB sires (**ADGP**) and their CB progeny (**ADGX**). Although F1 dams were individually identified at the farm level, no phenotypic or pedigree information was available on the female side. The definition for traits recorded on PB and CB animals differed slightly in recording methods. ADGP was defined as PB animals' live weight measured at the off-test divided by the animal's age, whereas ADGX was defined as the hot carcass for CB divided by the animal's age. Similarly, BFP was recorded as an ultrasonic measurement at off-test, whereas BFX was measured on the hot carcass following slaughter. In addition to phenotypic records, contemporary groups, litter code, off-test weight (for PB), and hot carcass weight (for CB) were also available and are summarized in Table 1.

Purebred sires were genotyped for a commercial 50K SNP panel (GGP-Porcine HD BeadChip; GeneSeek, Lincoln, NE), whereas all the CB animals were genotyped for a low-density SNP panel of 600 SNP markers and imputed to 50K. After quality control steps, 40,628 and 44,368 SNP markers were available for 46,760 (10,622) and 138,026 (32,893) PB (CB) animals in PP1 and PP2, respectively.

Variance components estimation

A multitrait model containing the two CB and PB traits, without genomic information, was used for variance component estimation for each population. The model description follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{l} + \mathbf{Z}\mathbf{a} + \mathbf{e},$$

where $\mathbf{y}$ is the vector of phenotypes; $\boldsymbol{\beta}$ is the vector containing the fixed effects of the contemporary group composed of farm, sex, year, and week of birth for PB traits and farm, sex, year, week of birth, and slaughter date for CB traits, and linear covariables of off-test weight (for BFP) and hot carcass weight (for BFX); $\mathbf{l}$ is the vector of random effects of common litter environment; $\mathbf{a}$ is the vector of additive genetic random effects; $\mathbf{e}$ is the vector of random residuals; and $\mathbf{X}$, $\mathbf{W}$ and $\mathbf{Z}$ are the incidence matrices for the effects contained in $\boldsymbol{\beta}$, $\mathbf{l}$, and $\mathbf{a}$, respectively. A uniform distribution (i.e., flat prior) was assumed for $\boldsymbol{\beta}$, whereas $\mathbf{l} \mid \boldsymbol{\Sigma}_l \sim \mathbf{N}(0, \mathbf{I}_l \otimes \boldsymbol{\Sigma}_l)$, $\mathbf{a} \mid \boldsymbol{\Sigma}_a \sim \mathbf{N}(0, \mathbf{A} \otimes \boldsymbol{\Sigma}_a)$, and $\mathbf{e} \mid \boldsymbol{\Sigma}_e \sim \mathbf{N}(0, \mathbf{I}_e \otimes \boldsymbol{\Sigma}_e)$, where $\mathbf{A}$ is the pedigree relationship matrix, $\mathbf{I}_l$ and $\mathbf{I}_e$ are identity matrices, $\otimes$ is the Kronecker product, and $\boldsymbol{\Sigma}_l$, $\boldsymbol{\Sigma}_a$, and $\boldsymbol{\Sigma}_e$ are the (co)variance matrices of common litter environment, additive genetic effects, and residuals, respectively. (Co)variance matrices were defined as follows:

Table 1. Descriptive statistics of populations 1 (PP1) and 2 (PP2)

Item[1]	PP1		PP2	
	Records	Mean (SD)	Records	Mean (SD)
ADGP, g/d	137,538	716.42 (75.97)	207,375	756.36 (83.05)
ADGX, g/d	10,622	520.74 (58.26)	32,893	541.17 (59.66)
BFP, mm	135,576	9.19 (2.62)	203,294	9.39 (2.76)
BFX, mm	10,621	13.50 (2.46)	32,886	13.46 (2.55)
OW, kg	137,627	113.39 (12.25)	207,499	118.47 (13.79)
HCW, kg	10,622	98.08 (9.61)	32,893	99.66 (9.46)

[1]Purebred information: ADGP (average daily gain; live off-test weight divided by age), BFP (backfat thickness; ultrasonic measurement), and OW (live off-test weight), and crossbred information: ADGX (average daily gain; hot carcass divided by age), BFX (backfat thickness; measured on the hot carcass), and HCW (hot carcass weight).

$$\Sigma_l = \begin{bmatrix} \sigma^2_{l_{ADGP}} & \sigma_{l_{ADGP}}, \sigma_{l_{ADGX}} & 0 & 0 \\ & \sigma^2_{l_{ADGX}} & 0 & 0 \\ & & \sigma^2_{l_{BFP}} & \sigma_{l_{BFP}}, \sigma_{l_{BFX}} \\ \text{sym.} & & & \sigma^2_{l_{BFX}} \end{bmatrix},$$

$$\Sigma_a = \begin{bmatrix} \sigma^2_{a_{ADGP}} & \sigma_{a_{ADGP}}, \sigma_{a_{ADGX}} & \sigma_{a_{ADGP}}, \sigma_{a_{BFP}} & \sigma_{a_{ADGP}}, \sigma_{a_{BFX}} \\ & \sigma^2_{a_{ADGX}} & \sigma_{a_{ADGX}}, \sigma_{a_{BFP}} & \sigma_{a_{ADGX}}, \sigma_{a_{BFX}} \\ & & \sigma^2_{a_{BFP}} & \sigma_{a_{BFP}}, \sigma_{a_{BFX}} \\ \text{sym.} & & & \sigma^2_{a_{BFX}} \end{bmatrix},$$

$$\Sigma_e = \begin{bmatrix} \sigma^2_{e_{ADGP}} & \sigma_{e_{ADGP}}, \sigma_{e_{ADGX}} & 0 & 0 \\ & \sigma^2_{e_{ADGX}} & 0 & 0 \\ & & \sigma^2_{e_{BFP}} & \sigma_{e_{BFP}}, \sigma_{e_{BFX}} \\ \text{sym.} & & & \sigma^2_{e_{BFX}} \end{bmatrix},$$

Variance component estimation was performed on the software GIBBS1F90 (Misztal et al., 2014b). A total of 10,000 Markov Chain Monte Carlo (**MCMC**) samples (i.e., a total of 120,000 samples generated with a burn-in of 20,000 and a thinning interval of 10) were used to obtain the posterior parameter distributions on POSTGIBBSF90 (Misztal et al., 2014b). The convergence was checked by visual inspection of posterior distributions and the Geweke criterion (Geweke, 1992), as implemented in POSTGIBBSF90 (Misztal et al., 2014b).

Definition of validation sires and genomic sets

The validation sets in PP1 and PP2 were composed of 66 and 163 sires that only had phenotyped and genotyped CB progeny born from 2018 to 2020 and from 2019 to 2020, respectively. On average, validation sires had 108.0 CB progeny in PP1 and 142.6 in PP2 in the complete dataset. In addition, two genomic sets were defined: a set containing only PB genomic information (**G_PB**) and a second one where genomic information of CB animals was added to the first genomic set (**G_PB_CB**). The genomic sets G_PB (G_PB _CB) were composed of 46,760 (57,382) genotypes in PP1, and 138,026 (170,119) in PP2.

Predictability and bias of PB breeding values for CB performance

Genomic-breeding values (**GEBV**) for validation sires were estimated using the two genomic sets and six phenotypic sets
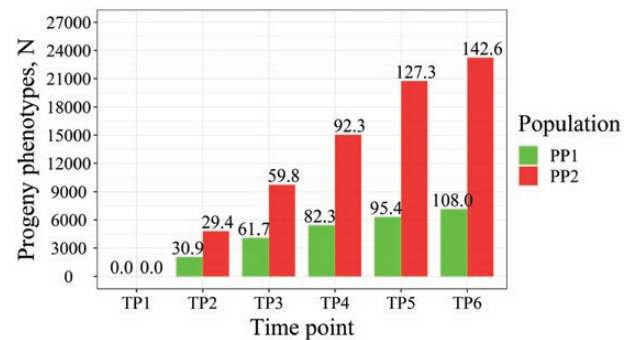
Figure 1. Total number of crossbred progeny phenotypes of validation sires at different time points. Bar labels show the average number of crossbred progeny phenotypes per validation sire in the given time point.

obtained from cutting off the complete phenotypic data in six-time points. For instance, in time point 1 (**TP1**), validation sires did not have CB progeny phenotypes available in the analysis. However, from time point 2 (**TP2**), CB progeny phenotypes were gradually added until time point 6 (**TP6**), which accounted for a complete phenotypic set. The average and the total number of CB progeny phenotypes per validation sire at each time point are shown in Figure 1.

The strategy of evaluating sires at different time points was used because, besides assessing the impacts in predictability and bias for PB sire's GEBV and whether or not adding CB genotypes (G_PB _CB vs. G_PB) could offset CB phenotyping, we also aimed to determine if results would differ for groups of sires with fewer CB progeny information. However, as a comprehensive investigation would not be possible with a small validation set (i.e., 66 sires in PP1 and 163 sires in PP2), sires were evaluated at multiple time points to increase the validation sample (i.e., instead of a single GEBV estimate per genomic set, each sire contributed with six GEBV estimates). This study used an empirical threshold of 300 CB phenotypes to define a proven sire; therefore, sires with less than 300 CB phenotypes in the analysis will be referred to as young sires hereafter.

Breeding values estimation was performed using the software BLUP90IOD2OMP1 (Misztal et al., 2014b) with the 12 unique datasets resultant from the combination of the six phenotypic datasets (TP1–TP6) and two genomic datasets (G_PB and G_PB _CB). After breeding values were estimated, predictability was separately calculated for each of

the two genomic sets and groups of validation sires. Validation groups were defined based on the maximum number of CB progeny phenotypes sires had at GEBV estimation. For instance, the first group was composed of sires with no CB progeny phenotypes in the analysis (maximum number of CB progeny = 0), whereas the second group was composed of sires with a maximum of 100 CB progeny phenotypes (maximum number of CB progeny = 100). Groups were formed following intervals of 100 CB progeny phenotypes, resulting in six groups for PP1 and 10 groups for PP2. As sires had GEBV estimated at six-time points, the same sire could be represented in the same validation group more than once, with GEBV estimated at different time points. The predictability of GEBV of validation sires was calculated as:

$$\rho_{GEBV_{ij}} = cor\left(\mathbf{DR},\ \mathbf{GEBV_{ij}}\right),$$

where $\mathbf{DR}$ is the vector of deregressed proofs (VanRaden et al., 2009) from sire's EBV (no genomics) calculated with a complete phenotypic dataset and $\mathbf{GEBV_{ij}}$ is the vector of sires' GEBV for CB performance given the *i*th validation group and *j*th genomic set. The dispersion bias was accessed by the linear regression of $\mathbf{DR}$ on $\mathbf{GEBV_{ij}}$, as follows:

$$\mathbf{DR} = \mu + \delta\mathbf{GEBV_{ij}} + \mathbf{e}$$

where μ is an overall mean, δ is the dispersion, and $\mathbf{e}$ is a vector of residuals. The slope of the linear regression (called dispersion bias hereafter) has an expectation of 1 in the absence of inflation/deflation of GEBV.

The single-step genomic BLUP method (**ssGBLUP**) (Legarra et al., 2009; Aguilar et al., 2010; Christensen and Lund, 2010) was used to incorporate genomic information into the mixed model equations. In ssGBLUP, the relationship between genotyped and nongenotyped animals is combined in the **H** matrix, which replaces the pedigree relationship matrix (**A**) in the traditional BLUP. The inverse of the **H** matrix (**H**⁻¹) was constructed as in Aguilar et al. (2010):

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix},$$

where $\mathbf{A}^{-1}$ is the inverse of the traditional pedigree relationship matrix, $\mathbf{A}_{22}^{-1}$ is the inverse of the pedigree relationship matrix for genotyped animals, and $\mathbf{G}^{-1}$ is the inverse of the genomic relationship matrix, as shown in VanRaden (2008). The **G** matrix was constructed as follows:

$$\mathbf{G} = \frac{\mathbf{ZZ}'}{2\sum p_j\ \left(1 - p_j\right)},$$

where $\mathbf{Z}$ is a matrix of SNP markers centered by twice the across-breed allele frequency (p) of the *j*th locus computed from the current genotyped population. To overcome singularity problems, **G** was blended with 5% of $\mathbf{A}_{22}$. For this first set of analyses, direct inverses of **G** were used in the mixed model equations (i.e., no APY).

APY core composition accounting for CB information

Accounting for CB genotypes in the genomic evaluation should greatly increase the size of the genotyped population.

In this situation, the direct inversion of the genomic relationship matrix might become computationally unfeasible. To overcome this problem, a sparse representation of the inverse of **G** can be computed with the APY algorithm ($\mathbf{G}_{APY}^{-1}$) (Misztal, 2016). The inverse of the $\mathbf{G}_{APY}^{-1}$ is constructed as:

$$\mathbf{G}_{APY}^{-1} = \begin{bmatrix} \mathbf{G}_{cc}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -\mathbf{G}_{cc}^{-1}\mathbf{G}_{cn} \\ \mathbf{I} \end{bmatrix} \mathbf{M}_{nn}^{-1} \begin{bmatrix} -\mathbf{G}_{nc}\mathbf{G}_{cc}^{-1}\mathbf{I} \end{bmatrix},$$

where $\mathbf{G}_{cc}^{-1}$ is the inverse of the genomic relationship matrix for core animals, $\mathbf{G}_{cn}$ is the genomic relationship matrix between core and noncore animals and $\mathbf{M}_{nn}^{-1}$ is the inverse of the relationship matrix for noncore animals, calculated as:

$$m_{nn,i} = diag\left\{ g_{ii} - \mathbf{g}_{ic}'\mathbf{G}_{cc}^{-1}\mathbf{g}_{ci} \right\}$$

where g_{ii} is the diagonal element of $\mathbf{G}_{nn}$ corresponding to the *i*th animal, and $\mathbf{g}_{ic}$ is the relationship between the noncore animal *i* with all the core animals.

In this study, the number of core animals was defined according to the dimensionality of the genomic information as the number of eigenvalues explaining 98% of the variation in **G** (Pocrnic et al., 2016). For easier calculations, the number of eigenvalues was obtained from the singular value decomposition of **Z** with the complete genotype set including PB and CB genotypes (i.e., G_PB_CB) on PREGSF90 (Misztal et al., 2014b). The number of eigenvalues explaining 98% of the variance in G_PB_CB was 7,239 in PP1 and 9,946 in PP2, and this was the number of core animals used in all scenarios.

For a genomic analysis accounting for CB genotypes (G_PB_CB), four core compositions based on different levels of CB genotype contribution were investigated. First, a core with no CB genotypes (**0%_CB**), then 50% of the core animals were CB (**50%_CB**), then 100% of the core animals were CB (**100%_CB**), and finally, a core set was chosen randomly from the entire genotyped population (**RANDOM**). In addition, a random core selection from a genomic set with PB genotypes only (G_PB) was investigated (**RANDOM_PB**). Note that in the 0%_CB and RANDOM_PB scenarios, no CB genotypes are included in the core set. However, in the first scenario, a genomic set accounting for CB and PB genotypes (G_PB_CB) was used for the construction of $\mathbf{G}_{APY}^{-1}$, whereas in the second one only PB genotypes were included (G_PB). Regardless of the core composition, core animals were always chosen randomly. Therefore, to account for the randomness of the selection, core sampling and further genetic analyses were repeated five times. The phenotypic set was constant for all APY analyses at time point TP1 (i.e., no CB progeny phenotypes of validation sires included).

The performance of analyses was measured based on the predictability and dispersion bias of the GEBV of PB sires for CB performance, as explained previously. In addition, the number of rounds to convergence (10⁻¹³) was compared across different core compositions. Predictability and dispersion bias were calculated, respectively, as follows:

$$\rho_{GEBV_{ij}} = cor\left(\mathbf{DR}, \mathbf{GEBV_k}\right),$$

$$\mathbf{DR} = \mu + \delta\mathbf{GEBV_k} + \mathbf{e},$$

where $\mathbf{GEBV_k}$ represents sires' GEBV calculated with the *k*th core composition and genomic set associated with it; other

parameters were previously described. For all APY analyses, breeding values estimation was performed on BLUP90I-OD2OMPP1 (Misztal et al., 2014b).

Results and discussion

Genetic parameters

Heritabilities and genetic correlations are shown in Table 2. The heritability for all traits had similar magnitudes in PP1 and PP2. The average daily gain recorded on CB (ADGX) had lower heritabilities (from 0.15 to 0.19) in comparison with PB average daily gain (ADGP; from 0.28 to 0.30) in both populations. Differently, slightly larger heritabilities were observed for backfat thickness recorded on CB animals (BFX; from 0.50 to 0.56) compared with the estimates for PB backfat thickness (BFP; at 0.43). For all the cases, genetic correlations were positive between and across trait definitions (i.e., ADGP vs. ADGX and ADGP vs. BFX or BFP). Even though positive, genetic correlations were stronger between trait definitions (from 0.51 to 0.89), whereas correlations across trait definitions ranged from weak to moderate (0.17 to 0.31). As expected, our results indicate that average daily gain and BFX recorded on CB and PB animals are genetically different traits (genetic correlations deviate considerably from 1), with the first presenting the higher genetic divergence. Genetic correlations that differ from unity are usually the norm in the pig industry (Hidalgo et al., 2015; Lourenco et al., 2016; Steyn et al., 2021). Such deviation reflects the different environmental and genetic backgrounds between PB and CB populations and enforces the importance of including CB information, particularly when the goal is to improve CB performance (Wei and van der Werf, 1994; See et al., 2020).

Predictability and bias of PB breeding values for CB performance

The predictability of PB sires' breeding values for CB traits given the two genotype sets and the number of CB progeny phenotypes is presented in Figure 2. The predictability for all traits and in both populations gradually increased with the number of CB progeny phenotypes up to a point where sires had around 300 CB phenotyped progeny. After this point, predictabilities remained constant with the further addition of CB progeny phenotypes. As shown in other studies, when the goal is to increase CB performance, including CB phenotypes in genetic evaluations is expected to be positive (Wei and van der Werf, 1994; Lutaaya et al., 2002; See et al., 2020). Although according to our results, the benefit of including those CB phenotypes is negligible after sires already have 300 CB progeny phenotypes included in the analyses.

Including CB genotypes did not improve the predictability of GEBV of PB validation sires for any evaluated CB traits (Figure 2), even for young (<300 CB progeny phenotypes) sires. On the contrary, except for BFX in PP2, for which the predictability remained similar, including CB genotypes marginally decreased the overall predictability. We hypothesized that young sires would benefit from CB progeny genotypes and possibly have compatible predictability as sires with less than 100 CB progeny phenotypes but no CB genotypes. In such a case, CB genotypes would offset CB phenotyping while keeping the predictability for PB selection candidates constant. However, our results showed that adding CB genotypes did not improve predictions for young or old sires and was of no benefit to offset CB phenotyping in those studied populations.

The observed results could be due to the moderate to high correlations (Table 2) between PB and CB traits in our study (0.51 to 0.89). A simulation study by See et al. (2020) showed that in comparison with an evaluation where CB phenotypes and genotypes are not available, including 10% of CB phenotypes per generation increased the CB performance from 21% to 134% when the magnitude of the genetic correlation between PB and CB trait was from strong (0.9) to weak (0.1), respectively. In the same study, adding genotypes for those animals provided an additional 109% increase in prediction accuracy when the PB and CB traits were lowly correlated. Still, when the correlations were moderate to high (0.3–0.9), there was no further improvement in prediction accuracy by adding CB genotypes. Therefore,

Table 2. Estimations of heritability (diagonals) and genetic correlations (off-diagonals) for all traits in populations 1 (PP1) and 2 (PP2)

Trait[1]	PP1			
	ADGP	ADGX	BFP	BFX
ADGP	$0.30^{(0.01)}$ [2]	$0.65^{(0.11)}$	$0.31^{(0.02)}$	$0.23^{(0.02)}$
ADGX		$0.17^{(0.04)}$	$0.17^{(0.13)}$	$0.27^{(0.14)}$
BFP			$0.43^{(0.01)}$	$0.89^{(0.05)}$
BFX				$0.50^{(0.04)}$

	PP2			
	ADGP	ADGX	BFP	BFX
ADGP	$0.28^{(0.01)}$	$0.51^{(0.08)}$	$0.29^{(0.02)}$	$0.23^{(0.02)}$
ADGX		$0.15^{(0.02)}$	$0.29^{(0.08)}$	$0.19^{(0.05)}$
BFP			$0.43^{(0.01)}$	$0.83^{(0.03)}$
BFX				$0.56^{(0.02)}$

[1]ADGP: purebred average daily gain (live off-test weight divided by age); ADGX: crossbred average daily gain (hot carcass divided by age); BFP: purebred backfat thickness (ultrasonic measurement); and BFX: crossbred backfat thickness (measured on the hot carcass).
[2]Standard deviations.
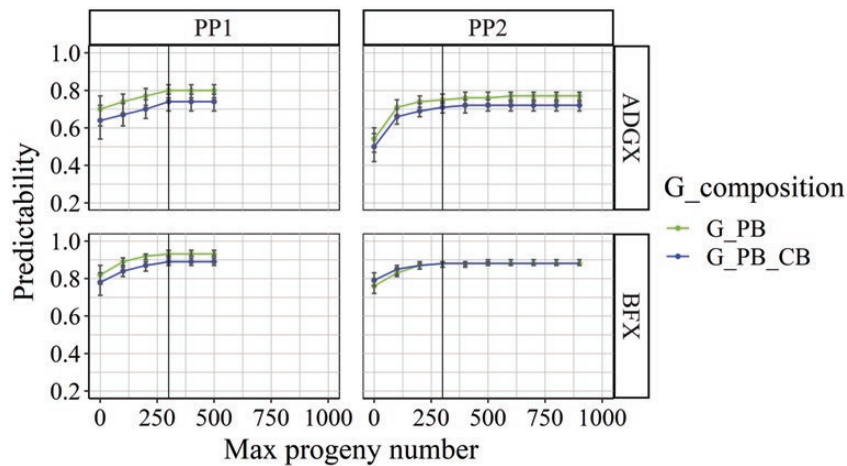
Figure 2. Predictability of breeding values of purebred sires for crossbred performance traits given two genotypes set compositions and the number of phenotype progeny information. The vertical line represents the threshold between young and proven sires (300 CB progeny phenotypes).

different results than those presented in our study might be expected for traits with lower PB–CB genetic correlations and for traits not recorded in the PB environment, such as mortality and disease traits. Another aspect that could have contributed to the small changes in predictability by adding CB genotypes observed herein is the already large size of the PB reference population. Using a reference population with up to 5,236 PB genotypes, Hidalgo et al. (2015) and Iversen et al. (2017) observed marginal increases in the accuracy of PB candidates for CB performance when CB genomic geno-types were added to the reference population. However, with a PB reference population of at least nine times bigger, no improvements were observed in our study. Given the smaller genotyped population used by the authors, that may sug-gest the observed benefits could also be a result of a general increase in the genotyped population rather than from the benefit of adding CB information.

The dispersion bias of PB sires' breeding values for CB per-formance traits given the two genotype sets compositions is presented in Figure 3. Overall, dispersion bias was associated with the trait and the number of CB progeny phenotypes but weakly associated with the inclusion of CB genotypes. PB breeding values were less biased for BFX than for ADGX, likely due to the higher heritability and stronger PB–CB genetic correlation of BFX. Dispersion bias decreased with an increase in CB progeny phenotypes; for BFX, the GEBV of young sires were slightly biased but almost unbiased for proven sires (>= 300 CB progeny phenotypes), whereas, for ADGX, GEBV were still biased, even after sires had more than 300 CB progeny phenotypes. Including CB genotypes did not impact GEBV bias, except for ADGX in PP2, where including CB genotypes reduced the overall bias by 0.10 points for proven sires on average. An unbiased evaluation avoids selecting the wrong set of candidates as it allows for proper comparison between individuals from different generations (Legarra and Reverter, 2017). In this study, the observed dispersion bias could be originated from different sources, namely, incompatibility between genomic and pedi-gree relationships (Tsuruta et al., 2021), selection (Vitezica et al., 2011), and heritability estimates with nongenomic models in the presence of genomic selection (Hidalgo et al., 2020). Some of this bias could be alleviated by modeling unknown parent groups (Legarra et al., 2015) and removing older data

(Lourenco et al., 2014). However, investigating sources of bias was outside the scope of this study.

All combined, our results indicate that incorporating CB genotypes imputed from a low-density panel (600 SNP) will result in minor to no improvements in predictability and dispersion bias of PB sires GEBV for CB performance and might not offset CB phenotyping. That was unexpected once many other studies have shown the benefits and support for the inclusion of CB genomic information when PB selection aims to improve CB performance (Dekkers, 2007; Hidalgo et al., 2015; Iversen et al., 2017; See et al., 2020). In our study, a possible factor that could have impacted both predictability and dispersion bias was the need for CB genotype imputation from a low-density panel. The imputation of CB genotypes is challenging and commonly relies on a PB reference popula-tion composed of two or more PB lines that are themselves frequently imputed (Leite et al., 2021). Moreover, the low-ini-tial density of the SNP panel, the differences in LD pattern between PB and CB populations (Xiang et al., 2015), and the decrease in relationship across lines may further hinder the imputation of CB genotypes. For example, Xiang et al. (2015) evaluated the imputation accuracy of PB and CB genotypes in a pig population and showed that the imputation of CB gen-otypes from a low-density SNP panel (425 SNP) had an accu-racy of around 0.7. The observed low-imputation accuracy from the low-density panel was associated with the weaker LD with few SNP and the presence of narrower haplotype segments in CB genotypes, which further challenges imputa-tion accuracy (Toosi et al., 2010; Xiang et al., 2015).

APY core composition accounting for CB information

The predictability, dispersion bias, and the number of rounds to convergence given different APY core compositions are shown in Figures 4, 5, and 6, respectively. Among APY cores accounting for CB genotypes, the predictability was gener-ally lowly impacted by different levels of CB contribution (Figure 4), presenting a maximum difference between sce-narios of 0.02 points. However, a marginal increase in dis-persion bias (Figure 5) and an increase in the number of rounds for convergence (Figure 6) were associated with core compositions with a higher proportion of CB genotypes (i.e., 50%_CB and 100%_CB) (Figure 5 and 6). On average,

RANDOM presented the smaller number of rounds to converge and was followed, respectively, by 0%, CB, 50%_CB, and 100%_CB cores compositions showing a roughly linear increase in rounds with an increase in CB genotype contribution (Figure 6).

The poorer performance of APY with higher CB contribution in the core could be associated with CB animals' age and inadequate representation of the genotyped population. In both studied populations, PB genotypic information was available for animals born from 2000 to 2020, whereas for CB animals, genotypes were only available for younger animals born from 2017 to 2020. Therefore, the APY cores mainly composed of CB genotypes investigated herein can also be seen as APY cores composed of younger animals. This was confirmed by a deviation of the average relationship between core individuals (**Gcc**) from the population expectation (i.e., 0.0), suggesting a core sampled from a stratified group of genotyped animals. Fragomeni et al. (2015) stated that the APY algorithm presents better properties when core and noncore sets are well related; however, that might not be achieved when the core is mainly composed of animals born in recent generations. Generally, our results indicate that if CB genotypes are incorporated in genomic analyses using APY, a core composed at random (RANDOM) or with no CB contribution (0%_CB) will generally result in a smaller bias and a smaller number of rounds to convergence. In contrast, predictability should be lowly affected by core composition.

To further exploit our findings, we compared the performance of previous APY scenarios that accounted for CB genotypes in the genomic set with a random core chosen from a genomic set with PB genotypes only (RANDOM_PB). In practice, CB animals are rarely genotyped since they are not candidates for selection; therefore, a genetic
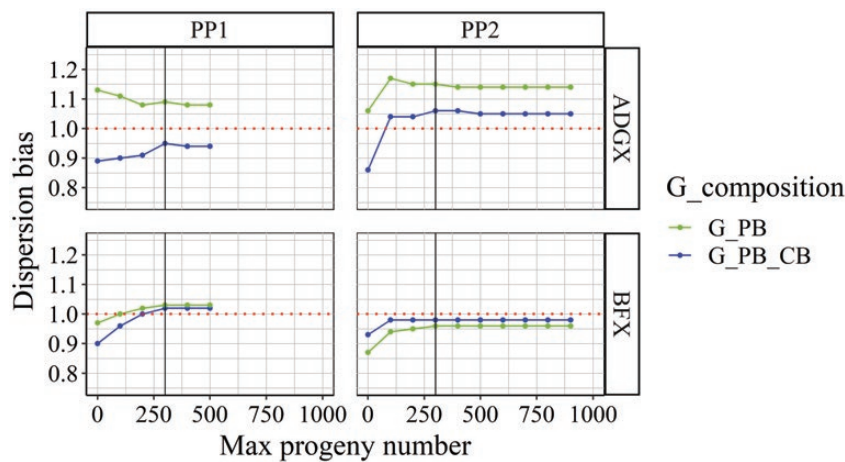

Figure 3. Dispersion bias of breeding values of purebred sires for crossbred performance traits given two genotypes set compositions and the number of CB phenotype progeny information. The vertical line represents the threshold between young and proven sires (300 CB progeny phenotypes), and the dotted line indicates unbiasedness.
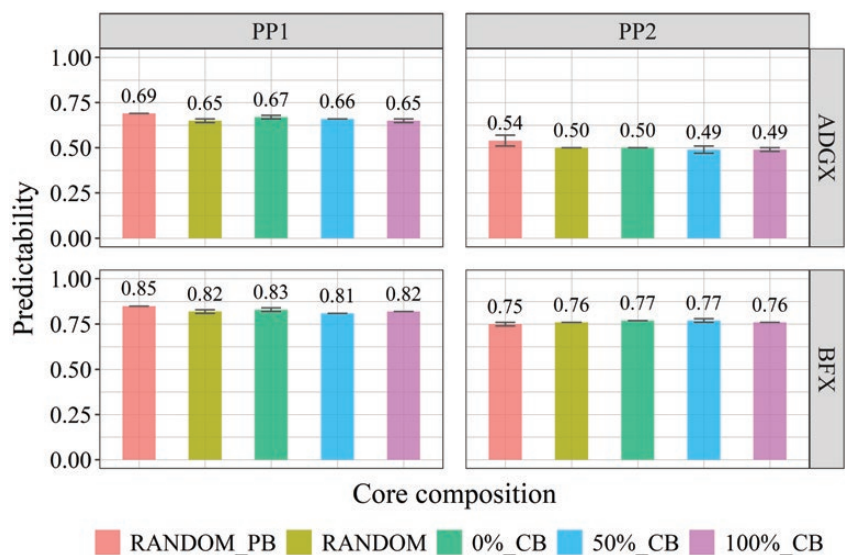
Figure 4. Predictability of breeding values of purebred sires for crossbred performance traits given different APY core compositions. Bars indicate the range within one standard deviation between five replicates. Scenarios 0%_CB, 50%_CB, and 100%_CB, and RANDOM used the full genotyped set (G_PB_CB) with APY cores composed of 0%, 50%, and 100% CB genotypes, or of genotypes of PB and CB entirely chosen at random, respectively. The scenario RANDOM_PB used the genotype set only accounting for PB genotypes (G_PB) with a core composed of PB genotypes chosen randomly.
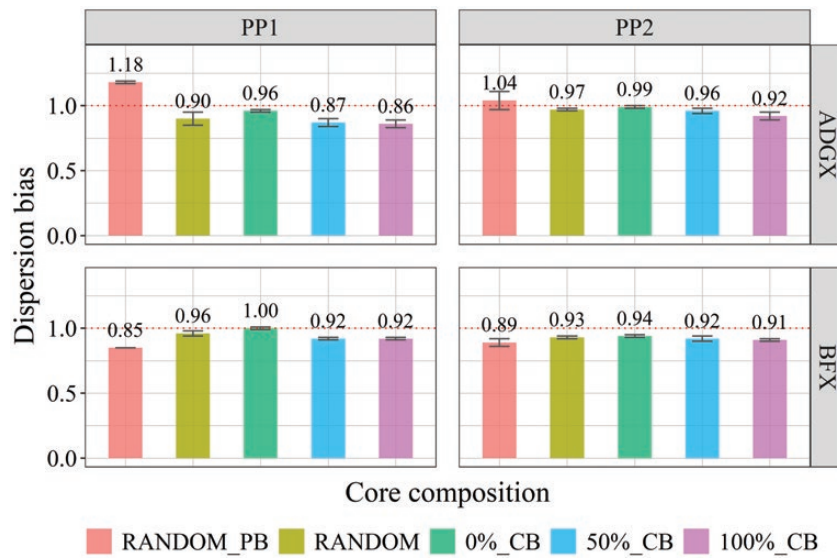
Figure 5. Dispersion bias of breeding values of purebred sires for crossbred performance traits given different APY core compositions. Bars indicate the range within one standard deviation between five replicates. Scenarios 0%_CB, 50%_CB, and 100%_CB, and RANDOM used the full genotyped set (G_PB_CB) with APY cores composed of 0%, 50%, and 100% CB genotypes, or of genotypes of PB and CB entirely chosen at random, respectively. The scenario RANDOM_PB used the genotype set only accounting for PB genotypes (G_PB) with a core composed of PB genotypes chosen randomly.
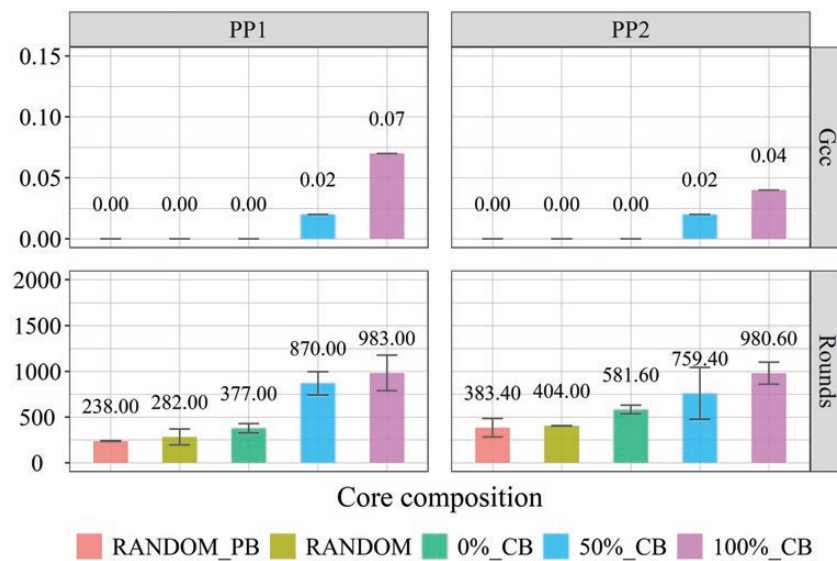
Figure 6. Average relationship between core animals and average number of rounds to BLUP convergence (10⁻¹³) given different APY core compositions. Bars indicate the range within one standard deviation between five replicates. Scenarios 0%_CB, 50%_CB, and 100%_CB, and RANDOM used the full genotyped set (G_PB_CB) with APY cores composed of 0%, 50%, and 100% CB genotypes, or of genotypes of PB and CB entirely chosen at random, respectively. The scenario RANDOM_PB used the genotype set only accounting for PB genotypes (G_PB) with a core composed of PB genotypes chosen randomly.

evaluation exclusively based on a PB reference population might represent a more usual analysis in the pig industry. Compared to core scenarios accounting for CB genotypes, RANDOM_PB presented the highest predictability, with relative increases ranging from 2.4% to 10.2% (Figure 4). An exception occurred for BFX in PP2, where RANDOM_PB performed similarly to other core compositions. On the other hand, an overall increase in bias was observed with the RANDOM_PB core (Figure 5). This difference was more evident for traits in PP1, in which the increase in bias with RANDOM_PB ranged from 0.04 to 0.15 points in absolute values compared with other core scenarios. Moreover, the number of rounds to convergence

with RANDOM_PB was, on average, the smallest among all core scenarios, indicating that $\mathbf{G}_{\mathbf{APY}}^{-1}$ from a genomic set with no CB genotypes might present a better numerical condition (Fragomeni et al., 2015). Pocrnic et al. (2019) and Vandenplas et al. (2018) also evaluated the inclusion of CB genotypes in the APY core using a dataset on a real and simulated pig population, respectively. Differently than observed in our study, the authors suggested that CB performance was better predicted when CB animals contribute with genotypes in the APY cores, whereas the differences in bias were generally similar for different core setups. The different results observed in our study could be due to the potential lower imputation accuracy of CB genotypes from

a low-density panel. As pointed out by Misztal (2016), the optimum performance of APY might also rely on the quality of genotypes being used in the core set.

Conclusions

Adding CB genotypes imputed from a low-density panel (600 SNP) to genomic analyses using the direct inverse of **G** neither improves predictability nor reduces the dispersion bias of PB sires' GEBV for CB performance, even for sires with fewer phenotypic progeny information in the analysis. That further indicates that the inclusion of those imputed CB genotypes is of no benefit to offset CB phenotyping while keeping constant predictability for PB sires' GEBV. When CB imputed genotypes are incorporated in analyses with APY, a random core composition or a core with no CB genotypes will reduce dispersion bias and the number of rounds to convergence but will not affect predictability. Still, a random core composition from a genomic set with only PB genotypes will result in the highest predictability and the smallest number of rounds to convergence, although dispersion bias will be increased. Genotyping CB individuals for a low-density panel (600 SNP) is a valuable identification tool for linking CB phenotypes to pedigree; however, the inclusion of those CB genotypes imputed from a low-density panel (600 SNP) might not benefit genomic predictions for PB individuals or offset CB phenotyping. Further studies will help understand the usefulness of those imputed CB genotypes for traits with lower PB–CB genetic correlations and traits not recorded in the PB environment, such as mortality and disease traits.

Acknowledgments

Conflict of Interest Statement

The authors declare no conflicts of interest.

Literature Cited

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci.* 93:743–752. doi:10.3168/jds.2009-2730.

Cesarani, A., D. Lourenco, S. Tsuruta, A. Legarra, E. Nicolazzi, P. VanRaden, and I. Misztal. 2022. Multibreed genomic evaluation for production traits of dairy cattle in the United States using single-step genomic best linear unbiased predictor. *J. Dairy Sci.* 105:5141–5152. doi:10.3168/jds.2021-21505.

Christensen, O. F., and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. *Genet. Sel. Evol.* 42:1–8. doi:10.1186/1297-9686-42-2.

Dekkers, J. C. M. 2007. Marker-assisted selection for commercial crossbred performance. *J. Anim. Sci.* 85:2104–2114. doi:10.2527/jas.2006-683.

Fragomeni, B., D. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, A. Legarra, T. Lawlor, and I. Misztal. 2015. Hot topic: use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes. *J. Dairy Sci.* 98:4090–4094. doi:10.3168/jds.2014-9125.

Garrick, D. J. 2017. The role of genomics in pig improvement. *Anim. Prod. Sci.* 57:2360–2365. doi:10.1071/an17277.

Geweke, J. 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Stat.* 4:641–649. doi:10.21034/sr.148.

Hidalgo, A. M., J. W. M. Bastiaansen, M. S. Lopes, B. Harlizius, M. A. M. Groenen, and D. J. de Koning. 2015. Accuracy of Predicted Genomic Breeding Values in Purebred and Crossbred Pigs. *G3-Genes Genom. Genet.* 5:1575–1583. doi:10.1534/g3.115.018119.

Hidalgo, J., S. Tsuruta, D. Lourenco, Y. Masuda, Y. Huang, K. A. Gray, and I. Misztal. 2020. Changes in genetic parameters for fitness and growth traits in pigs under genomic selection. *J. Anim. Sci.* 98:1–12. doi:10.1093/jas/skaa032.

Hollifield, M. K., D. Lourenco, S. Tsuruta, M. Bermann, J. T. Howard, and I. Misztal. 2021. Impact of including the cause of missing records on genetic evaluations for growth in commercial pigs. *J. Anim. Sci.* 99:1–5. doi:10.1093/jas/skab226.

Iversen, M. W., O. Nordbø, E. Gjerlaug-Enger, E. Grindflek, M. S. Lopes, and T. H. E. Meuwissen. 2017. Including crossbred pigs in the genomic relationship matrix through utilization of both linkage disequilibrium and linkage analysis. *J. Anim. Sci.* 95:5197–5207. doi:10.2527/jas2017.1705.

Knol, E. F., B. Nielsen, and P. W. Knap. 2016. Genomic selection in commercial pig breeding. *Anim. Front.* 6:15–22. doi:10.2527/af.2016-0003.

Legarra, A., and A. Reverter. 2017. Can we frame and understand cross-validation results in animal breeding. In: Proceedings of the 22nd Conference Association for the Advancement of Animal Breeding and Genetics. p 2–5.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.* 92:4656–4663. doi:10.3168/jds.2009-2061.

Legarra, A., O. F. Christensen, Z. G. Vitezica, I. Aguilar, and I. Misztal. 2015. Ancestral relationships using metafounders: finite ancestral populations and across population relationships. *Genetics* 200:455–468. doi:10.1534/genetics.115.177014.

Leite, N. G., E. F. Knol, A. L. S. Garcia, M. S. Lopes, L. Zak, S. Tsuruta, F. F. e. Silva, and D. Lourenco. 2021. Investigating pig survival in different production phases using genomic models. *J. Anim. Sci.* 99:1–11. doi:10.1093/jas/skab217.

Lourenco, D. A. L., I. Misztal, S. Tsuruta, I. Aguilar, T. J. Lawlor, S. Forni, and J. I. Weller. 2014. Are evaluations on young genotyped animals benefiting from the past generations? *J. Dairy Sci.* 97:3930–3942. doi:10.3168/jds.2013-7769.

Lourenco, D., S. Tsuruta, B. Fragomeni, C. Chen, W. Herring, and I. Misztal. 2016. Crossbreed evaluations in single-step genomic best linear unbiased predictor using adjusted realized relationship matrices. *J. Anim. Sci.* 94:909–919. doi:10.2527/jas.2015-9748.

Lutaaya, E., I. Misztal, J. Mabry, T. Short, H. Timm, and R. Holzbauer. 2002. Joint evaluation of purebreds and crossbreds in swine. *J. Anim. Sci.* 80:2263–2266. doi:10.1093/ansci/80.9.2263.

Maiorano, A., A. Assen, P. Bijma, C. Chen, J. Silva, W. Herring, S. Tsuruta, I. Misztal, and D. Lourenco. 2019. Improving accuracy of direct and maternal genetic effects in genomic evaluations using pooled boar semen: a simulation study. *J. Anim. Sci.* doi:10.1093/jas/skz207.

Misztal, I. 2016. Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. *Genetics: genetics* 115:182089. doi: 10.1534/genetics.115.182089.

Misztal, I., A. Legarra, and I. Aguilar. 2014a. Using recursion to compute the inverse of the genomic relationship matrix. *J. Dairy Sci.* 97:3943–3952. doi:10.3168/jds.2013-7752.

Misztal, I., S. Tsuruta, D. Lourenco, I. Aguilar, A. Legarra, and Z. Vitezica. 2014b. Manual for BLUPF90 family of programs – (accessed October 1, 2022) http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all2.pdf.

Pocrnic, I., D. A. Lourenco, Y. Masuda, A. Legarra, and I. Misztal. 2016. The dimensionality of genomic information and its effect on genomic prediction. *Genetics* 203:573–581. doi:10.1534/genetics.116.187013.

Pocrnic, I., D. A. Lourenco, C. -Y. Chen, W. O. Herring, and I. Misztal. 2019. Crossbred evaluations using single-step genomic BLUP and algorithm for proven and young with different sources of data. *J. Anim. Sci.* 97:1513–1522. doi:10.1093/jas/skz042.

See, G. M., B. E. Mote, and M. L. Spangler. 2020. Impact of inclusion rates of crossbred phenotypes and genotypes in nucleus selection programs. *J. Anim. Sci.* 98:1–13. doi:10.1093/jas/skaa360.

See, G. M., B. E. Mote, and M. L. Spangler. 2021. Bias in variance component estimation in swine crossbreeding schemes using selective genotyping and phenotyping strategies. *J. Anim. Sci.* 99:1–13. doi:10.1093/jas/skab293.

Steyn, Y., D. A. Lourenco, C. Y. Chen, B. D. Valente, J. Holl, W. O. Herring, and I. Misztal. 2021. Optimal definition of contemporary groups for crossbred pigs in a joint purebred and crossbred genetic evaluation. *J. Anim. Sci.* 99:1–8. doi:10.1093/jas/skaa396.

Toosi, A., R. Fernando, J. Dekkers, and R. Quaas. 2010. Genomic selection in admixed and crossbred populations. *J. Anim. Sci.* 88:32. doi:10.2527/jas.2009-1975.

Tsuruta, S., T. J. Lawlor, D. A. L. Lourenco, and I. Misztal. 2021. Bias in genomic predictions by mating practices for linear type traits in a large-scale genomic evaluation. *J. Dairy Sci.* 104:662–677. doi:10.3168/jds.2020-18668.

Vandenplas, J., M. P. Calus, and J. Napel. 2018. Sparse single-step genomic BLUP in crossbreeding schemes. *J. Anim. Sci.* 96:2060–2073. doi:10.1093/jas/sky136.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91:4414–4423. doi:10.3168/jds.2007-0980.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.* 92:16–24. doi:10.3168/jds.2008-1514.

Vitezica, Z., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in genomic predictions for populations under selection. *Genet. Res.* 93:357–366. doi:10.1017/S001667231100022X.

Wei, M., and J. H. J. van der Werf. 1994. Maximizing genetic response in crossbreds using both purebred and crossbred information. *Anim. Sci.* 59:401–413. doi:10.1017/S0003356100007923.

Xiang, T., P. Ma, T. Ostersen, A. Legarra, and O. F. Christensen. 2015. Imputation of genotypes in Danish purebred and two-way crossbred pigs using low-density panels. *Genet. Sel. Evol.* 47:54. doi:10.1186/s12711-015-0134-4.