

# Crossbred evaluations using single-step genomic BLUP and algorithm for proven and young with different sources of data.



# Crossbred evaluations using single-step genomic BLUP and algorithm for proven and young with different sources of data[1]

**Ivan Pocrnic,[†,2] Daniela A. L. Lourenco,[†] Ching-Yi Chen,[‡] William O. Herring,[‡] and Ignacy Misztal[†]**

[†]Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602; and [‡]Genus PIC, Hendersonville, TN 37075

**ABSTRACT:** Genomic selection (**GS**) is routinely applied to many purebreds and lines of farm species. However, this method can be extended to predictions across purebreds as well as for crossbreds. This is useful for swine and poultry, for which selection in nucleus herds is typically performed on purebred animals, whereas the commercial product is the crossbred animal. Single-step genomic BLUP (**ssGBLUP**) is a widely applied method that can explore the recently developed algorithm for proven and young (**APY**). The APY allows for greater efficiency in computing BLUP solutions by exploiting the theory of limited dimensionality of genomic information and chromosome segments (**Me**). This study investigates the predictivity as a proxy for accuracy across and within 2 purebred pig lines and their crosses, under the application of APY in ssGBLUP setup, and different levels of Me overlapping across populations. The data consisted of approximately 210k phenotypic records for 2 traits (T1 and T2) with moderate heritability. Genotypes for 43k SNP markers were available for approximately 46k animals, from which 26k and 16k belong to 2 pure lines (L1 and L2), and approximately 4k are crossbreds. The complete pedigree had more than 720k animals. Different multivariate ssGBLUP models were applied, either with the regular or APY inverse of the genomic relationship matrix (**G**). The models included a standard bivariate animal model with 3 lines evaluated as 1 joint line, and for each trait individually, a 3-trait animal model with each line treated as a separate trait. Both models provided the same predictivity across and within the lines. Using either of the pure lines data as a training set resulted in a similar predictivity for the crossbreed animals (0.18 to 0.21). Across-line predictive ability was limited to less than half of the maximum predictivity for each line (L1T1 0.33, L1T2 0.25, L2T1 0.35, L2T2 0.36). For crossbred predictions, APY performed equivalently to regular **G** inverse when the number of core animals was equal to the number of eigenvalues explaining between 98% and 99% of the variance of **G** (4k to 8k) including all lines. Predictivity across the lines is achievable because of the shared Me between them. The number of those shared segments can be obtained via eigenvalue decomposition of genomic information available for each line.

**Key words:** across-breed prediction, chromosome segments, genomic selection, multibreed evaluation



[1]This research was partly supported by Genus PIC (Hendersonville, TN). Editing by Taylor M. McWhorter and Bruno Valente is gratefully acknowledged.

[2]Corresponding author: ipocrnic@uga.edu

Received November 19, 2018.
Accepted January 28, 2019.

J. Anim. Sci. 2019.97:1513–1522
doi: 10.1093/jas/skz042

## INTRODUCTION

Genomic selection (**GS**) is routinely applied within many purebreds and lines of livestock species. In certain cases, joint evaluations are desirable, e.g., when GS is mainly on purebreds, but the commercial production is based on crossbreds, a joint analysis allows evaluating the impact of purebred selection on crossbred performance (Dekkers, 2007). Breeds/lines originating from common ancestors may share chromosome segments (**Me**) or causative variants; therefore, interbreed prediction may be possible if those can be identified and their allele substitution effects and frequencies are similar, which may be more likely for recently separated breeds/lines. However, predictions may be poor if dominance or epistasis effects cause substitution effects to differ (Esfandyari et al., 2015). The GS relies on linkage disequilibrium (**LD**) between SNP and causative variants, and on the ability to accurately estimate Me effects (Meuwissen et al., 2001), which requires large number of phenotypes. The expected number of Me for a randomly mated population was given by Stam (1980) as 4NeL, where Ne is the effective population size and L is genome length (Morgan). Me seems to be the key parameter of the algorithm for proven and young (**APY**) (Misztal et al., 2014), which reduces the computational cost of the inversion of genomic relationship matrix (**G**) by shrinking the dimensionality of genomic information (Misztal, 2016). For large populations, such dimensionality was close to 4NeL and corresponded to the number of eigenvalues explaining 98% variability of **G** (Pocrnic et al., 2016a, 2016b). The same study showed that Me varies from about 4k for pigs and chickens to over 10k for cattle. The main purpose of this study was to examine predictivity across and within 2 recently separated pig lines and their crosses using the Me concepts via APY, and secondly, to determine the degree of overlap across the Me in the populations.

## MATERIALS AND METHODS

Animal Care and Use Committee approval was not needed as data were obtained from preexisting databases.

### *Data*

Data were provided by PIC (a Genus company, Hendersonville, TN) for 2 traits (**T1** and **T2**) measured on 2 purebred terminal sire pig lines (**L1** and **L2**) and their F1 crosses (**C**). Both traits were moderately heritable, with heritability estimates 0.28 and 0.35, respectively, and genetic correlation of 0.27. The number of recorded phenotypes were 211,987 for T1 and 209,260 for T2. Specifically, L1 had 181,030 phenotypes for T1 and 178,796 for T2, consequently, L2 had 25,318 phenotypes for T1 and 25,028 for T2, and C had 5,639 phenotypes for T1 and 5,436 for T2. Genomic information consisted of genotypes for 43,456 SNP markers from 46,488 (38,535) animals (with phenotypes), representing 26,543 (22,812) for L1, 15,976 (13,166) for L2, and 3,969 (2,557) for C. Initial pedigree consisted of 727,303 animals, with the number of animals reduced depending on the amount of data available.

### *Analyses and Computations*

Four different scenarios were considered: 1) using phenotypic records for L1, L2, and C jointly, 2) L1 and L2 jointly, 3) only L1, and 4) only L2. Same genomic information, including L1, L2, and C, was used across scenarios. Data together with pedigree were processed with the renumbering software (RENUMF90) that is part of the BLUPF90 family of programs (Misztal et al., 2018). Pedigree was included for all animals with phenotypes or genotypes and up to 3 generations of their ancestors. After reduction, pedigree data consisted from only 62k individuals for scenario 4 to approximately 220k individuals for scenario 1. Predictions were computed using single-step genomic BLUP (**ssGBLUP**), implemented with the BLUP90IOD2 software (Tsuruta et al., 2001) either with the standard (direct) inverse of **G** or the APY inverse (Misztal et al., 2014). The initial **G** ($G_0$) was constructed as in VanRaden (2008):

$$\mathbf{G}_0 = \mathbf{MM}'/2\Sigma p_j(1-p_j),$$

where **M** is a matrix of allele content centered for allele frequencies and $p_j$ is the allele frequency for marker j. The allele frequencies were computed directly from the complete genotyped population, i.e., L1, L2, and C jointly.

We tested 2 different statistical models. In the first model, traits were treated as a bivariate variable. All lines were included, but line distinctions

were ignored. The second model was fitted within a trait, but records of different lines were treated as different variables in a multiple-trait formulation.

The first model can be represented as a 2-trait animal model:

$$\mathbf{y}_t=\mathbf{X}_t\mathbf{b}_t+\mathbf{Z}_t\mathbf{u}_t+\mathbf{W}_t\mathbf{c}_t+\mathbf{e}_t,$$

where $\mathbf{y}_t$ is a vector of phenotypes for trait t (t = T1 and T2); **b**, **u**, **c**, and **e** are vectors of fixed effects, additive genetic effects, common litter environment effects, and random residuals, respectively, with **X**, **Z**, and **W** being the assigned incidence matrices. Variance–covariance structures for random effects were

$$\text{Var}(\mathbf{u}) = \begin{bmatrix} \sigma^2_{uT1} & \sigma_{uT1,uT2} \\ \sigma_{uT2,uT1} & \sigma^2_{uT2} \end{bmatrix} \otimes \mathbf{H},$$

$$\text{Var}(\mathbf{c}) = \begin{bmatrix} \sigma^2_{pT1} & 0 \\ 0 & \sigma^2_{pT2} \end{bmatrix} \otimes \mathbf{I},$$

$$\text{Var}(\mathbf{e}) = \begin{bmatrix} \sigma^2_{eT1} & \sigma_{eT1,eT2} \\ \sigma_{eT2,eT1} & \sigma^2_{eT2} \end{bmatrix} \otimes \mathbf{I}.$$

In all cases, diagonal elements correspond to additive genetic, common environmental, and residuals variances, whereas off-diagonal elements represent corresponding covariances between the trait-specific random terms when present. **I** is the identity matrix and **H** is a matrix that combines pedigree and genomic relationships, with its inverse as in Aguilar et al. (2010), i.e.,

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

where $\mathbf{A}^{-1}$ is the inverse of a pedigree-based relationship matrix for all animals included in the analysis and $\boldsymbol{A}_{22}^{-1}$ is the inverse of the pedigree-based relationship matrix for genotyped animals alone ($\boldsymbol{A}_{22}$). The **G** matrix was constructed by blending $0.95\boldsymbol{G}_0$ with $0.05\boldsymbol{A}_{22}$ to avoid singularity problems (VanRaden, 2008) and then tuned for compatibility with $\mathbf{A}_{22}$ using the default options in BLUPF90 family of programs (e.g., Chen et al., 2011; Vitezica et al., 2011).

The second model, as applied within each trait, can be written as a similar 3-trait animal model:

$$\mathbf{y}_l=\mathbf{X}_l\mathbf{b}_l+\mathbf{Z}_l\mathbf{u}_l+\mathbf{W}_l\mathbf{c}_l+\mathbf{e}_l,$$

where $\mathbf{y}_l$ is a vector of phenotypes for line l (l = L1, L2, and C); **b**, **u**, **c**, and **e** are vectors of fixed effects, additive genetic effects, common litter environment effects, and random residuals, respectively, with **X**, **Z**, and **W** being the assigned incidence matrices. Variances were

$$\text{Var}(\mathbf{u}) = \begin{bmatrix} \sigma^2_{uL1} & \sigma_{uL1,uL2} & \sigma_{uL1,uC} \\ \sigma_{uL2,uL1} & \sigma^2_{uL2} & \sigma_{uL2,uC} \\ \sigma_{uC,uL1} & \sigma_{uC,uL2} & \sigma^2_{uC} \end{bmatrix} \otimes \mathbf{H},$$

$$\text{Var}(\mathbf{c}) = \begin{bmatrix} \sigma^2_{pL1} & 0 & 0 \\ 0 & \sigma^2_{pL2} & 0 \\ 0 & 0 & \sigma^2_{pC} \end{bmatrix} \otimes \mathbf{I},$$

$$\text{Var}(\mathbf{e}) = \begin{bmatrix} \sigma^2_{eL1} & 0 & 0 \\ 0 & \sigma^2_{eL2} & 0 \\ 0 & 0 & \sigma^2_{eC} \end{bmatrix} \otimes \mathbf{I}.$$

Variances and matrices **H** and **I** are defined similarly to analogous parameters of the first model. For both models, variance components were estimated via Bayesian inference. We applied a Gibbs sampler algorithm as implemented in GIBBS2F90 program (Misztal et al., 2018).

To investigate independence between the lines and to find the number of core animals needed for the APY inverse, we applied singular value decomposition (**SVD**) to the matrix **M** using subroutine DGESVD in LAPACK (Anderson et al., 1999). This is equivalent to eigenvalue decomposition of **G**, but with a lower cost. The SVD of matrix **M** is **M** = **UDV**′, where **D** is a diagonal matrix of singular values that correspond to the square root of the nonzero eigenvalues of **M**′**M** and **MM**′. The columns of **U** are left singular vectors (**U**′**U** = **UU**′ = **I**), and the columns of **V** are right singular vectors (**V**′**V** = **I**). They correspond to eigenvectors of **MM**′ and **M**′**M**, respectively. Therefore, $\mathbf{M}'\mathbf{M}=\mathbf{VD}'\mathbf{U}'\mathbf{UDV}'=\mathbf{VD}^2\mathbf{V}'$ and $(\mathbf{M}'\mathbf{M})\mathbf{V}=\mathbf{VD}^2$, where $\boldsymbol{D}^2$ is a diagonal matrix of eigenvalues of **M**′**M** (squares of singular values of matrix **M**) and the columns of **V** are eigenvectors of **M**′**M**. Same follows for **MM**′, as $\mathbf{MM}'=\mathbf{UD}^2\mathbf{U}'$. Based on the proof, singular values were sorted in descending order, squared to get eigenvalues, and counted until they explained 50%, 80%, 90%, 95%, 98%, and 99% of the variance in **G** (Pocrnic et al., 2016a). This was applied for 5 groups of genotyped animals: L1, L2, and C each separately; L1 and L2 jointly; and L1, L2, and, C jointly.

Finally, the number of largest eigenvalues that explained 90%, 98%, or 99% of variation from the combined group (L1, L2, and C jointly) was used as the number of core animals in APY. For this purpose, core animals for APY were randomly selected

from L1 solely, L2 solely, or from all the genotyped animals (L1, L2, and C). When the core animals were randomly selected from all the genotypes (L1, L2, and C), contribution of each line to the core subset was following contribution of each line to the total number of genotypes, i.e., approximately 57% from L1, 34% from L2, and 9% from C. Analyses were repeated using the first model, with the only difference being the type of inverse of **G**.

### *Validation*

Validation population consisted of genotyped animals born in 2017 that had their phenotypes removed from the analysis. Accuracy was defined as the correlation between genomic EBV (**GEBV**) and phenotypes adjusted for the fixed effects in the model (y*), which is similar to the method proposed by Legarra et al. (2008) if correlations would be divided by the square root of heritability. Potential GEBV inflation and bias was measured by the regression:

$$y^* = b_0 + b_1 GEBV + e,$$

where a regression coefficient ($b_1$) smaller or greater than 1 indicates GEBV inflation or deflation, respectively, and intercept ($b_0$) indicates bias. In addition, correlations between GEBV obtained by APY inverse and direct inverse of **G** were calculated. Validation was separate for each group of animals (L1, L2, and C), resulting in up to 2,770 animals from L1, 2,623 from L2, and 2,557 from C.

## RESULTS AND DISCUSSION

Correlations between GEBV and y* for the first model are shown in Table 1. Compared with using all available phenotypes (L1, L2, and C), removing crossbred phenotypes did not reduce predictive ability for young animals from any group (L1, L2, and C) for both traits evaluated. This was surprising, as in several studies, impact of adding crossbred phenotypes was mostly positive (e.g., Bijma and van Arendonk, 1998; Bijma et al., 2001; Lutaaya et al., 2002). Possible reason could be relatively small number of crossbred phenotypes in comparison to phenotypes coming from L1 and L2. In all the analyses, complete genotyped data (L1, L2, and C) was used as previous studies indicated that inclusion of crossbreed genotypes is beneficial for prediction accuracies (Lourenco et al., 2016; Iversen et al., 2017; Sewel et al., 2018). When phenotypes for only L1 were available, predictive ability for L2 and C was affected, whereas for L1 it stayed the same as when using all phenotypes. In the scenario where phenotypes only from L2 were used to fit the model, predictive ability for L1 and C was affected, whereas for L2 it stayed the same.

**Table 1.** Correlations between genomic EBV and phenotypes adjusted for fixed effects, for different groups of validation animals (purebred animals L1 and L2, and their crosses C) with a different source of phenotypes available, shown for traits 1 (T1) and 2 (T2), under the first model (2-trait animal model without the distinction between the lines)

| | T1 | | | T2 | | |
|---|---|---|---|---|---|---|
| Phenotypes[1] | L1 | L2 | C | L1 | L2 | C |
| L1 + L2 + C | 0.33 | 0.34 | 0.26 | 0.24 | 0.36 | 0.25 |
| L1 + L2 | 0.33 | 0.34 | 0.26 | 0.24 | 0.36 | 0.25 |
| L1 | 0.33 | 0.15 | 0.19 | 0.25 | 0.14 | 0.19 |
| L2 | 0.18 | 0.35 | 0.21 | 0.11 | 0.36 | 0.18 |

[1]Phenotypes coming from L1, L2, and C jointly, L1 and L2 jointly, L1 solely, or L2 solely.

For the second model and T1, heritability estimates (posterior standard deviations) were 0.31 (0.01) for L1, 0.26 (0.02) for L2, and 0.33 (0.05) for C. When applied to T2, heritability estimates (posterior standard deviations) were 0.39 (0.01) for L1, 0.44 (0.02) for L2, and 0.30 (0.04) for C. Therefore, the second model provided comparable, but slightly greater heritability estimates relative to the first model. Typically, moderate to high genetic correlation between purebreds and crossbred is needed to achieve good predictions for the latter group. When the second model was applied to T1, genetic correlations (posterior standard deviations) were 0.51 (0.13) between L1 and C, 0.90 (0.04) between L2 and C, and 0.29 (0.16) between L1 and L2. T2 showed a different pattern of genetic correlations: 0.92 (0.03), 0.57 (0.09), and 0.80 (0.07), respectively, for the same pairs of groups. Genetic correlations between purebreds and crossbreds in pigs have been reported in several studies. The correlations for lifetime daily gain were 0.99 with one purebred line and 0.62 with the second purebred line (Lutaaya et al., 2001). The same correlations for backfat were 0.32 and 0.70, respectively. Lourenco et al. (2016) reported genetic correlations for the number of stillborn and litter size below 0.8, Xiang et al. (2016) reported values from 0.57 to 0.79 for the total number of piglets born, and Tusell et al. (2016) reported values ranging from 0.69 to 0.91 for several traits (i.e., growth rate, feed conversion ratio, lean meat, pH measured in the *longissimus dorsi*, drip loss, and intramuscular fat). Genetic correlations between pure breeds are less

commonly estimated and reported, e.g., between Landrace and Yorkshire breeds correlations were reported to be from 0.2 to 0.3 (Xiang et al., 2017).

Predictive abilities for the second model are shown in Tables 2 and 3 for the biological traits T1 and T2, respectively. When phenotypes from all lines were used, predictive abilities for validation animals were comparable to the ones from the first model. Specifically, for crossbred predictions, correlations between GEBV and y* were 0.26 for T1 and 0.25 for T2 using the first model, and 0.24 for T1 and 0.22 for T2 using the second model. For the L1 and L2, it was in the opposite direction, and correlations between GEBV and y* were slightly greater when the second model was used. These small variations are probably due to the re-estimated variance components used in the second model. When only L1 or L2 phenotypes were considered in the training data set, correlations between GEBV and y* using the second model were almost identical to those from the first model. Using phenotypes for either of the pure lines solely resulted in similar predictive abilities for the crossbred animals. When this scenario was applied to predictions on the other pure line, correlations between GEBV and y* were roughly halved both for L1 to L2 and L2 to L1 predictions. Both models performed equally well for all different sources of available phenotypes for fitting. In the terms of GEBV inflation (Table 4), models were comparable as well, and the inflation was smallest when all available phenotypes were included. Inflation was more extreme for T2, and especially for L1. Presence of actual inflation is questionable due to the method used (regression), i.e., sensitivity to adjustments in y*, specific population structure (combined lines), and specific selection. This could potentially be solved by use of unknown parent groups in the models or metafounders approach (Legarra et al., 2015). Statistical models used in this study are arguably simple, and more complex models that consider alleles breed of origin and breed-specific relationship matrix are available (e.g., Ibanez-Escriche et al., 2009; Christensen et al., 2014). Nevertheless, the application of these methods might be cumbersome in the APY setup. The breed-specific model was tested with mixed success; e.g., several studies (Lopes et al., 2017; Ibanez-Escriche et al., 2009; Lourenco et al., 2016) obtained similar prediction accuracies as with the simple model, whereas Xiang et al. (2016) found better prediction with the more complex model. Model used by Xiang et al. (2016) was based on the construction of separate, breed-specific pedigree and genomic relationship matrices via 3-step process: 1) accounting for breed of origin-specific genetic effects for crossbreds, 2) construction of breed-specific partial relationship matrices for each breed of origin genetic effects, and 3) combining pedigree-based and adjusted marker-based partial relationship matrices to a combined partial relationship matrix. Additional implementation challenge for these models is the assumption of known alleles for the breed of origin, and this assumption can be relaxed by approximations (Vandenplas et al., 2016).

Analyses of eigenvalues are useful when dealing with large multivariate genomic data (Cavalli-Sforza et al., 2003; Patterson et al., 2006; Solberg et al., 2009; Macciotta et al., 2010). In this work, SVD was applied to **M** as equivalence to eigenvalue decomposition of **G** to assess the independence of the lines. The number of largest eigenvalues that explain certain percentages of variance in **G** was calculated from different combinations of lines (Table 5). If we assume there is complete independence of Me across the lines, i.e., no shared segments across them, we would expect the resulting number of eigenvalues obtained from a combined

**Table 2.** Correlations between genomic EBV and phenotypes adjusted for fixed effects, for different groups of validation animals (purebred animals L1 and L2, and their crosses C) with a different source of phenotypes available, shown for trait 1 (T1) when that trait was separated into 3 traits based on the line of the animals (second model)

| Phenotypes[1] | L1 | L2 | C |
|---|---|---|---|
| L1 + L2 + C | 0.33 | 0.35 | 0.24 |
| L1 | 0.33 | 0.15 | 0.19 |
| L2 | 0.18 | 0.35 | 0.20 |

[1]Phenotypes coming from L1, L2, and C jointly, L1 solely, or L2 solely.

**Table 3.** Correlations between genomic EBV and phenotypes adjusted for fixed effects, for different groups of validation animals (purebred animals L1 and L2, and their crosses C) with a different source of phenotypes available, shown for trait 2 (T2) when that trait was separated into 3 traits based on the line of the animals (second model)

| Phenotypes[1] | L1 | L2 | C |
|---|---|---|---|
| L1 + L2 + C | 0.25 | 0.38 | 0.22 |
| L1 | 0.25 | 0.15 | 0.20 |
| L2 | 0.12 | 0.38 | 0.19 |

[1]Phenotypes coming from L1, L2, and C jointly, L1 solely, or L2 solely.

**Table 4.** Regression coefficients (b1) of adjusted phenotypes on genomic EBV, for different groups of validation animals (purebred animals L1 and L2, and their crosses C) with a different source of phenotypes available, shown for traits 1 (T1) and 2 (T2), under the first model (M1; 2-trait animal model without the distinction between the lines) and second model (M2; when that trait was separated into 3 traits based on the line of the animals)

| | M1–T1 | | | M1–T2 | | |
|---|---|---|---|---|---|---|
| Phenotypes[1] | L1 | L2 | C | L1 | L2 | C |
| L1 + L2 + C | 0.82 | 0.99 | 0.78 | 0.48 | 0.97 | 0.56 |
| L1 + L2 | 0.82 | 0.99 | 0.78 | 0.48 | 0.97 | 0.56 |
| L1 | 0.81 | 0.59 | 0.60 | 0.50 | 0.59 | 0.52 |
| L2 | 0.64 | 1.03 | 0.76 | 0.32 | 0.93 | 0.48 |
| | M2–T1 | | | M2–T2 | | |
| | L1 | L2 | C | L1 | L2 | C |
| L1 + L2 + C | 0.79 | 1.06 | 0.77 | 0.49 | 0.94 | 0.79 |
| L1 | 0.79 | 2.08 | 1.09 | 0.51 | 0.98 | 0.80 |
| L2 | 2.11 | 1.07 | 0.78 | 0.31 | 0.92 | 0.90 |

[1]Phenotypes coming from L1, L2 and C jointly, L1 and L2 jointly, L1 solely, or L2 solely

**Table 5.** Numbers of largest eigenvalues (Eig) explaining 50%, 80%, 90%, 95%, 98%, and 99% of the variance in the genomic relationship matrix with a different source of genotypes available

| Genotypes[1] | Number genotyped | Eig50 | Eig80 | Eig90 | Eig95 | Eig98 | Eig99 |
|---|---|---|---|---|---|---|---|
| L1 | 26,543 | 119 | 531 | 1,068 | 1,944 | 3,888 | 5,957 |
| L2 | 15,976 | 125 | 602 | 1,209 | 2,112 | 3,884 | 5,601 |
| L1 + L2 | 42,519 | 126 | 728 | 1,528 | 2,763 | 5,381 | 8,137 |
| C | 3,969 | 105 | 479 | 864 | 1,315 | 1,968 | 2,459 |
| L1 + L2 + C | 46,488 | 130 | 735 | 1,533 | 2,759 | 5,368 | 8,141 |

[1]Purebred animals L1 and L2, and their crosses C.

population (L1 + L2 + C) would equal the sum of eigenvalues obtained from individual lines L1, L2, and C. The number of eigenvalues explaining 99% of the variance of **G** (**Eig99**) was 5,957 for L1, 5,601 for L2, and 8,137 when L1 and L2 were evaluated together. This indicates the lines are not completely independent of each other, and the genetic connectedness probably arose in the historical generations through coalescent ancestry before artificial selection started. One of the additional metrics used to assess connectedness between the lines is Wright's $F_{ST}$ statistic (Wright, 1965). McVean (2009) showed that the $F_{ST}$ for the 2 populations can be obtained as the fraction of total variance explained by the first principal component (**PC**). In this study, first PC explained 14.83% of the variance of **G** based on L1 and L2, and therefore, $F_{ST}$ for the L1 and L2 was 0.15, which confirms certain degree of connectedness and nonindependence between the lines. When 3,969 crossbred genotypes (Eig99 = 2,459) were added to the genomic data, the number of eigenvalues for the combined population (L1 + L2 + C) was 8,141, which is almost the same number obtained by combining L1 and L2 populations (L1 + L2). This indicates that the genomic information added from crossbred animals was already accounted for by the L1 and L2 genotypes. The number of eigenvalues for lower percentages of variance explained, e.g., 50%, is similar across the lines and their combinations. This suggests that a limited number of segments, with the same origin, could be shared across the lines. Eigenvalues are one of the indicators for the number of Me in a population, and the largest eigenvalues cluster a number of segments across the genome, comparable to developments described in Patterson et al. (2006). This can be interconnected with the results of a PC analysis of **G**, which are presented in Figure 1. Projections of genomic relationships from **G** into the first and second PC are clearly indicating stratification of the genotyped population. PC showed separation between the lines as expected, with crossbred population positioned centrally between the 2 purebred populations.

Eigenvalue decomposition is currently used to define the dimensionality of genomic data and select the number of core animals for APY. This was successfully demonstrated for several purebred

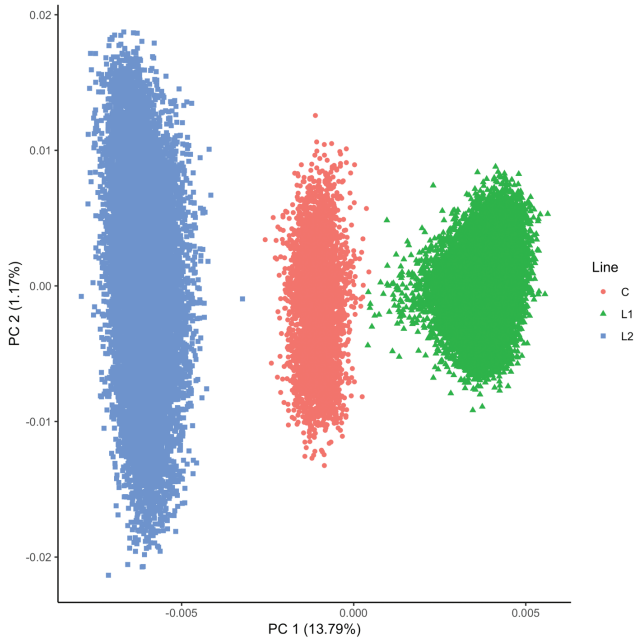


**Figure 1.** Projection of genomic relationships into first 2 principal components (PC), showing purebred animals (L1 and L2), and their crosses (C). The percentage of variance explained by each PC is shown in parentheses.

populations in different species (e.g., Pocrnic et al., 2016b); however, the application of this concept for crossbred/multibreed contexts was unclear. Bradford et al. (2017) found, by simulating a purebred population, that any core definition is robust in populations with complete pedigree; otherwise, selecting core animals randomly across multiple generations gives desirable accuracies. This is attributed to a random sample that increases the likelihood of all generations of genotyped animals being represented in the core group. Since the theory behind APY is based on the utilization of the number of effective Me in the population and selection is working through eigenvalues that cluster those segments, core definition in a mixed population should include animals from all pure lines for better representation of various segments in the population. In a recent simulation study based on a 3-way crossbreeding system with genotypes available for 3 different breeds and their F1 and F2 crosses, Vandenplas et al. (2018) showed that APY was as accurate as the direct inverse in ssGBLUP, when the core animals were randomly selected from all different breed compositions available and had a size between the number of eigenvalues explaining 98% (**Eig98**) and 99% of the variance of **G**.

One of the metrics used to assess approaches regarding the composition of the core in crossbred/multibreed populations is across line predictivity using APY with different core configurations. These results are shown in Table 6. When core animals were randomly selected from the L1 only, predictive ability was the same for L1 validation animals as when the regular **G** inverse was used, whereas for the L2 and C validation animals, it was lower especially when the number of core animals was less than the number of eigenvalues explaining 98% of the variance of **G**. When the core group was based on L2 animals only, predictive ability was greater for the L2 validation animals. Line 2-based core resulted in better predictions of L1 and crossbred when the number of core animals was at Eig98. This could be reflecting different selection pressure on these lines. When the core animals were randomly selected from all genotyped animals (L1 + L2 + C) and their number was either Eig98 or Eig99, predictive ability for all groups of animals was the same as the scenario with the regular **G** inverse. Similar results were obtained in a simulation study by Vandenplas et al. (2018). These findings can be confirmed by looking at the correlations between GEBV obtained by APY and regular **G** inverse, as shown in Table 7. Correlations between GEBV obtained by APY and regular **G** inverse were greater than 0.99 for both traits and all groups of validation animals when the core animals were selected in beforementioned way. When the core group was selected from L1 or L2 animals only, correlations between GEBV obtained by APY and regular **G** inverse were greater than 0.99 only for the validation animals coming from the line that core group was made of, and number in core being either Eig98 or Eig99. Core at Eig98 and based on L2 solely produced greater correlations between GEBV obtained by APY and regular **G** inverse for L1 and C validation animals (0.96 to 0.97) in comparison to core based on L1 solely (0.91 to 0.95).

Results from this study could be interpreted from the perspective of the overlapping number of eigenvalues between the lines and epistatic interactions. Epistasis interactions might be one of the causes of breed-specific SNP effects, with other causes being dominance, a difference in LD between SNP and QTL, or different QTL allele frequencies across breeds (Ibanez-Escriche et al., 2009; Esfandyari et al., 2015). When allele substitution effects between the breeds are different and only one line's information is used, (G)EBV may not accurately predict crossbred performances (Lopes et al., 2017). Although many causal variants have been identified, they seem to add little predictive power for multibreed predictions. One good example is the DGAT1 variant, as it is highly

**Table 6.** Correlations between genomic EBV and phenotypes adjusted for fixed effects, for different groups of validation animals (purebred animals L1 and L2, and their crosses C) shown for traits 1 and 2 (T1 and T2) under the first model, obtained by the algorithm for proven and young (APY) inverse of genomic relationship matrix (**G**) where the number of core animals was based on the largest eigenvalues (Eig) explaining 90%, 98%, and 99% variance of **G** and was randomly sampled from different groups of genotyped animals (L1, L2, or L1 + L2 + C)

| | T1 | | | T2 | | |
|---|---|---|---|---|---|---|
| Core | L1 | L2 | C | L1 | L2 | C |
| L1_Eig90 | 0.33 | 0.22 | 0.22 | 0.23 | 0.27 | 0.21 |
| L1_Eig98 | 0.33 | 0.29 | 0.24 | 0.23 | 0.31 | 0.23 |
| L1_Eig99 | 0.33 | 0.30 | 0.25 | 0.24 | 0.32 | 0.23 |
| L2_Eig90 | 0.31 | 0.32 | 0.24 | 0.20 | 0.35 | 0.20 |
| L2_Eig98 | 0.32 | 0.34 | 0.25 | 0.23 | 0.36 | 0.24 |
| L2_Eig99 | 0.33 | 0.34 | 0.26 | 0.23 | 0.36 | 0.24 |
| L1L2C_Eig90 | 0.32 | 0.30 | 0.24 | 0.23 | 0.33 | 0.22 |
| L1L2C_Eig98 | 0.33 | 0.33 | 0.25 | 0.24 | 0.36 | 0.24 |
| L1L2C_Eig99 | 0.33 | 0.34 | 0.26 | 0.24 | 0.36 | 0.24 |

**Table 7.** Correlations between genomic EBV obtained by the algorithm for proven and young (APY) inverse and regular (direct) inverse of genomic relationship matrix (**G**), for different groups of validation animals (purebred animals L1 and L2, and their crosses C) shown for traits 1 and 2 (T1 and T2) under the first model

| | T1 | | | T2 | | |
|---|---|---|---|---|---|---|
| Core | L1 | L2 | C | L1 | L2 | C |
| L1_Eig90 | 0.97 | 0.77 | 0.87 | 0.96 | 0.81 | 0.88 |
| L1_Eig98 | 0.99 | 0.91 | 0.95 | 0.99 | 0.92 | 0.95 |
| L1_Eig99 | 0.99 | 0.93 | 0.97 | 0.99 | 0.94 | 0.97 |
| L2_Eig90 | 0.89 | 0.93 | 0.89 | 0.89 | 0.94 | 0.90 |
| L2_Eig98 | 0.97 | 0.99 | 0.97 | 0.96 | 0.99 | 0.97 |
| L2_Eig99 | 0.98 | 0.99 | 0.98 | 0.97 | 0.99 | 0.98 |
| L1L2C_Eig90 | 0.96 | 0.92 | 0.94 | 0.95 | 0.93 | 0.94 |
| L1L2C_Eig98 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| L1L2C_Eig99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |

For the APY inverse the number of core animals was based on the largest eigenvalues (Eig) explaining 90%, 98%, and 99% variance of **G** and was randomly sampled from different groups of genotyped animals (L1, L2, or L1 + L2 + C).

associated with milk production traits across bovine breeds, but with different allele frequencies and substitution effects among them (Spelman et al., 2002; Thaller et al., 2003). As epistatic interactions can result in changes of substitution effects over generations, predictive ability may be more likely across breeds or lines that show recent divergence. Examples include breeds independently selected from multiple farms, or even within the same farm for different breeding goals. Since the prediction across lines was possible in the dataset used, it is implied that these populations share certain overlapping segments and have similar allele substitution effects, suggesting similar genetic backgrounds. They may have been connected or may have influenced one another in recent history. Additionally, individuals sharing identical chromosome segments by coancestry present greater phenotypic similarity than a random sample of individuals from a given population (Thompson, 2013).

## CONCLUSIONS

The APY can be applied to crossbred datasets if the core subset is selected randomly with consideration of all available breeds or lines. Limited prediction between the lines is possible, due to the shared Me between them. Increased predictivity requires the availability of phenotypes to accurately estimate effects for nonoverlapping segments. The number of overlapping segments can possibly be derived from the difference between eigenvalue decomposition of all lines/breeds separately and jointly.

## 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

Anderson, E., Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, et al.1999. LAPACK users' guide. 3rd ed. Society for Industrial and Applied Mathematics, Philadelphia, PA. doi:10.1137/1.9780898719604

Bijma, P., and J. A. M. van Arendonk. 1998. Maximizing genetic gain for the sire line of a crossbreeding scheme utilizing both purebred and crossbred information. Anim. Sci. 66:529–542. doi:10.1017/S135772980000970X

Bijma, P., J. A. Woolliams, and J. A. M. Van Arendonk. 2001. Genetic gain of pure line selection and combined crossbred purebred selection with constrained inbreeding. Anim. Sci. 72:225–232. doi:10.1017/S1357729800055715

Bradford, H. L., I. Pocrnić, B. O. Fragomeni, D. A. L. Lourenco, and I. Misztal. 2017. Selection of core animals in the algorithm for proven and young using a simulation model. J. Anim. Breed. Genet. 134:545–552. doi:10.1111/jbg.12276

Cavalli-Sforza, L. L., and M. W. Feldman. 2003. The application of molecular genetic approaches to the study

of human evolution. Nat. Genet. 33 (Suppl):266–275. doi:10.1038/ng1113

Chen, C. Y., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2011. Effect of different genomic relationship matrices on accuracy and scale. J. Anim. Sci. 89:2673–2679. doi:10.2527/jas.2010-3555

Christensen, O. F., P. Madsen, B. Nielsen, and G. Su. 2014. Genomic evaluation of both purebred and crossbred performances. Genet. Sel. Evol. 46:23. doi:10.1186/1297-9686-46-23

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

Esfandyari, H., A. C. Sørensen, and P. Bijma. 2015. Maximizing crossbred performance through purebred genomic selection. Genet. Sel. Evol. 47:16. doi:10.1186/s12711-015-0099-3

Ibanez-Escriche, N., R. L. Fernando, A. Toosi, and J. C. Dekkers. 2009. Genomic selection of purebreds for crossbred performance. Genet. Sel. Evol. 41:12. doi:10.1186/1297-9686-41-12

Iversen, M. W., Ø. 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

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

Legarra, A., C. Robert-Granié, E. Manfredi, and J. M. Elsen. 2008. Performance of genomic selection in mice. Genetics 180:611–618. doi:10.1534/genetics.108.088575

Lopes, M. S., H. Bovenhuis, A. M. Hidalgo, J. A. M. van Arendonk, E. F. Knol, and J. W. M. Bastiaansen. 2017. Genomic selection for crossbred performance accounting for breed-specific effects. Genet. Sel. Evol. 49:51. doi:10.1186/s12711-017-0328-z

Lourenco, D. A., S. Tsuruta, B. O. Fragomeni, C. Y. Chen, W. O. 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. W. Mabry, T. Short, H. H. Timm, and R. Holzbauer. 2001. Genetic parameter estimates from joint evaluation of purebreds and crossbreds in swine using the crossbred model. J. Anim. Sci. 79:3002–3007. doi:10.2527/2001.79123002x

Lutaaya, E., I. Misztal, J. W. Mabry, T. Short, H. 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

Macciotta, N. P., G. Gaspa, R. Steri, E. L. Nicolazzi, C. Dimauro, C. Pieramati, and A. Cappio-Borlino. 2010. Using eigenvalues as variance priors in the prediction of genomic breeding values by principal component analysis. J. Dairy Sci. 93:2765–2774. doi:10.3168/jds.2009-3029

McVean, G. 2009. A genealogical interpretation of principal components analysis. Plos Genet. 5:e1000686. doi:10.1371/journal.pgen.1000686

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829. http://www.genetics.org/content/157/4/1819.long

Misztal, I. 2016. Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. Genetics 202:401–409. doi:10.1534/genetics.115.182089

Misztal, I., A. Legarra, and I. Aguilar. 2014. 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. A. L. Lourenco, Y. Masuda, I. Aguilar, A. Legarra, and Z. Vitezica. 2018. Manual for BLUPF90 family of programs. http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all7.pdf (Accessed 31 October 2018.)

Patterson, N., A. L. Price, and D. Reich. 2006. Population structure and eigenanalysis. Plos Genet. 2:e190. doi:10.1371/journal.pgen.0020190

Pocrnic, I., D. A. Lourenco, Y. Masuda, A. Legarra, and I. Misztal. 2016a. 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, Y. Masuda, and I. Misztal. 2016b. Dimensionality of genomic information and performance of the algorithm for proven and young for different livestock species. Genet. Sel. Evol. 48:82. doi:10.1186/s12711-016-0261-6

Sewel, A., H. Li, C. Schwab, C. Maltecca, and F. Tiezzi. 2018. On the value of genotyping terminal crossbred pigs for nucleus genomic selection for carcass traits. In: Proc. World Congr. Genet. Appl. Livest. Prod.; Auckland, New Zealand. 11.775.

Solberg, T. R., A. K. Sonesson, J. A. Woolliams, and T. H. E. Meuwissen. 2009. Reducing dimensionality for prediction of genome-wide breeding values. Genet. Sel. Evol. 41:29. doi:10.1186/1297-9686-41-29

Spelman, R. J., C. A. Ford, P. McElhinney, G. C. Gregory, and R. G. Snell. 2002. Characterization of the DGAT1 gene in the New Zealand dairy population. J. Dairy Sci. 85:3514–3517. doi:10.3168/jds.S0022-0302(02)74440-8

Stam, P. 1980. The distribution of the fraction of the genome identical by descent in finite random mating populations. Genet. Res. 35:131–155. doi:10.1017/S0016672300014002

Thaller, G., W. Krämer, A. Winter, B. Kaupe, G. Erhardt, and R. Fries. 2003. Effects of DGAT1 variants on milk production traits in German cattle breeds. J. Anim. Sci. 81:1911–1918. doi:10.2527/2003.8181911x

Thompson, E. A. 2013. Identity by descent: variation in meiosis, across genomes, and in populations. Genetics 194:301–326. doi:10.1534/genetics.112.14882

Tsuruta, S., I. Misztal, and I. Strandén. 2001. Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed-model equations in animal breeding applications. J. Anim. Sci. 79:1166–1172. doi:10.2527/2001.7951166x

Tusell, L., H. Gilbert, J. Riquet, M. J. Mercat, A. Legarra, and C. Larzul. 2016. Pedigree and genomic evaluation of pigs using a terminal-cross model. Genet. Sel. Evol. 48:32. doi:10.1186/s12711-016-0211-3

Vandenplas, J., M. P. Calus, C. A. Sevillano, J. J. Windig, and J. W. Bastiaansen. 2016. Assigning breed origin to alleles

in crossbred animals. Genet. Sel. Evol. 48:61. doi:10.1186/s12711-016-0240-y

Vandenplas, J., M. P. L. Calus, and J. Ten 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

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

Wright, S. 1965. The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution 19:395–420. doi:10.2307/2406450

Xiang, T., O. F. Christensen, and A. Legarra. 2017. Technical note: genomic evaluation for crossbred performance in a single-step approach with metafounders. J. Anim. Sci. 95:1472–1480. doi:10.2527/jas.2016.1155

Xiang, T., B. Nielsen, G. Su, A. Legarra, and O. F. Christensen. 2016. Application of single-step genomic evaluation for crossbred performance in pig. J. Anim. Sci. 94:936–948. doi:10.2527/jas.2015-9930