Selection of core animals in the Algorithm for Proven and Young using a simulation model.

Received: 5 January 2017 | Accepted: 25 March 2017

DOI: 10.1111/jbg.12276

ORIGINAL ARTICLE

WILEY Journal of Animal Breeding and Genetics

Selection of core animals in the Algorithm for Proven and Young using a simulation model

H.L. Bradford | **I. Pocrnić** | **B.O. Fragomeni** | **D.A.L. Lourenco** | **I. Misztal**

Department of Animal and Dairy Science, University of Georgia, Athens, GA, USA

Correspondence
H.L. Bradford, Department of Animal and Dairy Science, University of Georgia, Athens, GA, USA.
Email: heather.bradford25@uga.edu

Funding information
USDA National Institute of Food and Agriculture, Grant/Award Number: 2015-67015-22936

Summary

The Algorithm for Proven and Young (APY) enables the implementation of single-step genomic BLUP (ssGBLUP) in large, genotyped populations by separating genotyped animals into core and non-core subsets and creating a computationally efficient inverse for the genomic relationship matrix (**G**). As APY became the choice for large-scale genomic evaluations in BLUP-based methods, a common question is how to choose the animals in the core subset. We compared several core definitions to answer this question. Simulations comprised a moderately heritable trait for 95,010 animals and 50,000 genotypes for animals across five generations. Genotypes consisted of 25,500 SNP distributed across 15 chromosomes. Genotyping errors and missing pedigree were also mimicked. Core animals were defined based on individual generations, equal representation across generations, and at random. For a sufficiently large core size, core definitions had the same accuracies and biases, even if the core animals had imperfect genotypes. When genotyped animals had unknown parents, accuracy and bias were significantly better ($p \leq .05$) for random and across generation core definitions.

KEYWORDS

APY, genetic evaluation, genomic selection, imputation, single-step genomic BLUP

1 | INTRODUCTION

Breeders have implemented genomic selection using single-step genomic BLUP (ssGBLUP) in many species worldwide (Aguilar et al., 2010; Christensen & Lund, 2010). The ssGBLUP combined pedigree, phenotypes and genotypes into an analysis using the same framework as historical genetic evaluations. Traditionally, the blended genomic relationship matrix (**G**) was directly inverted; however, this matrix was dense and had dimensions equal to the number of genotyped animals. Inverting **G** was computationally feasible when the most advanced livestock populations had up to 150,000 genotyped animals. With increasing adoption of genotyping globally, the ssGBLUP methodology was adapted to efficiently incorporate millions of genotyped animals into genetic evaluations. Misztal, Legarra, and Aguilar (2014) solved this problem by developing the Algorithm for Proven and Young animals (APY).

To implement APY, the genotyped population is divided into core and non-core animals such that core animals contain most of the genomic information, and **G** is partitioned into core and non-core animals. For APY $\mathbf{G}^{-1}$, only the core animals' partition is inverted directly. The APY $\mathbf{G}^{-1}$ also includes relationships between core and non-core animals and diagonal elements for non-core animals. These other components are linear functions of the inverse for the core animals' partition, genomic

relationships between core and noncore animals, and diagonal elements of **G** for noncore animals.

The dimensionality of the genomic information is limited by the minimum of number of SNP, number of effective SNP markers (or independent chromosome segments) and number of genotyped animals. This dimensionality is related to the core size used in the implementation of APY (Pocrnic, Lourenco, Masuda, Legarra, & Misztal, 2016; Pocrnic, Lourenco, Masuda, & Misztal, 2016). To assess dimensionality, eigenvalue decomposition of the original **G** without blending ($\mathbf{G_0}$) is used to determine the number of largest eigenvalues explaining most of the variation in $\mathbf{G_0}$, and this number of eigenvalues is used as the core size in APY. With a core size based on 98% of the variation in $\mathbf{G_0}$, APY was at least as accurate as the traditional $\mathbf{G^{-1}}$ in ssGBLUP. Thus, APY can replace traditional $\mathbf{G^{-1}}$ in large, genotyped populations because of the limited dimensionality of the genomic information.

According to theory, the choice of core animals is generally unimportant because of the limited dimensionality. With adequate core size, the true breeding values (TBV) of core animals are functions of the effects of independent chromosome segments, and the TBV of noncore animals are functions of the TBV for core animals. This concept can be extended to ssGBLUP using proven sires as core animals (Misztal et al., 2014). The estimated breeding values (EBV) for young animals are then functions of the EBV for proven sires. When proven sires were used as core animals, APY was as accurate as traditional ssGBLUP (Fragomeni, Lourenco, Tsuruta, Masuda, Aguilar, & Misztal 2015). More recently, Ostersen, Christensen, Madsen, and Henryon (2016) proposed that core definitions may have different accuracies. When selection is occurring, prediction accuracies for direct genomic values are known to decrease as the prediction and predicted populations become more distantly related (Muir, 2007; Saatchi et al., 2011). Thus, better understanding the importance of individual generations in APY is important for theoretical understanding and practical implementation. As APY became the choice for large-scale genomic evaluations in BLUP-based methods, a common question is how to choose animals to be part of the core subset.

A limited number of core definitions have been investigated. Using proven animals (many progeny) as core resulted in nearly identical EBV and accuracies as using random core definitions in cattle (Fragomeni et al., 2015 Lourenco et al., 2015; Masuda et al., 2016). In addition, random core definitions provided the same accuracy as the young core definition, which indicated that the core definition may be arbitrary (Fragomeni et al., 2015). Recently, different EBV were reported for swine when using old or young core definitions (Ostersen et al., 2016). Our objectives for the current study were to investigate different core definitions, to quantify accuracy changes when core animals were older and less related to the youngest generation and to ascertain why the random core definition worked well in implementation.

2 | MATERIALS AND METHODS

Animal care and use committee approval was not needed because data were simulated.

2.1 | Simulation

The population structure started with a founder population to generate initial linkage disequilibrium between SNP and QTL. The founder population began with 5,000 individuals and steadily decreased to 1,000 individuals after 1,000 generations. Then, the population size steadily increased for 250 generations to 5,010 individuals, 10 males and 5,000 females. Individuals in the last generation were parents for the first generation of the current population.

We simulated 10 non-overlapping generations for the current population undergoing selection on males. Selection was only for males to control the effective population size and to have a manageable number of genotyped animals. Individuals were randomly mated with two full-sibling offspring per mating (10,000 offspring per generation; equal sex ratio). From these offspring, 10 males were selected based on BLUP EBV along with all 5,000 females to be parents for the next generation. This process generated a pedigree with 105,010 individuals. Generations 0 to 9 had phenotypes (n = 95,010) for a moderately heritable trait (h^2 = 0.30), and generation 10 was used for validation. Five replicates were simulated using QMSim (Sargolzaei & Schenkel, 2009).

The simulations had small effective population sizes. The theoretical effective population size was 40 based on the formula given by Wright (1931). Mean realized effective population size (*SE*) was 26 (7.6) based on the amount of inbreeding per generation and defined by Falconer and Mackay (1996). The realized and theoretical effective population sizes differed because selection violated the assumptions of an idealized population, but both estimates indicated a small effective population size.

Generations 6 to 10 had genotypes (n = 50,000) based on the following assumptions. While less realistic, all animals in these generations were assumed to be genotyped; this simplification allowed for a better theoretical understanding of how to select core animals. The simulated genomes contained fifteen 1 M long chromosomes, 25,500 biallelic SNP and 2550 biallelic QTL. The SNP and QTL were randomly positioned on the chromosomes with equal numbers per chromosome. The simulations created a similar number of SNP per chromosome as medium-density

genotyping typical in cattle. The QTL effects were simulated from the Gamma distribution (shape = 0.40, scaled internally for a genetic variance of 0.30) resulting in QTL with small effects and accounted for all the genetic variation in the trait. All SNP and QTL had 0.5 allele frequencies to begin the founder population. On average, 1 crossover occurred per chromosome with no interference, and the recurrent mutation was 2.5×10^{-5} mutations per meiosis per loci. Allele frequencies and linkage disequilibrium changed throughout the simulation. For linkage disequilibrium, mean (*SE*) pooled r^2 per chromosome was 0.38 (0.01) based on default calculations in QMSim (Sargolzaei & Schenkel, 2009).

2.2 | Methodology

We constructed $\mathbf{G_0}$ following VanRaden (2008):

$$\mathbf{G_0} = \frac{\mathbf{ZZ'}}{2\sum p_i(1-p_i)},$$

in which $\mathbf{Z}$ was a centred gene content matrix and p_i was the minor allele frequency of SNP i. Allele frequencies were calculated from all observed genotypes. A blended $\mathbf{G_0}$ was used in implementation and was defined as follows:

$$\mathbf{G} = 0.95\mathbf{G_0} + 0.05\mathbf{A_{22}},$$

in which $\mathbf{A_{22}}$ was the partition of the numerator relationship matrix corresponding to genotyped animals.

The traditional ssGBLUP involved replacing $\mathbf{A^{-1}}$, the inverse of the numerator relationship matrix, with $\mathbf{H^{-1}}$ defined by Aguilar et al. (2010) as

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

in which $\mathbf{G^{-1}}$ was calculated directly. This $\mathbf{G^{-1}}$ becomes more computationally challenging as more animals are genotyped. Alternatively, a sparse $\mathbf{G^{-1}}$ was created using APY (Misztal et al., 2014). For APY, animals were categorized as either core (c) or non-core (n) animals. Thus, $\mathbf{G}$ was partitioned as follows:

$$\mathbf{G} = \begin{bmatrix} \mathbf{G_{cc}} & \mathbf{G_{cn}} \\ \mathbf{G_{nc}} & \mathbf{G_{nn}} \end{bmatrix}.$$

The APY inverse was calculated as follows:

$$\mathbf{G_{APY}^{-1}} = \begin{bmatrix} \mathbf{G_{cc}^{-1}} + \mathbf{G_{cc}^{-1}G_{cn}M^{-1}G_{nc}G_{cc}^{-1}} & -\mathbf{G_{cc}^{-1}G_{cn}M^{-1}} \\ -\mathbf{M^{-1}G_{nc}G_{cc}^{-1}} & \mathbf{M^{-1}} \end{bmatrix},$$

with

$$M_{ii} = G_{ii} - \mathbf{g_{ic}G_{cc}^{-1}g_{ci}},$$

in which $\mathbf{M}$ was a diagonal matrix with dimensions equal to the number of non-core animals. Thus, the inverted matrices were a diagonal matrix and a small subset of $\mathbf{G}$. Misztal (2016) presented complete derivations and theory for APY. We analysed all data using the BLUPF90 family of programs (Misztal, et al. 2016).

2.3 | Scenarios

The core size has been linked to the dimensionality of the genomic information. A limited number of effective SNP markers or independent chromosome segments exist in livestock populations; so, adding more genotyped animals contributes less and less new information about the population. Enough core animals were needed to account for most of the variation in $\mathbf{G}$ and to ultimately obtain accurate EBV. The core size was determined through eigenvalue decomposition of $\mathbf{G_0}$ (Pocrnic, Lourenco, Masuda, Legarra, et al. 2016). Core sizes were the numbers of largest eigenvalues explaining 98, 95 or 90% of the variation in $\mathbf{G_0}$. The core size was calculated for each simulation replicate, and the same core size (98, 95, or 90%) was used for scenarios within the replicate. Hence, core sizes differed across replicates but were based on the same proportion of variation in $\mathbf{G_0}$. We focused on the effect of core size for one group of analyses, and all remaining analyses used a core size equal to the number of largest eigenvalues explaining 98% of the variation in $\mathbf{G_0}$ for each replicate. This value was selected based on previously reported accuracies (Pocrnic, Lourenco, Masuda, Legarra, et al. 2016, Pocrnic, Lourenco, Masuda, & Misztal, 2016).

The core definition was investigated by analysing the same data set with ssGBLUP but using different core animals in APY, and the core animals were selected based on specific subsets of the genotyped animals (Table 1). The core animals were randomly selected from parents in one generation (generations 6 to 9) and from young animals (generation 10). In addition, an across-generation core was defined by randomly selecting 20% of core animals from each of the five genotyped generations (only parents in generations 6 to 9). Core animals were also randomly selected from all genotyped animals (random) to make comparisons with previous studies. The restriction of using parents when selecting core animals from specific generations maintained consistency in the type of core animals among generations and replicates.

We considered additional factors to assess the utility of core definitions in less ideal situations. We investigated genotype accuracy as a source of variation potentially affecting the best core definition because genotype errors may impact the dimensionality of $\mathbf{G}$. Genotypes were modified to be 98% accurate to emulate imputed genotypes for all animals in generations 9 and 10. These modified genotypes were referred to as imputed genotypes throughout this study. Thus, imputed genotypes were core animals for some scenarios and non-core animals for others. The original genotypes were used in the eigenvalue decomposition to select

TABLE 1 Criteria for randomly selecting core animals for different core definitions in the Algorithm for Proven and Young

Core	Criteria for selection as core animals
Gen 6	Born in generation 6 with offspring in generation 7
Gen 7	Born in generation 7 with offspring in generation 8
Gen 8	Born in generation 8 with offspring in generation 9
Gen 9	Born in generation 9 with offspring in generation 10
Gen 10	Born in generation 10
All Gen	20% from each of generations 6 to 10, meets above criteria
Random	All animals in generations 6 to 10

the core size resulting in a smaller core size than using the imputed genotypes for eigenvalue decomposition.

For another scenario, we evaluated pedigree completeness for any interaction with the core definition. To investigate different ancestral pedigree depths for genotyped animals, 25% of animals were randomly selected from generations 1 to 5, and we removed their sires. These animals with unknown sires had phenotypes, and progeny were the closest possible genotyped relatives. In addition, we considered the consequences of genotyping animals with no pedigree information. We randomly removed both parents from 80% of genotyped animals.

2.4 | Validation

We modelled the simulated phenotype using an animal model with the overall mean as a fixed effect and direct additive genetic and residual as random effects. For validation, we assessed accuracy and bias for animals born in generation 10; these 10,000 animals had genotypes but no phenotypes. We measured accuracy as the correlation between TBV and EBV and bias as the regression of TBV on EBV. Also, we considered rounds to convergence using a convergence criterion of 10^{-12}. Within each analysis, we compared pairwise means for eight core definitions using Tukey's honest significant difference test (Tukey, 1949) to detect differences in accuracy, bias and rounds to convergence.

3 | RESULTS AND DISCUSSION

3.1 | Number of core animals

Accuracy and bias are presented in Figure 1 for different core sizes (numbers of largest eigenvalues for $\mathbf{G_0}$) and core definitions. Core definitions included individual generations (6 to 10), equal representation across generations and random. The core size is approximately 75% smaller when 90% instead of 98% of the variation in $\mathbf{G_0}$ is used. All scenarios were very accurate, and the accuracy may have resulted from the strong selection and corresponding large linkage disequilibrium in the simulation. For the larger core size (98%), accuracy and bias for APY are no different from traditional $\mathbf{G^{-1}}$ ($p > .05$) meaning solutions are robust to core definition. Within the APY core definitions, accuracies differ by <0.01, and biases differ by <0.03. The more recent single-generation core definitions typically had numerically greater accuracy than core definitions with older generations, and the random core definition was more accurate than any single-generation core definitions. For the smaller core size (90%), validation accuracies significantly decrease when core definitions are based on a single generation (6 to 9) or across generations when compared with traditional $\mathbf{G^{-1}}$ ($p \leq .05$). On average, accuracies are 0.06 less for APY with the smaller core size (90%) than for traditional $\mathbf{G^{-1}}$. A decrease in accuracy is expected because the smaller core size accounts for less variation in $\mathbf{G_0}$. A few core definitions do not differ from traditional $\mathbf{G^{-1}}$, but we expect them to differ with more replicates and greater power. Although accuracy is less for the smaller core size (90%), accuracies do not differ across the core definitions for APY with a range in accuracy of 0.02, and the greatest accuracy was for the random core definition. The smaller core size has no bias differences across the core definitions ($p > .05$) with a range of 0.06. Results are intermediate to those presented in Figure 1 when core size is associated with 95% of the variation in $\mathbf{G_0}$. Using fewer core animals in APY decreases accuracy but may not affect bias.

The mean (*SE*) numbers of largest eigenvalues (core sizes) explaining 98, 95 and 90% of the variation in $\mathbf{G_0}$ were 2521 (107), 1194 (69) and 603 (44), respectively. Each replicate used the number of eigenvalues calculated from the $\mathbf{G_0}$ for that specific replicate. Most of the genomic variation was contained in 2,000 of the 50,000 genotyped animals. Thus, instead of directly inverting a $\mathbf{G}$ with dimensions of 50,000, a small matrix can be inverted when calculating an APY $\mathbf{G^{-1}}$. The APY $\mathbf{G^{-1}}$ substantially reduces computing time and memory compared with $\mathbf{G^{-1}}$ (Masuda et al., 2016). When using APY in large, genotyped populations, breeders can implement ssGBLUP for a reasonable computational requirement.

Pocrnic, Lourenco, Masuda, Legarra, et al. (2016) suggested modifying formulas from Stam (1980) to make the core size a function of genome length and effective population size. Combining an effective population size (N_e) of 40 and a genome length (L) of 15 Morgans with their formulas, predictions are 2,400 (98%; $4N_eL$), 1,200 (95%; $2N_eL$) and 600 (90%; N_eL) largest eigenvalues in $\mathbf{G_0}$ depending on the amount of variation explained. The predictions are similar to the actual numbers of eigenvalues if theoretical effective population size is used. Realized effective population size underestimates the numbers of eigenvalues because the approximations were derived from

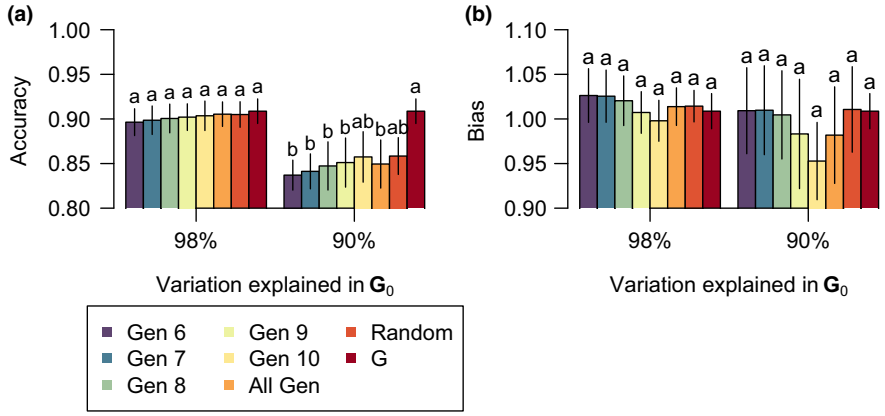

FIGURE 1 Accuracy (a) and bias (b) for traditional single-step genomic BLUP (G) and for different Algorithm for Proven and Young core definitions based on core sizes equal to the numbers of largest eigenvalues explaining 98 or 90% of the variation in $\mathbf{G_0}$. Accuracy was defined as the correlation between true and estimated breeding values. Bias was measured as the regression of true on estimated breeding value. Results with the same core size and no common letters differed significantly ($p \leq .05$). Error bars were ±2 *SE*

random mating populations but we included strong selection. In populations undergoing selection, theoretical effective population size is the better measure for predicting the numbers of eigenvalues to use as the core size.

EBV comparisons are important for practical implementation. For all replicates and core sizes, EBV correlations for all animals were >0.99 between ssGBLUP with traditional $\mathbf{G^{-1}}$ and APY with different core definitions. These outcomes differ from a previous study in which the EBV correlation decreased for some core definitions (Ostersen et al., 2016). These differences can result from the strong, single-trait selection in the simulation. On a population-wide scale, EBV from APY are comparable to traditional ssGBLUP. For validation animals, EBV correlations between methods follow the same pattern as accuracies. For sufficient core size (98%), correlations between APY and traditional ssGBLUP were >0.99 for all core definitions. Correlations for the smaller core size (90%) range from 0.91 to 0.94 and are slightly weaker ($r \geq .89$) than a simulation by Pocrnic, Lourenco, Masuda, Legarra, et al. (2016). Livestock populations are typically selected for multiple traits; so, correlations may be stronger because of less intensive selection in those populations.

The numbers of rounds to convergence were presented in Figure 2 for the core size associated with 98% of the variation in $\mathbf{G_0}$. Most core definitions had similar numbers of rounds as traditional $\mathbf{G^{-1}}$, but the number of rounds began to increase for generation 9 and doubled for the generation 10 core definition. In all analyses, the number of rounds displayed a similar pattern. Animals in generation 10 are young animals with genotypes, no phenotypes and no progeny. Animals in generation 9 have genotypes, phenotypes, genotyped progeny and no phenotyped progeny. The number of rounds also increased when young dairy cattle were used as core animals (Fragomeni, Lourenco, Tsuruta, Masuda, Aguilar, Legarra, et al., 2015). To avoid convergence problems in practice, core animals should not primarily consist of animals without phenotypes. In a previous study, all animals had genotypes and phenotypes, and the number of rounds was actually less for the young core definition (Ostersen et al., 2016). Possibly, numerical stability improves when the core includes animals with phenotypes and phenotyped progeny. In addition, convergence differences could be caused by slight changes in scaling of $\mathbf{G}$ with different core subsets in relation to the scaling of $\mathbf{A_{22}}$ in the default implementation in BLUPF90 (Misztal, et al. 2016).

3.2 | Imputation

Because genotype accuracy affects the dimensionality of genomic information (results not shown), we considered imputation as a contributing factor for selecting the core definition. When genotypes imputed with 98% accuracy are included, accuracy and bias did not differ ($p > .05$) across core definitions. Accuracies (*SE*) ranged from 0.89 (0.01) to 0.90 (0.01), and biases (*SE*) ranged from 0.99 (0.01) to 1.04 (0.02). No differences occur despite generations 9 and 10 ($n = 20{,}000$) having accurately imputed genotypes and being used as core animals. Thus, the best core definition is not affected by the presence of accurately imputed genotypes in simulation. In practice, any core differences will be smaller because eigenvalue decomposition will be used for the imputed not the actual genotypes and core size will increase. The small amount of genotype errors do not dramatically affect the dimensionality of the genomic information or independent chromosome segments based on the accuracies. If imputation accuracy is <98%, including those imputed genotypes as core animals might affect EBV. Imputation is increasing because of the cost-effectiveness

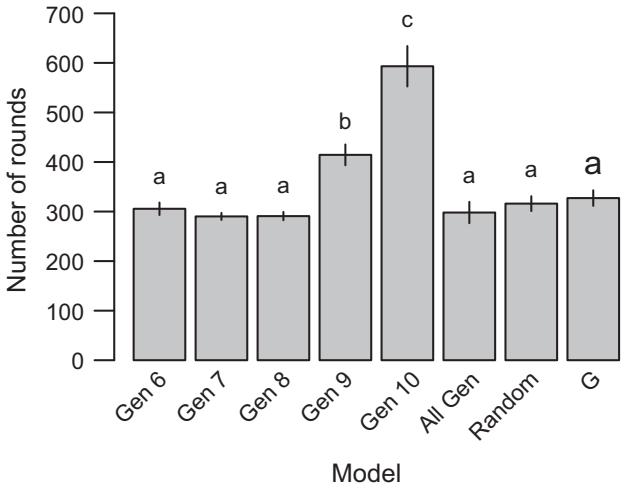

FIGURE 2 Numbers of rounds to convergence (10^{-12}) for traditional single-step genomic BLUP (G) and different Algorithm for Proven and Young core definitions. Results with no common letters differed significantly ($p \leq .001$). Error bars were ±2 *SE*

of low-density genotyping panels. The importance of imputation needs to be studied in livestock populations because including imputed genotypes as young, core animals previously affected EBV (Ostersen et al., 2016). Nonetheless, the current study finds no effect of imputation on the core definition.

3.3 | Incomplete pedigree

We examined two scenarios with incomplete pedigree information and found different conclusions. The first scenario was incomplete ancestral pedigrees that created different pedigree depths for genotyped animals. We altered pedigree depths by removing sires for 25% of non-genotyped animals. Incomplete ancestral pedigrees do not affect accuracy or bias for different core definitions ($p > .05$). Accuracies (*SE*) ranged from 0.90 (0.01) to 0.91 (0.01), and biases (*SE*) ranged from 0.99 (0.01) to 1.02 (0.02). Thus, incomplete ancestral pedigrees do not affect EBV when using different core definitions. Core definitions should be robust across species with different degrees of pedigree depth.

Conversely, the core definition matters when most genotyped animals have unknown parents (Figure 3). Accuracy is less and bias is greater than traditional $\mathbf{G}^{-1}$ for single-generation core definitions ($p \leq .05$). Random core definitions perform well as expected from previous research (Fragomeni, Lourenco, Tsuruta, Masuda, Aguilar, Legarra, et al., 2015; Lourenco et al., 2015; Masuda et al., 2016; Ostersen et al., 2016). In addition, the across-generation core definition is as accurate as the random core definition and traditional $\mathbf{G}^{-1}$. The mean accuracy for random and across-generation cores is 0.76 compared with a mean

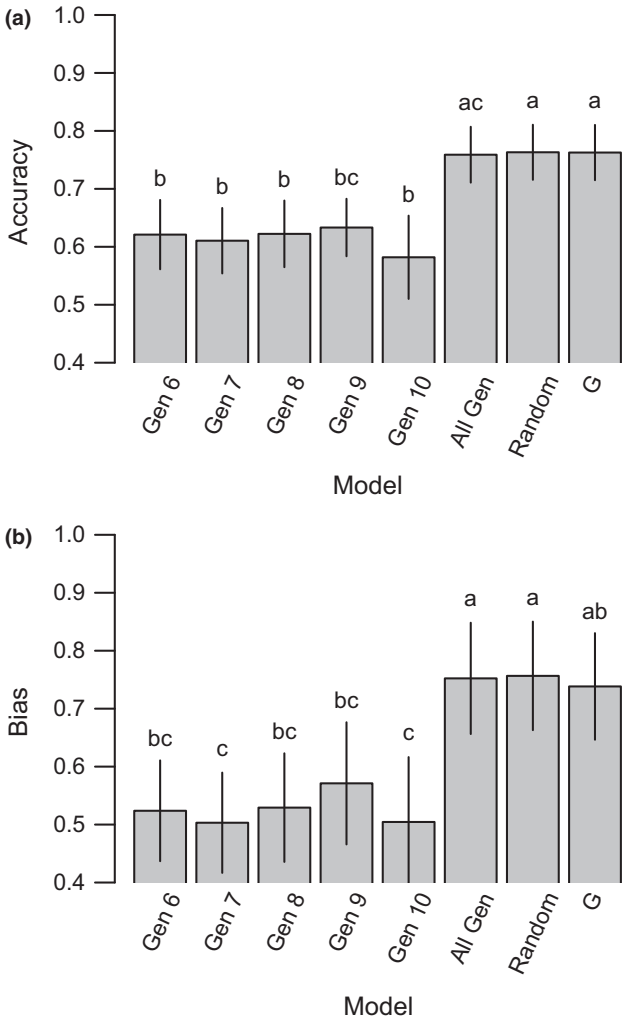

FIGURE 3 Accuracy (a) and bias (b) for traditional single-step genomic BLUP (G) and different Algorithm for Proven and Young core definitions when genotyped animals had unknown parents. Accuracy was defined as the correlation between true and estimated breeding values. Bias was measured as the regression of true on estimated breeding value. Results with no common letters differed significantly ($p \leq .05$). Error bars were ±2 *SE*

accuracy of 0.61 for single-generation cores (range 0.05). Correlations between EBV from the two methods follow a similar pattern with random and across-generation cores >0.99 and single-generation cores ranging from 0.89 to 0.93. We consider this accuracy difference to be meaningful and recommend the use of multigenerational core definitions (random or equal representation across generations). These results indicate that the random core definition is effective because the core animals represent multiple generations. Interestingly, core definitions including 2 or 3 generations increase accuracy but are still numerically less accurate than using all generations. The across-generation core definition would be applicable for species with multi-sire breeding cohorts or no pedigrees for genotyped

animals. Again, differences between core definitions can be attributed to differences in scaling of **G**.

3.4 | Interpretation

The simulation assumptions affect the dimensionality of $\mathbf{G_0}$ as the simulation has more genotyped animals (50,000) than the number of SNP (<25,500). In livestock populations, medium-density genotyping is common (~50,000 SNP), and APY is needed when the number of genotyped animals (~100,000 to 150,000) is at least twice the number of SNP for these populations. We expect 2,400 independent chromosome segments (Stam, 1980) in this population. Our number of SNP is 9 to 10 times greater than the number of independent chromosome segments, which is less than the 12 times needed to capture all the junctions between segments (MacLeod, Haley, Woolliams, & Stam, 2005). Thus, either the number of SNP or the number of independent chromosome segments limits the dimensionality of $\mathbf{G_0}$. Doubling the genome size would cause a smaller proportional increase in the number of largest eigenvalues. Our conclusions are not expected to change with different simulation parameters because our core sizes would account for a large percentage of the variation in $\mathbf{G_0}$.

Given the simulated scenario with selection, the generational core definitions are robust even for smaller core size. For the five generational core definitions, no pairwise comparisons differ for accuracy or bias in any scenario. Accuracy does not decrease as the relationships between core and validation animals decrease as previously proposed (Ostersen et al., 2016). Potentially, the independent chromosome segments present in generation 6 are applicable for generation 10. The accuracies indicate that the same core definition can be used for multiple generations unless pedigrees are incomplete. With incomplete pedigrees, across-generation core definitions may better represent the independent chromosome segments in the core animals. Because these differences are not seen in the other scenarios, the results are more likely caused by the genomic relationships between core animals correcting for the lack of pedigree connectedness across generations. In data sets with incomplete pedigree, metafounders can be used to better account for the missing pedigree relationships and need to be investigated (Legarra, Christensen, Vitezica, Aguilar, & Misztal, 2015).

Accuracy differences are expected for generational core definitions based on the research by Ostersen et al. (2016). When comparing traditional and APY ssGBLUP, the EBV correlations were least with old or young core definitions. The core size can affect their conclusions as the study was published concurrently to the implementation of eigenvalue decomposition for core size. Their core size was approximately 90 or 95% of the variation in a different commercial swine population with similar number of genotyped animals and SNP (Pocrnic, Lourenco, Masuda, & Misztal, 2016). The EBV correlations were similar for the two studies when comparing traditional and APY ssGBLUP with a random core definition (Ostersen et al., 2016; Pocrnic, Lourenco, Masuda, & Misztal, 2016). In practice, more core animals can increase correlations because computation time was reasonable and the core size was smaller than the optimal number of eigenvalues explaining 98% of the variation in $\mathbf{G_0}$.

4 | CONCLUSIONS

The core definition is robust to the core size, accurate imputation and incomplete ancestral pedigree. The core definitions become more important when genotyped animals have incomplete pedigrees. When genotyped animals have unknown parents, the core definition is more important, and the core needs to include multiple generations to maintain accuracy and unbiasedness. In this scenario, random or across generation core definitions are appropriate to include all generations. These ideas need to be applied to livestock populations, particularly those with incomplete pedigrees to assess accuracy changes with different core definitions.

REFERENCES

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

Christensen, O. F., & Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. *Genetics Selection Evolution*, *42*, 1. https://doi.org/10.1186/1297-9686-42-2

Falconer, D. S., & Mackay, T. F. C. (1996). *Introduction to quantitative genetics*, 4th ed. Harlow, Essex, UK: Longmans Green.

Fragomeni, B., Lourenco, D., Tsuruta, S., Masuda, Y., Aguilar, I., Legarra, A., … Misztal, I. (2015). Hot topic: Use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes. *Journal of Dairy Science*, *98*, 4090-4094. https://doi.org/10.3168/jds.2014-9125

Fragomeni, B., Lourenco, D., Tsuruta, S., Masuda, Y., Aguilar, I., & Misztal, I. (2015). Use of genomic recursions and algorithm for proven and young animals for single-step genomic BLUP analyses–a simulation study. *Journal of Animal Breeding and Genetics*, *132*, 340–345. https://doi.org/10.1111/jbg.12161

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

Lourenco, D. A., Misztal, I., Tsuruta, S., Fragomeni, B. D., Aguilar, I., Masuda, Y., & Moser, D. (2015). Direct and indirect genomic

evaluations in beef cattle. Orlando, FL: Interbull Bulletin No. 49. (pp. 80–84).

MacLeod, A., Haley, C., Woolliams, J., & Stam, P. (2005). Marker densities and the mapping of ancestral junctions. *Genetical Research*, *85*, 69–79. https://doi.org/10.1017/s0016672305007329

Masuda, Y., Misztal, I., Tsuruta, S., Legarra, A., Aguilar, I., Lourenco, D., . . . Lawlor, T. (2016). Implementation of genomic recursions in single-step genomic best linear unbiased predictor for US Holsteins with a large number of genotyped animals. *Journal of Dairy Science*, *99*, 1968–1974. https://doi.org/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. https://doi.org/10.1534/genetics.115.182089

Misztal, I., Legarra, A., & Aguilar, I. (2014). Using recursion to compute the inverse of the genomic relationship matrix. *Journal of Dairy Science*, *97*, 3943–3952. https://doi.org/10.3168/jds.2013-7752

Misztal, I., Tsuruta, S., Lourenco, D. A. L., Masuda, Y., Aguilar, I., Legarra, A., & Vitezica, Z. (2016). *Manual for BLUPF90 family of programs*. Retrieved from http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all5.pdf (Accessed 9 November 2016).

Muir, W. (2007). Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *Journal of Animal Breeding and Genetics*, *124*, 342–355. https://doi.org/10.1111/j.1439-0388.2007.00700.x

Ostersen, T., Christensen, O. F., Madsen, P., & Henryon, M. (2016). Sparse single-step method for genomic evaluation in pigs. *Genetics Selection Evolution*, *48*, 48. https://doi.org/10.1186/s12711-016-0227-8

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

Pocrnic, I., Lourenco, D. A., Masuda, Y., & Misztal, I. (2016). Dimensionality of genomic information and performance of the Algorithm for Proven and Young for different livestock species. *Genetics Selection Evolution*, *48*, 82. https://doi.org/10.1186/s12711-016-0261-6

Saatchi, M., McClure, M. C., McKay, S. D., Rolf, M. M., Kim, J., Decker, J. E., . . . Taylor J. F. (2011). Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. *Genetics Selection Evolution*, *43*, 40. https://doi.org/10.1186/1297-9686-43-40

Sargolzaei, M., & Schenkel, F. S. (2009). QMSim: A large-scale genome simulator for livestock. *Bioinformatics*, *25*, 680–681. https://doi.org/10.1093/bioinformatics/btp045

Stam, P. (1980). The distribution of the fraction of the genome identical by descent in finite random mating populations. *Genetical Research*, *35*, 131–155. https://doi.org/10.1017/s0016672300014002

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

VanRaden, P. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*, 4414–4423. https://doi.org/10.3168/jds.2007-0980

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, *16*, 97–159.

How to cite this article: Bradford HL, Pocrnić I, Fragomeni BO, Lourenco DAL, Misztal I. Selection of core animals in the Algorithm for Proven and Young using a simulation model. *J Anim Breed Genet*. 2017;134:545–552. https://doi.org/10.1111/jbg.12276