Journal of
Animal Breeding and Genetics

J. Anim. Breed. Genet. ISSN 0931-2668

ORIGINAL ARTICLE

Accounting for genetic architecture in single- and multipopulation genomic prediction using weights from genomewide association studies in pigs[a]

R. Veroneze[1,2], P.S. Lopes[1], M.S. Lopes[2,3], A.M. Hidalgo[2,4], S.E.F. Guimarães[1], B. Harlizius[3], E.F. Knol[3], J.A.M. van Arendonk[2], F.F. Silva[1] & J.W.M. Bastiaansen[2]

1 Departamento de Zootecnia, Universidade Federal de Viçosa, Viçosa, Brazil
2 Animal Breeding and Genomics Centre, Wageningen University, Wageningen, the Netherlands
3 Topigs Norsvin Research Center, Beuningen, the Netherlands
4 Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden

Keywords

accuracy; genetic variance; genomic relationship; single nucleotide polymorphisms.

Correspondence

R. Veroneze, Departamento de Zootecnia,
Universidade Federal de Viçosa, Viçosa,
Brazil.
Tel: +55 31 3899 2273;
Fax: +5531 3899-2275;
E-mail: veronezerenata@gmail.com

Received: 24 July 2015;
accepted: 18 December 2015

[a]We thank the CAPES/Nuffic program, CNPq and FAPEMIG for financial support, and Topigs Norsvin for providing data.

Summary

We studied the effect of including GWAS results on the accuracy of single- and multipopulation genomic predictions. Phenotypes (backfat thickness) and genotypes of animals from two sire lines (SL1, n = 1146 and SL3, n = 1264) were used in the analyses. First, GWAS were conducted for each line and for a combined data set (both lines together) to estimate the genetic variance explained by each SNP. These estimates were used to build matrices of weights (D), which was incorporated into a GBLUP method. Single population evaluated with traditional GBLUP had accuracies of 0.30 for SL1 and 0.31 for SL3. When weights were employed in GBLUP, the accuracies for both lines increased (0.32 for SL1 and 0.34 for SL3). When a multipopulation reference set was used in GBLUP, the accuracies were higher (0.36 for SL1 and 0.32 for SL3) than in single-population prediction. In addition, putting together the multipopulation reference set and the weights from the combined GWAS provided even higher accuracies (0.37 for SL1, and 0.34 for SL3). The use of multipopulation predictions and weights estimated from a combined GWAS increased the accuracy of genomic predictions.

Introduction

Genomewide association studies (GWAS) have been conducted to disclose the genetic architecture of complex traits. These studies have generated a considerable amount of information for many traits in livestock species, but this information has not been extensively exploited in genomic prediction.

The traditional GBLUP assumes that quantitative traits are controlled by a large number of genes that contribute equally to the trait (infinitesimal model); thus, the same variance is, *a priori*, attributed to all markers (Goddard 2009). Nevertheless, it has been shown that a finite number of genes control quantitative traits (Hayes & Goddard 2001); therefore, models that represent the true underlying genetic architecture of the trait may have higher accuracy than the GBLUP.

For single-population genomic prediction, Zhang *et al.* (2010) proposed a trait-specific, marker-derived relationship matrix (TA-matrix), which had a greater predictive ability than the traditional genomic best linear unbiased prediction (GBLUP) method. Those authors attributed the greater predictive ability to the fact that the TA-matrix emphasized markers that contributed to the genetic variance of the trait.

Multipopulation genomic prediction emerged as an alternative for implementing genomic selection in

doi:10.1111/jbg.12202

small populations. Combining populations from different breeds or lines increases the number of animals available, which might contribute to a more accurate prediction of genomic breeding values (GEBVs) than a single-population prediction.

Multipopulation genomic predictions have been studied in cattle (Olson *et al.* 2012), pigs (Hidalgo *et al.* 2015), chickens (Simeone *et al.* 2012) and sheep (Legarra *et al.* 2014). However, the results showed that multiple populations sometimes increased and other times decreased the accuracy of genomic predictions. The variability in these previous results reflected the fact that predictions in multipopulation analyses were more complex than single-population predictions. First, increasing the reference population by adding unrelated animals will decrease the average relationship between animals in the reference and validation sets. Second, differences in the population history, such as demography, inbreeding, genetic background, and selection, could lead to divergences in linkage disequilibrium (LD), allele frequencies and genetic architecture. Genetic differences between populations depend on the number of generations since their last common ancestor, the size of the populations, and the degree of exchange of genetic material between populations.

According to Harris & Johnson (2010), differences in allele frequency between populations should be considered in a multipopulation prediction. In a study on beef cattle, Chen *et al.* (2013) proposed a two-population genomic relationship matrix, which considered differences in allele frequencies between populations. However, they did not find increased accuracy in the multipopulation prediction compared to the single-population prediction. Based on the same approach with experimental data from several different pig populations, Veroneze *et al.* (2015) found similar or lower accuracies for the multipopulation prediction compared to single-population prediction. Those studies suggested that considering differences in allele frequency was not sufficient to improve the accuracy of multipopulation predictions.

Another potential improvement that addressed differences between populations was the incorporation of GWAS results from these populations. This approach could benefit multipopulation genomic predictions, because including GWAS results could emphasize markers that explain genetic variance in the target population. In addition, differences in allele frequency between populations could be accounted for simultaneously in the two-population genomic relationship matrix. However, the value of using GWAS results in multipopulation predictions has not been studied.

The objective of this study was to develop weighted genomic relationships incorporating GWAS results and to investigate their effect on the accuracy of single- and multipopulation genomic predictions.

Material and methods

Data recording and sample collection were conducted strictly in line with the Dutch law on the protection of animals.

Data

Data used in this study consisted of phenotypes (backfat thickness measured on live animals using ultrasound) and genotypes of animals from two purebred pig populations (SL1, n = 1146; SL3, n = 1264). SL1 is a synthetic sire line, and it is a combination of Duroc and Belgian Landrace created in about 1980, and SL3 is a Pietrain sire line. Animals were genotyped using the Illumina Porcine SNP60 Beadchip. The GenABEL package, implemented in R software (Aulchenko *et al.* 2007), was used to perform individual sample and single nucleotide polymorphisms (SNP) quality control. Animals with call rates <95% and SNPs with call rates <95%, with minor allele frequencies <0.01, or with deviations from Hardy–Weinberg equilibrium ($p < 10^{-7}$) were excluded. All SNPs located on sex chromosomes were also excluded. After quality control, missing genotypes of SNPs were imputed with BEAGLE 3.3.2 software (Browning & Browning 2013), using the default parameters.

Estimates of the fixed effects used for precorrecting the phenotypes were obtained by fitting a single trait, pedigree-based linear model ASReml v3.0 (Gilmour *et al.* 2009) across lines as described by Veroneze *et al.* (2015). In this analysis, a larger data set (706 023 animals) that included all contemporaneous animals of the genotyped animals were used to more accurately account for the contemporary group effects. The model included fixed effects of sex, herd-year-month, and the covariate body weight at the time of measuring backfat. The animal additive genetic, litter and residual were included as random effects.

Model and genomic relationship matrices

The genomic best linear unbiased prediction (GBLUP) method was used for genomic prediction. The general model was: $\mathbf{y} = \mathbf{1}\mu + \mathbf{Zg} + \mathbf{Wc} + \mathbf{e}$, where $\mathbf{y}$ is the phenotype corrected for fixed effects; μ is the overall mean; $\mathbf{g}$ is the vector of breeding values, $\mathbf{g} \sim N(0, \sigma_g^2\mathbf{G})$; $\mathbf{c}$ is the vector of random litter effect, $\mathbf{c} \sim N(0, \sigma_c^2\mathbf{I})$; $\mathbf{e}$ is the vector of residuals, $\mathbf{e} \sim N(0, \sigma_e^2\mathbf{I})$. $\mathbf{Z}$ and $\mathbf{W}$ are the incidence matrices for $\mathbf{g}$ and $\mathbf{c}$, respectively. In the multipopulation scenarios, a fixed effect of population was added to the model.

In the genomic relationship matrix, differences in allele frequencies between populations were accounted for with the method described by Chen *et al.* (2013). Summarizing, X was a matrix with genotype values coded as −1, 0 and 1 for the three SNP genotypes (A_1A_1, A_1A_2 and A_2A_2, respectively); the matrix had dimensions, n × m, where n was the number of animals and m was the number of SNPs. Matrix $\mathbf{X}$ included all animals, from both the reference and validation sets. Matrix $\mathbf{X}$ was organized into two blocks: $\mathbf{X} = [\mathbf{X_1}\ \ \mathbf{X_2}]'$ where $\mathbf{X_1}$ represents the genotypes of line 1, and $\mathbf{X_2}$ represents the genotypes of line 2. $\mathbf{P}$ was a matrix of allele frequencies $\mathbf{P} = [\mathbf{P_1}\ \ \mathbf{P_2}]'$ that corresponded to $\mathbf{X}$; each row in $\mathbf{P_1}$ (or $\mathbf{P_2}$) was a replicate row vector, $\mathbf{p_1}$ (or $\mathbf{p_2}$), with the frequency of allele A_2 for SNP *k* in line 1 (or line 2). The matrix $\mathbf{M}$ was computed to set the mean values of the allele effects to 0: $\mathbf{M} = [\mathbf{M_1}\ \ \mathbf{M_2}]' = \mathbf{X} - \mathbf{2P} + \mathbf{1}$, where $\mathbf{1}$ represents a matrix of ones. The matrix, $\mathbf{G}$, was computed as follows:

$$\mathbf{G} = \begin{bmatrix} \mathbf{M_1DM_1'}/2\sum p_{1k}(1-p_{1k}) & \mathbf{M_1DM_2'}\Big/2\sum\left[p_{1k}(1-p_{1k})p_{2k}(1-p_{2k})\right]^{1/2} \\ \mathbf{M_2DM_1'}\Big/2\sum\left[p_{1k}(1-p_{1k})p_{2k}(1-p_{2k})\right]^{1/2} & \mathbf{M_2DM_2'}/2\sum p_{2k}(1-p_{2k}) \end{bmatrix}$$

Here, $\mathbf{D}$ is a diagonal matrix of weights for the SNPs, which will be described in detail in the next section. In the traditional GBLUP, $\mathbf{D}$ is an identity matrix.

Diagonal matrices

First, a single-population and multipopulation GBLUP analysis was carried out with a $\mathbf{G}$ matrix, computed as described by VanRaden (2008), which included all animals: $\mathbf{G} = \mathbf{MM}'/2\sum p_iq_i$.

This $\mathbf{G}$ matrix was entered as a user defined matrix (*grm* option) in the software ASREML (Gilmour *et al.* 2009) to predict the GEBVs. The predicted GEBVs ($\hat{g}$) were used to compute the diagonal elements of $\mathbf{D}$, according to the method proposed by Wang *et al.* (2012). In that method, SNP effects ($\hat{\mathbf{u}}$) were estimated with the equation: $\hat{\mathbf{u}} = \lambda\mathbf{M}'\mathbf{G}^{-1}\hat{\mathbf{g}}$, where $\lambda = 1\big/(\sum_{i=1}^{m} 2p_i(1-p_i))$; in the latter equation, m is the number of SNPs, and p_i is the allele frequency of the second allele of the *i*th SNP. Then, the variance of each SNP effect was estimated as described by Falconer & Mackay (1996): $\hat{\sigma}_{u_i}^2 = \hat{u}_i^2 2p_i(1-p_i)$. These variances were used to build the diagonal elements of the $\mathbf{D_0}$ matrix. The $\mathbf{D_0}$ matrix was normalized with $\mathbf{D} = (\mathbf{tr}(\mathbf{I})/\mathbf{tr}(\mathbf{D_0}))*\mathbf{D_0}$, where $\mathbf{I}$ is an identity matrix.

Four different $\mathbf{D}$ matrices were used in this study: one was an identity matrix (traditional GBLUP), and three were $\mathbf{D}$ matrices obtained with the three data sets used to estimate the weights (Fig. 1). These diagonal $\mathbf{D}$ matrices contained weights for the SNPs. These weights were included in the genomic relationship matrix, which resulted in four different $\mathbf{G}$ matrices ($\mathbf{G^{identity}}$, $\mathbf{G^{D\_comb}}$, $\mathbf{G^{D1}}$ and $\mathbf{G^{D3}}$). Each $\mathbf{G}$ matrix was used to predict GEBVs for animals from SL1 and SL3, with both single- and multipopulation reference sets.

The predictions were conducted according to four strategies: first, the traditional GBLUP, where all markers had the same weight (scenarios 1, 5 and 9 in Table 1); second, markers were weighted, and the weights were from the same population that will be predicted (scenarios 2, 6 and 10); third, the markers were weighted, but the weights were from an unrelated population (scenarios 3, 7 and 11); and fourth, the markers were weighted, and the weights were from a combined GWAS, which used both pig lines together (scenarios 4, 8 and 12). In scenarios 1–4, the

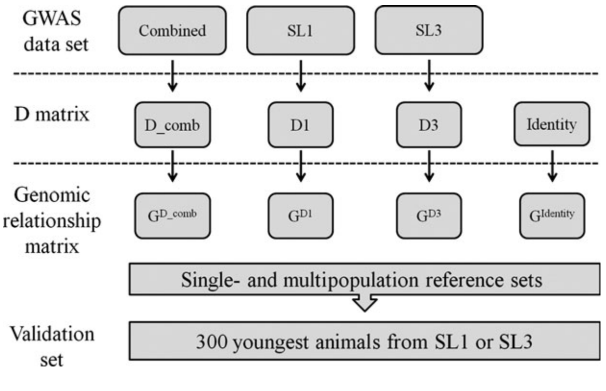
GWAS data set
Combined
SL1
SL3
D matrix
D_comb
D1
D3
Identity
Genomic relationship matrix
G^D_comb
G^D1
G^D3
G^Identity
Single- and multipopulation reference sets
Validation set
300 youngest animals from SL1 or SL3

Figure 1 Schematic representation of the scenarios evaluated. Combined refers to a data set composed by the combination of SL1 and SL3.

Table 1 Accuracy, genomic heritability (h^2) and slope for backfat thickness in the scenarios evaluated

Validation population	Scenarios	Reference population	Genomic relationship matrix	Number of animals in the reference population	Accuracy	h^2	Slope
SL1	1	SL1	$G^{Identity}$	800	0.30	0.49	0.89
	2	SL1	G^{D1}	800	0.32	0.76	0.47
	3	SL1	G^{D3}	800	0.29	0.45	0.87
	4	SL1	G^{D_comb}	800	0.32	0.40	0.88
	5	SL1 + SL3	$G^{Identity}$	800	0.20	0.43	0.78
	6	SL1 + SL3	G^{D1}	800	0.29	0.57	0.62
	7	SL1 + SL3	G^{D3}	800	0.25	0.52	0.73
	8	SL1 + SL3	G^{D_comb}	800	0.32	0.36	1.07
	9	SL1 + SL3	$G^{Identity}$	1600	0.36	0.38	1.16
	10	SL1 + SL3	G^{D1}	1600	0.33	0.53	0.58
	11	SL1 + SL3	G^{D3}	1600	0.36	0.42	0.98
	12	SL1 + SL3	G^{D_comb}	1600	0.37	0.30	1.09
SL3	1	SL3	$G^{Identity}$	800	0.31	0.40	1.22
	2	SL3	G^{D3}	800	0.24	0.65	0.47
	3	SL3	G^{D1}	800	0.33	0.42	1.27
	4	SL3	G^{D_comb}	800	0.34	0.37	1.12
	5	SL1 + SL3	$G^{Identity}$	800	0.08^{ns}	0.43	0.30
	6	SL1 + SL3	G^{D3}	800	0.15	0.52	0.32
	7	SL1 + SL3	G^{D1}	800	0.04^{ns}	0.57	0.11
	8	SL1 + SL3	G^{D_comb}	800	0.24	0.36	0.69
	9	SL1 + SL3	$G^{Identity}$	1600	0.32	0.38	1.12
	10	SL1 + SL3	G^{D3}	1600	0.29	0.42	0.64
	11	SL1 + SL3	G^{D1}	1600	0.23	0.53	0.63
	12	SL1 + SL3	G^{D_comb}	1600	0.34	0.30	1.10

predictions were performed with a single-population reference set. In scenarios 5–8, half of the animals were substituted with individuals from another population to calculate multipopulation predictions. In scenarios 9–12, two unrelated populations were combined, which doubled the number of animals, to calculate multipopulation predictions.

The validation sets consisted of the 300 youngest animals of each population. The reference sets consisted of 800 animals for single-population predictions and 800 (400 animals of each line) or 1600 (800 animals of each line) animals for multipopulation predictions.

The accuracy of the GEBVs was computed as the Pearson correlation between the predicted GEBV and the corrected phenotype. To measure the bias of the GEBV, the regression coefficient (slope) was calculated for each scenario regressing the corrected phenotypes on the GEBVs.

Results

SNP effects

Our aim was to improve genomic predictions, based on single- and multipopulation reference sets. The first step of our methodology was to estimate the effects of each SNP on backfat thickness that were used to compute the weights for each SNP in building the genomic relationship matrices. Manhattan plots (Fig. 2) show these SNP effects for the SL1 and SL3 lines, separately, and for the combined SL1 + SL3 data set. The estimated SNP effects in the SL3 data set were, on average, lower and less variable than the SNP effects in the SL1 data set. When the two pig lines were combined (SL1 + SL3), the average effects became larger than those observed for a single-population analysis. In addition to the Manhattan plots, the dispersion plots (Fig. 3) clearly showed differences in the genetic architecture between these two pig lines. The correlation between the SNP effects in the two lines was only 0.02. The dispersion plots of SNP effects estimated for the combined data set and the SNP effects for the single populations showed higher correlations between the effects (0.36 for SL1 and 0.38 for SL3).

Genomic relationships

The effects of the different weights for the SNPs in the **G** matrix were visualized in multidimensional scaling plots of the populations (Fig. 4). In both lines, the use

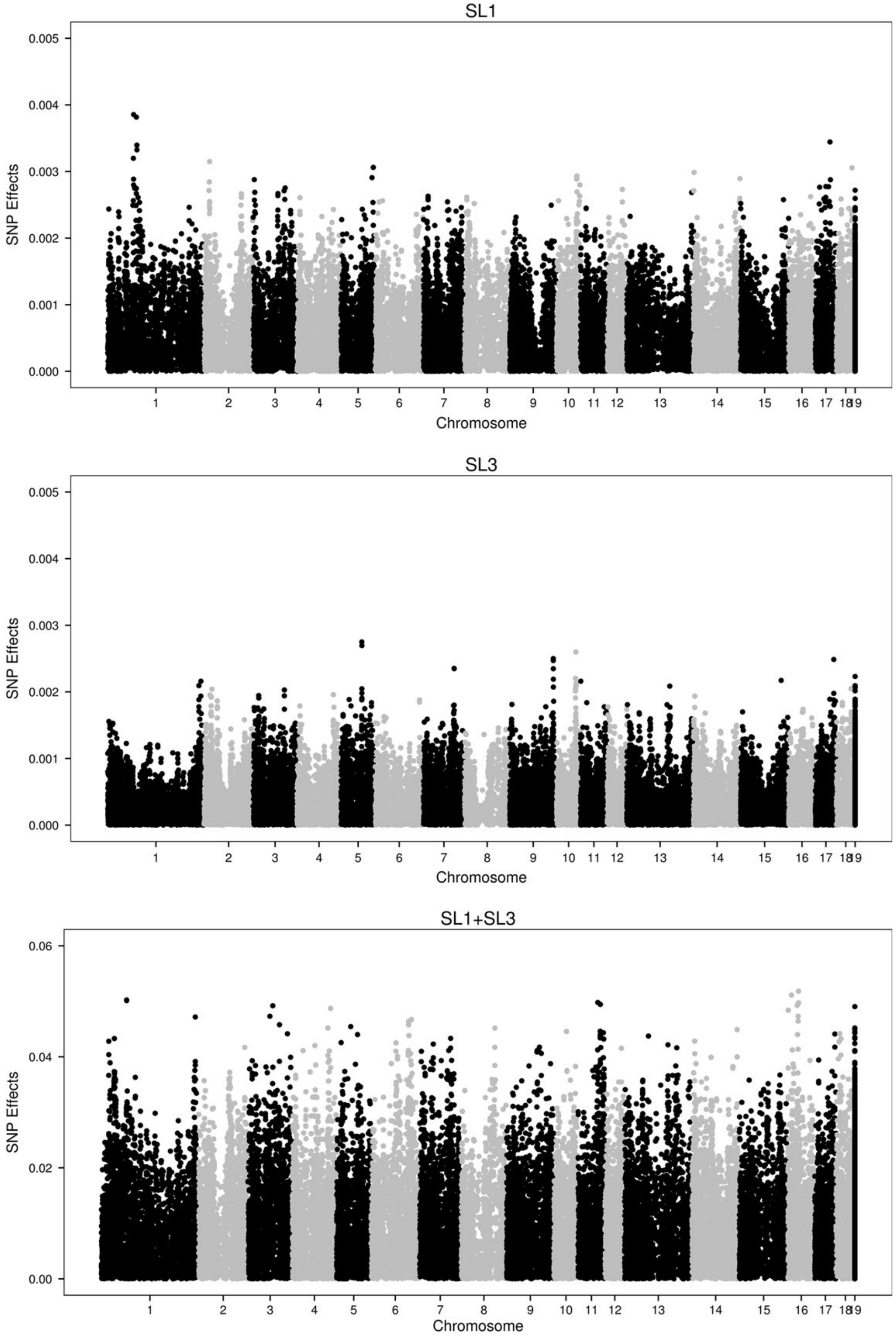

Figure 2 Manhattan plots of the SNPs effects for backfat thickness for SL1, SL3 and the combined data set (SL1 + SL3).

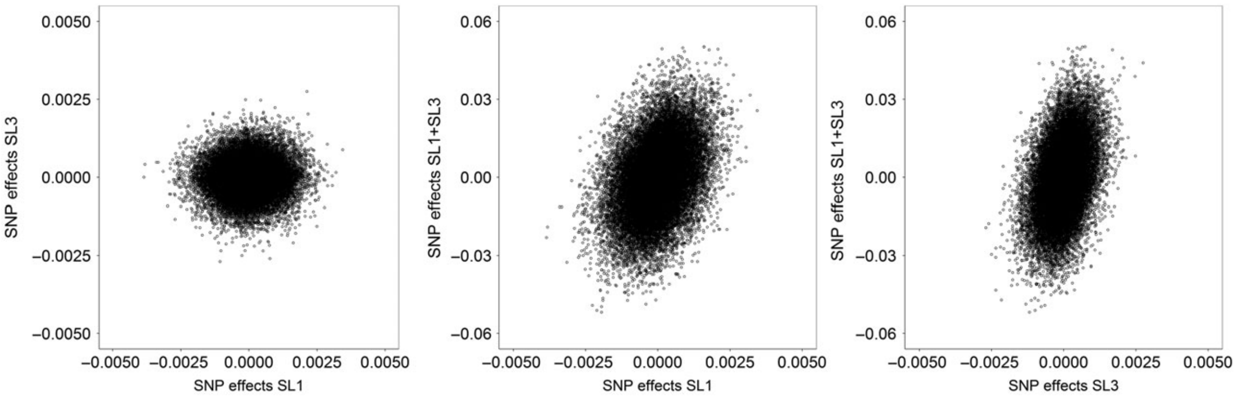

Figure 3 Dispersion plot for backfat thickness for SL1, SL3 and a combined data set (SL1 + SL3).

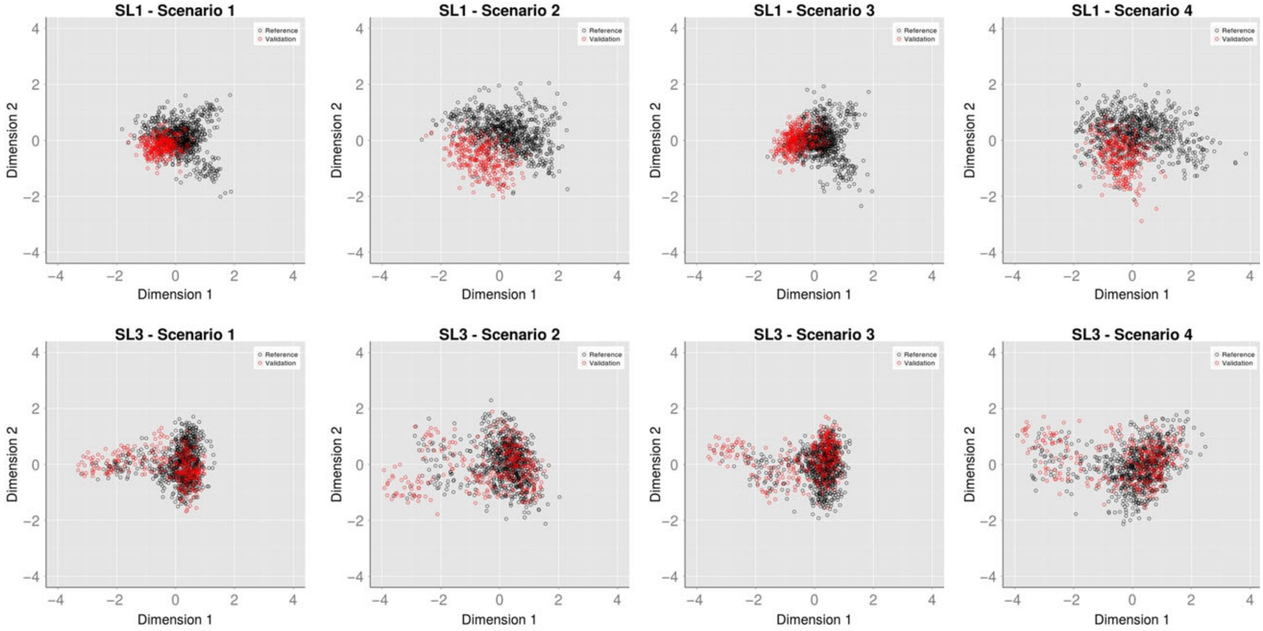

Figure 4 Multidimensional scaling plot showing a two dimensional projection of the populations using different weighted relationships. Scenario 1 corresponds to traditional GBLUP; in scenario 2, the weights were obtained using the same population that was predicted; in scenario 3, the weights were computed using a unrelated population; and in scenario 4, the weights were computed combining the two populations.

of weights that were estimated from the data set of the other line (scenario 3) only slightly modified the projection of reference and validation populations. With weights that were estimated from the data set of the same line (scenario 2) or weights estimated from the combined data set (scenario 4), the projection of the animals became more dispersed.

GEBVs accuracy

In the first 4 scenarios, the reference and validation populations were from the same pig line. For these *within-population* scenarios, in a traditional GBLUP ($\mathbf{G}^{\mathbf{Identity}}$), where all markers had the same weight, the accuracies (Table 1) were 0.30 for SL1 and 0.31 for SL3. When weighted markers were used in the genomic relationship matrix, the accuracy for SL1 increased (0.32) when the weights were obtained from the data set of the same line ($\mathbf{G}^{\mathbf{D1}}$) or from the combined data set ($\mathbf{G}^{\mathbf{D\_comb}}$), but the accuracy decreased when the weights were obtained from the data set of the other line ($\mathbf{G}^{\mathbf{D3}}$). For SL3, a different pattern was observed. The accuracy increased when the weights were obtained from the data set of the other line ($\mathbf{G}^{\mathbf{D1}}$) or from the combined data set ($\mathbf{G}^{\mathbf{D\_comb}}$), but the accuracy decreased when the weights were obtained from the data set of the same line ($\mathbf{G}^{\mathbf{D3}}$).

In scenarios 5–8, half the animals in the reference set were replaced with animals from the other pig line. Thus, this *multi-population* prediction was conducted with 400 animals from each line. In scenarios 5–8, the reference population was always the same, and only the genomic relationship matrix was changed (Table 1). For SL1, adding weights to the markers increased the accuracy, and the highest accuracy was obtained with $\mathbf{G}^{\mathbf{D\_comb}}$ (0.32). This accuracy was equal to that obtained in scenario 4, where 800 animals from the same population were included, and the $\mathbf{G}^{\mathbf{D\_comb}}$ weighted matrix was used. For SL3, scenarios 6 ($\mathbf{G}^{\mathbf{D3}}$) and 8 ($\mathbf{G}^{\mathbf{D\_comb}}$) showed increased accuracy compared to scenario 5 ($\mathbf{G}^{\mathbf{Identity}}$), but the accuracy was lower than the single-population predictions obtained in scenarios 1–4.

Scenarios 9–12 were also *multi-population* predictions, but the reference sets included an extra 800 animals from a different line, which doubled the reference set to 1600 animals (Table 1). For SL1, this increase in the reference population resulted in greater accuracies with all the **G** matrices compared to all the single-population predictions (scenarios 1–4).

The use of $\mathbf{G}^{\mathbf{D\_comb}}$ (scenario 12) resulted in the highest accuracy (0.37); the second highest accuracy was achieved with $\mathbf{G}^{\mathbf{identity}}$ (scenario 9; 0.36). For SL3, the increase in the reference population only increased the accuracy compared to the single-population predictions with $\mathbf{G}^{\mathbf{identity}}$ (0.32 versus 0.31) and $\mathbf{G}^{\mathbf{D3}}$ (0.29 versus 0.24), but the accuracy was reduced with $\mathbf{G}^{\mathbf{D1}}$ (0.23 versus 0.33) and it remained the same with $\mathbf{G}^{\mathbf{D\_comb}}$ (0.34 versus 0.34).

Weighted genomic relationship matrices affected the estimates of genomic heritability (h^2). The size of this effect was found to depend on which population was used to obtain the weights. For both lines, the h^2 was largest in scenario 2, where a single population was used, and the weights were obtained from the data set of the reference population. The lowest h^2 was observed when $\mathbf{G}^{\mathbf{D\_comb}}$ was used (Table 1). The slope coefficient of the regression of the corrected phenotypes on GEBVs was in most of the cases away from 1, indicating bias in the GEBVs, mainly when using $\mathbf{G}^{\mathbf{D1}}$ and $\mathbf{G}^{\mathbf{D3}}$. The use of $\mathbf{G}^{\mathbf{D\_comb}}$ resulted in less biased predictions, in most scenarios, compared to predictions calculated with the other **G** matrices.

Discussion

The Manhattan plots revealed differences in the SNP effects on backfat between the two populations evaluated. Differences observed in both the peak sizes and in the distributions indicated that the two lines had different genetic architectures. We explored whether these different genetic architectures affected genomic predictions. Extending on the methodology proposed by Wang *et al.* (2012), we computed weights for the SNPs, based on GWAS analyses. These weights were subsequently used to build different **G** matrices for the GBLUP analyses. GBLUP was then applied with either single- or multipopulation reference sets. We found that, when the **G** matrix was built with weights based on GWAS information of the two populations combined ($\mathbf{G}^{\mathbf{D\_comb}}$), the prediction accuracy was increased with both single and multiple reference populations. Moreover, in addition to using $\mathbf{G}^{\mathbf{D\_comb}}$, doubling the multipopulation data sets resulted in even higher prediction accuracies than when a single population was used.

SNP effects

The plots of the SNP effects on backfat thickness for SL1 and SL3 pigs showed that, despite the fact that the same trait was evaluated, the populations differed in the distribution and size of the effects. These divergences can be explained by differences between the two lines in LD, population history (initial variants, bottlenecks and allele frequencies) and gene interactions. Previously, Veroneze *et al.* (2014) showed differences in the LD patterns of SL1 and SL3.

In the present study, when SL1 and SL3 data were combined into one data set, the SNP effects were, on average, greater than the effects observed in the single-population estimation. In a study on dairy cattle, Raven *et al.* (2014) suggested that a multibreed GWAS resulted in more precise mapping of the QTL, due to the lower level of LD between markers, in comparison with single-breed LD. In a study on German Holstein cattle, Liu *et al.* (2011) showed that, when the number of reference bulls increased from 735 to 5025, the SNP with the largest effect showed a 4.13-fold increase in effect size. In addition to increasing the precision of finding the QTL peaks, the use of a larger number of animals in the multipopulation GWAS might increase the statistical power of the analysis (Stranger *et al.* 2011).

Accuracy

Two populations can exhibit distinct genetic architectures (QTL number, distribution and effects), due to divergences in breeding goals or in the initial allele frequency. For example, allele substitution in the

DGAT1 locus caused different effects in Jersey and Holstein–Friesian populations in New Zealand (Spelman *et al.* 2002) and between Fleckvieh and Holstein–Friesian populations in Germany (Thaller *et al.* 2004). Hence, differences in genetic architecture between populations could be exploited by emphasizing the markers that explain more genetic variance in the target population. This notion introduces the possibility of using prior knowledge of genetic architecture for making single- and multipopulation predictions.

In the present study, we evaluated three strategies for computing weights of markers to be used in the GBLUP: (i) the weights were obtained using the same population that was predicted, (ii) the weights were obtained in an unrelated population and (iii) the weights were obtained in the two populations combined.

When the weights were obtained using the same population as the reference set (scenario 2), the prediction accuracy increased for SL1, but decreased for SL3. It has been shown that the accuracy of genomic prediction can be improved by reducing distances between the reference and validation animals and by increasing distances between animals within the reference population (Pszczola *et al.* 2012). In the present study, for both pig lines, including the marker weights resulted in increasing the distances between animals within the reference population (Fig. 4). This effect was more pronounced for SL1 than for SL3 pigs. However, before the inclusion of weights, compared to SL1, the SL3 pigs showed greater distances between individuals within the reference set and smaller distance between the validation and reference animals. These initial differences in the distances between the SL1 and SL3 groups may be one cause for the different changes in prediction accuracies when marker weights were included in the matrix. The use of weights obtained in an unrelated population (scenario 3) increased the accuracy of predicted GEBVs for SL3. This might be due to SNP effects for alleles with a rather low frequency in one line that might be estimated more precisely in the other line. This finding is important, because it suggests that, if a large number of genotyped animals is available in the other line, conducting a GWAS in that line will have greater power and result in better SNP effect estimates which leads to better marker weights for the smaller population.

Indeed, we found that using the $\mathbf{G}^{\mathbf{D\_comb}}$ resulted in higher prediction accuracy than the traditional GBLUP ($\mathbf{G}^{\mathbf{Identity}}$) for both pig lines, in single- and multipopulation scenarios. This improvement could be attributed to a better estimation of the marker effects, due to the large number of animals. Moreover, when two lines are pooled, the LD is reduced, and this increases the QTL mapping resolution. Thus, in the combined data set, the SNPs closest to the QTL would be identified, and consequently, they would be better able to track the effects of interest, both within and across populations.

Previous studies (Zhang *et al.* 2010, 2014; Tiezzi & Maltecca 2015) have indicated that the accuracy gained using weighted relationships matrices depended on how the trait was controlled; they showed that more efficiency was likely to be gained for traits controlled by small number of QTLs. The trait we evaluated was backfat thickness, which is apparently controlled by a large number of QTLs. Therefore, our results may not have been favoured by the trait analysed.

Evaluating dairy cattle and rice data, Zhang *et al.* (2014) also incorporated GWAS results in genomic predictions by adding weights for the markers. This strategy increased the prediction accuracy for two of three traits in dairy cattle and for nine of 11 traits in rice. In a study on Holstein cattle, Tiezzi & Maltecca (2015) concluded that weighted relationship matrices yielded higher accuracy and less bias in predictions for traits regulated by a few QTLs.

In scenarios 9–12, we added extra animals from an unrelated population to the reference set. With this approach, even when all markers were equally weighted (scenario 9), the accuracy was increased compared to that of the single-population prediction; the highest accuracies were obtained with $\mathbf{G}^{\mathbf{D\_comb}}$ (scenario 12) for both populations. Chen *et al.* (2014) found that, with pooled data, the accuracy of genomic prediction may be reduced when the analysis used weakly correlated QTL effects or a relatively low-density SNP panel. In our study, the SNP effects of both lines were very weakly correlated; therefore, for a different trait with highly correlated SNP effects across lines, a large increase in prediction accuracy might be found with this approach. Alternatively, the prediction accuracy might be further improved for the same trait using a higher density SNP panel. The SNP panel used in the present study did not show a consistent LD phase across the evaluated lines (Veroneze *et al.* 2014).

This was the first study to compute weights based on GWAS results from combined cohorts, and then, to use them in multipopulation predictions. The methodology presented here should be evaluated with additional traits and populations. Weights have been used in single-step GBLUP to predict within-population GEBVs (Wang *et al.* 2012). Single-step GBLUP has the advantage of including all information on genotyped as well as non-genotyped individuals. This is important because the use of only the genotyped animals may introduce bias in the GEBV prediction (Vitezica *et al.* 2011). An evaluation of applying single-step GBLUP to multipopulation prediction while also including the weighted relationship matrix may lead to further improvement of accuracies and should be evaluated. An even further extension to multitrait prediction using both the single step and GWAS weights methodology would also be of interest. However, for combining multiple traits new strategies on how to combine the weights from the different traits will need to be developed.

Differences in SNP effects can be accommodated in genomic selection by applying the widely described Bayesian methods. GBLUP has the advantage of being a more straightforward methodology (in terms of statistical complexity and computational demand) in comparison with Bayesian methods. In addition, the GBLUP methodology has the advantage to fit directly in the existing routines for estimating breeding values and their accuracies that are applied in current breeding programmes. Our approach made it possible to include genotyped and phenotyped individuals from multiple populations, and it emphasized the similarities between the populations. In addition, the method for computing weights in the multipopulation approach addressed two important differences between groups that can affect the accuracy of genomic predictions: (i) allele frequency and (ii) genetic architecture.

References

Aulchenko Y.S., Ripke S., Isaacs A., van Duijn C.M. (2007) GenABEL: an R library for genome-wide association analysis. *Bioinformatics*, **23**, 1294–1296.

Browning B.L., Browning S.R. (2013) Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, **194**, 459–471.

Chen L., Schenkel F., Vinsky M., Crews D.H., Li C. (2013) Accuracy of predicting genomic breeding values for residual feed intake in Angus and Charolais beef cattle. *J. Anim. Sci.*, **91**, 4669–4678.

Chen L., Li C., Miller S., Schenkel F. (2014) Multi-population genomic prediction using a multi-task Bayesian learning model. *BMC Genet.*, **15**, 53.

Falconer D.S., Mackay T.F.C. (1996) Introduction to Quantitative Genetics. 4th edn. Longmans Green, Harlow, Essex, UK.

Gilmour A.R., Gogel B.J., Cullis B.R., Thompson R. (2009) ASReml User Guide Release 3.0.

Goddard M. (2009) Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*, **136**, 245–257.

Harris B.L., Johnson D.L. (2010) Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. *J. Dairy Sci.*, **93**, 1243–1252.

Hayes B., Goddard M.E. (2001) The distribution of the effects of genes affecting quantitative traits in livestock. *Genet. Sel. Evol.*, **33**, 209–229.

Hidalgo A.M., Bastiaansen J.W.M., Lopes M.S., Harlizius B., Groenen M.A.M., de Koning D.J. (2015) Accuracy of predicted genomic breeding values in purebred and crossbred pigs. *G3*, **5**(8), 1575–1583.

Legarra A., Baloche G., Barillet F., Astruc J.M., Soulas C., Aguerre X., Arrese F., Mintegi L., Lasarte M., Maeztu F., Beltrán de Heredia I., Ugarte E. (2014) Within- and across-breed genomic predictions and genomic relationships for Western Pyrenees dairy sheep breeds Latxa, Manech, and Basco-Béarnaise. *J. Dairy Sci.*, **97**, 1–13.

Liu Z., Seefried F.R., Reinhardt F., Rensing S., Thaller G., Reents R. (2011) Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. *Genet. Sel. Evol.*, **43**, 19.

Olson K.M., VanRaden P.M., Tooker M.E. (2012) Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. *J. Dairy Sci.*, **95**, 5378–5383.

Pszczola M., Strabel T., Mulder H.A., Calus M.P.L. (2012) Reliability of direct genomic values for animals with different relationships within and to the reference population. *J. Dairy Sci.*, **95**, 389–400.

Raven L.A., Cocks B.G., Hayes B.J. (2014) Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle. *BMC Genom.*, **15**, 62.

Simeone R., Misztal I., Aguilar I., Vitezica Z.G. (2012) Evaluation of a multi-line broiler chicken population using a single-step genomic evaluation procedure. *J. Anim. Breed. Genet.*, **129**, 3–10.

Spelman R.J., Ford C.A., McElhinney P., Gregory G.C., Snell R.G. (2002) Characterization of the DGAT1 gene in the New Zealand dairy population. *J. Dairy Sci.*, **85**, 3514–3517.

Stranger B.E., Stahl E.A., Raj T. (2011) Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, **187**, 367–383.

Thaller G., Kramer W., Winter A., Kaupe B., Erhardt G., Fries R. (2004) Effects of DGAT1 variants on milk production traits in Jersey cattle. *J. Anim. Sci.*, **81**, 1911–1918.

Tiezzi F., Maltecca C. (2015) Accounting for trait architecture in genomic predictions of US Holstein cattle using a weighted realized relationship matrix. *Genet. Sel. Evol.*, **1**, 13.

VanRaden P.M. (2008) Efficient methods to compute genomic predictions. *J. Dairy Sci.*, **91**, 4414–4423.

Veroneze R., Bastiaansen J.W.M., Knol E.F., Guimarães S.E.F., Silva F.F., Harlizius B., Lopes M.S., Lopes P.S. (2014) Linkage disequilibrium patterns and persistence of phase in purebred and crossbred pig (Sus scrofa) populations. *BMC Genet.*, **15**, 126.

Veroneze R., Lopes M.S., Hidalgo A.M., Guimarães S.E.F., Silva F.F., Harlizius B., Lopes P.S., Knol E.F., van Arendonk J.A.M., Bastiaansen J.W.M. (2015) Accuracy of genome-enabled prediction exploring purebred and crossbred pig populations. *J. Anim. Sci.*, **93**, 4684–4691.

Vitezica Z.G., Aguilar I., Misztal I., Legarra A. (2011) Bias in genomic predictions for populations under selection. *Genet. Res. (Camb)*, **93**, 357–366.

Wang H., Misztal I., Aguilar I., Legarra A., Muir W.M. (2012) Genome-wide association mapping including phenotypes from relatives without genotypes. *Genet. Res. (Camb)*, **94**, 73–83.

Zhang Z., Liu J., Ding X., Bijma P., de Koning D.J., Zhang Q. (2010) Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PLoS One*, **5**, 1–8.

Zhang Z., Ober U., Erbe M., Zhang H., Gao N., He J., Li J., Simianer H. (2014) Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. *PLoS One*, **9**, e93017.