Sparse single-step genomic blup in crossbreeding schemes

Vandenplas, J., Calus, M. P. L., & ten Napel, J.

Sparse single-step genomic BLUP in crossbreeding schemes[1]

J. Vandenplas,[*2] M.P.L. Calus, [*] J. ten Napel[*]

[*]Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, P.O. Box 338, 6700 AH Wageningen, the Netherlands

[1]This study was financially supported by the Dutch Ministry of Economic Affairs (TKI-Toeslag; Public-private partnership "Breed4Food" code BO-22.04-011-001-ASG-LR-3) and the Breed4Food partners Cobb Europe, CRV, Hendrix Genetics and Topigs Norsvin.

[2]Corresponding author: jeremie.vandenplas@wur.nl

We declare that we do not have any competing interest in the matter and results covered by this manuscript.

ABSTRACT

The algorithm for Proven and Young animals (**APY**) efficiently computes an approximated inverse of the genomic relationship matrix, by dividing genotyped animals in so-called core and non-core animals. The APY leads to computationally feasible single-step genomic Best Linear Unbiased Prediction (**ssGBLUP**) with a large number of genotyped animals, and was successfully applied to real single breed or line datasets. This study aimed to assess the quality of genomic breeding values (**GEBV**) when using the APY (**$\mathrm{GEBV}_{\mathrm{APY}}$**), in comparison to GEBV when using the directly inverted genomic relationship matrix (**$\mathrm{GEBV}_{\mathrm{DIRECT}}$**), for situations based on crossbreeding schemes, including F1 and F2 crosses, such as the ones for pigs and chickens. Based on simulations of a three-way crossbreeding program, we compared different approximated inverses of a genomic relationship matrix, by varying the size and the composition of the core group. We showed that $\mathrm{GEBV}_{\mathrm{APY}}$ were accurate approximations of $\mathrm{GEBV}_{\mathrm{DIRECT}}$ for multivariate ssGBLUP involving different breeds and their crosses. $\mathrm{GEBV}_{\mathrm{APY}}$ as accurate as $\mathrm{GEBV}_{\mathrm{DIRECT}}$ were obtained when the core groups included animals from different breed compositions, and when the core groups had a size between the numbers of the largest eigenvalues explaining 98% and 99% of the variation in the raw genomic relationship matrix.

Key words: single-step, genomic evaluation, APY

INTRODUCTION

Single-step genomic Best Linear Unbiased Prediction (**ssGBLUP**) is currently the method of choice to predict genomic breeding values in many species (Legarra et al., 2014). The main reason is that ssGBLUP enables simultaneous use of phenotypes from genotyped and non-genotyped animals by combining genomic and pedigree relationship matrices. An inconvenience of ssGBLUP is that the inverse of a dense genomic relationship matrix ($\mathbf{G}$) is required, leading to a soft limit of approximately 100,000 genotyped animals for the currently available computers (Misztal et al., 2014).

Recently, Misztal et al. (2014, 2016) proposed the so-called Algorithm for Proven and Young animals (**APY**) to compute an approximated inverse of $\mathbf{G}$ ($\mathbf{G}_{APY}^{-1}$) for a large number of genotyped animals. The computation of $\mathbf{G}_{APY}^{-1}$ involves the inversion of a genomic relationship submatrix among a limited number of genotyped animals, called core animals, and the recursive computation of other coefficients for non-core animals. The APY was successfully applied on (large) real datasets with animals originating from a single breed or line (Fragomeni et al., 2015; Lourenco et al., 2015; Masuda et al., 2016; Ostersen et al., 2016; Pocrnic et al., 2016b; Strandén et al., 2017). However, several livestock production systems, such as the ones for pigs and chickens, are based on well-structured crossbreeding schemes, generating production animals with a specific breed composition. In these cases, the ssGBLUP may include non-genotyped and genotyped animals from different breeds, as well as their crossbred progeny. Using the APY with such datasets is desirable for implementing ssGBLUP in crossbreeding schemes efficiently.

The aim of this study was to assess the quality of genomic estimated breeding values (**GEBV**) when using $\mathbf{G}_{APY}^{-1}$, in comparison to GEBV when using the direct inversion of $\mathbf{G}$ ($\mathbf{G}_{direct}^{-1}$), for situations based on well-structured crossbreeding schemes that include genotyped animals from a few different breeds and their F1 and F2 crosses. Influence of the selection strategy of the core animals and of the number of core animals, were also investigated. All analyses were based on simulated data.

MATERIALS AND METHODS

Single-step genomic Best Linear Unbiased Prediction

The ssGBLUP method replaces the inverse of the pedigree relationship matrix for all animals ($\mathbf{A}^{-1}$) with the inverse of the combined pedigree-genomic relationship matrix ($\mathbf{H}^{-1}$), defined as (Aguilar et al., 2010; Christensen and Lund, 2010):

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

where $\mathbf{A}_{22}$ is the pedigree relationship matrix for the genotyped animals, $\mathbf{G} = (1 - w)\mathbf{G}_a + w\mathbf{A}_{22}$ with $\mathbf{G}_a$ being a genomic relationship matrix adjusted to be on the same scale as $\mathbf{A}_{22}$, and w being the weight on the pedigree relationship matrix. Several approaches for computing $\mathbf{G}_a$ by adjusting a raw genomic relationship matrix $\mathbf{G}^*$ towards $\mathbf{A}_{22}$ were proposed in the literature (Powell et al., 2010; Vitezica et al., 2011; Christensen, 2012; Lourenco et al., 2016).

Highest computational costs for creating $\mathbf{H}^{-1}$ are the creation and the inversion of the dense matrices $\mathbf{G}$ and $\mathbf{A}_{22}$. Additional computational costs also appear during solving of the mixed model equations due to an increase of non-zero elements in $\mathbf{H}^{-1}$, increasing the number of operations per iteration, e.g., of the preconditioned conjugate gradient used to solve the mixed model equations (Ostersen et al., 2016).

Sparse inversion of G

The matrix $\mathbf{G}$ can be divided into four submatrices as:

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{cc} & \mathbf{G}_{cn} \\ \mathbf{G}'_{cn} & \mathbf{G}_{nn} \end{bmatrix}$$

where the subscript *c* refers to a group of genotyped animals called hereafter “core animals”, and the subscript *n* refers to a second group of genotyped animals called hereafter “noncore animals”.

Following Misztal (Misztal et al., 2014; Misztal, 2016), the inverse of $\mathbf{G}$, $\mathbf{G}^{-1}$, can be approximated using the APY as follows:

$$\mathbf{G}_{APY}^{-1} = \begin{bmatrix} \mathbf{G}_{cc}^{-1} + \mathbf{G}_{cc}^{-1}\mathbf{G}_{cn}\mathbf{M}^{-1}\mathbf{G}'_{cn}\mathbf{G}_{cc}^{-1} & -\mathbf{G}_{cc}^{-1}\mathbf{G}_{cn}\mathbf{M}^{-1} \\ -\mathbf{M}^{-1}\mathbf{G}'_{cn}\mathbf{G}_{cc}^{-1} & \mathbf{M}^{-1} \end{bmatrix}$$

where the matrix $\mathbf{M}$ is a diagonal matrix of size of the number of noncore animals and with a diagonal element for the *i*th noncore animal equal to $\mathbf{M}_{ii} \ = diag\left(\mathbf{G}_{nn_{ii}} - \mathbf{G}'_{ci}\mathbf{G}_{cc}^{-1}\mathbf{G}_{ci}\right)$ with $\mathbf{G}_{ci}$ being the *i*th column of $\mathbf{G}_{cn}$. It is worth noting that the matrix $\mathbf{M}$ is an approximation of the Schur complement of $\mathbf{G}_{cc}$, i.e., $\mathbf{S} = \mathbf{G}_{nn} - \mathbf{G}'_{cn}\mathbf{G}_{cc}^{-1}\mathbf{G}_{cn}$. Replacing $\mathbf{M}$ by $\mathbf{S}$ in the formula of $\mathbf{G}_{APY}^{-1}$ would lead to the computation of the inverse of $\mathbf{G}$, $\mathbf{G}^{-1}$.

The APY only requires the computation of the submatrices $\mathbf{G}_{cc}$, $\mathbf{G}_{cn}$ and of the diagonal elements of $\mathbf{G}_{nn}$, in addition to the inversion of the submatrix $\mathbf{G}_{cc}$. Thus, the computational costs of the APY are reduced in comparison to the setting up and the direct inversion of $\mathbf{G}$. Also, the memory costs of the APY are reduced because only submatrices, $\mathbf{G}_{cc}$ and $\mathbf{G}_{cn}$, must be stored and the matrix $\mathbf{G}_{APY}^{-1}$ is sparse thanks to the diagonal matrix $\mathbf{M}^{-1}$.

Simulated data

Populations. The assessment of the quality of the genomic predictions from a sparse ssGBLUP in crossbreeding schemes was achieved by simulating a three-way crossbreeding program with random selection (Figure 1). Simulations of historic, purebred and crossbred recent populations were performed using the QMSim software (Sargolzaei and Schenkel, 2009). For the historic population, 70 discrete random mating generations (i.e., generations 1 to 70) with a constant size of 18,840 individuals with equal number of individuals from each sex were simulated, followed by 10 generations (i.e., generations 71 to 80) in which the effective population size was gradually reduced to 390 individuals. The next 20 generations (i.e. generations 81 to 100) were simulated to gradually expand the population size to 18,840. The last generation (i.e. generation 100) included 90 males and 18,750 females. Matings for all generations were based on the random union of gametes, which were randomly sampled from the pools of male and female gametes. To simulate the three breed populations (hereafter referred to as breeds A, B, and C), three random samples were drawn from the generation 100 of the historic population, each including 30 males and 6,250 females. Subsequently, within each breed, 100 generations (i.e. generations 101 to 200) of random mating were simulated before starting the three-way crossbreeding program (Figure 1). In each of the simulated 100 generations of random mating, each female had one male and one female offspring.

In the second step, a three-way crossbreeding program was simulated (Figure 1). Purebred (i.e., A, B, and C) animals that were used as founders of the pedigree (i.e., the first generation of the pedigree) were from generations 200. For each breed, A, B, and C, the next 9 discrete generations (i.e. generations 201 to 209) of purebred animals were simulated by means of random selection and matings while maintaining a constant size of 30 males and 6,250 females. For mimicking a three-way crossbreeding program, from the generation 205 until the generation 208, B and C purebred animals were randomly crossed to produce four generations (i.e. generations 206 to 209) of F1 animals, that is 30 BC crossbred males and 6,250 BC crossbred females. These BC crossbred animals were then randomly mated to males from breed A to produce four generations (i.e. generations 206 to 209) of F2 animals, called A(BC) crossbred animals. For each generation, 6,280 A(BC) crossbred animals were simulated (Figure 1). Purebred animals that were used as parents of crossbred animals could also be parents of purebred animals in the next generation. A total of 5 replicates were simulated using the QMSim software.

Genotypes. The genome was simulated using the QMSim software, simultaneously with the simulation of the historic, purebred and crossbred recent populations. The genome consisted of 18 chromosomes designed to resemble the Sus Scrofa genome with a SNP density that was comparable to that of a 60k SNP chip. The SNP positions were randomized across the genome and a recurrent mutation rate of 2.5x10^{-5}, as well as 1 mean crossover per 1 Morgan, were assumed. All SNPs that segregated in the last historical generation (i.e., generation 100) and with a minor allele frequency (MAF) higher than or equal to 0.05 were selected and used to simulate the genotypes of the purebred and crossbred animals. In addition to the SNPs, 4,500 QTL were simulated, and their positions were also randomized across the genome. Mutation rate and MAF of the QTL were the same as the ones for the simulated SNPs.

Phenotypes. For all purebred and crossbred animals, phenotypes for the breed composition to which they belonged were simulated under additive gene action using a custom Fortran program. This resulted in five traits: one trait for each of the purebred performances A, B and C, and one trait for each of the crossbred performances BC and A(BC). Genetic correlations between traits were randomly sampled in the range [0.2-0.8] from a uniform distribution. Simulated genetic correlations between purebred and crossbred traits were in the lowest range of reported values in the literature as reviewed by Wientjes and Calus (2017) (Table 1). Heritabilities (h_i^2) were randomly sampled in the range of [0.2-0.4] from a uniform distribution. Residual covariances were set to zero, as they would be in practice, because each animal has a phenotype for one of the five traits only. The same genetic correlations and heritabilities were used in all replicates, and are reported in Table 1.

For each animal and for each of the five traits, a true breeding value (TBV) was simulated by summing a polygenic effect and the multiplication of the simulated allele substitution effects with the genotypes of the 4,500 QTL coded as 0, 1 and 2. This genotype multiplication allowed different genetic levels across breeds for the same trait because QTL allele frequencies differ across breeds. For each trait, the polygenic effect of each individual was equal to the sum of the average of polygenic effects of the parent and a Mendelian sampling term. The Mendelian sampling terms for the five traits were sampled from a multinormal distribution with means of 0 and variances equal to the Mendelian sampling variances (Mrode, 2005). Correlations between the simulated Mendelian sampling terms were assumed to be equal to the genetic correlations. The variance of the polygenic effect of each ith trait was assumed to be equal to 5% of the total additive genetic variance (σ_{Ai}^2).

The allele substitution effects of QTLs were sampled from a multinormal distribution with means of 0, and variances of 1. The correlations between allele substitution effects of the QTL underlying the 5 traits were equal to the genetic correlations. For each trait, the genetic variance explained by all QTLs was computed as the sum of the variances across all QTLs, assuming no correlation between the QTLs. The simulated additive genetic variance of each jth QTL was calculated as $\sigma_{gj}^2 = 2p_j(1 - p_j)a_j^2$, where p_j is the allele frequency and a_j is the allele substitution effect of jth QTL. For each trait, the allele substitution effects were rescaled to obtain an additive genetic variance explained by the QTLs (σ_g^2) equal to 1. The part of the total additive genetic variance explained by the QTLs was assumed to be equal to 95% for each i^{th} trait. Finally, the phenotypes for each trait for each animal were generated by summing the TBV and a residual error sampled from a normal distribution with a mean 0 and a variance equal to $\left(\frac{1}{h_i^2} - 1\right) * \sigma_{Ai}^2$.

Datasets. For all the analyses, the pedigree included all the animals simulated for the creation of the three-way crossbreeding program. The phenotype dataset included 126,000 records. Among all records, 100,000 records were associated with purebred (i.e. A, B, and C) animals randomly sampled among all purebred animals from generations 204 until 208. A total of 16,000 records were associated with A(BC) crossbred animals randomly sampled among all A(BC) crossbred animals from generations 206 until 209. Finally, 10,000 records were associated with BC crossbred dams. Average numbers of purebred and crossbred animals per generation with a phenotype are given in the E-Supplements Table S1.

The genotype dataset included 89,000 genotypes. This included all 26,000 phenotyped BC and A(BC) crossbred animals. A total of 48,000 genotypes were from purebred (i.e. A, B, and C) animals randomly sampled among all purebred animals from the generations 205 until 208, regardless whether they had a phenotype or not. A total of 15,000 genotypes were from purebred (i.e. A, B, and C) animals randomly sampled among all purebred animals from generation 209. These 15,000 animals did not have phenotypes and are hereafter considered as selection candidates. Average numbers of purebred and crossbred animals per generation with a phenotype and a genotype are given in the E-Supplements Table S2.

Model and scenarios evaluated

Five-trait ssGBLUP was performed. The model for the *i*th trait (*i = A, B, C, BC, A(BC)*) was as follows:

$$\mathbf{y}_i = \mathbf{1}\mu_i + \mathbf{W}_i\mathbf{a}_i + \mathbf{e}_i$$

where, for the *i*th trait, $\mathbf{y}_i$ is the vector of records, μ_i is the general mean, $\mathbf{a}_i$ is the vector of additive genetic effects, $\mathbf{e}_i$ is the vector of residuals, the vector $\mathbf{1}$ is a vector of 1's relating the records to the general mean, and $\mathbf{W}_i$ is an incidence matrix relating the records to the animals. The variance components used for the simulations were used for the five-trait ssGBLUP. The vector of additive genetic effects $\mathbf{a} = \begin{bmatrix}\mathbf{a}'_A & \mathbf{a}'_B & \mathbf{a}'_C & \mathbf{a}'_{BC} & \mathbf{a}'_{A(BC)}\end{bmatrix}'$ followed a multivariate normal (MVN) distribution $MVN(\mathbf{0}, \mathbf{H}^{-\mathbf{1}} \otimes \mathbf{\Gamma})$ where $\otimes$ is the Kronecker product, $\mathbf{\Gamma}$ is the additive genetic (co)variance matrix, and the vector of residuals $\mathbf{e} = \begin{bmatrix}\mathbf{e}'_A & \mathbf{e}'_B & \mathbf{e}'_C & \mathbf{e}'_{BC} & \mathbf{e}'_{A(BC)}\end{bmatrix}'$ followed a MVN distribution $MVN(\mathbf{0}, \mathbf{I} \otimes \mathbf{R})$ where $\mathbf{R}$ is the residual (co)variance matrix.

Using all the 89,000 genotypes, the matrix $\mathbf{G}$ required for the computation of $\mathbf{H}^{-\mathbf{1}}$ was computed without breed-specific adjustments, as suggested by Lourenco et al. (2016). This matrix was equal to $\mathbf{G} = 0.95\mathbf{G}_a + 0.05\mathbf{A}_{22}$ with the adjusted genomic relationship matrix $\mathbf{G}_a$ computed as follows:

$$\mathbf{G}_a = \left(1 - \overline{f_p}\right)\mathbf{G}^* + 2\overline{f_p}\mathbf{J}$$

where $\mathbf{G}^*$ is a raw genomic relationship matrix computed following the first method of VanRaden (2008) using current allele frequencies computed from all genotyped animals, $\mathbf{J}$ is a matrix of ones, and $\overline{f_p}$ is the average pedigree inbreeding coefficient across (core) genotyped animals. The matrix $\mathbf{H}^{-1}$ was constructed in two different ways. First, the complete $\mathbf{G}$ was directly inverted to obtain $\mathbf{G}_{direct}^{-\mathbf{1}}$. Second, $\mathbf{G}_{direct}^{-\mathbf{1}}$ was replaced by $\mathbf{G}_{APY}^{-1}$. Because the APY relies on the size and the composition of the set of core animals (Misztal et al., 2014), we investigated different numbers of core animals and different strategies to select the core animals. For all the strategies, the selection candidates were allowed to be considered as core animals. The number of core animals were 4,000, 6,000, 8,000, 10,000, and 13,000. For each size, four different strategies were applied to select the core animals. The core animals were randomly sampled 1) among all breed A genotyped animals (called “Breed A”), 2) among all purebred genotyped animals (called “Purebred”), or 3) among all purebred and crossbred genotyped animals (called “Purebred + Crossbred”). For the fourth strategy, a QR decomposition with pivoting of the transposed genotype matrix was applied to the animals. The QR decomposition with pivoting returns a permutation matrix such that the diagonal elements of the upper triangular matrix **R** are decreasing (Golub and Van Loan, 1996). The genotyped animals corresponding to the highest diagonal elements of the matrix **R** were chosen as core animals (called “QR”). The aim of this fourth strategy was to select core animals such that the conditioning of the mixed model equations was improved, resulting in faster convergence, in comparison to the other three strategies (Fernando et al., 2016). All computations and analyses were run using our own custom programs for QR decomposition and statistical analyses, calc_grm (Calus and Vandenplas, 2016) for the computation of the different relationship matrices (i.e., $\mathbf{G}_{direct}^{-\mathbf{1}}$, $\mathbf{G}_{APY}^{-\mathbf{1}}$, and $\mathbf{A}_{22}^{-1}$), and MiXBLUP (ten Napel et al., 2016) for predicting the different GEBV. The matrices $\left(\mathbf{G}_{direct}^{-\mathbf{1}} - \mathbf{A}_{22}^{-1}\right)$ and $\left(\mathbf{G}_{APY}^{-\mathbf{1}} - \mathbf{A}_{22}^{-1}\right)$ were provided to MiXBLUP as external matrices.

Criteria

We evaluated the prediction of GEBV of genotyped selection candidates for the purebred A, B, and C performances and the crossbred A(BC) performances, for each set of core animals and each breed separately. Three criteria were computed from the GEBV of the selection candidates. First, the ratios between the accuracies of $\mathrm{GEBV_{APY}}$ from alternative core groups and the accuracies of $\mathrm{GEBV_{DIRECT}}$ (i.e., from $\mathbf{G}_{direct}^{-1}$), were computed. Accuracies were computed as the Pearson correlation between GEBV and TBV. A ratio of accuracies smaller than 1 means that $\mathrm{GEBV_{APY}}$ is less accurate than $\mathrm{GEBV_{DIRECT}}$. Second, regression coefficients of TBV on $\mathrm{GEBV_{APY}}$ and on $\mathrm{GEBV_{DIRECT}}$ (hereafter called bias) were computed. Third, ratios between mean squares errors (MSE) of $\mathrm{GEBV_{APY}}$ and MSE of $\mathrm{GEBV_{DIRECT}}$, were computed. The MSE were computed as the mean of the squared differences between GEBV and TBV. All results were averaged across five replicates. Tukey's honest significant difference test (Tukey, 1949) was used to assess significance of differences between scenarios at a 5% significance level.

For situations with single breeds, the number of required core animals that gives accurate GEBV, can be determined as the number of largest eigenvalues explaining 98-99% of the variation in $\mathbf{G}^*$ (Misztal, 2016; Pocrnic et al., 2016a; b). For investigating this relationship in situations involving multiple breeds and their F1 and F2 crosses, we computed the numbers of eigenvalues that explained 98% and 99% of the variation in $\mathbf{G}^*$ that included all the 89,000 genotyped purebred and crossbred animals. Computations were performed with calc_grm (Calus and Vandenplas, 2016). For each scenario, the number of eigenvalues were compared to the number of core animals needed such that the accuracies of $\mathrm{GEBV_{APY}}$ were equal to or higher than 99% of the accuracy for $\mathrm{GEBV_{DIRECT}}$ for both purebred and crossbred performance traits.

RESULTS

Characteristics of simulated data

The simulation yielded three breeds, A, B, and C, that were highly separated, as shown by the projections of genomic relationships into the two first eigenvectors for the first replicate (Figure 2). The estimated global Wright's F_{st} statistics, that is a measure to quantify the level of genetic differentiation between the breeds, was equal to 0.35 on average across the five replicates. The global Wright's F_{st} statistics were estimated from the genotypes of all purebred animals of the generation 204 with the software Genepop (4.2) (Raymond and Rousset, 1995; Rousset, 2008). The mean absolute difference in allele frequencies between breeds was about 0.34 on average across the five replicates. All these observations suggest three genetically divergent populations. The average linkage disequilibrium, expressed as r^2 (Hill and Robertson, 1968), between adjacent SNP pairs with MAF > 0.05 and across chromosomes, was 0.25 for the three breeds on average across the five replicates. Genomic relationship matrices required for the singular value decomposition and genomic predictions were based on 52,518 SNPs on average across the five replicates.

Composition of the core groups

Four selection strategies were applied to compose the core groups: (1) the core animals were randomly selected among only breed A animals, (2) the core animals were randomly selected among purebred animals of breed A, B, and C, (3) the core animals were randomly selected among purebred animals of breed A, B, and C, and crossbred BC and A(BC) animals, and (4) the core animals were selected based on a QR decomposition of the genotype matrix. For the four selection strategies, Figure 3 shows the proportions of core animals across the generations and across the breed compositions of a randomly chosen replicate for the scenario with 8,000 core animals. Similar results were obtained for the other replicates and sizes of core groups. Proportions of core animals were similar across the generations, and across the breed compositions for the first three selection strategies. For the selection strategy based on

QR decomposition, core animals were unequally spread across all generations and breed compositions: the highest proportions of core animals selected within a generation and a breed composition were observed among the crossbred A(BC) animals and the first generation of genotyped purebred animals (Figure 3).

Quality of GEBV with G_{direct}^{-1}

On average 5,000 genotyped selection candidates per breed were considered for computing accuracy, bias, and MSE (Table 2). For purebred performance, the accuracies were between 0.79 and 0.81. For crossbred performance, the accuracies were between 0.63 and 0.71. All sets of GEBV were almost unbiased (i.e., values for bias were close to 1) and had values of MSE close to 0 (Table 2).

Quality of GEBV with only breed A core animals

When the core groups included only breed A animals, $GEBV_{APY}$ were predicted as accurately as $GEBV_{DIRECT}$ for the breed A selection candidates for both purebred and crossbred performance traits, as shown by the ratios between the accuracies of $GEBV_{APY}$ and of $GEBV_{DIRECT}$ (Figure 4). In addition, $GEBV_{APY}$ were unbiased, and MSE was close to 0 (Figure 4; Table 3; Table 4; E-Supplements Tables S3-S6). However, $GEBV_{APY}$ were less accurate and more biased than $GEBV_{DIRECT}$ for the breed B and breed C selection candidates, as shown by low ratios of accuracies, and high values for bias and ratios of MSE of $GEBV_{APY}$ (Figure 4; Table 3; Table 4; E-Supplements Tables S3-S6). Across core groups, $GEBV_{APY}$ were from 18% to 40% less accurate than $GEBV_{DIRECT}$, and MSE of $GEBV_{APY}$ were between 16 and 81% higher than the corresponding MSE of $GEBV_{DIRECT}$.

Quality of GEBV with core animals of different breed compositions

Based on the three performance criteria, ratios of accuracies, bias, and ratios of MSE, the scenarios with core animals of different breed compositions outperformed the scenarios with only breed A core animals for both purebred and crossbred performance traits. Use of core groups with core animals of different breed compositions allowed the prediction of $\mathrm{GEBV_{APY}}$ that were unbiased, and (almost) as accurate as $\mathrm{GEBV_{DIRECT}}$, for all selection candidates and performance traits. Indeed, the regression coefficients of TBV on $\mathrm{GEBV_{APY}}$ were close to 1 (Table 3); the ratios of accuracies were higher than 0.97 for the purebred performance trait, and higher than 0.94 for the crossbred performance trait (Figure 5; Figure 6; E-Supplements Table S3); and the MSE of $\mathrm{GEBV_{APY}}$ were similar to MSE of $\mathrm{GEBV_{DIRECT}}$ (Table 4; E-Supplements Table S6). Ratios of accuracies close to, or higher than, 0.99 were then obtained for both traits when at least 8,000 core animals were used. The corresponding Pearson correlations between $\mathrm{GEBV_{APY}}$ and $\mathrm{GEBV_{DIRECT}}$, which is usually used as criteria in studies on real datasets (e.g., Ostersen et al., 2016; Strandén et al., 2017), were about 0.995 (E-Supplements). It is worth noting that the core size of 8,000 animals is between the numbers of eigenvalues that explained 98% and 99% of the variation in $\mathbf{G}^*$, that is about 6,498 and 9,213 eigenvalues on average across the five replicates, respectively (Figure 4-Figure 6).

Comparison of the three performance criteria for the purebred performance trait showed no difference among the three core selection strategies involving core animals of different breed compositions (Figure 5;Table 3; Table 4; E-Supplements Tables S3-S6). For the crossbred performance trait, the scenarios with purebred and crossbred core animals, either randomly chosen or chosen based on a QR decomposition, slightly outperformed the scenarios with only purebred core animals (Figure 6). However, these outperformances were not always significant (E-Supplements).

Quality of GEBV for core and non-core selection animals

Table 5 shows ratios of accuracies and of MSE, and the regression coefficients for the scenario using 8000 core animals randomly selected among purebred and crossbred animals. The regression coefficients and ratios of MSE for $\mathrm{GEBV}_{\mathrm{APY}}$ of core selection candidates and of non-core selection candidates were similar. Ratios of accuracies for non-core selection candidates were slightly lower than the corresponding ratios for the core selection candidates, meaning that $\mathrm{GEBV}_{\mathrm{APY}}$ of non-core selection candidates were slightly less accurate than those of core selection candidates, in comparison to $\mathrm{GEBV}_{\mathrm{DIRECT}}$. However, the differences between accuracies of $\mathrm{GEBV}_{\mathrm{APY}}$ of core and of non-core selection candidates were not significant following a Welch's t-test (Welch, 1947) with a 5% significance level.

Convergence of ssGBLUP with alternative core groups

Convergence of ssGBLUP with alternative core groups of 8,000 animals were compared against ssGBLUP using $\mathbf{G}_{direct}^{-1}$. Number of iterations of ssGBLUP using $\mathbf{G}_{APY}^{-1}$ were expressed as the ratio to the number of iterations of ssGBLUP using $\mathbf{G}_{direct}^{-1}$. Average values of this ratio across the 5 replicates (SD within brackets), were 0.85 (0.39) using breed A core animals, 1.05 (0.33) using purebred core animals, 0.95 (0.31) using purebred and crossbred animals, and 0.94 (0.30) using core animals selected based on a QR decomposition of the genotype matrix. In comparison to ssGBLUP with $\mathbf{G}_{direct}^{-1}$, use of the APY led to similar number of iterations to reach convergence. The selection strategy based on the QR decomposition led to similar convergence as the other selection strategies.

DISCUSSION

In this study, we showed that $\mathrm{GEBV}_{\mathrm{APY}}$ were accurate approximations of $\mathrm{GEBV}_{\mathrm{DIRECT}}$ for multivariate ssGBLUP involving multiple breeds and their crosses. $\mathrm{GEBV}_{\mathrm{APY}}$ as accurate as

$\mathrm{GEBV}_{\mathrm{DIRECT}}$ were obtained when the core groups included animals from different breed compositions, and when the core groups had a size between the numbers of the largest eigenvalues explaining 98% and 99% of the variation in the raw (i.e., before blending with the pedigree relationship matrix) genomic relationship matrix ($\mathbf{G}^*$).

Composition of the core groups and selection strategies

The quality of the $\mathrm{GEBV}_{\mathrm{APY}}$ for both purebred and crossbred performance traits was close to the $\mathrm{GEBV}_{\mathrm{DIRECT}}$ as long as all classes of purebred and crossbred animals were well represented in the core group. This was not the case if not all breeds were included in the core group. Such a situation where core animals are only from one breed, could be obtained with a naive random selection strategy on a large genotype dataset that is dominated by one breed. Due to the properties of the simulated datasets, e.g, similar numbers of genotyped animals per breed and per generation, a random selection of core animals across the full dataset led to similar proportions of core animas per breed composition and per generation. Based on a study involving single breed ssGBLUP, Ostersen et al. (2016) advised that core groups should represent all generations. Including animals from each generation in the core group was also recommended by Bradford et al. (2017), especially when genotyped animals had incomplete pedigree, such as unknown parents. Incomplete pedigree could be common in crossbreeding schemes, because pedigree data for crossbred animals in field conditions is difficult to collect (Ibáñẽz-Escriche et al., 2009). From our results with the selection strategy based on QR decomposition with pivoting, it seems that all generations, and all breed compositions, do not have to be similarly represented in core groups. Indeed, in comparison to a random selection, the selection strategy based on QR decomposition included higher proportions of crossbred A(BC) animals and of the first generation of genotyped purebred animals selected as core animals. One possible explanation is that genotypes of the crossbred A(BC) animals and of the first generation of genotyped purebred animals include a large proportion of the independent chromosome segments from all the genotyped purebred and crossbred animals. However, core groups including animals that were randomly selected and that represented similarly all generations and all breed composition gave results similar to the numerical strategy based on QR decomposition, which is computationally expensive. Therefore, a random selection of core animals by ensuring that core animals represent similarly all generations and all breed compositions is advisable for the implementation of the APY in well-structured crossbreeding schemes as investigated in this study. More complex situations, such as multibreed (beef) cattle populations with a large variation in the observed breed compositions, would probably benefit from more advanced APY core selection approaches (Mäntysaari et al., 2017), such as the proposed numerical strategy based on QR decomposition.

Size of the core groups

For single breed ssGBLUP, Pocrnic et al. (2016a; b) showed that the size of the core groups required to predict $\mathrm{GEBV}_{\mathrm{APY}}$ at least as accurate as $\mathrm{GEBV}_{\mathrm{DIRECT}}$ was related to the dimensionality of the genomic information. In their studies, the most accurate $\mathrm{GEBV}_{\mathrm{APY}}$ were obtained when the core size was at least equal to the number of largest eigenvalues that explained 98% of the variation in the raw genomic relationship matrix $\mathbf{G}^*$. In this study, $\mathrm{GEBV}_{\mathrm{APY}}$ as accurate as $\mathrm{GEBV}_{\mathrm{DIRECT}}$ (i.e., with correlations between them ≥ 0.995) were obtained when the core sizes were between the numbers of largest eigenvalues that explained 98% and 99% of the variation in the raw genomic relationship matrix $\mathbf{G}^*$, provided that the composition of the core group represented the variation in all the breeds and crosses. Using a multibreed beef cattle population, Mäntysaari et al. (2107) also showed that a core size larger than the number of largest eigenvalues that explained 98% of the variation in $\mathbf{G}^*$ was needed to get correlations between GEBV_{APY} and $\text{GEBV}_{\text{DIRECT}}$ close to 1. Furthermore, Mäntysaari et al. (2107) observed that the correlation between GEBV_{APY} and $\text{GEBV}_{\text{DIRECT}}$ depended on the composition of the core groups, even with a core size close to the number of largest eigenvalues that explained 98% of the variation in $\mathbf{G}^*$. All these results suggest that the core size involving multiple breeds and crosses can be also approximated based on the dimensionality of the genomic information of all breeds and crosses together to ensure that the core size is optimal. It should be noted, however, that in crossbreeding situations relationships between the core size, the dimensionality of the genomic information, and some population parameters (e.g., number of independent segments, effective population size) is not as straightforward in as in single breed situations (Pocrnic et al., 2016a; b).

CONCLUSIONS

We showed that the APY algorithm gives results equivalent to those obtained with the direct inversion of the genomic relationship matrix when genotyped animals belong to a few different breeds and their F1 and F2 crosses, such as commonly observed in pig and poultry breeding programs. For such situations, we suggest that core animals could be randomly selected among all purebred and crossbred genotyped animals, while ensuring that they represent all generations and all breed compositions. It was also shown that selecting a number of core animals equal to the number of largest eigenvalues needed to explain 98-99% of the variation on the raw genomic relationship matrix, is sufficient to achieve good quality of GEBV in crossbreeding schemes.

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.

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.

Calus, M.P.L., and J. Vandenplas. 2016. Calc_grm – a Program to Compute Pedigree, Genomic, and Combined Relationship Matrices. ABGC, Wageningen UR Livestock Research.

Christensen, O.F. 2012. Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. Genet. Sel. Evol. 44:37.

Christensen, O.F., and M.S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42:2.

Fernando, R.L., H. Cheng, and D.J. Garrick. 2016. An efficient exact method to obtain GBLUP and single-step GBLUP when the genomic relationship matrix is singular. Genet. Sel. Evol. 48:80. doi:10.1186/s12711-016-0260-7.

Fragomeni, B.O., D.A.L. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, A. Legarra, T.J. 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.

Golub, G., and C.F. Van Loan. 1996. Matrix Computations. third ed. Johns Hopkins University Press, Baltimore, MD, USA.

Hill, W.G., and A. Robertson. 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38:226–231. doi:10.1007/BF01245622.

Ibánẽz-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.

Legarra, A., O.F. Christensen, I. Aguilar, and I. Misztal. 2014. Single Step, a general approach for genomic selection. Livest. Sci. 166:54–65.

Lourenco, D.A.L., 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. 94:909–19. doi:10.2527/jas.2015-9748.

Lourenco, D.A.L., S. Tsuruta, B.O. Fragomeni, Y. Masuda, I. Aguilar, A. Legarra, J.K. Bertrand, T.S. Amen, L. Wang, D.W. Moser, and I. Misztal. 2015. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. J. Anim. Sci. 93:2653–2662. doi:10.2527/jas.2014-8836.

Mäntysaari, E.A., R.D. Evans, and I. Strandén. 2017. Efficient single-step genomic evaluation for a multibreed beef cattle population having many genotyped animals. J. Anim. Sci. 95:4728–4737. doi:10.2527/jas2017.1912.

Masuda, Y., I. Misztal, S. Tsuruta, A. Legarra, I. Aguilar, D.A.L. Lourenco, B.O. Fragomeni, and T.J. Lawlor. 2016. Implementation of genomic recursions in single-step genomic best linear unbiased predictor for US Holsteins with a large number of genotyped animals. J. Dairy Sci. 99:1968–1974. doi:10.3168/jds.2015-10540.

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

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.

Mrode, R.A. 2005. Linear Models for the Prediction of Animal Breeding Values. 2nd ed. CABI Publishing, Wallingford, UK.

ten Napel, J., M.P.L. Calus, M. Lidauer, I. Stradén, E.A. Mäntysaari, H.A. Mulder, and R.F. Veerkamp. 2016. MiXBLUP, User-Friendly Software for Large Genetic Evaluations Systems. Version 2.0. Wageningen, the Netherlands.

Ostersen, T., O.F. Christensen, P. Madsen, and M. Henryon. 2016. Sparse single-step method for genomic evaluation in pigs. Genet. Sel. Evol. 48:48. doi:10.1186/s12711-016-0227-8.

Pocrnic, I., D.A.L. Lourenco, Y. Masuda, A. Legarra, and I. Misztal. 2016a. The dimensionality of genomic information and its effect on genomic prediction. Genetics 203:573–581.

Pocrnic, I., D.A.L. 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.

Powell, J.E., P.M. Visscher, and M.E. Goddard. 2010. Reconciling the analysis of IBD and IBS in complex trait studies. Nat. Rev. Genet. 11:800–805. doi:10.1038/nrg2865.

Raymond, M., and F. Rousset. 1995. GENEPOP (Version 1.2): Population genetics software for exact tests and ecumenicism. J. Hered. 86:248–249.

Rousset, F. 2008. genepop’007: a complete re-implementation of the genepop software for Windows and Linux. Mol. Ecol. Resour. 8:103–106. doi:10.1111/j.1471-8286.2007.01931.x.

Sargolzaei, M., and F.S. Schenkel. 2009. QMSim: a large-scale genome simulator for livestock. Bioinformatics 25:680–681.

Strandén, I., K. Matilainen, G.P. Aamand, and E.A. Mäntysaari. 2017. Solving efficiently large single-step genomic best linear unbiased prediction models. J. Anim. Breed. Genet. 134:264–274. doi:10.1111/jbg.12257.

Tukey, J.W. 1949. Comparing individual means in the analysis of variance. Biometrics 5:99–114.

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

Welch, B.L. 1947. THE GENERALIZATION OF “STUDENT”S’ PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARLANCES ARE INVOLVED. Biometrika 34:28–35. doi:10.1093/biomet/34.1-2.28.

Wientjes, Y.C.J., and M.P.L. Calus. 2017. BOARD INVITED REVIEW: The purebred-crossbred correlation in pigs: A review of theory, estimates, and implications. J. Anim. Sci. 95:3467–3478.

E-Supplements

Table S1 Number of purebred and crossbred animals with a phenotype per generation (average for the 5 replicates; SD within brackets).

Table S2 Number of purebred and crossbred animals with a phenotype and a genotype per generation (average for the 5 replicates; SD within brackets).

Table S3. Relative accuracies (average for the 5 replicates; SD within brackets) of GEBV from alternative core groups for the purebred (PB) and crossbred (CB) performance for genotyped selection candidates.

Table S4. Pearson correlations (average for the 5 replicates; SD within brackets) between GEBV for genotyped selection candidates from alternative core groups1 and GEBV from the direct inversion of **G**.

Table S5. Regression coefficients (average for the 5 replicates; SD within brackets) of TBV on GEBV from alternative core groups and the direct inversion of G for genotyped selection candidates.

Table S6. Relative mean squares errors (average for the 5 replicates; SD within brackets) of GEBV from alternative core groups for genotyped selection candidates.

Figures

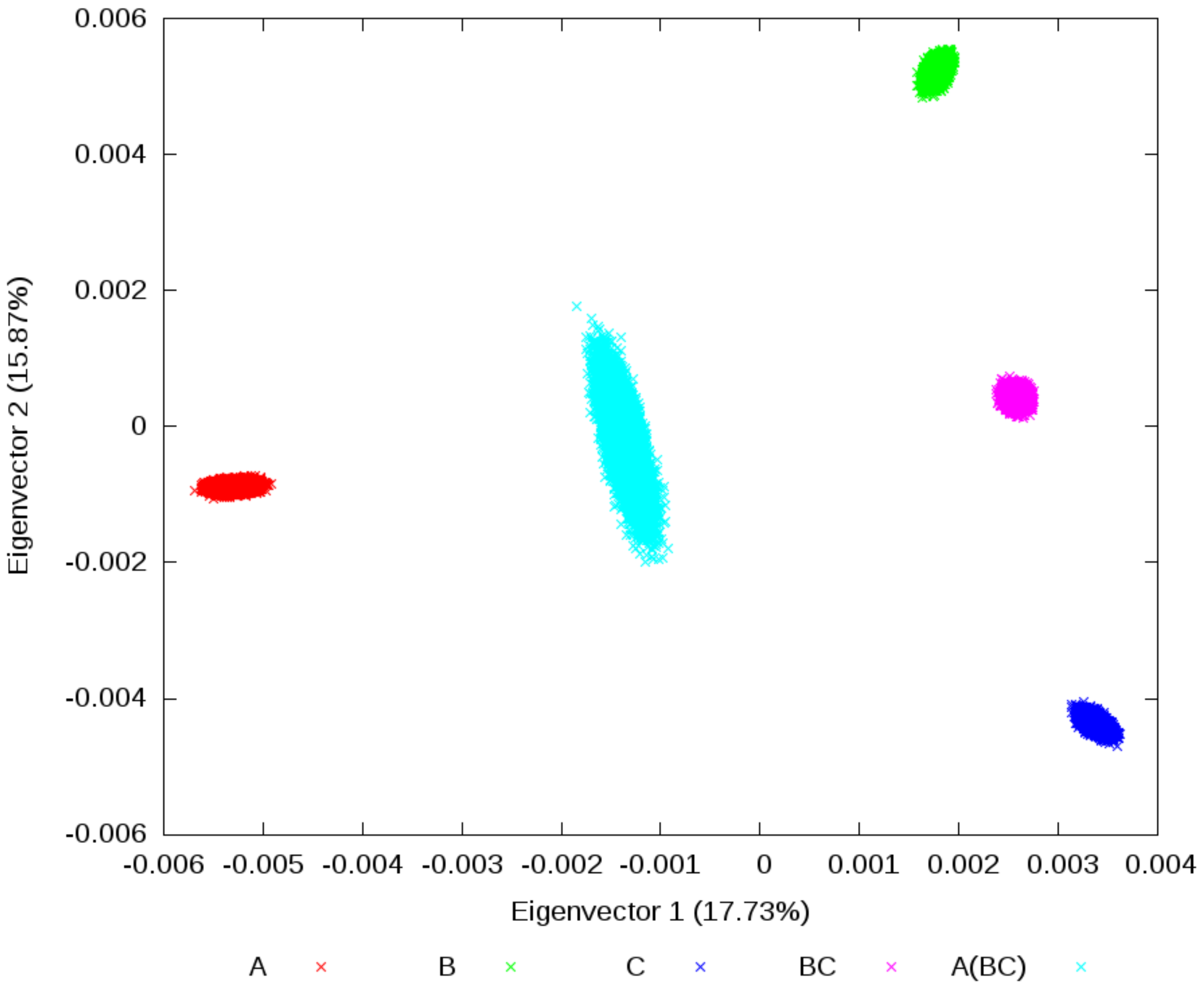
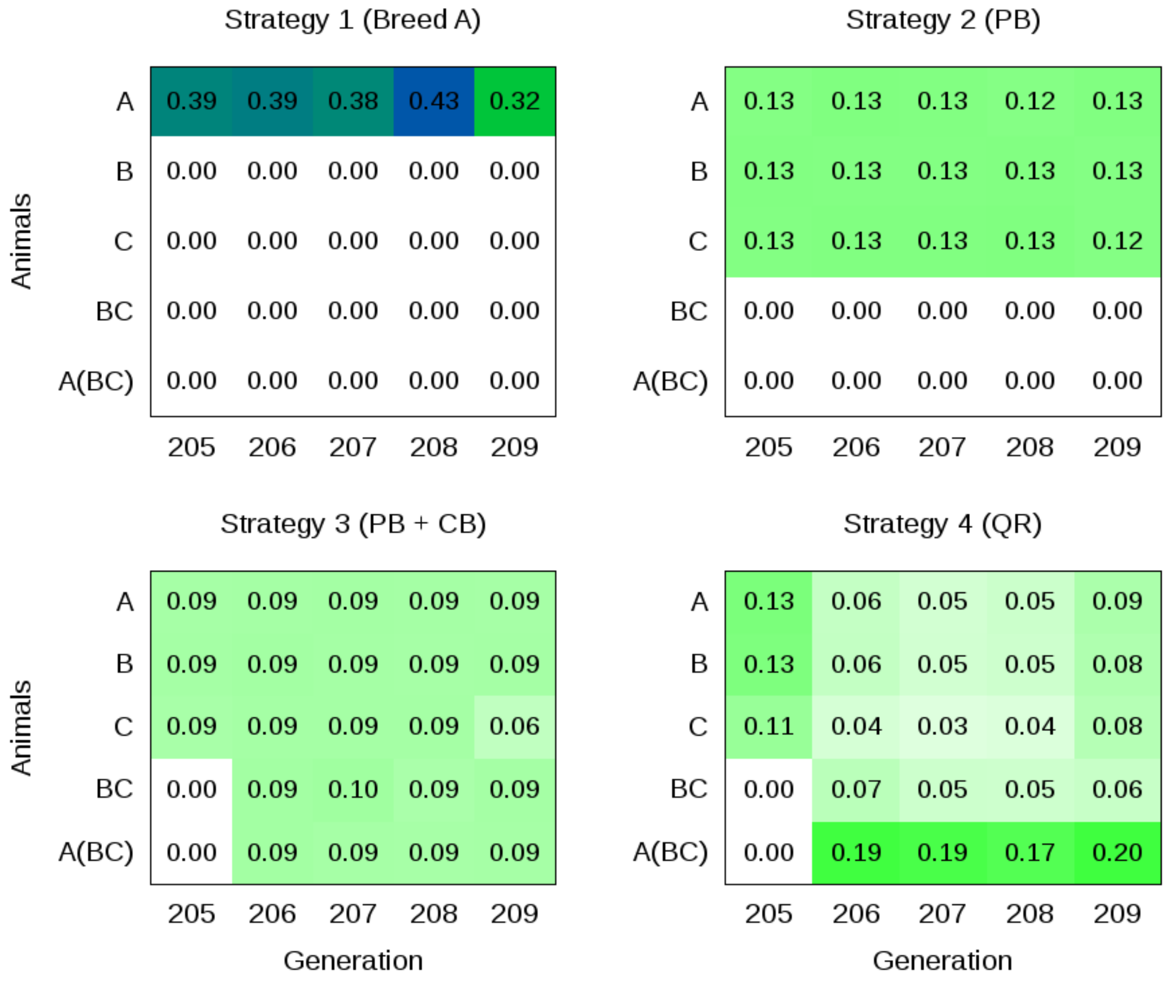


Figure 2.



Strategy 1 (Breed A)	205	206	207	208	209
A	0.39	0.39	0.38	0.43	0.32
B	0.00	0.00	0.00	0.00	0.00
C	0.00	0.00	0.00	0.00	0.00
BC	0.00	0.00	0.00	0.00	0.00
A(BC)	0.00	0.00	0.00	0.00	0.00

Strategy 2 (PB)	205	206	207	208	209
A	0.13	0.13	0.13	0.12	0.13
B	0.13	0.13	0.13	0.13	0.13
C	0.13	0.13	0.13	0.13	0.12
BC	0.00	0.00	0.00	0.00	0.00
A(BC)	0.00	0.00	0.00	0.00	0.00

Strategy 3 (PB + CB)	205	206	207	208	209
A	0.09	0.09	0.09	0.09	0.09
B	0.09	0.09	0.09	0.09	0.09
C	0.09	0.09	0.09	0.09	0.06
BC	0.00	0.09	0.10	0.09	0.09
A(BC)	0.00	0.09	0.09	0.09	0.09

Strategy 4 (QR)	205	206	207	208	209
A	0.13	0.06	0.05	0.05	0.09
B	0.13	0.06	0.05	0.05	0.08
C	0.11	0.04	0.03	0.04	0.08
BC	0.00	0.07	0.05	0.05	0.06
A(BC)	0.00	0.19	0.19	0.17	0.20

Figure 3.

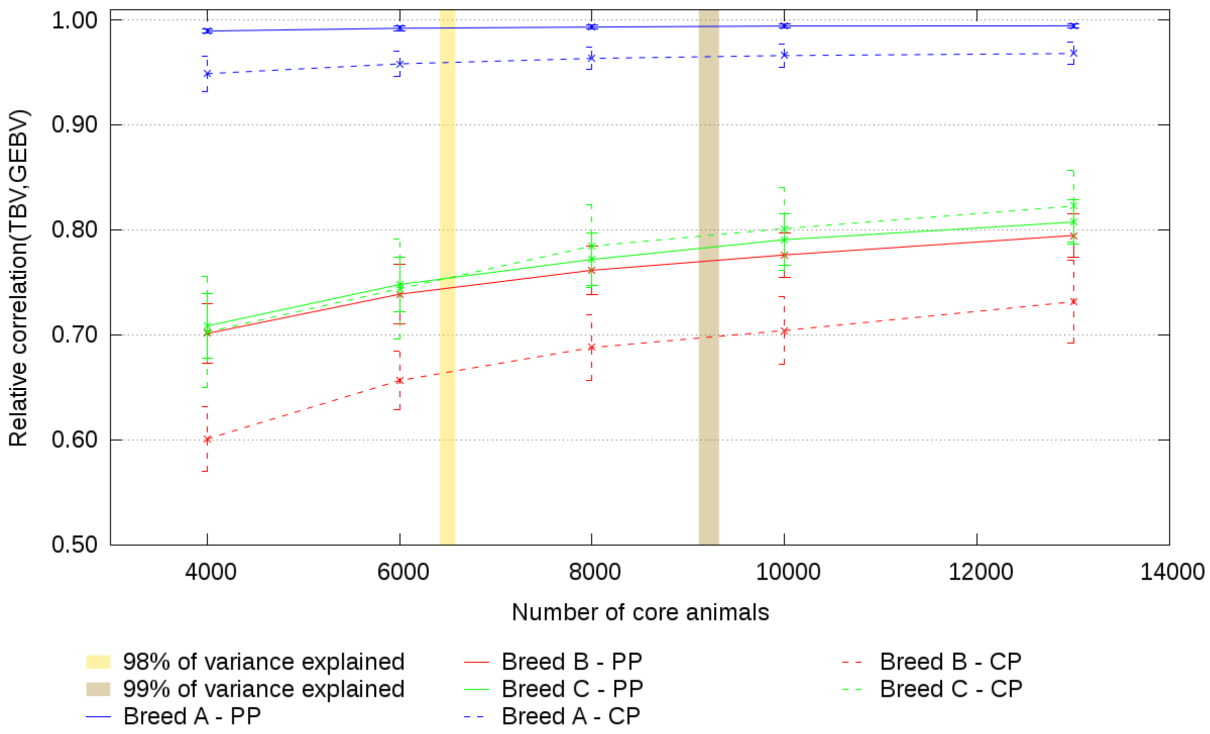
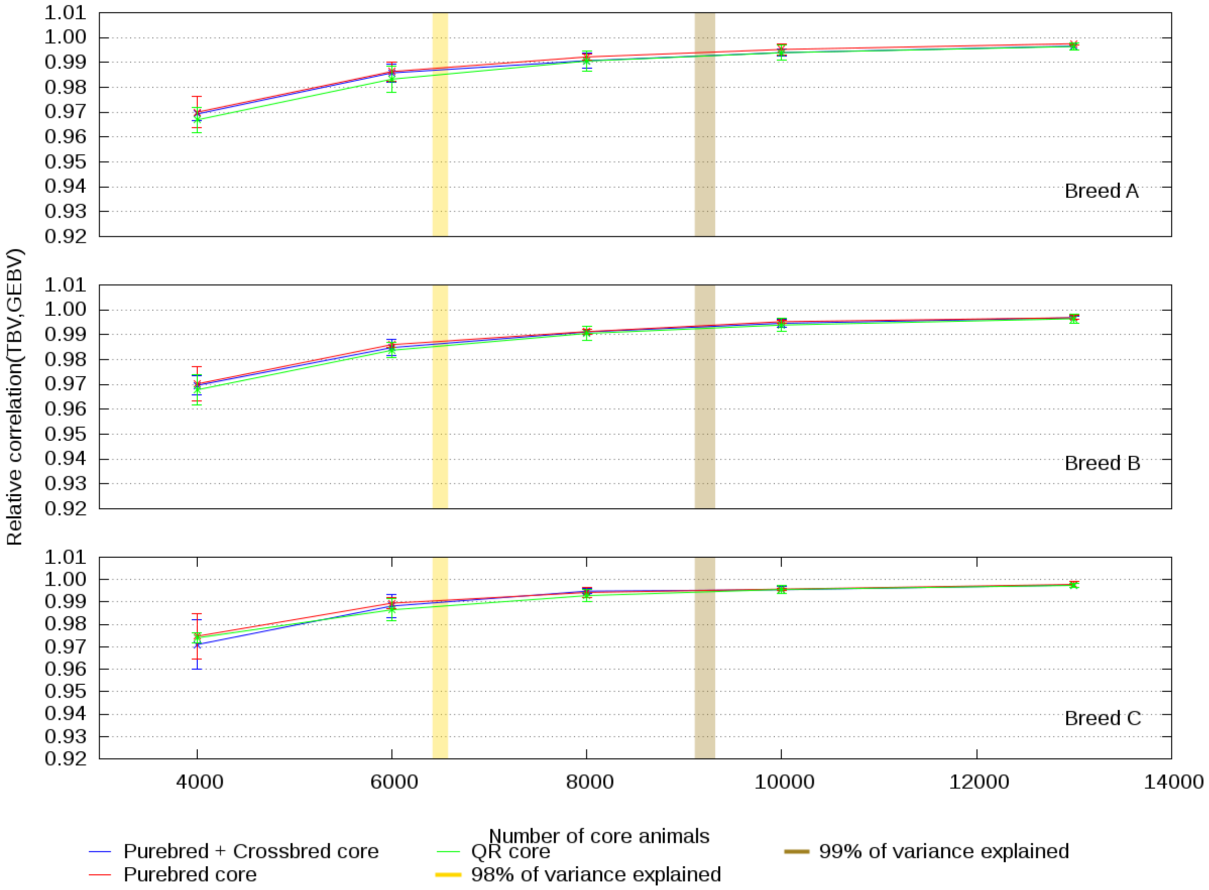

Relative correlation(TBV,GEBV)
1.00
0.90
0.80
0.70
0.60
0.50
4000
6000
8000
10000
12000
14000
Number of core animals
98% of variance explained
99% of variance explained
Breed A - PP
Breed B - PP
Breed C - PP
Breed A - CP
Breed B - CP
Breed C - CP

Figure 4.



1.01
1.00
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
Breed A
Breed B
Breed C
Relative correlation(TBV,GEBV)
4000
6000
8000
10000
12000
14000
Number of core animals
Purebred + Crossbred core
Purebred core
QR core
98% of variance explained
99% of variance explained

Figure 5.

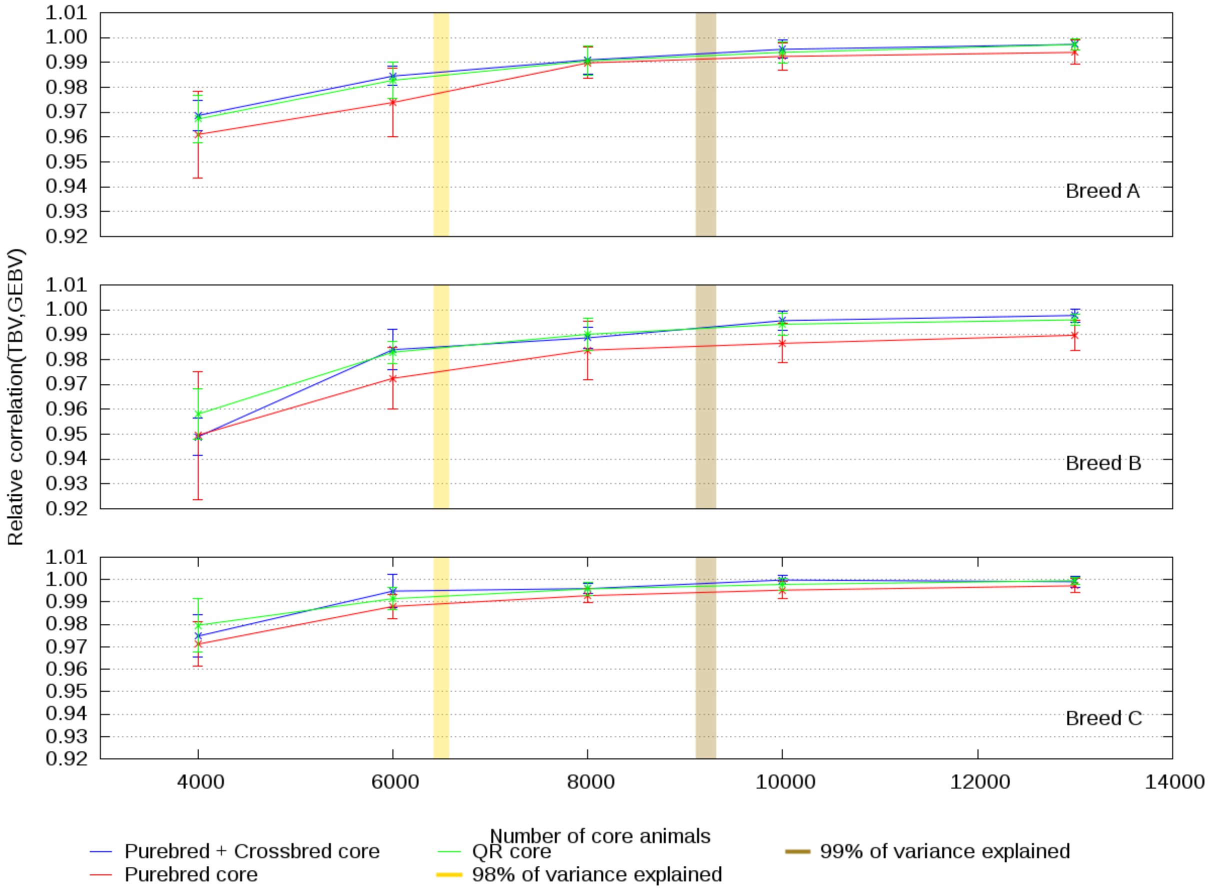

Relative correlation(TBV,GEBV)
1.01
1.00
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
Breed A
Breed B
Breed C
4000
6000
8000
10000
12000
14000
Number of core animals
Purebred + Crossbred core
Purebred core
QR core
98% of variance explained
99% of variance explained

Figure 6.

Figure 1. Schematic representation of the simulation. The crossbreeding program started at generation 200 (generation numbers in bold). The number of males (M) and females (F) per generation and per breed (A, B, and C), or per cross (BC, and A(BC)), are reported within brackets. Blue arrows denote the sires and dams of the next generation; red arrows denote the dams of the next generation; green arrows denote the sires of the next generation.

Figure 2. Projections of genomic relationships for purebred (A, B, and C) and crossbred (BC and A(BC)) genotyped animals into the two first eigenvectors for the first replicate.

Figure 3. Proportions of core animals per generation and breed composition of one replicate for the scenario using 8,000 core animals. Core animals were selected using four different strategies: 1) only from breed A animals (Breed A), 2) from purebred animals of breed A, B and C (PB), 3) from purebred animals of breed A, B and C, and crossbred BC and A(BC) animals (PB + CB), and (4) chosen based on a QR decomposition of the genotype matrix (QR). Darker colours represent higher proportions of core animals per generation and breed composition.

Figure 4. Relative correlations of GEBV from different sizes of core groups with only breed A animals. Relative correlations for the purebred performance (PP) and crossbred performance (CP) traits are defined as the ratio between the accuracies of GEBV from alternative core groups and the corresponding accuracies of GEBV from $\mathbf{G}_{direct}^{-1}$. Vertical columns depict the number of eigenvalues that explained 98% and 99% of the variation in $\mathbf{G}^{*}$. Results are averages for the 5 replicates.

Figure 5. Relative correlations of GEBV from alternative core groups for the purebred performance traits. Core groups include randomly selected purebred and crossbred animals (Purebred + Crossbred core), randomly selected purebred animals (Purebred core), and animals selected based on a QR decomposition of the genotype matrix (QR core). Relative correlations are defined as the ratio between the accuracies of GEBV from alternative core groups and the corresponding accuracies of GEBV from $\mathbf{G}_{direct}^{-1}$. Vertical columns depict the number of eigenvalues that explained 98% and 99% of the variation in $\mathbf{G}^{*}$. Results are averages for the 5 replicates.

Figure 6. Relative correlations of GEBV from alternative core groups for the crossbred performance trait. Core groups include randomly selected purebred and crossbred animals (Purebred + Crossbred core), randomly selected purebred animals (Purebred core), and animals selected based on a QR decomposition of the genotype matrix (QR core). Relative correlations are defined as the ratio between the accuracies of GEBV from alternative core groups and the corresponding accuracies of GEBV from $\mathbf{G}_{direct}^{-1}$. Vertical columns depict the number of eigenvalues that explained 98% and 99% of the variation in $\mathbf{G}^{*}$. Results are averages for the 5 replicates.

Tables

Table 1. Heritabilities (diagonal) and genetic correlations (off-diagonal) among the five simulated traits.

Trait	Purebred A	Purebred B	Purebred C	Crossbred BC	Crossbred A(BC)
Purebred A	0.28				
Purebred B	0.46	0.39			
Purebred C	0.27	0.80	0.22		
Crossbred BC	0.33	0.58	0.30	0.36	
Crossbred A(BC)	0.55	0.31	0.26	0.69	0.23

Table 2. Accuracies, bias, and mean square errors (MSE) of GEBV from the direct inversion of **G** (average for the 5 replicates; SD within brackets).

Selection candidates	**Number**	**Purebred performance**			**Crossbred performance**		
		Accuracy	**Bias**	**MSE**	**Accuracy**	**Bias**	**MSE**
Breed A	5010 (24)	0.81 (0.02)	1.04 (0.05)	1.11 (0.69)	0.68 (0.04)	0.98 (0.08)	0.68 (0.51)
Breed B	4975 (30)	0.85 (0.01)	1.06 (0.03)	1.16 (0.81)	0.63 (0.02)	0.95 (0.04)	0.90 (0.43)
Breed C	5016 (45)	0.79 (0.04)	1.04 (0.03)	1.42 (0.74)	0.71 (0.04)	1.04 (0.07)	1.35 (1.18)

Table 3. Regression coefficients (average for the 5 replicates; SD within brackets) of TBV on GEBV from alternative core groups[1] for genotyped selection candidates.

	Purebred performance				**Crossbred performance**			
Number of core animals	**Breed A**	**PB**	**PB+CB**	**QR**	**Breed A**	**PB**	**PB+CB**	**QR**
Breed A selection candidates								
4000	1.04	1.05	1.06	1.06	0.90	0.96	0.99	0.99
	(0.06)	(0.06)	(0.06)	(0.06)	(0.08)	(0.08)	(0.09)	(0.08)
8000	1.04	1.05	1.05	1.05	0.92	0.97	0.99	0.99
	(0.05)	(0.06)	(0.06)	(0.06)	(0.08)	(0.08)	(0.08)	(0.08)
13000	1.04	1.05	1.05	1.05	0.93	0.97	0.98	0.98
	(0.05)	(0.05)	(0.05)	(0.05)	(0.08)	(0.08)	(0.08)	(0.08)
Breed B selection candidates								
4000	1.49	1.06	1.06	1.07	1.62	0.91	0.93	0.94
	(0.08)	(0.03)	(0.02)	(0.03)	(0.11)	(0.08)	(0.05)	(0.06)
8000	1.46	1.06	1.06	1.06	1.58	0.93	0.95	0.95
	(0.08)	(0.02)	(0.03)	(0.03)	(0.13)	(0.06)	(0.05)	(0.04)
13000	1.43	1.06	1.06	1.06	1.54	0.93	0.95	0.95
	(0.09)	(0.03)	(0.03)	(0.03)	(0.14)	(0.05)	(0.05)	(0.04)
Breed C selection candidates								
4000	1.69	1.05	1.06	1.05	2.41	1.07	1.09	1.08
	(0.15)	(0.04)	(0.04)	(0.03)	(0.30)	(0.08)	(0.07)	(0.07)
8000	1.62	1.04	1.05	1.04	2.27	1.06	1.06	1.06
	(0.12)	(0.04)	(0.04)	(0.03)	(0.19)	(0.07)	(0.07)	(0.07)
13000	1.58	1.04	1.04	1.04	2.14	1.06	1.05	1.05
	(0.11)	(0.03)	(0.04)	(0.03)	(0.14)	(0.07)	(0.07)	(0.07)

1 Core groups include 1) randomly selected breed A animals only (Breed A), 2) randomly selected purebred animals (PB), 3) randomly selected purebred and crossbred animals (PB+CB), and 4) animals selected based on a QR decomposition of the genotype matrix (QR).

Table 4. Relative mean squares errors[1] (average for the 5 replicates; SD within brackets) of GEBV from alternative core groups[2] for genotyped selection candidates.

	Purebred performance				**Crossbred performance**			
Number of core animals	**Breed A**	**PB**	**PB+CB**	**QR**	**Breed A**	**PB**	**PB+CB**	**QR**
Breed A selection candidates								
4000	1.09	1.05	1.05	1.04	1.14	1.14	1.04	1.03
	(0.33)	(0.08)	(0.04)	(0.06)	(0.29)	(0.20)	(0.04)	(0.04)
8000	1.11	0.98	0.98	0.99	1.20	1.11	1.11	1.11
	(0.34)	(0.05)	(0.06)	(0.07)	(0.28)	(0.22)	(0.23)	(0.22)
13000	1.12	1.04	0.98	0.98	1.20	1.11	1.12	1.10
	(0.33)	(0.19)	(0.06)	(0.06)	(0.28)	(0.22)	(0.22)	(0.23)
Breed B selection candidates								
4000	1.75	1.12	1.05	1.08	1.34	0.99	1.04	1.02
	(0.93)	(0.16)	(0.06)	(0.08)	(0.30)	(0.15)	(0.08)	(0.07)
8000	1.81	1.09	1.09	1.09	1.27	0.99	0.95	0.95
	(1.04)	(0.18)	(0.17)	(0.18)	(0.21)	(0.15)	(0.10)	(0.10)
13000	1.77	1.15	1.08	1.08	1.24	1.00	0.93	0.94
	(0.95)	(0.23)	(0.18)	(0.18)	(0.21)	(0.14)	(0.11)	(0.09)
Breed C selection candidates								
4000	1.29	0.97	1.02	1.00	1.20	1.00	1.00	1.02
	(0.43)	(0.08)	(0.05)	(0.05)	(0.44)	(0.16)	(0.08)	(0.07)
8000	1.28	0.96	0.97	0.96	1.19	0.99	0.94	0.95
	(0.37)	(0.08)	(0.09)	(0.09)	(0.38)	(0.16)	(0.09)	(0.09)
13000	1.24	0.88	0.96	0.96	1.16	1.00	0.95	0.95
	(0.34)	(0.16)	(0.08)	(0.09)	(0.35)	(0.16)	(0.09)	(0.08)

[1] Results are expressed as the ratio between MSE of GEBV from alternative core groups and MSE of GEBV from the direct inversion of **G**.

[2] Core groups include 1) randomly selected breed A animals only (Breed A), 2) randomly selected purebred animals (PB), 3) randomly selected purebred and crossbred animals (PB+CB), and 4) animals selected based on a QR decomposition of the genotype matrix (QR).

Table 5. Quality of GEBV using APY for the core and non-core selection candidates. [1]

	Number	Purebred performance			Crossbred performance		
Selection candidates		**Accuracy[2]**	**Reg. coef.**	**MSE[2]**	**Accuracy[2]**	**Reg. coef.**	**MSE[2]**
A core	453 (19)	0.999 (0.001)	1.02 (0.05)	0.960 (0.058)	0.997 (0.001)	0.98 (0.07)	1.087 (0.203)
A non-core	4557 (32)	0.990 (0.003)	1.06 (0.06)	0.980 (0.066)	0.990 (0.006)	0.99 (0.08)	1.112 (0.232)
B core	456 (23)	0.998 (0.002)	1.02 (0.06)	1.076 (0.169)	0.995 (0.004)	0.88 (0.09)	0.946 (0.091)
B non-core	4519 (37)	0.991 (0.001)	1.07 (0.03)	1.093 (0.170)	0.988 (0.004)	0.95 (0.04)	0.949 (0.096)
C core	322 (43)	0.998 (0.001)	1.07 (0.07)	0.961 (0.082)	0.999 (0.004)	1.03 (0.06)	0.939 (0.089)
C non-core	4694 (27)	0.994 (0.001)	1.04 (0.04)	0.966 (0.087)	0.996 (0.003)	1.06 (0.07)	0.942 (0.093)

[1] Results (average for the 5 replicates; SD within brackets) are shown for the scenario using 8000 core animals randomly selected among purebred and crossbred animals.

[2] Results for accuracies and mean square errors (MSE) are expressed as the ratio between accuracies (MSE) of GEBV using APY and accuracies (MSE) of GEBV using the direct inversion of **G**.