Journal of Animal Science, 2020, Vol. 98, No. 6, 1–9

doi:10.1093/jas/skaa154
Advance Access publication May 6, 2020
Received: 11 February 2020 and **Accepted:** 30 April 2020
Animal Genetics and Genomics

ANIMAL GENETICS AND GENOMICS

Indirect predictions with a large number of genotyped animals using the algorithm for proven and young

Andre L.S. Garcia,[†,1] Yutaka Masuda,[†] Shogo Tsuruta,[†] Stephen Miller,[‡] Ignacy Misztal,[†] and Daniela Lourenco[†]

[†]Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602, [‡]Angus Genetics Inc. St. Joseph, MO 64506

[1]Corresponding author: andre.garcia@uga.edu

ORCiD numbers: 0000-0001-9778-7978 (A.L.S. Garcia); 0000-0002-3428-6284 (Y. Masuda); 0000-0002-6897-6363 (S. Tsuruta); 0000-0002-0382-1897 (I. Misztal); 0000-0003-3140-1002 (D. Lourenco).

Abstract

Reliable single-nucleotide polymorphisms (**SNP**) effects from genomic best linear unbiased prediction BLUP (**GBLUP**) and single-step GBLUP (**ssGBLUP**) are needed to calculate indirect predictions (**IP**) for young genotyped animals and animals not included in official evaluations. Obtaining reliable SNP effects and IP requires a minimum number of animals and when a large number of genotyped animals are available, the algorithm for proven and young (**APY**) may be needed. Thus, the objectives of this study were to evaluate IP with an increasingly larger number of genotyped animals and to determine the minimum number of animals needed to compute reliable SNP effects and IP. Genotypes and phenotypes for birth weight, weaning weight, and postweaning gain were provided by the American Angus Association. The number of animals with phenotypes was more than 3.8 million. Genotyped animals were assigned to three cumulative year-classes: born until 2013 (N = 114,937), born until 2014 (N = 183,847), and born until 2015 (N = 280,506). A three-trait model was fitted using the APY algorithm with 19,021 core animals under two scenarios: 1) core 2013 (random sample of animals born until 2013) used for all year-classes and 2) core 2014 (random sample of animals born until 2014) used for year-class 2014 and core 2015 (random sample of animals born until 2015) used for year-class 2015. GBLUP used phenotypes from genotyped animals only, whereas ssGBLUP used all available phenotypes. SNP effects were predicted using genomic estimated breeding values (**GEBV**) from either all genotyped animals or only core animals. The correlations between GEBV from GBLUP and IP obtained using SNP effects from core 2013 were ≥0.99 for animals born in 2013 but as low as 0.07 for animals born in 2014 and 2015. Conversely, the correlations between GEBV from ssGBLUP and IP were ≥0.99 for animals born in all years. IP predictive abilities computed with GEBV from ssGBLUP and SNP predictions based on only core animals were as high as those based on all genotyped animals. The correlations between GEBV and IP from ssGBLUP were ≥0.76, ≥0.90, and ≥0.98 when SNP effects were computed using 2k, 5k, and 15k core animals. Suitable IP based on GEBV from GBLUP can be obtained when SNP predictions are based on an appropriate number of core animals, but a considerable decline in IP accuracy can occur in subsequent years. Conversely, IP from ssGBLUP based on large numbers of phenotypes from non-genotyped animals have persistent accuracy over time.

Key words: interim evaluations, persistence, genomic selection

Abbreviations

APY	algorithm for proven and young
BLUP	best linear unbiased prediction
BW	birth weight
DNA	deoxyribonucleic acid
GBLUP	genomic best linear unbiased prediction
GEBV	genomic estimated breeding values
IP	indirect predictions
PWG	postweaning gain
SNP	single nucleotide polymorphism
SNP-BLUP	single-nucleotide polymorphism best linear unbiased prediction
ssGBLUP	single-step genomic best linear unbiased prediction
WW	weaning weight

Introduction

The availability of dense single-nucleotide polymorphism (**SNP**) panels has allowed the implementation of genomic selection in many livestock species. Once the deoxyribonucleic acid (**DNA**) markers are available, methods such as SNP-best linear unbiased prediction (**SNP-BLUP**), genomic BLUP (**GBLUP**), and single-step genomic BLUP (**ssGBLUP**) can be used to obtain genomic predictions (Meuwissen et al., 2001; Aguilar et al., 2010; Christensen and Lund, 2010).

The number of genotyped animals has steadily increased as genomic selection became widespread and genotyping costs decreased. The U.S. dairy industry now has more than 3 million genotyped animals (queries.uscdcb.com/Genotype/cur_density.html) and the American Angus Association has more than 750,000 genotyped animals (Steve Miller, Angus Genetics Inc., personal communication). When GBLUP and ssGBLUP are used for such large genomic datasets, the computing cost becomes an issue because inverting the genomic relationship matrix (**G**) has a cubic cost with the number of genotyped animals, making it unfeasible for datasets with more than 150,000 genotyped animals (Fragomeni et al., 2015). To solve this problem, Misztal et al. (2014a) proposed the algorithm for proven and young (**APY**). The APY algorithm divides the set of genotyped animals into core and noncore animals such that direct inversion is needed only for core animals in the genomic relationship matrix and the remaining components are obtained recursively, dramatically reducing the computing costs.

Even with appropriate tools, the addition of newly genotyped animals will increase computing time on routine evaluations which can increase the time between collection of DNA samples and obtaining genomic predictions (Wiggans et al., 2015). This time period is important because most genotypes come from young animals and producers rely on genomic predictions for culling purposes. Being able to quickly decide which animals to keep and which ones to cull will potentially decrease rearing costs at the farm level (Nicolazzi et al., 2018). Genomic predictions are also important for commercial producers to make more accurate management decisions. One example is the utilization of genomic predictions for commercial non-registered Angus females marketed as "GeneMax Advantage."

One common issue in the genomic era is that often animals are genotyped before phenotypes are collected, and sometimes pedigree information is missing. These animals may not contribute with information to official genomic estimated breeding values (**GEBV**) and their inclusion may decrease the accuracy and increase inflation of GEBV (Bradford et al., 2017, 2019). If SNP predictions are available, indirect predictions (**IP**) can be used as interim GEBV providing quick genomic predictions for newly genotyped and non-registered animals, without affecting routine evaluations (Lourenco et al., 2015).

Two commonly used procedures for genomic evaluation are SNP-BLUP and GBLUP. While SNP-BLUP fits SNP effects as random effects with a diagonal (co)variance structure, GBLUP fits breeding values as random effects and uses the genomic relationship matrix as the (co)variance structure. Both procedures assume that SNP markers account for all the genetic variance and the random effects are assumed to be normally distributed with mean zero and (co)variance structure as described above for each model. Because of these assumptions, SNP-BLUP and GBLUP yield identical GEBV.

Because of the equivalence between SNP-BLUP and GBLUP, SNP effects can be predicted based on GEBV and the inverse of **G** ($\mathbf{G}^{-1}$) for genotyped animals both in GBLUP (VanRaden, 2008; Strandén and Garrick, 2009) and in ssGBLUP (Wang et al., 2012). Because backsolving for SNP effects from GEBV involves $\mathbf{G}^{-1}$, using all genotyped animals to predict SNP may be prohibitive, thus tools such as APY (Misztal et al., 2014a) can help to overcome this limitation. Lourenco et al. (2018) investigated IP from ssGBLUP using almost 81,000 genotyped animals from the American Angus Association. Their results show that accurate IP can be obtained from ssGBLUP with $\mathbf{G}^{-1}$ calculated using APY or with only a set of core animals. Although their study showed the feasibility of obtaining IP from ssGBLUP with APY, the number of genotyped animals used was small compared with the current database, and the impact of adding new genotypes was not investigated. Therefore, the objectives of this study were: 1) to test the stability of IP and check if the core group should be updated when large numbers of genotyped animals are added to the database; 2) to investigate the choice of core animals used to obtain SNP effects used to compute IP, that is, whether all genotyped animals or only core genotyped animals should be used; and 3) to determine the minimum number of genotyped animals needed to obtain reliable IP based on SNP effects from GBLUP and ssGBLUP.

Materials and Methods

Animal care and use committee approval was not needed because data were obtained from an existing database.

Data and model

The dataset used in the study was provided by the American Angus Association. Phenotypes were available for birth weight (**BW**; N = 7,574,765), weaning weight (**WW**; N = 8,302,222), and postweaning gain (**PWG**; N = 4,145,166). The pedigree file included 9,145,109 animals, from which 280,506 animals born until 2015 had 39,774 genotyped SNP markers after quality control.

The following three-trait model was used:

$$\mathbf{y}_t = \mathbf{X}\mathbf{b}_t + \mathbf{W}_1\mathbf{u}_t + \mathbf{W}_2\mathbf{mat}_t + \mathbf{W}_3\mathbf{mpe}_t + \mathbf{e}_t \quad (1)$$

Where **t** refers to traits (BW, WW, and PWG); $\mathbf{y}_t$ is the vector of phenotypes for trait **t**; $\mathbf{b}_t$ is the vector of fixed contemporary group effects; $\mathbf{u}_t$, $\mathbf{mat}_t$, and $\mathbf{mpe}_t$ are the vectors of random effects for additive genetic direct, maternal, and maternal permanent environmental effects, respectively; $\mathbf{e}_t$ is the vector of residuals. The $\mathbf{X}$, $\mathbf{W}_1$, $\mathbf{W}_2$, and $\mathbf{W}_3$ are incidence matrices for the effects in $\mathbf{b}_t$, $\mathbf{u}_t$, $\mathbf{mat}_t$, and $\mathbf{mpe}_t$, respectively. All random effects were

present for WW, but only $\mathbf{u}_t$, $\mathbf{mat}_t$, and $\mathbf{e}_t$ for BW, and $\mathbf{u}_t$ and $\mathbf{e}_t$ for PWG.

Genomic analyses

Genomic BLUP provides a simple framework to test IP because when SNP effects are predicted using GEBV from GBLUP, the IP will be on the same scale as the GEBV. The same is true for ssGBLUP, although a mean has to be added to IP to account for the tuning of **G** to match **A** (Lourenco et al., 2018). Genomic analyses were performed using GBLUP and ssGBLUP procedures, although the process for obtaining IP on the same scale as GEBV with ssGBLUP is still under investigation. A constant reflecting the average GEBV in the population used to predict SNP effects can be added to IP to match the scale of GEBV (Legarra et al., 2018; Lourenco et al., 2018); hence, correlations were not affected.

The ssGBLUP inverse of the relationship matrix combining pedigree and genomic relationships ($\mathbf{H}^{-1}$) was constructed as in Aguilar et al. (2010):

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix} \quad (2)$$

where $\mathbf{G}^{-1}$ is the inverse of the genomic relationship matrix, and $\mathbf{A}_{22}^{-1}$ is the inverse pedigree relationship matrix for genotyped animals.

The initial genomic relationship matrix ($\mathbf{G}_0$) for GBLUP and ssGBLUP was constructed following VanRaden (2008):

$$\mathbf{G}_0 = \frac{\mathbf{ZZ}'}{2\sum p_i \quad (1-p_i)} \quad (3)$$

where **Z** is a matrix of centered gene contents and $\mathbf{p}_i$ is the minor allele frequency of SNP i. Allele frequencies were calculated based on current genotypes. To avoid singularity problems, the matrix **G** was computed as $\mathbf{G} = 0.99\mathbf{G}_0 + 001\mathbf{I}$ for GBLUP and ssGBLUP. The impact of other blending proportions, such as $\mathbf{G} = 0.01\mathbf{A}_{22}$, $0.05\mathbf{A}_{22}$ and $0.1\ \mathbf{A}_{22}$, was also investigated for ssGBLUP.

The large number of genotyped animals used in the present study made direct inversion of **G** unfeasible. Thus, APY was used to compute the inverse of **G** ($\mathbf{G}_{APY}^{-1}$) as proposed by Misztal et al. (2014a) and Misztal (2016). In APY, the matrix of genomic relationships among genotyped animals is partitioned based on core animals (c) and noncore animals (n):

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{cc} & \mathbf{G}_{cn} \\ \mathbf{G}_{nc} & \mathbf{G}_{nn} \end{bmatrix} \quad (4)$$

And $\mathbf{G}_{APY}^{-1}$ is calculated as follows:

$$\mathbf{G}_{APY}^{-1} = \begin{bmatrix} \mathbf{G}_{cc}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -\mathbf{G}_{cc}^{-1}\mathbf{G}_{cn} \\ \mathbf{I} \end{bmatrix} \mathbf{M}_{nn}^{-1} \begin{bmatrix} -\mathbf{G}_{nc}\mathbf{G}_{cc}^{-1} & \mathbf{I} \end{bmatrix} \quad (5)$$

where each element of $\mathbf{M}_{nn}$ for the *ith* noncore animal is computed as follows:

$$m_{nn,i} = g_{ii} - \mathbf{G}_{ic}\mathbf{G}_{cc}^{-1}\mathbf{G}_{ci} \quad (6)$$

The only direct inversion needed for APY is the part of **G** containing relationships among core animals. All other components are obtained through recursions.

Pocrnic et al. (2016b) showed that the number of core animals can be determined as the number of eigenvalues explaining 98% to 99% of the variance of $\mathbf{G}_0$. The eigenvalue decomposition of $\mathbf{G}_0$ is computationally more expensive than the equivalent singular value decomposition of **Z**, thus eigenvalues were obtained as the square of singular values of **Z**. The resulting number of core animals corresponding to 99% of the variance was 19,021 animals. This number of core animals was used in this study. This number also corresponds to the number of core animals used in routine evaluations by the American Angus Association.

SNP effects, IP, and validation

To evaluate the impact of increasing the number of genotyped animals in the prediction of SNP effects used to compute IP, the genotyped animals were divided into three year-classes mimicking the increase of the number of animals each year: 1) genotyped animals born until 2013 (N = 114,937); 2) genotyped animals born until 2014 (N = 183,847; includes 68,910 genotyped animals born in 2014); and 3) genotyped animals born until 2015 (N = 280,506; all genotyped animals; includes 96,659 genotyped animals born in 2015). The heritabilities and numbers of records per year-class for BW, WW, and PWG included in GBLUP and ssGBLUP are presented in Table 1.

While the number of core animals remained the same in all analyses (19,021), two core definitions for APY were tested:

1) Core 2013: core animals were randomly sampled from animals born up to 2013 and remained the same across all year-classes;
2) Core 2014 and core 2015: core animals were randomly sampled from animals born until 2014 and 2015 for year-classes 2014 and 2015, respectively.

After the core groups were defined, GEBV were calculated using GBLUP and ssGBLUP with APY for each year-class dataset. Then, SNP effects were obtained with either $\mathbf{G}_{APY}^{-1}$ or $\mathbf{G}^{-1}$ only for core animals ($\mathbf{G}_{core}^{-1}$), using the formula derived by Wang et al. (2012):

$$\hat{\mathbf{a}}_{Full} = \lambda\ \mathbf{DZ}'\mathbf{G}_{APY}^{-1}\hat{\mathbf{u}} \quad (7)$$

$$\hat{\mathbf{a}}_{core} = \lambda\ \mathbf{DZ}'_{core}\mathbf{G}_{core}^{-1}\hat{\mathbf{u}}_{core} \quad (8)$$

where $\hat{\mathbf{a}}$ is a vector of SNP effects; $\hat{\mathbf{u}}$ is a vector of GEBV for all genotyped animals; $\hat{\mathbf{u}}_{core}$ is a vector of GEBV for core animals; λ

Table 1. Heritabilities and number of records per year-class for each trait included in ssGBLUP and GBLUP

		ssGBLUP			GBLUP		
Trait	H[1]	2013	2014	2015	2013	2014	2015
BW	0.42	6,944,152	7,250,456	7,574,765	73,850	120,389	188,241
WW	0.20	7,659,259	7,972,273	8,302,222	75,428	122,838	191,792
PWG	0.24	3,835,752	3,985,075	4,145,166	56,254	91,422	140,975

[1]H, heritability.

is the ratio of SNP variance to additive genetic variance, **D** is a diagonal matrix of SNP weights (**D** = **I** in this study), **Z** and $\mathbf{Z}_{core}$ are matrices of centered gene contents for all genotyped animals and for core animals only, respectively, and $\mathbf{G}_{APY}^{-1}$ and $\mathbf{G}_{core}^{-1}$ are genomic relationship matrices for all genotyped animals (computed using APY) and for core animals only, respectively.

Once SNP effects became available, IP were calculated as follows:

$$\mathbf{IP}_{Full} = \mathbf{Z}\hat{a}_{Full} \quad (9)$$

$$\mathbf{IP}_{core} = \mathbf{Z}\,\hat{a}_{core} \quad (10)$$

Where **Z** is the centered gene content matrix for all genotyped animals within each year-class.

In the GBLUP context, $\hat{\mathbf{u}}|\hat{\mathbf{a}} = \mathbf{Z}\,\hat{\mathbf{a}}$. Thus, to assess whether we were able to retrieve GEBV, given that SNP effects are known, the correlation between IP (i.e., $\mathbf{IP}_{Full}$ and $\mathbf{IP}_{core}$) and GEBV was calculated for each year-class and core definition.

Although these correlations measure the ability to retrieve the GEBV using SNP effects for the same animals, IP are typically computed for young genotyped animals not included in the genomic evaluations. Thus, we also performed a validation study using genotyped animals born in 2016 (*N* = 54,997), as validation animals. These animals had genotypes and phenotypes for all traits; however, their data were not included in previous analyses. IP for validation animals were computed for each year-class and core definition using SNP effects previously obtained with GBLUP and ssGBLUP procedures. The predictive ability was calculated as the correlation between adjusted phenotypes (based on traditional BLUP with full data) and IP for validation animals.

Minimum number of genotyped animals needed to compute reliable IP

To investigate the minimum number of genotyped animals needed to predict SNP effects while keeping correlations between IP and GEBV >0.99, we randomly assigned genotyped animals from the complete set (280,506) to subsets ranging from 500 to 40,000 animals (i.e., 500, 1k, 2k, 3k, 4k, 5k, 10k, 15k, 20k, 30k, and 40k). Once the subsets were created, SNP effects were computed as follows:

$$\hat{\mathbf{a}}_{subset} = \lambda\,\mathbf{D}\mathbf{Z}'_{subset}\mathbf{G}^{-1}_{subset}\hat{\mathbf{u}}_{subset} \quad (11)$$

where $\mathbf{G}^{-1}_{subset}$ is the direct $\mathbf{G}^{-1}$ computed for each subset of genotyped animals and $\hat{\mathbf{u}}_{subset}$ is a vector of GEBV for the subset animals. GEBV were computed using ssGBLUP with APY based on all genotyped animals and core 2013. IP were then calculated for all genotyped animals as $\mathbf{IP}_{subset} = \mathbf{Z}_{subset}\,\hat{\mathbf{a}}_{subset}$. Subsequently, the correlations between IP and GEBV were obtained for each subset.

All analyses were performed using software from the BLUPF90 family of programs (Misztal et al., 2014b) and in-house bash and R (R Core Team, 2019) scripts.

Results and Discussion

Genomic estimated breeding values

The correlations between GEBV across core definitions using all genotyped animals (year-class 2015) were ≥0.99 for all traits, which indicates that the changes in GEBV arising from APY computations with different core definitions were minimal. Previous studies with simulated and actual datasets investigated the changes in GEBV when using APY found that as long as the number of core animals reflects the dimensionality of the genomic information (i.e., number of eigenvalues explaining at least 98% of the variance of **G**), the choice of core animals is arbitrary (Fragomeni et al., 2015; Masuda et al., 2016; Bradford et al., 2017).

IP with G_{APY}^{-1} and G_{core}^{-1}

When $\mathbf{G}_{APY}^{-1}$ was used, the correlations between $\mathbf{IP}_{Full}$ and GEBV computed with ssGBLUP (Table 2) and GBLUP (Table 3) were ≥0.96 for all traits and scenarios.

Although the number of core animals remained constant, the number of noncore animals in $\mathbf{G}_{APY}^{-1}$ increased with the addition of new genotyped animals (i.e., different year-classes). Our results show that as long as the number of core animals represents the dimensionality of the genomic information, APY yields robust IP under ssGBLUP and GBLUP, regardless of core definition and addition of a large number of genotyped animals.

However, when SNP effects were predicted using $\mathbf{G}_{core}^{-1}$, the results from ssGBLUP and GBLUP differed. While the correlations between $\mathbf{IP}_{core}$ and GEBV were ≥0.99 for ssGBLUP regardless of core definition (Table 2), there was a dramatic decrease in the correlations values between $\mathbf{IP}_{core}$ and GEBV from GBLUP when core 2013 was used (Table 3). The correlations decreased from 0.99 to 0.64 for BW, from 0.99 to 0.12 for WW, and from 0.99 to 0.07 for PWG. When core animals were chosen from the recent population (i.e., core 2014 and core 2015), correlations were restored to 0.99 (Table 3). Although in both cases GEBV were computed using APY with all genotyped animals, SNP effects and IP were computed based on $\mathbf{G}^{-1}$ that contained only relationships for core animals. In this case, the backsolving process used only a portion of the equations. Core 2013 represented a population of 114,937 genotyped animals, whereas core 2015 was a random sample from the complete set of 280,506 genotyped animals.

Table 2. Correlations between IP and GEBV computed using ssGBLUP with all genotyped animals (IP_{Full}) and only core animals (IP_{core}) for all year-classes and core definitions

		BW		WW		PWG	
Core definition	Year-class	IP_{Full}	IP_{core}	IP_{Full}	IP_{core}	IP_{Full}	IP_{Core}
2013	2013[1]	0.98	0.99	0.99	1.00	0.99	1.00
	2014	0.97	0.99	0.99	1.00	0.99	1.00
	2015	0.96	0.99	0.99	1.00	0.99	1.00
2014	2014	0.96	0.99	0.99	1.00	0.99	1.00
2015	2015	0.98	0.99	0.99	1.00	0.99	1.00

[1]Results from year-class 2013 are the same in both core definitions.

Using core 2013 to compute IP for all animals born until 2015 may not reflect the current state of the population under GBLUP. On the other hand, the fact that ssGBLUP uses much more data than GBLUP may have contributed to a more robust GEBV and, therefore, more reliable SNP effects and IP.

Pocrnic et al. (2019) investigated the accuracy of GBLUP in terms of the number of eigenvalues of the genomic relationship matrix. Using simulated populations with 3k, 6k, or 12k genotyped animals, they found that in smaller populations (i.e., small effective population size), eliminating 90% of the smallest eigenvalues from **G** did not reduce the accuracy, because the 10% largest eigenvalues were capable of explaining 90% of the variation in **G**. However, in larger populations with more phenotypic records, including more eigenvalues from **G** increased the accuracy of GEBV. Further, these authors obtained similar accuracies using **G** with a restricted number of eigenvalues or an equivalent number of core animals in the APY algorithm (from 15 to 1,215), confirming that the number of eigenvalues of **G** can be used as a proxy for the number of core animals in APY as previously reported (Pocrnic et al., 2016a, 2016b). Pocrnic et al. (2019) added that the 10% largest eigenvalues of **G** represent many chromosome segments and once they are accounted for, accuracies are similar across different population sizes. In a small genotyped population, only a small number of large eigenvalues could be estimated which may be sufficient to explain a reasonable large fraction of the genetic variation in **G** and yield intermediate accuracies but further increases in accuracy would require additional genotyped animals to estimate the remaining smaller eigenvalues. The clusters of chromosome segments accounted for in small datasets may differ in future generations, leading to low persistence of predictions. On the other hand, when datasets are sufficiently large to estimate nearly all eigenvalues and indirectly, chromosome segments, prediction persistency is likely to increase. Similar accuracies with the same number of eigenvalues or core animals suggest that such groups of animals contain information on almost the same chromosome segments as those captured by the largest eigenvalues of **G**.

The decrease in correlation between GEBV and IP with $\mathbf{G}^{-1}_{APY}$ and $\mathbf{G}^{-1}_{core}$ was also reflected by the correlations between SNP effects. The correlations between SNP effects under core 13 decreased from year-class 2013 to 2015 for both genomic procedures, but the decrease was much smaller for ssGBLUP than for GBLUP (Table 4). For example, the correlations for PWG decreased from 0.92 to 0.88 (0.04 points) in ssGBLUP, but from 0.95 to 0.73 (0.22 points) in GBLUP. Under core 2014 and core 2015 scenarios, the correlations between SNP effects were very similar between the two genomic procedures. There was a small decrease across year-classes, but this decrease was much smaller than the one observed with the core 2013 scenario, especially for GBLUP (Table 4). The correlations in Table 4 show that GBLUP prediction of SNP effects with core 2013 did not improve with the addition of more genotyped animals.

Although a decrease in the values of correlations between $\mathbf{IP}_{core}$ and GEBV using core 2013 under GBLUP was observed for all traits, BW seemed to be more persistent (Table 3). This could be because of heritability and selection intensity. The heritability for BW was almost twice the heritability values of WW and PWG (Table 1). With higher heritabilities, more eigenvalues of small effect are accounted for and their information contributes to higher accuracy of GEBV (Pocrnic et al., 2019), or as in our study, to higher persistence of IP.

In a study with layer chickens, Wolc et al. (2011) showed that traits with higher heritability had more persistent accuracies across generations as opposed to lowly heritable traits.

Regarding selection, in a simulation study with a population under selection, zeroing on the first eigenvalues of **G** and using the reconstructed matrix for genomic evaluations decreased the selection response by almost 40%, indicating a strong effect of selection on the persistence of GEBV, especially if the dataset is

Table 3. Correlations between IP and GEBV computed using GBLUP with all genotyped animals (IP_{Full}) and only core animals (IP_{core}) for all year-classes and core definitions

		BW		WW		PWG	
Core definition	Year-class	IP_{Full}	IP_{core}	IP_{Full}	IP_{core}	IP_{Full}	IP_{core}
2013	2013[1]	0.99	0.99	0.99	0.99	0.99	0.99
	2014	0.98	0.82	0.99	0.34	0.99	0.31
	2015	0.97	0.64	0.99	0.12	0.99	0.07
2014	2014	0.98	0.99	0.99	0.99	0.99	0.99
2015	2015	0.97	0.99	0.99	0.99	0.99	0.99

[1]Results from year-class 2013 are the same in both core definitions.

Table 4. Correlations between predicted SNP effects computed with all genotyped animals and only core animals in different year-classes within the same core definition

		BW		WW		PWG	
Core definition	Year-class	ssGBLUP	GBLUP	ssGBLUP	GBLUP	ssGBLUP	GBLUP
2013	2013[1]	0.86	0.88	0.92	0.92	0.92	0.95
	2014	0.82	0.83	0.90	0.85	0.90	0.86
	2015	0.78	0.78	0.87	0.75	0.88	0.73
2014	2014	0.82	0.84	0.89	0.90	0.90	0.93
2015	2015	0.78	0.79	0.86	0.88	0.88	0.91

[1]Results from year-class 2013 are the same in both core definitions.

limited (Yvette Stein, University of Georgia, Athens GA, personal communication). Figure 1 shows genetic trends standardized by the additive genetic standard deviation for all traits. Although there is a genetic improvement for all traits, selection pressure on BW is different from WW and PWG. Low BW is desirable to avoid calving problems. However, BW is positively correlated with WW and PWG; therefore, selecting for increased WW and PWG while decreasing BW requires extra selection pressure on the latter. In this way, the persistence of predictions for WW and PWG is expected to be different from BW because of lower heritabilities and different selection pressures.

Impact of blending and tuning

In ssGBLUP, **G** has to be blended and tuned to make it invertible and compatible with the pedigree relationships in **A** (VanRaden, 2008; Vitezica et al., 2011). If these steps are not performed, IP will be affected by changes in blending parameters. Preliminary analyses using different blending strategies (1% $\mathbf{A}_{22}$, 5% $\mathbf{A}_{22}$, and 10% $\mathbf{A}_{22}$) showed that the higher the blending percentage with $\mathbf{A}_{22}$, the lower the correlation between IP and GEBV. Additionally, the more animals used, the bigger the impact of blending ($\mathbf{IP_{Full}}$ vs. $\mathbf{IP_{core}}$) (Table 5). Table 2 shows that the correlations between $\mathbf{IP_{core}}$ and GEBV are slightly higher than $\mathbf{IP_{Full}}$ which was likely due to the impact of blending.

Lourenco et al. (2018) investigated the impact of not accounting for tuning on IP and showed that under GBLUP $\mathbf{E(u)} = 0$ and $\hat{u}|\hat{a} = Z\hat{a}$, but in ssGBLUP, this assumption does not hold because genotyping is more recent than the entire pedigree, which creates a difference between genetic bases from pedigree and genomic data. The authors recommended adding the average GEBV to IP such that $\hat{u}|\hat{a} = \hat{\mu} + Z\hat{a}$, which makes the two predictions comparable. More recently, Legarra et al. (2018) derived formulas taking blending and tuning parameters into account when computing SNP effects from ssGBLUP:

$$\hat{\mathbf{a}} = b\alpha\lambda\mathbf{DZ}'\mathbf{G}^{-1}\hat{\mathbf{u}} \quad (12)$$

where α and b are the blending and tuning parameters, with b as in Vitezica et al. (2011).

Validation

Our validation study represents a more realistic scenario of IP in which young genotyped animals are predicted based only on their genotypes without being part of the routine evaluations. The same patterns of the previous results (Tables 2 and 3) were observed in our validation results (Tables 6 and 7). The predictive abilities for ssGBLUP were very similar with either $\mathbf{IP_{Full}}$ or $\mathbf{IP_{core}}$. Further, as the number of genotyped animals increased from 2013 to 2015, the predictive abilities increased slightly. The predictive abilities increased from 0.38 to 0.43 for BW, from 0.35 to 0.38 for WW, and from 0.28 to 0.31 for PWG during these 3 years (Table 6). On the other hand, when $\mathbf{G}_{\mathbf{core}}^{-1}$ was used to compute the SNP effects and IP, GBLUP and ssGBLUP behaved differently with a fixed set of core animals (core 2013). The predictive ability for GBLUP decreased from 0.38 to 0.30 for BW, from 0.33 to 0.05 for WW, and from 0.26 to 0.04 for PWG as more genotyped animals were added. Conversely, when an updated set of core animals was used (core 2014 and core 2015), the predictive ability for GBLUP was restored to the same levels as ssGBLUP (Tables 6 and 7).

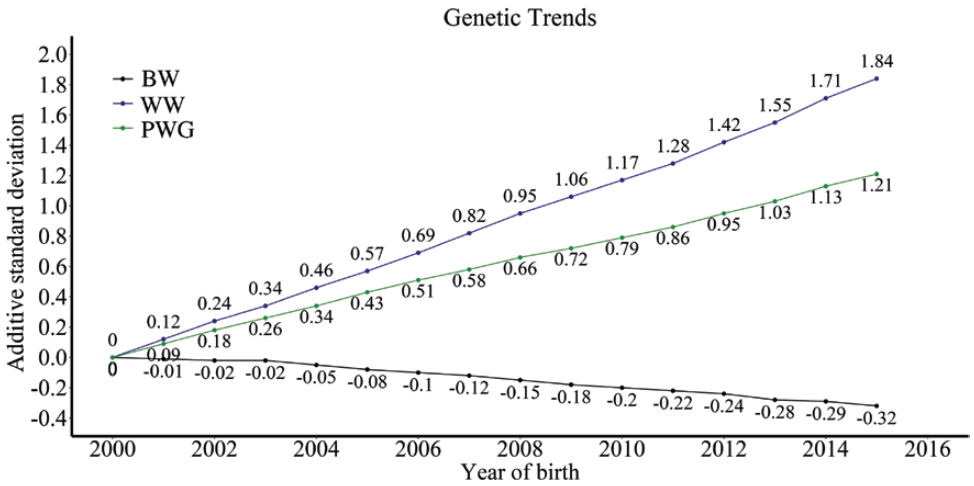

Figure 1. Genetic trends for BW, WW, and PWG. Genetic trends are presented as additive genetic standard deviations and the genetic base is adjusted to 2000.

Table 5. Correlation between IP and GEBV with different blending strategies computed using ssGBLUP with all genotyped animals (IP_{Full}) and only core animals (IP_{core})

	BW		WW		PWG	
Blending[1]	IP_{Full}	IP_{core}	IP_{Full}	IP_{core}	IP_{Full}	IP_{core}
1% A_{22}	0.96	0.99	0.98	0.99	0.99	1.00
5% A_{22}	0.94	0.98	0.97	0.99	0.98	0.99
10% A_{22}	0.92	0.97	0.95	0.98	0.96	0.98

[1]Year-class 2015 and core 2015 definition; A_{22}, pedigree relationship matrix for genotyped animals.

Table 6. Predictive ability of IP for validation animals born in 2016 computed using ssGBLUP with all genotyped animals (IP_{Full}) and only core animals (IP_{core})

		BW		WW		PWG	
Core definition	Year-class	IP_{Full}	IP_{Core}	IP_{Full}	IP_{Core}	IP_{Full}	IP_{Core}
2013	2013[1]	0.38	0.39	0.35	0.35	0.28	0.28
	2014	0.40	0.41	0.36	0.36	0.30	0.30
	2015	0.43	0.44	0.38	0.38	0.31	0.31
2014	2014	0.40	0.41	0.36	0.36	0.30	0.30
2015	2015	0.43	0.44	0.38	0.38	0.31	0.31

[1]Results from year-class 2013 are the same in both core definitions.

Table 7. Predictive ability of IP for validation animals born in 2016 computed using GBLUP with all genotyped animals (IP_{Full}) and only core animals (IP_{core})

		BW		WW		PWG	
Core definition	Year-class	IP_{Full}	IP_{Core}	IP_{Full}	IP_{Core}	IP_{Full}	IP_{Core}
2013	2013[1]	0.37	0.38	0.33	0.33	0.26	0.26
	2014	0.39	0.34	0.34	0.14	0.28	0.11
	2015	0.42	0.30	0.36	0.05	0.30	0.04
2014	2014	0.39	0.40	0.34	0.35	0.28	0.28
2015	2015	0.42	0.43	0.36	0.37	0.30	0.30

[1]Results from year-class 2013 are the same in both core definitions.

The behavior of IP under the validation setting was similar to the correlations between GEBV and IP, showing that in the scenarios where GEBV were successfully retrieved using SNP effects, the predictive ability of IP was higher. Accordingly, when the correlations between GEBV and IP decreased, the predictive ability also decreased as the number of genotyped animals increased.

Another interesting aspect of our validation was that when $\mathbf{G}_{APY}^{-1}$ was used, the predictive abilities for ssGBLUP (Table 6) and GBLUP (Table 7) were very similar. This confirms that once there is enough information available to estimate most of the chromosome segments, accuracies are similar regardless of the procedure used for the computation of GEBV (Pocrnic et al., 2019). Karaman et al. (2016) investigated the accuracies of genomic prediction using different models and concluded that when the reference population was big enough, different genomic procedures (GBLUP, BayesB, and BayesC) "converged" to the same accuracy.

The results from Lourenco et al. (2018) and from this study show that the APY algorithm can be used to compute SNP effects with ssGBLUP and GBLUP that yield reliable IP in large genotyped populations. Furthermore, with the current implementation of APY in the BLUPF90 family of programs (Misztal et al., 2014b), SNP effects and IP can be obtained using a large number of genotyped animals without constraints in computing time and memory usage. Additionally, the use of a subset of core animals to compute IP is also a viable option when ssGBLUP is the procedure of choice for official evaluations.

Number of animals used to compute IP

Although all genotyped animals could be used to predict the SNP effects from GBLUP or ssGBLUP using tools such as APY, we also investigated the minimum number of animals needed to obtain reliable predictions of SNP effects and IP, assuming that a representative set of genotyped animals had GEBV available from a previous evaluation. IP were calculated for all genotyped animals. The correlations between GEBV and IP are presented in Figure 2 for BW, Figure 3 for WW, and Figure 4 for PWG as a function of the number of genotyped animals used in each subset.

The correlations followed an exponential trend as the number of genotyped animals increased, indicating that more than 5,000 animals would be needed to obtain reliable SNP effects and IP in beef cattle populations. The correlations between GEBV and IP were ≥0.97 for all traits when the number of animals reached 10,000, and they increased to ≥0.98 for all traits when 15,000 or more genotyped animals were used. Subsequently, they reached a plateau at what seems to be a minimum number of animals needed. Interestingly, the optimal number of animals needed to reach correlations ≥ 0.98 is close to the number of eigenvalues explaining 98% of the variance of **G** (Figures 2–4). Thus, the theory of limited dimensionality of genomic information (Misztal, 2016) seems to play a role in the amount of information needed for the prediction of SNP effects.

These results agree with Lourenco et al. (2015) who investigated reference populations with 2k, 8k, and 33k genotyped animals to compute IP with ssGBLUP. The authors suggested the use of approximately 33k genotyped animals to obtain reliable IP. In this study, we examined a wider range of reference group sizes, which permitted us to obtain a more precise number of genotyped animals needed to obtain stable IP. Thus, results here confirm the previous finding by Lourenco et al. (2015) that when the number of animals used to predict SNP effects is large enough and their GEBV are available from previous official evaluations (Wiggans et al., 2015), it is possible to obtain reliable IP from both ssGBLUP and GBLUP. Assuming that the ideal number of genotyped animals used to predict SNP effects depends on the dimensionality of genomic information, this number will likely vary across populations as indicated by Pocrnic et al. (2016b). In their study, the number of eigenvalues explaining 98% of the variance of **G** was 14k for Holsteins, 11.5k for Jerseys, 10.6k for Angus, and 4.1k for pigs and chickens. A smaller subset of genotyped animals could be a feasible

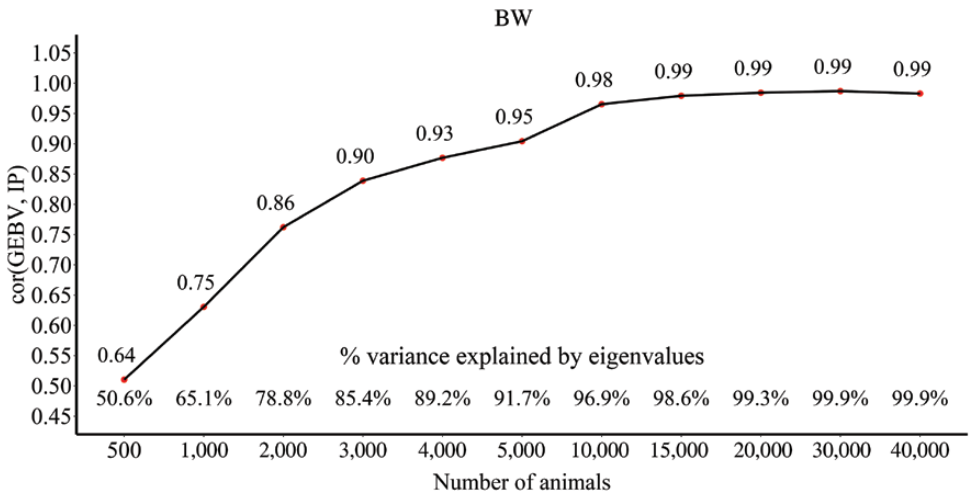

Figure 2. Correlations between GEBV and IP for BW with an increasing number of genotyped animals used to predict SNP effects.

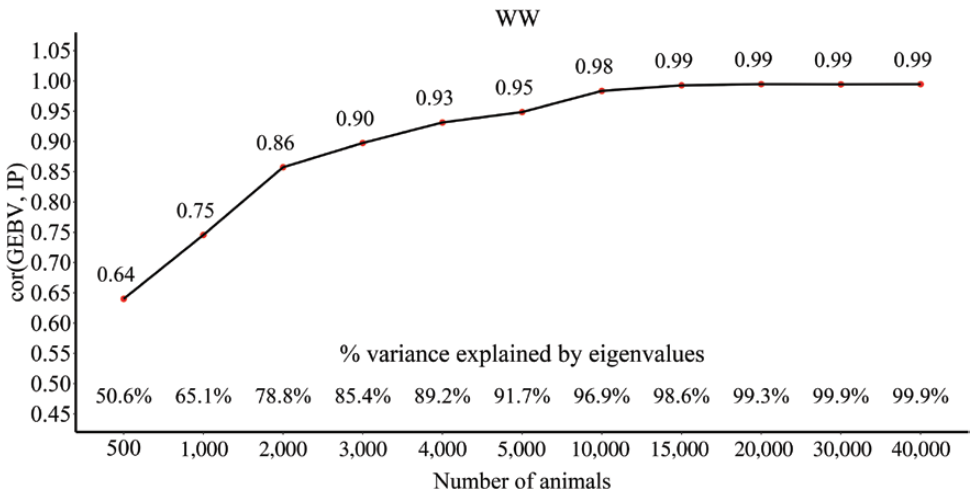

Figure 3. Correlations between GEBV and IP for WW with an increasing number of genotyped animals used to predict SNP effects.

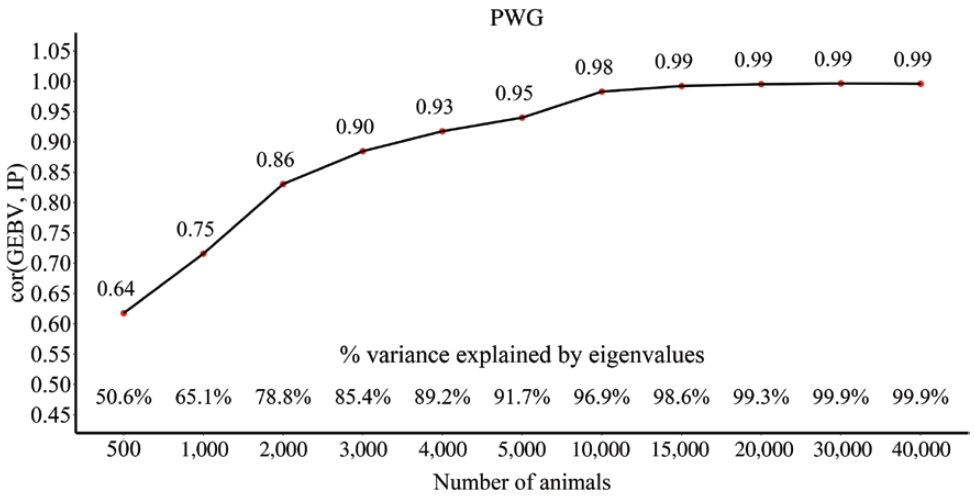

Figure 4. Correlations between GEBV and IP for PWG with an increasing number of genotyped animals used to predict SNP effects.

alternative if these animals represented the dimensionality of the genomic information, and if this subset were a fair representation of the genotyped population.

As pointed out by Wiggans et al. (2015), IP are computed much faster than official evaluations and they permit weekly or even daily evaluations, shortening the interval between the

DNA sampling and genomic prediction. Additionally, they can be used as genomic predictions for non-registered animals without having to include them in official evaluations, an advantage because their inclusion could potentially lead to lower accuracy and increased inflation of GBEV due to lack of phenotypes and missing pedigrees (Bradford et al., 2017, 2019). In these scenarios, IP may become a useful tool to provide quick and reliable genomic predictions for young and non-registered genotyped animals.

Conclusion

With increasing numbers of genotyped animals, using all available genotypes and GEBV from previous official evaluations to predict SNP effects is a practical approach to ensure that IP are stable and reliable. The APY algorithm is a feasible option to predict SNP effects from GBLUP and ssGBLUP when the number of genotyped animals is large. However, if a subset of genotyped animals is used to predict SNP effects with GBLUP, the number and the choice of animals have a considerable impact on IP and predictive abilities. A sample of at least 15,000 animals representing the complete genotyped population would provide reliable predictions of SNP effects and IP in purebred beef cattle populations; however, using the information on all genotyped animals from the previous official evaluation is the usual procedure. In large datasets, ssGBLUP provides more persistent GEBV and IP than GBLUP because it is less sensitive to the time interval between core animals and the most recent genotyped animals.

Acknowledgments

We gratefully acknowledge the editing of the manuscript by Mauricio Elzo. This study was supported by Angus Genetics Inc. (St. Joseph, MO).

Conflict of interest statement

The authors declare that they have no conflicts of interest.

Literature Cited

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot Topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci.* **93**(2):743–752. doi:10.3168/jds.2009–2730

Bradford, H. L., Y. Masuda, P. M. VanRaden, A. Legarra, and I. Misztal. 2019. Modeling missing pedigree in single-step genomic BLUP. *J. Dairy Sci.* **102**(3):2336–2346. doi:10.3168/jds.2018–15434

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

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

Fragomeni, B. O., D. A. 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. doi:10.3168/jds.2014-9125

Karaman, E., H. Cheng, M. Z. Firat, D. J. Garrick, and R. L. Fernando. 2016. An upper bound for accuracy of prediction using GBLUP. *PLoS One*. **11**:e0161054. doi:10.1371/journal.pone.0161054

Legarra, A., D. A. Lourenco, and Z. Vitezica. 2018. Bases for genomic prediction. Available from http://genoweb.toulouse.inra.fr/~alegarra/GSIP.pdf

Lourenco, D. A. L., A. Legarra, S. Tsuruta, D. Moser, S. Miller, and I. Misztal. 2018. Tuning indirect predictions based on SNP effects from single-step GBLUP. *Interbull Bulletin No.*: 53; Uppsala (Sweden): Interbull.

Lourenco, D. A., S. Tsuruta, B. O. Fragomeni, Y. Masuda, I. Aguilar, A. Legarra, J. K. Bertrand, T. S. Amen, L. Wang, D. W. Moser, *et al*. 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

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

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**:1819–1829.

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

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

Misztal, I., S. Tsuruta, D. A. L. Lourenco, Y. Masuda, I. Aguilar, A. Legarra, and Z. Vitezica. 2014b. *Manual for BLUPF90 family of programs*. Available from http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all2.pdf

Nicolazzi, E. L., J. W. Durr, and G. R. Wiggans. 2018. Genomics in the US dairy industry: current and future challenges. *Interbull Bulletin No.* **53**; Uppsala (Sweden): Interbull.

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

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

Pocrnic, I., D. A. L. Lourenco, Y. Masuda, and I. Misztal. 2019. Accuracy of genomic BLUP when considering a genomic relationship matrix based on the number of the largest eigenvalues: a simulation study. *Genet. Sel. Evol.* **51**:75. doi:10.1186/s12711-019-0516-0

R Core Team. 2019. *R: a language and environment for statistical computing*. Vienna (Austria): R Foundation for Statistical Computing.

Strandén, I., and D. J. Garrick. 2009. Technical Note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *J. Dairy Sci.* **92**(6):2971–2975. doi:10.3168/jds.2008-1929

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

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

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. *Genet. Res. (Camb).* **94**:73–83. doi:10.1017/S0016672312000274

Wiggans, G. R., G. Su, T. A. Cooper, U. S. Nielsen, G. P. Aamand, B. Guldbrandtsen, M. S. Lund, and P. M. VanRaden. 2015. Short Communication: Improving accuracy of Jersey genomic evaluations in the United States and Denmark by sharing reference population bulls. *J. Dairy Sci.* **98**:3508–3513. doi:10.3168/jds.2014-8874

Wolc, A., J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, R. Preisinger, D. Habier, R. Fernando, D. J. Garrick, and J. C. Dekkers. 2011. Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. *Genet. Sel. Evol.* **43**:23. doi:10.1186/1297-9686-43-23