Received: 12 August 2020 | Revised: 21 November 2020 | Accepted: 9 December 2020

DOI: 10.1111/jbg.12533

ORIGINAL ARTICLE

Journal of Animal Breeding and Genetics WILEY

Avoiding preselection bias in subsequent single-step genomic BLUP evaluations of genomically preselected animals

Ibrahim Jibrila | Jeremie Vandenplas | Jan ten Napel | Roel F. Veerkamp | Mario P. L. Calus

Wageningen University & Research Animal Breeding and Genomics, Wageningen, The Netherlands

Correspondence
Ibrahim Jibrila, Wageningen University & Research Animal Breeding and Genomics, Droevendaalsesteeg 1, Wageningen 6708PB, The Netherlands.
Email: ibrahim.jibrila@wur.nl

Funding information
Dutch Ministry of Economic Affairs (TKI Agri & Food project), Grant/Award Number: 16022; Breed4Food consortium (with Cobb Europe, CRV, Hendrix Genetics and Topigs Norsvin in addition to Wageningen University and Research as partners)

Abstract

In animal breeding, parents of the next generation are usually selected in multiple stages, and the initial stages of this selection are called preselection. Preselection reduces the information available for subsequent evaluation of preselected animals and this sometimes leads to bias. The objective of this study was to establish the minimum information required to subsequently evaluate genomically preselected animals without bias arising from preselection, with single-step genomic best linear unbiased prediction (ssGBLUP). We simulated a nucleus of a breeding program in which a recent population of 15 generations was produced. In each generation, parents of the next generation were selected in a single-stage selection based on pedigree BLUP. However, in generation 15, 10% of male and 15% of female offspring were preselected on their genomic estimated breeding values (GEBV). These GEBV were estimated using ssGBLUP, including the pedigree of all animals in generations 0–15, genotypes of all animals in generations 13–15 and phenotypes of all animals in generations 11–14. In subsequent ssGBLUP evaluation of these preselected animals, genotypes and phenotypes from various groups of animals were excluded one after another. We found that GEBV of the preselected animals were only estimated without preselection bias when genotypes and phenotypes of all animals in generations 13 and 14 and of the preselected animals were included in the subsequent evaluation. We also found that genotypes of the animals discarded at preselection only helped in reducing preselection bias in GEBV of their preselected sibs when genotypes of their parents were absent or excluded from the subsequent evaluation. We concluded that to prevent preselection bias in subsequent ssGBLUP evaluation of genomically preselected animals, information representative of the reference data used in the evaluation at preselection and genotypes and phenotypes of the preselected animals are needed in the subsequent evaluation.

KEYWORDS
bias, genomic preselection, multi-stage selection, single-step genomic BLUP

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2020 The Authors. *Journal of Animal Breeding and Genetics* published by John Wiley & Sons Ltd.

1 | BACKGROUND

In animal breeding programs, parents of the next generation are usually selected in multiple stages (e.g., Árnason et al., 2012; Meyer & Thompson, 1984; Xu et al., 1995), and the initial stages of this selection are referred to as preselection (e.g., Janhunen et al., 2014; Masuda et al., 2018; Patry & Ducrocq, 2011). Impact of preselection on subsequent genetic evaluations has been a subject of research for a long time in the field of animal breeding (e.g., Appel et al., 1998; Henderson, 1975; Masuda et al., 2018; Patry & Ducrocq, 2011a). Traditionally, preselection for target traits has mostly been based on correlated indicator traits that are easily and cheaply measurable early in lives of selection candidates. For example, piglets could be preselected based on weaning weight as an indicator trait for average daily gain during performance testing. In such situations, multi-trait evaluations are performed including both the target traits and the indicator traits based on which animals are preselected (e.g., Henderson, 1975; Janhunen et al., 2014; Pollak et al., 1984), to prevent preselection bias in the subsequent evaluations. In this case, animals retained at preselection (preselected animals) will have better phenotypes for the indicator traits compared to their discarded (preculled) siblings, and this informs the subsequent evaluations using pedigree-based best linear unbiased prediction (PBLUP) model that for the target traits, preselected animals are better-than-average sets of offspring of their parents. In other words, including the indicator traits in subsequent evaluations provides the PBLUP model in the subsequent evaluations with data to better estimate the (on-average-positive) Mendelian sampling terms of preselected animals. In the genomic era, preselection is mostly based on genomic estimated breeding values (GEBV) of young selection candidates, and this form of preselection is called genomic preselection (GPS, e.g., Masuda et al., 2018; Patry & Ducrocq, 2011a; Sullivan et al., 2019). Although GPS is practiced in several livestock species, including pigs and poultry, reports on GPS in the literature so far are all focussing on dairy cattle. Subsequent PBLUP evaluations after GPS, such as the Interbull and national dairy cattle evaluations, have been reported to be biased (e.g., Masuda et al., 2018; Patry et al., 2013; Sullivan, 2018). It has been shown that this preselection bias can be prevented by including genomic information in form of genomic pseudoperformances (e.g., deregressed proofs) of both preselected and preculled animals in the subsequent PBLUP evaluations (Patry & Ducrocq, 2011b). The genomic pseudoperformances of preculled animals help to inform the PBLUP model in subsequent evaluations that preselected animals are better-than-average subsets of offspring of their parents (Patry & Ducrocq, 2011b). Jibrila et al. (2020) showed that using ssGBLUP in subsequent evaluations prevents GEBV of preselected animals from becoming biased due to preselection, even if genotypes of preculled animals are excluded. This suggests that, in contrast with subsequent PBLUP evaluations, information from preculled animals is not strictly needed to prevent preselection bias in subsequent ssGBLUP evaluation of their preselected sibs. Based on the literature and our previous work (Jibrila et al., 2020; Koivula et al., 2018; Shabalina et al., 2017), we hypothesize that the impact of genotypes of preculled animals in subsequent ssGBLUP evaluations depends on whether genotypes of their parents are included in the subsequent evaluations. The objective of this study was to establish, through simulation, the minimum information required in subsequent ssGBLUP evaluations to estimate GEBV of genomically preselected animals without bias associated with preselection. We also investigated under which circumstances the use of genotypes of preculled animals is beneficial in subsequent evaluations of their preselected sibs. And finally, we evaluated the accuracy realized with each of the implemented scenarios of subsequent evaluation.

2 | MATERIALS AND METHODS

2.1 | Data simulation

Before designing this study, we had discussions with the industrial partners of the Breed4Food consortium (https://breed4food.com/), which are breeding companies in dairy cattle (CRV), pigs (Topigs Norsvin and Hendrix Genetics), poultry (Hendrix Genetics and Cobb Europe) and Aquaculture (Hendrix Genetics). During these discussions, it became clear that breeding practices for the different species are relatively similar and can be represented by a general breeding program as simulated in our study. Therefore, we used inputs from these breeding companies and simulated a nucleus of a general breeding program. We used QMSim (Sargolzaei & Schenkel, 2009) to simulate the datasets, and the details of the simulation can be found in Jibrila et al. (2020).

Briefly, at the end of a historical population of 3,000 generations of random mating, we randomly selected 100 males and 1,000 females and used them as founders. From these founders, we produced a recent population of 15 generations. In each of these recent generations, 100 males and 1,000 females were selected in a single-stage PBLUP-based selection to produce the next generation of 16,000 animals. Within sex, all selected parents had equal contribution to the next generation. Across a simulated genome consisting of 30 chromosomes, 60,000 single nucleotide polymorphisms (SNP) and 3,000 quantitative trait loci (QTL) were evenly distributed. The breeding goal was made up of a single quantitative trait that was measured in both sexes, with heritability of 0.1. For every individual, the phenotype of the trait was simulated as the summation of random additive genetic and residual effects (so no fixed effects). The additive genetic variance was made

TABLE 1 Implementation and results of subsequent genetic evaluations with varying sources and amounts of genomic information[a]

	Genotypes included? Yes (✓) or No (×)						Measures of bias and accuracy		
Scenario	g13	Selected g14	Culled g14	Preselected g15	Preculled g15	Number of animals with both genotypes and phenotypes (as a proxy for reference population)	Absolute bias (in genetic *SD* units)	Dispersion bias ($b_{\mathrm{TBV,GEBV}}$)	Accuracy
1 (control)	✓	✓	✓	✓	✓	34,000	0.03 (0.01)	1.01 (0.03)	0.48 (0.01)
2	✓	✓	✓	✓	×	34,000	0.03 (0.01)	1.01 (0.03)	0.48 (0.02)
3	✓	✓	✓	×	×	32,000	0.36 (0.02)	0.39 (0.03)	0.23 (0.02)
4	×	✓	✓	✓	✓	18,000	0.12 (0.01)	0.67 (0.03)	0.38 (0.02)
5	×	✓	✓	✓	×	18,000	0.12 (0.01)	0.67 (0.03)	0.38 (0.02)
6	×	✓	×	✓	✓	3,100	0.32 (0.01)	0.51 (0.03)	0.32 (0.02)
7	×	✓	×	✓	×	3,100	0.32 (0.01)	0.51 (0.03)	0.32 (0.02)
8	×	×	×	✓	✓	2,000	0.32 (0.02)	0.53 (0.03)	0.32 (0.02)
9	×	×	×	✓	×	2,000	0.45 (0.02)	0.45 (0.03)	0.28 (0.02)

[a] Pedigree of all animals from generation 0 up to preselected generation 15 and phenotypes of all animals from generation 11 up to preselected generation 15 included in every scenario. Pedigree of preculled animals included whenever their genotypes were included and excluded whenever their genotypes were excluded.

TABLE 2 Implementation and results of subsequent genetic evaluations with varying sources and amounts of phenotypic information[a]

	Phenotypes included? Yes (✓) or No (×)					Measures of bias and accuracy		
Scenario	g11 & g12	g13	g14	Preselected g15	Number of animals with both genotypes and phenotypes (as a proxy for reference population)	Absolute bias (in genetic *SD* units)	Dispersion bias ($b_{\mathrm{TBV,GEBV}}$)	Accuracy
2 (control)[b]	✓	✓	✓	✓	34,000	0.03 (0.01)	1.01 (0.03)	0.48 (0.02)
10	✓	✓	✓	×	32,000	0.10 (0.03)	1.04 (0.04)	0.41 (0.01)
11	✓	✓	×	✓	18,000	0.20 (0.01)	0.62 (0.03)	0.39 (0.02)
12	✓	×	×	✓	2,000	0.28 (0.02)	0.41 (0.02)	0.30 (0.02)
13	×	✓	✓	✓	34,000	0.01 (0.01)	0.99 (0.03)	0.48 (0.02)

[a] Pedigree of all animals from generation 0 up to preselected generation 15 and genotypes of all animals from generation 13 up to preselected generation 15 included in every scenario.

[b] Scenario 2 was used as control here, as opposed to scenario 1 in Table 1. This means that preculled animals were completely excluded from all the scenarios in this Table.

up of QTL variance (90%) and polygenic variance (10%). In this study, we used the entire pedigree (i.e., consisting of all animals in generations 0–15), genotypes of the three most recent generations (i.e., consisting of all animals in generations 13–15) and phenotypes of the five most recent generations (i.e., consisting of all animals in generations 11–15).

2.2 | Preselection and subsequent genetic evaluations

We implemented preselection only in the most recent generation (i.e., generation 15). From the selection candidates (i.e., all animals in generation 15), 10% of males and 15% of females were preselected based on their individual GEBV. These GEBV were obtained using ssGBLUP, including the pedigree of animals in generations 0–15, genotypes of animals in generations 13–15, and phenotypes of all the animals in the generations 11–14.

In subsequent (second stage) evaluation, we implemented 13 scenarios, with varying amounts and sources of genomic and phenotypic information. Scenarios 1–9 included either the entire genomic information available, or a subset of it, in addition to the phenotypic information available. Details of the information included in each of these scenarios are in Table 1. Similarly, each of the last four scenarios included either the entire phenotypic information available or a subset of it, in addition to all the genomic information available (except for the genotypes of preculled animals). Details of the information included in each of these scenarios are in Table 2. Available sources of genomic and phenotypic information for subsequent evaluation of the preselected animals, based on their closeness to the preselected animals, can be grouped as follows:

- *Sources of genomic information*: (a) the preselected animals themselves, (b) the preculled animals, (c) parents of the selection candidates (i.e., selected animals in generation 14) and (d) other animals with genotypes, which were, respectively, the unselected sibs of parents of the selection candidates (i.e., the rest of generation 14 animals) and the selection candidates' grandparental generation (i.e., generation 13 animals).
- *Sources of phenotypic information*: (a) the preselected animals themselves, (b) animals with both genotypes and phenotypes at the time of preselection (i.e., selection candidates' parental and grand parental generations/animals in generations 13 and 14) and (c) animals with phenotypes but no genotypes at the time of preselection (i.e., selection candidates' (great) great grandparental generations/animals in generations 11 and 12).

All the genetic evaluations (including at the preselection stage) were performed using the ssGBLUP procedure implemented in MiXBLUP (ten Napel et al., 2017). Every step of this study (data simulation, preselection and subsequent evaluations) was replicated 10 times.

2.3 | Implementation of single-step GBLUP

For each replicate, we used a pedigree-based animal model in ASReml (Gilmour et al., 2009) to estimate the additive genetic and residual variances that we later supplied to MiXBLUP. Pedigree information from all animals in generations 0–14 and phenotypic information from all animals in generations 11–14 were used to estimate these variances. The model used in both ASReml (for estimation of variance components) and MiXBLUP (for breeding value estimation) was as follows:

$$\mathbf{y} = \mathbf{x}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

where $\mathbf{y}$ was the vector of phenotypes; $\mathbf{x}$ and $\mathbf{Z}$ were incidence vector and matrix linking phenotypes to overall mean and random animal effects, respectively; $\mathbf{b}$ was the overall mean; $\mathbf{u}$ was the vector of breeding values; and $\mathbf{e}$ was the vector of random residuals.

For estimation of breeding values, genetic relationships among animals were accounted for by the inverse of the combined pedigree-genomic relationship ($\mathbf{H}^{-1}$), obtained as follows (Aguilar et al., 2010; Christensen & Lund, 2010):

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \left(0.9\mathbf{G_t} + 0.1\mathbf{A}_{22}\right)^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix},$$

where $\mathbf{A}^{-1}$ was the inverse of the pedigree relationship matrix, and $\mathbf{A}_{22}$ was the pedigree relationship matrix among genotyped animals. We considered inbreeding in setting up both $\mathbf{A}^{-1}$ and $\mathbf{A}_{22}$ to avoid the bias caused by ignoring inbreeding (Tsuruta et al., 2019). The matrix $\mathbf{G_t}$ was the genomic relationship matrix calculated as follows:

$$\mathbf{G_t} = \left(1 - \bar{f}_{\mathrm{p}}\right)\mathbf{G_r} + 2\bar{f}_{\mathrm{p}}11',$$

where $\bar{f}_{\mathrm{p}}$ was the average pedigree inbreeding coefficient across genotyped animals, $\mathbf{G_r}$ was the raw genomic relationship matrix computed following the first method of VanRaden (VanRaden, 2008), and $11'$ is a matrix of 1 s. The transformation of $\mathbf{G_r}$ to $\mathbf{G_t}$ was done to make the average genomic inbreeding equal to the average pedigree inbreeding, that is to have $\mathbf{G}$ and $\mathbf{A}_{22}$ on the same scale so that they are compatible. This formula is similar to the formula $\mathbf{G_t} = \mathbf{G_r} + \alpha 11'$ of Vitezica et al. (2011), and it is equivalent to the F_{st}-based formula proposed by Powell et al. (2010), which can be rewritten to $\mathbf{G_t} = \left(1 - \frac{1}{2}\alpha\right)\mathbf{G_r} + \alpha 11'$ (Vitezica et al., 2011). The connection between these three formulas can be seen as follows. In these notations, α is the difference in average pedigree and genomic relationships (Vitezica et al., 2011). Using current allele frequencies to compute $\mathbf{G_r}$, the mean genomic relationship is expected to be zero, such that α reduces to the mean pedigree relationship, and in this case, assuming random mating, $\mathrm{E}\left(\bar{f}_p\right) = \frac{1}{2}\alpha$ (Falconer & Mackay, 1996). As there were scenarios in this study in which selective genotyping was introduced (scenarios 6 and 7), this transformation made sure that its impact was taken care of (Hsu et al., 2017; Vitezica et al., 2011). In computing $\mathbf{G_r}$, we calculated (current) allele frequencies using all available genomic data (i.e., using all the available genomic data, per scenario), and only used SNP with minor allele frequency of at least 0.005. We gave the weights of 0.9 to $\mathbf{G_t}$ and 0.1 to $\mathbf{A}_{22}$ to account for polygenic variance (which was simulated to be 10% of the genetic variance) and to ensure that $\mathbf{G}$ was invertible (Aguilar et al., 2010; Christensen & Lund, 2010). The MiXBLUP instruction file for the ssGBLUP analysis can be found in Jibrila et al. (2020).

2.4 | Measures of bias and accuracy in the subsequent evaluations

Bias was calculated in two ways. Firstly, absolute bias was calculated as the difference between mean True Breeding Value (TBV) and mean GEBV of all preselected animals and expressed in genetic standard deviation (*SD*) units. Absolute bias is a measure of whether estimated genetic gain is equal to true genetic gain. Therefore, if there is no absolute bias (i.e., when mean difference is 0), estimated genetic gain is equal to true genetic gain. A negative difference means that on average GEBV overestimate TBV, and therefore genetic gain is overestimated and vice versa. To have TBV on the same scale as GEBV, we subtracted mean TBV and mean GEBV of the animals in generations 11–14 from TBV and GEBV of each preselected animal, respectively. Secondly, dispersion bias was calculated as the regression coefficient of TBV on GEBV ($b_{\mathrm{TBV,GEBV}}$) of all preselected animals. Dispersion bias is a measure of how well differences in (G)EBV of animals represent the differences in their TBV. If $b_{\mathrm{TBV,(G)EBV}}$ is 1, then there is no dispersion bias. A value of $b_{\mathrm{TBV,(G)EBV}} <1$ means that variance of (G)EBV of animals is inflated compared with variance of their TBV, and so differences in (G)EBV of the animals overestimate differences in their TBV, and vice versa. Accuracy was calculated as the Pearson's correlation coefficient between TBV and GEBV of all preselected animals.

3 | RESULTS

Results of the subsequent evaluations conducted in this study are presented in Tables 1 and 2. Results in Table 1 are for the scenarios with varying amounts and sources of genomic information, and those in Table 2 are for the scenarios with varying amounts and sources of phenotypic information.

3.1 | Impact of genomic information from various groups of animals on bias and accuracy

With all available phenotypes included in the subsequent evaluation, negligible absolute and dispersion biases were observed when all available genotypes were included (scenario 1) and even when genotypes of preculled animals were excluded (scenario 2). For these two scenarios, absolute bias was only 0.03 genetic SDs and $b_{\mathrm{TBV,GEBV}}$ was 1.01. The highest accuracy of GEBV of the preselected animals was achieved when all genotypes and phenotypes available after preselection were included (0.48, scenario 1), and when genotypes of the preculled animals were excluded (scenario 2). This means that just like with bias, accuracy too was not affected moving from scenario 1–2.

When genotypes of the preselected animals, of the selection candidates' parents, of the selection candidates' parents' culled sibs, or of the selection candidates' grandparental generation were excluded from the subsequent evaluation, both absolute and dispersion biases and accuracy loss were observed (scenarios 3–9). In scenarios 1–9, both absolute and dispersion biases increased and accuracy decreased with decreasing number of animals with both genotypes and phenotypes. The only exception is scenario 3, because this was the only scenario where genotypes of the preselected animals were excluded from the subsequent evaluation (see Tables 1 and 2). We also observed that across all scenarios where preculled animals were included, their genotypes only helped in minimizing bias and accuracy loss when genotypes of the selection candidates' parents were excluded. This can be seen by comparing scenario 1 against 2, 4 against 5, and 6 against 7 on the one hand, and scenario 8 against 9 on the other hand, as described next.

In scenario 1, there was no bias, and the highest accuracy was achieved. Excluding genotypes of preculled animals, that is moving from scenario 1–2, did not cause any change. There was bias and accuracy loss in scenario 4, in which genotypes of the selection candidates' grandparental generation were excluded. Further exclusion of genotypes of the preculled animals, that is moving from scenario 4–5, again did not make any difference. The bias increased and accuracy dropped further, in scenario 6, as a result of exclusion of genotypes of culled sibs of the selection candidates' parents, in addition to excluding genotypes of the selection candidates' grandparental generation. Here, further exclusion of genotypes of the preculled animals, that is moving from scenario 6–7, did not make any difference either, because genotypes of the preselected animals and of all the selection candidates' parents were still in the model.

In scenario 8, in which only genotypes of all selection candidates were included in the subsequent evaluation, absolute bias was 0.32, $b_{\mathrm{TBV,GEBV}}$ was 0.53 and accuracy was 0.32. These values are similar to those observed in scenarios 6 and 7, where the only genotypes included were those of the selection candidates and their parents. However, when the only genotypes included were those of the preselected animals (scenario 9), absolute bias increased to 0.45, $b_{\mathrm{TBV,GEBV}}$ decreased to 0.45 and accuracy decreased to 0.28. In summary, with all available phenotypes included, including genotypes of the preselected animals and of the preselected animals' parental and grandparental generations in the subsequent evaluation appeared to be sufficient to prevent preselection bias and minimize accuracy loss due to preselection.

3.2 | Impact of phenotypic information from various groups of animals on bias and accuracy

Because in the previous section we observed that genotypes of the preculled animals were not needed in our subsequent evaluation, we ignored them in this section. Compared to scenario 2, which is the control scenario in Table 2, both absolute and dispersion biases increased and accuracy decreased according to the number of animals with both genotypes and phenotypes. With genotypes of the preselected animals and of the preselected animals' parental and grandparental generations included in the subsequent evaluation, excluding phenotypes of the preselected animals (moving from scenario 2–10) resulted in some absolute bias (0.10 genetic *SD*) and some accuracy loss (from 0.48 to 0.41). A tendency towards deflation was also observed in scenario 10 ($b_{\mathrm{TBV,GEBV}}$ of 1.04). Both absolute and dispersion biases and accuracy loss were also observed when phenotypes of the selection candidates' parental generation were excluded (scenario 11) and even more when additionally phenotypes of the selection candidates' grandparental generation were excluded (scenario 12). Although phenotypes of animals three or four generations before the generation of the selection candidates (i.e., generations 11 and 12) were included in the evaluation at preselection stage, excluding these phenotypes from the subsequent evaluation (scenario 13) did not cause any bias or accuracy loss in GEBV of the preselected animals.

4 | DISCUSSION

In this study, our objective was to investigate the roles of genotypes and phenotypes from various groups of animals in preventing bias due to preselection, when estimating GEBV of genomically preselected animals in subsequent ssGBLUP evaluation. To achieve this objective, we performed simulations involving several simplifying assumptions, as discussed hereafter, that helped to assess the impact of these different sources of information, which may not be possible in data resembling the full complexity of breeding programs in practice. One of the assumptions was to have discrete generations, to enable assessing the impact of using data of different groups of ancestors of the preselected animals. We also modelled only one step of (genomic) preselection, although in reality preselection usually takes place in more than one step. For example, it is common to genotype only members of families preselected based on parent average, and then genomic preselection takes place within these families. Nevertheless, the impact of multiple steps of (different types of) low-intensity preselection is expected to be similar to that of one step of high-intensity preselection. In both cases for the subsequent ssGBLUP evaluation, phenotypes are only available for the animals that survived the last step of preselection. Although we simulated a trait whose phenotypes can be measured on both sexes, the results of our study are (in most instances) applicable to sex-limited traits as well. The main difference between the trait we simulated and sex-limited traits is availability of records on males. In practice, for sex-limited traits, progeny information serves as phenotype for males. So in our study, the fact that we performed the subsequent evaluation after preselected animals had their own records is comparable to the subsequent evaluation that takes place, in dairy cattle for example, when preselected young bulls have daughter information (though performance of many daughters is more valuable, at least for accuracy of breeding values, than a single own performance; e.g., Mrode, 2014). Overall, although the characteristics of the simulated trait more closely resemble some traits in pigs and poultry, where phenotyping is across both sexes, our results are (in most instances) applicable to pig, poultry and dairy cattle breeding schemes.

4.1 | Implementation of single-step GBLUP

The variance components used in ssGBLUP were estimated from our current data, rather than using the simulated values. We did this to reflect what happens in practice, where base generation variance components are not known, but estimated from current data. The trait studied in this study was simulated with heritability of 0.1 and phenotypic variance of 100. Therefore, at the base generation (generation 0) additive genetic variance was 10 and residual variance was 90. When we estimated the variance components as described in the methodology section, additive genetic and residual variances across the 10 replicates were on average 7.41 and 90.24, respectively. Note that information from the selection candidates (generation 15 animals) was not used in estimating these variance components, so all the subsequent evaluation scenarios used the same values per replicate. Using the same data as used in this study, we studied the impact of decreasing or increasing the base generation additive genetic variance by 25% while keeping the residual variance the same. We found that that does not have any statistically significant impact on accuracy and bias of ssGBLUP evaluations (results not shown).

Our scenarios 6 and 7 introduced the problem of selective genotyping, which has been reported to cause bias and reduce accuracy of ssGBLUP evaluations (e.g., Christensen, 2012; Hsu et al., 2017; Vitezica et al., 2011). The implementation of our ssGBLUP model by default takes care of this problem by making the average genomic inbreeding equal to the average pedigree inbreeding, as indicated in the Methodology section. To verify whether this correction worked, we repeated scenario 6, this time without preselection (so with phenotypes of the preculled animals included). The results we found were statistically similar as the results obtained with all available information included (i.e., pedigree of all animals in generations 0–15, genotypes of all animals in generations 13–15 and phenotypes of all animals in generations 11–15). This confirms that the biases and accuracy loss we observed in this study were the result of excluding, from the subsequent ssGBLUP evaluation, either some of the information used as preselection reference data (scenarios 4–9, 11 and 12) or information from preselected candidates themselves (scenarios 3 and 10).

4.2 | The minimum information required in subsequent ssGBLUP evaluation to prevent preselection bias

Although phenotypes of animals three or four generations before the generation of the selection candidates (i.e., generations 11 and 12) were included in the evaluation at preselection stage, excluding these phenotypes from the subsequent evaluation did not cause any bias or accuracy loss in estimating GEBV of the preselected animals. The facts that these animals (the (great) great grandparents) were far from the preselected animals, and that the selection candidates' parental and grandparental generations had both genotypes and phenotypes may explain this. Lourenco et al. (2014) found that truncating phenotypic information to only two to three ancestral generations does not affect accuracy of predicting (G)EBV of young animals in dairy cattle and pig breeding programs. The findings of Lourenco et al. (2014) also mean that in our study, phenotypes of the selection candidates' (great) great grandparental generations did not contribute much in estimating GEBV of the selection candidates during the evaluation at preselection stage. Therefore, including genotypes and phenotypes of the selection candidates' parental and grandparental generations in our subsequent evaluation implies that the most relevant ancestral information used in our evaluation at preselection stage was considered.

Similarly, excluding genotypes of the preculled animals from the subsequent evaluation of their preselected sibs did not cause bias or accuracy loss in GEBV of the preselected animals. Because genotypes and phenotypes of the selection candidates' parents were already included in the subsequent evaluation, including genotypes and phenotypes of the preselected animals alone (without necessarily including genotypes of their preculled sibs) provided the ssGBLUP model in the subsequent evaluation with the remaining data it needed to estimate the positive average Mendelian sampling term of the preselected animals. In preventing bias and accuracy loss in subsequent ssGBLUP evaluation of the preselected animals, genotypes of preselected animals appear to be more important than phenotypes of the preselected animals. This can be seen by comparing scenarios 3 and 10. Although scenarios 3 and 10 have the same number of animals with both genotypes and phenotypes, results of scenario 3 (in which genotypes of preselected animals were excluded from the subsequent evaluation) were worse than those of scenario 10 (in which phenotypes of the preselected animals were excluded from the subsequent evaluation). The fact that genotypes of the preselected animals were included in the evaluation at preselection stage (and phenotypes of the preselected animals were not) may explain this. For preselected dairy sires, however, which usually have performance of many daughters instead of a single own performance, the relative importance of own genomic and phenotypic information in preventing bias and accuracy loss due to preselection may be different from what we saw in this study. This is because performance of many daughters is more valuable, at least for accuracy of breeding values, than own performance (e.g., Mrode, 2014). In summary, to prevent preselection bias in subsequent ssGBLUP evaluation of genomically preselected animals, it is sufficient to supply the model with (a) information representative of the reference data used in the evaluation at preselection stage and (b) genotypes and phenotypes of the preselected animals, which are the main source of information that informs ssGBLUP that the preselected animals are a better-than-average subset of offspring of their parents.

4.3 | Comparison to observations in dairy cattle

In scenario 10, GEBV of the preselected animals are effectively the same as their GEBV at preselection stage, and the measures of bias and accuracy in these two evaluations were statistically similar (results not shown for the evaluation at preselection stage). In dairy cattle breeding programs, it is nowadays common to select all young sires in one stage as soon as they are genotyped (e.g., Mäntysaari et al., 2020), though some form of preselection based on parent average is often applied. If such GEBV are later compared with deregressed proofs or daughter yield deviations of such sires, we expect that some positive absolute bias, some accuracy loss, and a tendency towards deflation would be observed just as in our scenario 10. However, in practice, dairy cattle breeding companies observe negative absolute bias (overestimated genetic trend) and inflation when they make such comparisons (e.g., Mäntysaari et al., 2020). The reason for this is unclear and should be investigated in future studies.

4.4 | Role of genotypes of preculled animals in subsequent ssGBLUP evaluations

In ssGBLUP evaluations, pedigree relationships among genotyped and non-genotyped animals guide the implicit imputation of genotypes of non-genotyped animals (Christensen & Lund, 2010; Legarra et al., 2009; Misztal et al., 2009). Similarly, in subsequent ssGBLUP evaluations, when some or all parents of selection candidates are not genotyped, more accurate imputation of genotypes of the non-genotyped parents and other non-genotyped animals in the pedigree is achieved by including genotypes of all (both preselected and preculled) offspring of the non-genotyped parents than by including genotypes of their preselected offspring alone (Shabalina et al., 2017). Because in our study all parents of the selection candidates had genotypes and these genotypes were included in the subsequent evaluation, genotypes of the preculled animals were no longer needed in the subsequent evaluation. However, just like the findings of Shabalina et al. (2017), our scenarios 8 and 9 show that including genotypes of preculled animals in subsequent ssGBLUP evaluations reduces bias and increases accuracy in estimating GEBV of their preselected sibs when their parents are not genotyped. In another ssGBLUP evaluation, Koivula et al. (2018) observed unexpected tendencies towards more dispersion bias and less accuracy in the GEBV of young selected bulls when genotypes of culled bulls were included compared to when they were excluded. Their results, however, were not statistically significant, and thus inconclusive. Our results show that when parents of selection candidates are genotyped, including genotypes of their preculled offspring in subsequent ssGBLUP evaluations neither improves nor deteriorates the quality of the evaluations. In current breeding programs for all livestock species, often not all dams are genotyped. If evaluations at preselection stage are done with ssGBLUP, dams that are not genotyped benefit from genotypes of all their offspring. Including genotypes of their preculled offspring in subsequent ssGBLUP evaluations ensures that the same levels of accuracy of imputing genotypes of such dams are achieved as in the evaluations at preselection stage. Therefore, in such situations, genotypes of preculled animals would be needed in subsequent ssGBLUP evaluations to estimate GEBV of preselected animals without preselection bias and accuracy loss.

5 | CONCLUSION

To prevent preselection bias in subsequent ssGBLUP evaluation of genomically preselected animals, it is sufficient to supply the model with (a) information representative of the reference data used in the evaluation at preselection stage and (b) genotypes and phenotypes of the preselected animals, which are the main source of information that informs ssGBLUP that the preselected animals are a better-than-average subset of offspring of their parents. When (some) parents of selection candidates are not genotyped, genotypes of preculled animals, together with genotypes of preselected animals, help in more accurately imputing genotypes of their ungenotyped parents in ssGBLUP evaluations at both preselection and subsequent evaluation stages. In such situations, genotypes of preculled animals are needed in subsequent ssGBLUP evaluations to estimate GEBV of their preselected sibs without preselection bias.

ACKNOWLEDGEMENTS

The authors thank Marco Bink, Katrijn Peeters, Jeroen Visscher and Abe Huisman from Hendrix Genetics; Egbert Knol, Rob Bergsma and Egiel Hanenberg from Topigs Norsvin; John Henshall and Randy Borg from Cobb; and André Hidalgo, Chris Schrooten, and Gerben de Jong from CRV for their inputs towards the data simulation and implementation of preselection. We also thank the Dutch Ministry of Economic Affairs (TKI Agri & Food project 16022) and the Breed4Food consortium (with Cobb Europe, CRV, Hendrix Genetics and Topigs Norsvin in addition to Wageningen University and Research as partners) for their financial support to this project.

CONFLICT OF INTERESTS

The authors declare that they have no conflict of interests.

DATA AVAILABILITY STATEMENT

The codes used to simulate the data used in this study are provided in Additional file 1 of Jibrila et al. (2020).

ORCID

Ibrahim Jibrila https://orcid.org/0000-0002-5683-1263
Jeremie Vandenplas https://orcid.org/0000-0002-2554-072X
Jan ten Napel https://orcid.org/0000-0002-1918-9080
Roel F. Veerkamp https://orcid.org/0000-0002-5240-6534
Mario P. L. Calus https://orcid.org/0000-0002-3213-704X

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

Appel, L. J., Strandberg, E., Danell, B., & Lundeheim, N. (1998). Adjusting for missing data due to culling before testing in genetic evaluations of swine. *Journal of Animal Science*, *76*, 1794–1802. https://doi.org/10.2527/1998.7671794x

Árnason, T., Albertsdóttir, E., Fikse, W. F., Eriksson, S., & Sigurdsson, Á. (2012). Estimation of genetic parameters and response to selection for a continuous trait subject to culling before testing. *Journal of Animal Breeding and Genetics*, *129*, 50–59. https://doi.org/10.1111/j.1439-0388.2011.00941.x

Christensen, O. F. (2012). Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. *Genetics Selection Evolution*, *44*, 1. https://doi.org/10.1186/1297-9686-44-37

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

Falconer, D. S., & Mackay, T. F. C. (1996). *Introduction to quantitative genetics* (4th ed.). Longman Group Ltd.

Gilmour, A. R., Gogel, B. J., Cullis, B. R., & Thompson, R. (2009). *ASReml user guide release 3.0*. Retrieved from https://asreml.kb.vsni.co.uk/wp-content/uploads/sites/3/2018/02/ASReml-3-User-Guide.pdf

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometris*, *31*, 423–447. https://doi.org/10.2307/2529430

Hsu, W.-L., Garrick, D. J., & Fernando, R. L. (2017). The accuracy and bias of single-step genomic prediction for populations under selection. *Genes|genomes|genetics*, *7*, 2685–2694. https://doi.org/10.1534/g3.117.043596

Janhunen, M., Kause, A., Vehviläinen, H., Nousiainen, A., & Koskinen, H. (2014). Correcting within-family pre-selection in genetic evaluation of growth-A simulation study on rainbow trout. *Aquaculture*, *434*, 220–226. https://doi.org/10.1016/j.aquaculture.2014.08.020

Jibrila, I., ten Napel, J., Vandenplas, J., Veerkamp, R. F., & Calus, M. P. L. (2020). Investigating the impact of preselection on subsequent single-step genomic BLUP evaluation of preselected animals. *Genetics Selection Evolution*, *52*(42). https://doi.org/10.1186/s12711-020-00562-6

Koivula, M., Strandén, I., Aamand, G. P., & Mäntysaari, E. A. (2018). Reducing bias in the dairy cattle single-step genomic evaluation by ignoring bulls without progeny. *Journal of Animal Breeding and Genetics*, *135*, 107–115. https://doi.org/10.1111/jbg.12318

Legarra, A., Aguilar, I., & Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science*, *92*, 4656–4663. https://doi.org/10.3168/jds.2009-2061

Lourenco, D. A. L., Misztal, I., Tsuruta, S., Aguilar, I., Lawlor, T. J., Forni, S., & Weller, J. I. (2014). Are evaluations on young genotyped animals benefiting from the past generations? *Journal of Dairy Science*, *97*, 3930–3942. https://doi.org/10.3168/jds.2013-7769

Mäntysaari, E. A., Koivula, M., & Strandén, I. (2020). Symposium review: Single-step genomic evaluations in dairy cattle. *Journal of Dairy Science*, *103*, 5314–5326. https://doi.org/10.3168/jds.2019-17754

Masuda, Y., VanRaden, P. M., Misztal, I., & Lawlor, T. J. (2018). Differing genetic trend estimates from traditional and genomic evaluations of genotyped animals as evidence of preselection bias in US Holsteins. *Journal of Dairy Science*, *101*, 5194–5206. https://doi.org/10.3168/jds.2017-13310

Meyer, K., & Thompson, R. (1984). Bias in variance and covariance component estimators due to selection on a correlated trait. *Zeitschrift Für Tierzüchtung Und Züchtungsbiologie*, *101*, 33–50. https://doi.org/10.1111/j.1439-0388.1984.tb00020.x

Misztal, I., Legarra, A., & Aguilar, I. (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *Journal of Dairy Science*, *92*, 4648–4655. https://doi.org/10.3168/jds.2009-2064

Mrode, R. (2014). *Linear models for the prediction of animal breeding values* (3rd ed.). CAB International.

Patry, C., & Ducrocq, V. (2011). Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. *Journal of Dairy Science*, *94*, 1011–1020. https://doi.org/10.3168/jds.2010-3804

Patry, C., & Ducrocq, V. (2011). Accounting for genomic pre-selection in national BLUP evaluations in dairy cattle. *Genetics Selection Evolution*, *43*, 30. https://doi.org/10.1186/1297-9686-43-30

Patry, C., Jorjani, H., & Ducrocq, V. (2013). Effects of a national genomic preselection on the international genetic evaluations. *Journal of Dairy Science*, *96*, 3272–3284. https://doi.org/10.3168/jds.2011-4987

Pollak, E. J., van der Werf, J., & Quaas, R. L. (1984). Selection bias and multiple trait evaluation. *Journal of Dairy Science*, *67*, 1590–1595. https://doi.org/10.3168/jds.S0022-0302(84)81481-2

Powell, J. E., Visscher, P. M., & Goddard, M. E. (2010). Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Reviews Genetics*, *11*, 800–805. https://doi.org/10.1038/nrg2865

Sargolzaei, M., & Schenkel, F. S. (2009). QMSim: A large-scale genome simulator for livestock. *Bioinformatics*, *25*(5), 680–681. https://doi.org/10.1093/bioinformatics/btp045

Shabalina, T., Pimentel, E. C. G., Edel, C., Plieschke, L., Emmerling, R., & Götz, K.-U. (2017). Short communication: The role of genotypes from animals without phenotypes in single-step genomic evaluations. *Journal of Dairy Science*, *100*, 8277–8281. https://doi.org/10.3168/jds.2017-12734

Sullivan, P. G. (2018). Mendelian Sampling variance tests with genomic preselection. *Interbull Bulletin*, *54*, 1–4. https://journal.interbull.org/index.php/ib/article/view/1463

Sullivan, P. G., Mäntysaari, E. A., Dejong, G., & Benhajali, H. (2019). Modifying MACE to accommodate genomic preselection effects. *Interbull Bulletin*, *55*, 77–80. https://journal.interbull.org/index.php/ib/article/view/1484

ten Napel, J., Vandenplas, J., Lidauer, M., Stranden, I., Taskinen, M., Mäntysaari, E., Calus, M. P. L., & Veerkamp, R. F. (2017). *MiXBLUP: A user-friendly softwarevfor large genetic evaluation systems*. Retrieved from https://www.mixblup.eu/documents/ManualMiXBLUP2.1_June2017_V2.pdf

Tsuruta, S., Lourenco, D. A. L., Masuda, Y., Misztal, I., & Lawlor, T. J. (2019). Controlling bias in genomic breeding values for young genotyped bulls. *Journal of Dairy Science*, *102*, 9956–9970. https://doi.org/10.3168/jds.2019-16789

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

Vitezica, Z. G., Aguilar, I., Misztal, I., & Legarra, A. (2011). Bias in genomic predictions for populations under selection. *Genetics Research*, *93*, 357–366. https://doi.org/10.1017/S001667231100022X

Xu, S., Martin, T. G., & Muid, W. M. (1995). Multistage Selection for Maximum Economic Return with an Application to Beef Cattle Breeding. *Journal of Animal Science*, *73*, 669–710. https://doi.org/10.2527/1995.733699x

How to cite this article: Jibrila I, Vandenplas J, ten Napel J, Veerkamp RF, Calus MPL. Avoiding preselection bias in subsequent single-step genomic BLUP evaluations of genomically preselected animals. *J Anim Breed Genet*. 2020;00:1–10. https://doi.org/10.1111/jbg.12533