Impact of genomic preselection on subsequent genetic evaluations with ssGBLUP - using real data from pigs

Ibrahim Jibrila[1], Jeremie Vandenplas[1], Jan ten Napel[1], Rob Bergsma[2], Roel F Veerkamp[1] and Mario P.L Calus[1]*

[1] Wageningen University and Research Animal Breeding and Genomics, PO Box 338
6700 AH Wageningen, the Netherlands

[2] Topigs Norsvin Research Center B.V., Schoenaker 6, 6641 SZ Beuningen, the Netherlands

*Corresponding author

Email addresses and ORCID:

IJ: ibrahim.jibrila@wur.nl; ORCID: 0000-0002-5683-1263

JV: jeremie.vandenplas@wur.nl; ORCID: 0000-0002-2554-072X

JtN: jan.tennapel@wur.nl; ORCID: 0000-00002-1918-9080

RB: rob.bergsma@topigsnorsvin.com; ORCID: 0000-0002-8254-5535

RFV: roel.veerkamp@wur.nl; ORCID: 0000-0002-5240-6534

MPLC: mario.calus@wur.nl; ORCID: 0000-0002-3213-704X

Abstract

Background

Empirically assessing the impact of preselection on subsequent genetic evaluations of preselected animals requires comparison of scenarios with and without preselection. However, preselection almost always takes place in animal breeding programs, so it is difficult, if not impossible, to have a dataset without preselection. Hence most studies on preselection used simulated datasets, concluding that subsequent genomic estimated breeding values (GEBV) from single-step genomic best linear unbiased prediction (ssGBLUP) are unbiased. The aim of this study was to investigate the impact of genomic preselection, using real data, on accuracy and bias of GEBV of validation animals.

Methods

We used data on four pig production traits from one sire-line and one dam-line, with more intense original preselection in the dam-line than in the sire-line. The traits are average daily gain during performance testing, average daily gain throughout life, backfat, and loin depth. Per line, we ran ssGBLUP with the entire data until validation generation and considered this scenario as the reference scenario. We then implemented two scenarios with additional layers of genomic preselection by removing all animals without progeny either i) only in the validation generation, or ii) in all generations. In computing accuracy and bias, we compared GEBV against progeny yield deviation of validation animals.

Results

Results showed only a limited loss in accuracy due to the additional layers of genomic preselection. This is true in both lines, for all traits, and regardless of whether validation animals had records or not. Bias too was largely absent, and did not differ greatly among corresponding scenarios with or without additional layers of genomic preselection.

Conclusion

We concluded that impact of recent and/or historical genomic preselection is minimal on subsequent genetic evaluations of selection candidates, if these subsequent genetic evaluations are done using ssGBLUP.

Background

In animal breeding, parents of the next generation are often selected in multiple stages, and the initial stages of this selection are called preselection [1–3]. Selection candidates that survive preselection are called preselected animals [1–3], and those that do not are called preculled animals [3,4]. Preselection aims to reduce costs and efforts spent on animals that are not interesting for the breeding program, and achieves this by avoiding phenotyping or further testing of the preculled animals. As preculled animals have neither progeny nor records for some or all breeding goal traits, they are generally not included in subsequent genetic evaluations (i.e. genetic evaluations that come after preselection). Preselection therefore decreases the amount of information available for subsequent genetic evaluations of preselected animals. Properly assessing the impact of preselection on subsequent genetic evaluation of preselected animals requires a scenario without preselection, against which scenarios with preselection can be compared. Because in animal breeding programmes preselection almost always takes place, it is difficult, if not impossible, to have a scenario without preselection. This is why most studies available on preselection used simulated datasets [e.g. 1–5]. Those studies have shown that when a subsequent genetic evaluation of preselected animals is done using pedigree-based best linear unbiased prediction (PBLUP), genomic preselection results in accuracy loss and bias in the estimated breeding values (EBV) of preselected animals [1,6–9]. Some of these studies [6–9] further showed that the accuracy loss and bias caused by genomic preselection can be avoided if the information on preculled animals that was utilized at preselection is included in the subsequent PBLUP evaluation. On the other hand, our previous works [3,4] have shown that when the subsequent genetic evaluation is done with single-step genomic BLUP (ssGBLUP), genomic EBV (GEBV) of preselected animals are estimated without bias. We [4] further showed that to avoid genomic preselection bias in subsequent ssGBLUP evaluation of preselected animals, genotypes of their preculled sibs are only needed if not all of their parents are genotyped.

In our previous works [3,4], being based on simulated datasets, preselection was the only possible source of bias in ssGBLUP evaluations. However, in real breeding programmes, other sources of bias in ssGBLUP evaluations may exist and are potentially difficult to control. Therefore, impact of preselection might be confounded by the impact of these other factors. These other possible sources of bias include, amongst others, inaccurate or incomplete pedigree [10], inaccurately estimated additive genetic (co)variances [10], and a reference population of selected genotyped animals [11,12]. Although some ways of reducing the bias caused by these factors have been developed, the bias is usually not completely eliminated in evaluations using real data (e.g. [10–12]). This may explain the observation that in practice GEBV obtained from ssGBLUP evaluations are sometimes biased. The aim of this study was to investigate the impact of genomic preselection on subsequent ssGBLUP evaluations, using real data from an ongoing pig breeding program in which preselection has taken place. To achieve this aim, we used the full dataset as control and retrospectively implemented additional layers of genomic preselection, and results from subsequent ssGBLUP evaluations after these additional layer of genomic preselection were compared against results from ssGBLUP evaluation of the full available data.

Methods

Data

In our analyses, additional layers of genomic preselection were implemented when the animals already had phenotypes, by discarding animals that did not have progeny in the data. Our subsequent genetic evaluations only involved reevaluating preselected animals, either with or without preculled animals in the reevaluations. We separated the available data in two parts, according to a cut-off birth date. Animals born before or on the cut-off birth date were used as reference population, and animals born after the cut-off birth date were used as validation population, from which animals were selected to be used for validation (these are hereafter referred to as "validation animals"). Only animals in the validation population that met the following two requirements were selected as validation animals: 1) none of their parents were included in the validation population, and 2) they had progeny associated with phenotypes. The first requirement ensured that validation animals represented the youngest generation of selection candidates in a breeding program in practice, and not multiple generations. The second requirement enabled validation of the GEBV of the validation animals against their progeny yield deviation (PYD) [13]. Meeting the second requirement was needed, because own phenotypes of the validation animals were used in our subsequent evaluations, and could thus not be used to validate their GEBV.

We obtained pig production traits data on one sire-line and one dam-line from Topigs Norsvin. These data were collected between 1970 and 2020, and the traits were average daily gain during performance testing, average daily gain throughout the lifetime, backfat, and loin depth. Topigs Norsvin (pre)selected both lines on these production traits. However, there was more emphasis on reproduction traits than on production traits in the dam-line. Details on the amount of data utilized in this study are in Table 1. The data were recorded on originally preselected animals (i.e. the animals preselected by Topigs Norsvin), with the sire-line being much more balanced than the dam-line, in terms of proportions males and females with records per generation (ratio of males with records to females with records is about 50:50 in the sire-line and about 20:80 in the dam-line). We studied impact of genomic preselection in the two lines separately, because the traits we studied had different weights in breeding goals of the two lines. The cut-off date to split the data into reference and validation populations was 31st January, 2017 for the sire line, and 31st December, 2015 for the dam-line. In the pedigree, animals with one or both parents missing were assigned to genetic groups, according to line and year of birth of each animal.

Genomic data and quality control

Our genomic data included genotypes of animals for about 21,000 SNP segregating in both lines, and distributed across the 18 autosomes in the pig genome. The SNP were genotyped using a custom SNP chip. We used Plink [14] for all quality control operations on our genomic data. Per genomic preselection scenario (as described later) and per line, animals and SNPs with call rates less than 90% were removed, as well as SNPs that deviated from Hardy-Weinberg equilibrium (Hardy-Weinberg equilibrium exact test p value = 10^{-15}), or had a minor allele frequency below 0.005. Table 1 contains the summary of the pedigree, genomic and phenotypic information utilized in the subsequent genetic evaluations following each genomic preselection scenario, per line.

Computation of pre-corrected phenotypes

In our genetic evaluations, we used pre-corrected phenotypes (rather than raw phenotypes) as records. Animals of different lines were sometimes raised together, so they shared some fixed and non-genetic random effects. Because we studied impact of genomic preselection within lines, it was necessary to correct phenotypes for all non-genetic effects before the data was divided into lines. Another motivation for using pre-corrected phenotypes was that some classes of these non-genetic effects could include only one or a few animals per class due to our implemented additional preselection. We used the following multi-trait pedigree-based animal model to compute pre-corrected phenotypes for all traits:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Wp} + \mathbf{Zu} + \mathbf{e}, \qquad (\mathbf{eq.}\ \mathbf{1}),$$

where **y** was the vector of phenotypes; **b** was the vector of fixed effects, with incidence matrix **X**; **p** was the vector of non-genetic random effects, with incidence matrix **W**; **u** was the vector of breeding values, with incidence matrix **Z**; and **e** was the vector of residuals. Then for every animal (i) with phenotype, precorrected phenotype (y_{ci}) was:

$$y_{ci} = \hat{u}_i + \hat{e}_i \qquad \textbf{(eq. 2).}$$

The (co)variance components used for this analysis were estimated, before separating the data into lines, from a multi-trait pedigree-based animal model in ASReml [15] using **eq. 1**. All computations of (G)EBV were performed using MiXBLUP [16].

Preselection

Per line, we implemented a reference scenario and two scenarios that added layers of genomic preselection. The reference scenario - against which other scenarios could be compared - only included the original genomic preselection implemented by Topigs Norsvin. Thus, the subsequent ssGBLUP evaluations following the reference scenario utilized the entire available data until the validation generation. The second scenario is called validation generation preselection (the VGP scenario). In this scenario, we only implemented additional genomic preselection in the validation generation, by discarding all animals in the validation generation that had no progeny in the data, but had genotypes and/or phenotypes. This scenario was implemented to study the impact of extreme genomic preselection in a single generation. The third scenario is called multi-generation preselection (the MGP scenario), in which we discarded any animal in the validation and previous generations with no progeny in the data. This scenario was implemented to study the carry-over impact of extreme genomic preselection in multiple generations. Animals kept after each of the genomic preselection scenarios are shown in Figure 1.

Subsequent genetic evaluations

Following every scenario of genomic preselection, we implemented a subsequent ssGBLUP evaluation with all animals that survived the genomic preselection. We call this evaluation subsequent because it came after the initial evaluation that provided the GEBV used in preselection. The ssGBLUP evaluations were conducted using MiXBLUP [16], with and without records (i.e. own phenotypes) on the animals in the validation generation (see Table 1). Progeny of validation animals were not included in the subsequent genetic evaluations. We estimated variance components after every preselection scenario, per line, using a pedigree-based multi-trait animal model in ASReml. We used these scenario-specific variance components in the subsequent genetic evaluations to ensure that the variance components used were appropriate for the pre-corrected phenotypes. At the subsequent genetic evaluations, the model used for the estimations of both variance components and breeding values was:

$$\mathbf{y} = \mathbf{xb} + \mathbf{Zu} + \mathbf{e} \qquad \textbf{(eq. 3)},$$

where **y** was the vector of pre-corrected phenotypes; **x** and **Z** were incidence vector and matrix linking pre-corrected phenotypes to overall mean and random animal effects, respectively; **b** was the overall mean; **u** was the vector of breeding values; and **e** was the vector of residuals. We also repeated all subsequent genetic evaluations using PBLUP, to verify the impact of using genotypes on the observed results.

Figure 1 here

Table 1 Data utilized in subsequent ssGBLUP[a] evaluations following each preselection scenario, after quality control

Data in the subsequent ssGBLUP evaluation/Preselection scenario	With records on animals in the validation generation			Without records on animals in the validation generation		
	Reference[b]	VGP[c]	MGP[d]	Reference[b]	VGP[c]	MGP[d]
The sire line						
Number of animals in the pedigree	81,875	60,950	12,777	81,875	60,950	12,777
Number of animals with record for at least one trait	75,129	54,217	6,065	52,846	52,846	4,694
Number of animals with genotypes	33,506	23,315	5,131	33,506	23,315	5,131
Number of SNP	20,550	20,963	20,926	20,550	20,963	20,926
The dam line						
Number of animals in the pedigree	160,426	124,031	33,485	160,426	124,031	33,485
Number of animals with record for at least one trait	139,403	103,018	12,514	100.710	100,710	10,206
Number of animals with genotypes	50,895	36,369	9,072	50,895	36,369	9,072
Number of SNP	19,199	19,256	20,647	19,199	19,256	20,647

[a] single-step genomic best linear unbiased prediction

[b] In the reference scenario, the subsequent ssGBLUP evaluation utilized the entire available data until the validation generation

[c] Validation generation preselection (VGP) scenario. In this scenario, additional genomic preselection was only implemented in the validation generation, by discarding all animals in the validation generation that did not have progeny in the data.

[d] Multi-generation preselection (MGP) scenario. In this scenario, any animal in the validation or reference generations with no progeny in the data was discarded.

Implementation of single-step GBLUP

The inverse of the combined pedigree-genomic relationship $(\mathbf{H^{-1}})$ was obtained as follows [17,18]:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & (0.95\mathbf{G_t} + 0.05\mathbf{A}_{22})^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix} \qquad \textbf{(eq. 4)},$$

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

$$\mathbf{G_t} = \left(1 - \bar{f}_p\right)\mathbf{G_r} + 2\bar{f}_p\mathbf{11}' \qquad \textbf{(eq. 5)},$$

where $\bar{f}_p$ 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 [19], and $\mathbf{11}'$ was a matrix of 1s. The scaling of $\mathbf{G_r}$ to $\mathbf{G_t}$ was done to make the average genomic inbreeding equal to the average pedigree inbreeding, i.e. to have $\mathbf{G}$ and $\mathbf{A}_{22}$ on the same scale so that they are compatible. As the animals with genotypes in this study were selectively genotyped, this transformation made sure that the impact of selective genotyping was taken care of [11,12]. In computing $\mathbf{G_r}$, we computed (current) allele frequencies using all available genomic data after quality control. We gave the weights of 0.95 to $\mathbf{G_t}$ and 0.05 to $\mathbf{A}_{22}$ to ensure that **G** was invertible [17,18].

Measures of accuracy and bias in the subsequent genetic evaluations

We used progeny yield deviation (PYD) [13] as a proxy for true breeding value (TBV), against which GEBV were compared when computing accuracy and bias. To compute PYD, we ran a multi-trait pedigree-based animal model per line in MiXBLUP, with precorrected phenotypes as records and an overall mean as the only fixed effect **(eq. 3)**. The (co)variance components used in this model were also estimated per line in ASReml, from precorrected phenotypes using a multi-trait pedigree-based animal model that only included a mean fixed effect **(eq. 3)**. From the output of this analysis, we computed PYD for each trait for all validation sires and dams as:

$$PYD_i = \frac{\sum_{p=1}^{n} y_{cp} - g_m}{n} \qquad \textbf{(eq. 6)},$$

where PYD_i was the progeny yield deviation of a sire or dam i, y_{cp} was the precorrected phenotype of a progeny p of the sire or dam i, g_m was the genetic contribution of the mate of sire or dam i to y_{cp}, and n was the number of phenotyped progeny of sire or dam i. Estimation of PYD was done before discarding progeny of validation animals from the data. Since progeny of validation animals were not included in subsequent genetic evaluations, comparing (G)EBV to PYD can be considered as a forward-in-time validation. To account for differences in number of progeny used in estimating PYD for different validation animals when estimating accuracy and bias, we approximated the reliability of PYD for each validation animal for each trait as:

$$\frac{{}^{1}/_{4}nh^2}{1+{}^{1}/_{4}(n-1)h^2} \qquad \textbf{(eq. 7)},$$

where n was the validation animal's number of half-sib progeny with records, and h^2 was the heritability of the trait [20]. For convenience, we assumed all progeny of a validation animal were half-sibs, though some of them were full-sibs.

Validation accuracy was computed as weighted Pearson's correlation coefficient between PYD and GEBV of all validation animals, with reliability of PYD used as the weight. We computed two types of bias. The first type is absolute bias, which is a measure of whether estimated genetic gain is equal to true genetic gain. Absolute bias was computed as the weighted mean difference between PYD and half of the (G)EBV of all validation animals, expressed in additive genetic standard deviation (SD) units of the trait. A negative difference means that GEBV are on average overestimated, and therefore genetic gain is overestimated, and vice versa. Before computing differences between PYD and half of the (G)EBV of validation animals, we made sure that PYD and (G)EBV were on the same scale. We did this in the following steps: from the model used in computing PYD, we computed average EBV across all animals in the first three reference generations. We then subtracted half of this average EBV from PYD of each validation animal. Then from each subsequent genetic evaluation, we computed the average (G)EBV of all animals in the first three reference generations. We then subtracted this average (G)EBV from (G)EBV of each validation animal. The second type of bias we computed is dispersion bias. Dispersion bias was measured by the weighted regression coefficient of PYD on (G)EBV of all validation animals. If the regression coefficient is equal to the expected value, then there is no dispersion bias. Note that the expected value is 0.5, because PYD only includes half of the breeding value of a parent. A regression coefficient less than the expected value means that variance of (G)EBV is inflated, and vice versa.

Results

Results of the subsequent genetic evaluations conducted with ssGBLUP are presented in Tables 2 and 3 for the sire-line and the dam-line, respectively. Results in Tables 4 and 5 are from subsequent genetic evaluations done with PBLUP, respectively for the sire-line and the dam-line. In addition to validation accuracy and bias, we also showed the estimated heritability for every subsequent genetic evaluation scenario, and number of validation animals.

Table 2 Performance of ssGBLUP[a] in the subsequent genetic evaluations in the sire-line

Measure/Preselection scenario	With records on animals in the validation generation			Without records on animals in the validation generation		
	Reference[b]	VGP[c]	MGP[d]	Reference[b]	VGP[c]	MGP[d]
Average daily gain during performance testing, number of validation animals = 1382						
Estimated heritability	0.24	0.25	0.33	0.24	0.24	0.35
Validation accuracy	0.51	0.51	0.50	0.47	0.47	0.44
Absolute bias	-0.09	-0.15	-0.01	-0.11	-0.11	-0.02
Dispersion bias	0.48	0.49	0.48	0.48	0.48	0.46
Average daily gain throughout life, number of validation animals = 1383						
Estimated heritability	0.26	0.28	0.33	0.27	0.27	0.35
Validation accuracy	0.57	0.56	0.55	0.52	0.52	0.48
Absolute bias	-0.10	-0.17	-0.06	-0.14	-0.14	-0.08
Dispersion bias	0.48	0.49	0.50	0.47	0.47	0.49
Backfat, number of validation animals = 1383						
Estimated heritability	0.58	0.58	0.58	0.58	0.58	0.60
Validation accuracy	0.69	0.68	0.67	0.63	0.63	0.56
Absolute bias	-0.02	-0.03	-0.03	-0.05	-0.05	-0.09
Dispersion bias	0.48	0.47	0.47	0.44	0.44	0.42
Loin depth, number of validation animals = 1383						
Estimated heritability	0.55	0.55	0.55	0.55	0.55	0.57
Validation accuracy	0.68	0.67	0.65	0.62	0.62	0.54
Absolute bias	0.01	0.00	0.00	0.00	0.00	-0.01
Dispersion bias	0.50	0.50	0.48	0.48	0.48	0.45

SEs were in the range 0.01-0.03 for estimated heritability and dispersion bias, and 0.01-0.02 for validation accuracy and absolute bias.

[a] single-step genomic best linear unbiased prediction

[b] In the reference scenario, the subsequent ssGBLUP evaluation utilized the entire available data until the validation generation

[c] Validation generation preselection (VGP) scenario. In this scenario, additional genomic preselection was only implemented in the validation generation, by discarding all animals in the validation generation that did not have progeny in the data.

[d] Multi-generation preselection (MGP) scenario. In this scenario, any animal in the validation or reference generations with no progeny in the data was discarded.

Table 3 Performance of ssGBLUP[a] in the subsequent genetic evaluations in the dam-line

Measure/Preselection scenario	With records on animals in the validation generation			Without records on animals in the validation generation		
	Reference[b]	VGP[c]	MGP[d]	Reference[b]	VGP[c]	MGP[d]
Average daily gain during performance testing, number of validation animals = 2323						
Estimated heritability	0.31	0.32	0.40	0.30	0.30	0.38
Validation accuracy	0.35	0.31	0.29	0.28	0.28	0.23
Absolute bias	-0.05	-0.14	0.04	0.03	0.03	0.14
Dispersion bias	0.46	0.43	0.41	0.44	0.44	0.43
Average daily gain throughout life, number of validation animals = 2405						
Estimated heritability	0.31	0.33	0.43	0.31	0.31	0.44
Validation accuracy	0.46	0.42	0.42	0.38	0.38	0.35
Absolute bias	-0.06	-0.16	-0.01	0.00	0.00	0.08
Dispersion bias	0.45	0.42	0.42	0.43	0.43	0.43
Backfat, number of validation animals = 2312						
Estimated heritability	0.51	0.51	0.51	0.51	0.51	0.53
Validation accuracy	0.52	0.50	0.50	0.45	0.45	0.42
Absolute bias	0.02	-0.01	-0.03	0.02	0.02	-0.01
Dispersion bias	0.43	0.41	0.41	0.42	0.42	0.41
Loin depth, number of validation animals = 1164						
Estimated heritability	0.50	0.50	0.55	0.49	0.49	0.53
Validation accuracy	0.62	0.60	0.59	0.55	0.56	0.49
Absolute bias	-0.02	-0.03	0.02	-0.04	-0.04	0.03
Dispersion bias	0.54	0.54	0.52	0.53	0.53	0.51

SEs were in the range 0.01-0.02 for estimated heritability, validation accuracy and absolute bias, and 0.01-0.04 for dispersion bias.

[a] single-step genomic best linear unbiased prediction

[b] In the reference scenario, the subsequent ssGBLUP evaluation utilized the entire available data until the validation generation

[c] Validation generation preselection (VGP) scenario. In this scenario, additional genomic preselection was only implemented in the validation generation, by discarding all animals in the validation generation that did not have progeny in the data.

[d] Multi-generation preselection (MGP) scenario. In this scenario, any animal in the validation or reference generations with no progeny in the data was discarded.

Table 4 Performance of PBLUP[a] in the subsequent genetic evaluations in the sire-line

Measure/Preselection scenario	With records on animals in the validation generation			Without records on animals in the validation generation		
	Reference[b]	VGP[c]	MGP[d]	Reference[b]	VGP[c]	MGP[d]
Average daily gain during performance testing, number of validation animals = 1382						
Estimated heritability	0.24	0.25	0.33	0.24	0.24	0.35
Validation accuracy	0.51	0.50	0.49	0.41	0.41	0.40
Absolute bias	-0.04	-0.11	0.01	-0.01	-0.01	0.01
Dispersion bias	0.53	0.54	0.48	0.55	0.55	0.49
Average daily gain throughout life, number of validation animals = 1383						
Estimated heritability	0.26	0.28	0.33	0.27	0.27	0.35
Validation accuracy	0.58	0.56	0.54	0.47	0.47	0.44
Absolute bias	-0.06	-0.14	-0.04	-0.05	-0.05	-0.05
Dispersion bias	0.55	0.55	0.51	0.56	0.56	0.54
Backfat, number of validation animals = 1383						
Estimated heritability	0.58	0.58	0.58	0.58	0.58	0.60
Validation accuracy	0.67	0.66	0.66	0.48	0.48	0.46
Absolute bias	-0.03	-0.03	-0.03	-0.09	-0.09	-0.10
Dispersion bias	0.50	0.50	0.50	0.46	0.46	0.43
Loin depth, number of validation animals = 1383						
Estimated heritability	0.55	0.55	0.55	0.55	0.55	0.57
Validation accuracy	0.66	0.65	0.64	0.49	0.49	0.46
Absolute bias	0.00	0.00	0.00	0.01	0.01	0.00
Dispersion bias	0.50	0.49	0.49	0.48	0.48	0.46

SEs were in the range 0.01-0.03 for estimated heritability and dispersion bias, and 0.01-0.02 for validation accuracy and absolute bias.

[a] Pedigree-based best linear unbiased prediction

[b] In the reference scenario, the subsequent PBLUP evaluation utilized the entire available data until the validation generation

[c] Validation generation preselection (VGP) scenario. In this scenario, additional genomic preselection was only implemented in the validation generation, by discarding all animals in the validation generation that did not have progeny in the data.

[d] Multi-generation preselection (MGP) scenario. In this scenario, any animal in the validation or reference generations with no progeny in the data was discarded.

Table 5 Performance of PBLUP[a] in the subsequent genetic evaluations in the dam-line

Measure/Preselection scenario	With records on animals in the validation generation			Without records on animals in the validation generation		
	Reference[b]	VGP[c]	MGP[d]	Reference[b]	VGP[c]	MGP[d]
Average daily gain during performance testing, number of validation animals = 2323						
Estimated heritability	0.31	0.32	0.40	0.30	0.30	0.38
Validation accuracy	0.35	0.30	0.30	0.24	0.24	0.21
Absolute bias	-0.04	-0.16	0.01	0.08	0.08	0.13
Dispersion bias	0.52	0.45	0.42	0.50	0.50	0.45
Average daily gain throughout life, number of validation animals = 2405						
Estimated heritability	0.31	0.33	0.43	0.31	0.31	0.44
Validation accuracy	0.48	0.43	0.43	0.34	0.34	0.31
Absolute bias	-0.05	-0.18	-0.03	0.05	0.05	0.07
Dispersion bias	0.51	0.47	0.44	0.51	0.51	0.44
Backfat, number of validation animals = 2312						
Estimated heritability	0.51	0.51	0.51	0.51	0.51	0.53
Validation accuracy	0.52	0.50	0.50	0.37	0.37	0.36
Absolute bias	0.02	0.00	-0.03	0.04	0.04	0.00
Dispersion bias	0.45	0.43	0.42	0.41	0.41	0.39
Loin depth, number of validation animals = 1164						
Estimated heritability	0.50	0.50	0.55	0.49	0.49	0.53
Validation accuracy	0.58	0.56	0.56	0.43	0.43	0.41
Absolute bias	0.00	-0.01	0.04	-0.02	-0.02	0.04
Dispersion bias	0.55	0.54	0.51	0.57	0.57	0.52

SEs were in the range 0.01-0.02 for estimated heritability, validation accuracy and absolute bias, and 0.01-0.04 for dispersion bias.

[a] Pedigree-based best linear unbiased prediction

[b] In the reference scenario, the subsequent PBLUP evaluation utilized the entire available data until the validation generation

[c] Validation generation preselection (VGP) scenario. In this scenario, additional genomic preselection was only implemented in the validation generation, by discarding all animals in the validation generation that did not have progeny in the data.

[d] Multi-generation preselection (MGP) scenario. In this scenario, any animal in the validation or reference generations with no progeny in the data was discarded.

Subsequent ssGBLUP evaluations with records on animals in the validation generation

With records on animals in the validation generation included in the subsequent ssGBLUP evaluations, estimated heritability for average daily gain traits in the sire-line increased from the reference to validation generation preselection (VGP) to multi-generation preselection (MGP) scenarios, with more increase from VGP to MGP than from reference to VGP. For backfat and loin depth, the heritability remained the same across all scenarios. For the dam-line, estimated heritability increased from reference to VGP to MGP scenarios, except for backfat, where it remained the same across all scenarios. Observed increases in estimated heritabilities were generally due to decreases in residual variances across the scenarios, while additive genetic variances generally remained similar (Tables S1 and S2). For both lines and for all traits, validation accuracy decreased from reference to VGP to MGP scenarios, albeit the differences were small. For both lines, absolute bias was largely absent for backfat and loin depth, and marginal for the average daily gain traits. The highest value of absolute bias recorded was -0.17 additive genetic SDs, under the VGP scenario for average daily gain throughout life in the sire-line (Table 2). Generally, the values of absolute bias for average daily gain traits moved further away from zero from reference to VGP, and then moved closest to zero with MGP. For the sire-line, regression coefficients of PYD on GEBV - an indicator of dispersion bias - showed no consistent pattern across preselection scenarios for all traits. For all traits and for all scenarios, they ranged from 0.47 to 0.50, being close to the expected value of 0.5. For the dam-line, the regression coefficients decreased or remained the same from reference to VGP to MGP scenarios. They were less than 0.5 for the two average daily gain traits and backfat. For loin depth, they were greater than 0.5.

Subsequent ssGBLUP evaluations without records on animals in the validation generation

Without records on animals in the validation generation in the subsequent ssGBLUP evaluations, all results for the reference and VGP scenarios were the same. Just like when records on animals in the validation generation were included, here too, estimated heritability increased from reference and VGP to MGP scenarios, and in this case for all traits in both lines. Validation accuracy also decreased from reference and VGP to MGP scenarios, and in this case with bigger decreases compared to when records on animals in the validation generation were included. Absolute bias was also largely absent for backfat and loin depth for both lines, and showed no particular pattern for average daily gain traits for the two lines. Even for the average daily gain traits, it was still small, with ±0.14 additive genetic SD being the highest value (Tables 2 and 3). Regression coefficients of PYD on GEBV were similar to their corresponding value when records on animals in the validation generation were included. The only exception were all scenarios for backfat in the sire-line, where the regression coefficients of PYD on GEBV appeared to be lower than their corresponding values when records on animals in the validation generation were included. For both lines, the regression coefficients ranged from 0.41 (for the MGP scenario for backfat in the dam-line) to 0.53 (for the reference and VGP scenarios for loin depth in the dam-line).

Subsequent genetic evaluations with PBLUP

With records on animals in the validation generation included, validation accuracies from subsequent PBLUP evaluations were similar in both magnitude and pattern across the preselection scenarios and lines, to their corresponding values from subsequent ssGBLUP evaluations. However, without records on animals in the validation generation in the subsequent genetic evaluations, validation accuracies were lower with PBLUP than with ssGBLUP for all scenarios in both lines. For both lines and with or without records on animals in the validation generation, absolute bias with PBLUP was always lower than or similar to its corresponding value with ssGBLUP. Regression coefficients of PYD on (G)EBV were also bigger than or similar to their corresponding values with ssGBLUP.

Discussion

In this study, we investigated the impact of genomic preselection on subsequent ssGBLUP evaluations of preselected animals, using real data from an ongoing pig breeding program in which preselection has taken place, by retrospectively implementing additional layers of preselection. The data was on production traits of pigs from one sire-line and one dam-line. Per line, we implemented three genomic preselection scenarios. We used pre-corrected phenotypes as records in the subsequent genetic evaluations, and progeny yield deviation (PYD) as the proxy for TBV. We did the subsequent genetic evaluations either with or without records on animals in the validation generation, and in all cases without progeny of validation animals. In both lines, for all traits and with or without records on validation animals, absolute bias was largely absent across the three genomic preselection scenarios, while with more preselection validation accuracy only showed small decreases and hardly any dispersion bias was induced.

In the two scenarios with additional genomic preselection (i.e. VGP and MGP scenarios), the preselected animals in every generation were the animals that in reality were selected and produced progeny in the next generation, and the preculled animals were those animals that were in reality culled after performance testing. Thus, these two scenarios represent either i) situations in which all the selection in a generation is done in only one stage, after selection candidates have own records, or ii) situations in which an additional selection stage is implemented after preselected animals have had progeny. While neither of these cases is true for the data we used, the scenarios we implemented enabled us to investigate the impact of genomic preselection on subsequent genetic evaluations of preselected animals using real data, by including different amounts of pedigree, genomic and phenotypic information in the subsequent genetic evaluations we implemented. The validation accuracy we computed as the correlation between (G)EBV and PYD is not numerically the same as the accuracy of predicting TBV, since variance of PYD has some non-genetic component, in addition to genetic component [13]. However, the two accuracies are proportional to each other, and this enabled us to make comparison among subsequent genetic evaluation scenarios [21].

Comparison of results across preselection scenarios and between ssGBLUP and PBLUP

With both ssGBLUP and PBLUP, validation accuracy decreased with more genomic preselection (i.e. from reference to VGP to MGP scenarios), and this could be explained by the fact that the amount of phenotypic information also reduced in that order (Table 1). In our previous study using simulated datasets [3], we found accuracy in subsequent ssGBLUP evaluations to be decreasing as amounts of phenotypic information decreased with more intense preselection. For most of the traits in the current study, estimated heritability increased with increase in genomic preselection, and this could have influenced, at least partly, the magnitude of decrease in accuracy with decrease in amount of phenotypic information due to preselection. This could also contribute to explaining why decrease in validation accuracy with more genomic preselection was small. We also observed that validation accuracy was higher with ssGBLUP than with PBLUP, in subsequent genetic evaluations when records on animals in the validation generation were excluded. However, when records on animals in the validation generation were included in subsequent genetic evaluations, validation accuracies were generally similar between corresponding ssGBLUP and PBLUP scenarios. The fact that heritabilities were all relatively high (ranging from 0.24 to 0.58, Tables 2 to 5) could, at least partly, explain the absence of significant differences between ssGBLUP and PBLUP evaluations when records on animals in the validation generation were included in the subsequent genetic evaluations. It is a common knowledge that the higher the heritability, the higher the importance of own record and the lesser the importance of genomic information in genetic evaluations (e.g. [13]).

In our previous study [3], we observed no absolute bias when ssGBLUP was used in subsequent genetic evaluations, irrespective preselection type or intensity. However, in [3], we found absolute bias to be increasing with intensity of preselection when we used PBLUP in subsequent genetic evaluations. Patry et al [1,6,7] also reported significant absolute bias when subsequent genetic evaluations of genomically preselected were done with PBLUP, except when some pseudo-phenotypic information on preculled animals was included in the subsequent PBLUP evaluations. As we did not include (pseudo) phenotypic information on preculled animals in our subsequent PBLUP evaluations, we expected to find significant absolute bias, which would increase from reference to VGP to MGP scenarios. However, in the current study absolute bias remained largely absent across all the three scenarios of genomic preselection, irrespective of whether ssGBLUP or PBLUP was used.

In the absence of selection, the expectation of regression coefficient of PYD on (G)EBV - an indicator of dispersion bias - is 0.5, because PYD only represents half of the breeding value of the parent. However, when validation animals are not a representative sample of all animals in their age group, the expectation of the regression coefficient decreases, depending on how much the validation animals deviate from a random sample of animals in their age group [22,23]. In the data used in this study, average daily gain traits had heavier weights in the breeding goals of the two lines than backfat and loin depth, so we expected that our genomic preselection would have a smaller impact on the regression coefficients for backfat and loin depth than for the two average daily gain traits. We however did not observe smaller regression coefficients or regression coefficients further away from 0.5 for average daily gain traits than for backfat and loin depth, neither with ssGBLUP nor with PBLUP.

Regression coefficient of PYD on (G)EBV generally decreased with more genomic preselection, but were in most cases only marginally different from the expected value of 0.5.

The decrease was more pronounced with PBLUP than with ssGBLUP. In many instances, the regression coefficients of reference scenarios with PBLUP were greater than 0.5, and they (the regression coefficients) became closer to 0.5 with more preselection. In our previous study with a simulated dataset [3], we found that regression coefficients of TBV on (G)EBV were bigger and closer to the expected value of 1 when ssGBLUP was used in the subsequent genetic evaluations compared to when PBLUP was used. In [3], we also found that the regression coefficient became smaller as preselection intensity increased when PBLUP was used, and remained similar irrespective of preselection intensity when ssGBLUP was used. The generally similar regression coefficients across the genomic preselection scenarios with ssGBLUP in this study further confirms that ssGBLUP is indeed able to prevent most of the impact of preselection on subsequent genetic evaluations, as we previously reported in [3]. We have no explanation as to why regression coefficients from PBLUP were greater than the expected value, and also greater than their corresponding values from ssGBLUP. In conclusion, absolute bias remained largely absent across the three genomic preselection scenarios, while with more preselection validation accuracy only showed small decreases and hardly any dispersion bias was induced.

Comparison of results across the two lines

Even in the dam-line where the original genomic preselection was more intense and ratio of males with records to females with records in any generation was about 20:80, we generally did not observe significantly greater biases with more genomic preselection. Although in both lines validation accuracy decreased with more genomic preselection for all traits and with or without records on animals in the validation generation, generally we did not find bigger decreases in the dam-line than in the sire-line. However, corresponding validation accuracies were always higher in the sire-line than in the dam-line, despite the corresponding estimated heritabilities being higher in the dam-line than in the sire-line for some traits. Corresponding regression coefficients of PYD on GEBV were also closer to the expected value of 0.5 in the sire-line than in the dam-line except for loin depth, where they were closer to 0.5 in the dam-line than in the sire-line. The observed higher accuracies and regression coefficients closer to the expected value in the sire-line than in the dam-line can most likely be explained by the higher phenotyping and genotyping rates in the sire-line than the dam-line (Table 1).

Genotypes of preculled animals did not affect the subsequent ssGBLUP evaluations

In the subsequent ssGBLUP evaluations without records on animals in the validation generation, results from corresponding reference and VGP scenarios were exactly the same, at least up to two decimal places (Tables 2 and 3). However, in terms of data content, reference scenarios contained genotypes of the animals preculled in the corresponding VGP scenarios, in addition to all the data contained in the corresponding VGP scenarios (Table 1). The fact that results from these two scenarios were the same means that genotypes of the preculled animals did not affect the reference scenarios. In this study, most (about 95%) of the validation animals and their parents had genotypes. This supports the conclusion from our previous study [4], that genotypes of preculled animals are only useful in subsequent ssGBLUP evaluations of their preselected sibs when their parents are not genotyped.

Potential additional sources of bias in ssGBLUP from our data

In practical datasets as used in this study, it is difficult to completely rule out some mistakes in pedigree recording and in genotyping. At our genomic data quality control stage, genotypes of a few thousand animals were discarded because the animals did not meet the genomic data quality standard (of being genotyped for at least 90% of the SNP). Genotyping mistakes could still not be completely ruled out in the genomic data that passed quality control. In Tables 2 to 5, we saw that for some traits, heritabilities were different for different preselection scenarios, even though the animals in the base generation were the same. This implies that different subsets of the same data gave rise to different estimated (co)variance components in the base generation, and that it is likely that after some of the genomic preselection scenarios were implemented, the estimated (co)variance components were different from their true values, at least for some of the traits. While these are all potential additional sources of bias in ssGBLUP evaluations, they are difficult to avoid in practice [10]. However, in general, we can say that these potential additional sources of bias did not cause significant bias in our ssGBLUP evaluations, as both absolute and dispersion biases were in most cases absent, and even when present they were only marginal.

Conclusions

When subsequent genetic evaluations of preselected animals are done with ssGBLUP, either with or without records on animals in the validation generation, realized accuracy reduces with genomic preselection in the validation generation, and even more with genomic preselection in multiple generations. On the other hand, absolute bias is largely absent, and dispersion bias only increases marginally with more genomic preselection in the current generation or in all generations. Impact of recent and/or historical genomic preselection is minimal on subsequent genetic evaluations of selection candidates, if these subsequent genetic evaluations are performed using ssGBLUP.

Declarations

Ethical approval

The data used for this study were collected as part of routine data recording in a commercial breeding program. Samples collected for DNA extraction were used for routine diagnostic purposes of the breeding program. Data recording and sample collection were conducted in line with local laws on protection of animals.

Availability of data

The data used in the present study were provided by Topigs Norsvin, and are not publicly accessible.

Funding

This study was financially supported by the Dutch Ministry of Economic Affairs (TKI Agri & Food project 16022) and the Breed4Food partners Cobb Europe, CRV, Hendrix Genetics and Topigs Norsvin. The use of the HPC cluster was made possible by CAT-AgroFood (Shared Research Facilities Wageningen UR).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors participated in the conception and the design of the study and of the analysis of the dataset. RB provided the dataset, IJ analysed the dataset and wrote the first draft of the manuscript, and the other authors revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Marco Bink and Katrijn Peeters from Hendrix Genetics, John Henshall from Cobb Europe, and Chris Schrooten and Gerben de Jong from CRV, for their inputs towards the design of this study.

References

1. Patry C, Ducrocq V. Evidence of biases in genetic evaluations due to genomic preselection

in dairy cattle. J Dairy Sci. 2011;94:1011–20.

2. Masuda Y, VanRaden PM, Misztal I, Lawlor TJ. Differing genetic trend estimates from traditional and genomic evaluations of genotyped animals as evidence of preselection bias in us holsteins. J Dairy Sci. 2018;101:5194–206.

3. Jibrila I, Napel J, Vandenplas J, Veerkamp RF, Calus MPL. Investigating the impact of preselection on subsequent single□step genomic blup evaluation of preselected animals. Genet Sel Evol. 2020;52.

4. 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. 2021;138: 432–41.

5. Shabalina T, Pimentel ECG, Edel C, Plieschke L, Emmerling R, Götz K-U. Short communication: the role of genotypes from animals without phenotypes in single-step genomic evaluations. J Dairy Sci. 2017;100:8277–81.

6. Patry C, Ducrocq V. Accounting for genomic pre-selection in national blup evaluations in dairy cattle. Genet Sel Evol. 2011;43.

7. Patry C, Jorjani H, Ducrocq V. Effects of a national genomic preselection on the international genetic evaluations. J Dairy Sci. 2013;96:3272–84.

8. Henderson CR. Best linear unbiased estimation and prediction under a selection model. Biometris. 1975;31:423–47.

9. Pollak EJ, van der Werf J, Quaas RL. Selection bias and multiple trait evaluation. J Dairy Sci. 1984;67:1590–5.

10. Tsuruta S, Lourenco DAL, Masuda Y, Misztal I, Lawlor TJ. Controlling bias in genomic breeding values for young genotyped bulls. J Dairy Sci. 2019;102:9956–70.

11. Vitezica ZG, Aguilar I, Misztal I, Legarra A. Bias in genomic predictions for populations under selection. Genet Res (Camb). 2011;93:357–66.

12. Hsu W-L, Garrick DJ, Fernando RL. The accuracy and bias of single-step genomic prediction for populations under selection. Genes|Genomes|Genetics. 2017;7:2685–94.

13. Mrode RA. Linear models for the prediction of animal breeding values. 3rd ed. 2014.

14. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.

15. Gilmour AR, Gogel BJ, Cullis BR, Thompson R. ASReml user guide release 3.0. VSN Int. Ltd. 2009. p. 275.

16. ten Napel J, Vandenplas J, Lidauer M, Stranden I, Taskinen M, Mäntysaari E, et al. MiXBLUP: a user-friendly software for large genetic evaluation systems. 2020. p. 62.

17. Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of holstein final score. J Dairy Sci. 2010;93:743–52.

18. Christensen OF, Lund MS. Genomic prediction when some animals are not genotyped. Genet Sel Evol. 2010;42.

19. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.

20. Cameron ND. Selection indices and prediction of genetic merit in animal breeding. 1997.

21. Duenk P, Calus MPL, Wientjes YCJ, Breen VP, Henshall JM, Hawken R, et al. Validation of genomic predictions for body weight in broilers using crossbred information and considering breed-of-origin of alleles. Genet Sel Evol. 2019;51.

22. Mäntysaari EA, Liu Z, VanRaden P. Interbull validation test for genomic evaluations. Interbull Bull. 2010;17.

23. Mäntysaari EA, Koivula M. GEBV validation test revisited. Interbull Bull. 2012;45.

Figures

Figure 1 Schematic representation of the animals included in the subsequent genetic evaluations following each genomic preselection scenario

Following the reference scenario, all animals in the figure were included in the subsequent evaluations. In the VGP scenario, only the culled animals in the validation generation were excluded from the subsequent evaluations. Finally, in the MGP scenario, all culled animals in all generations were excluded from the subsequent evaluations. Selection and culling here refer to those conducted by Topigs Norsvin as part of the company's routine practices.

Additional files

Additional file 1 Table S1

Format: .docx

Title: Estimated additive genetic and residual variances in the sire-line

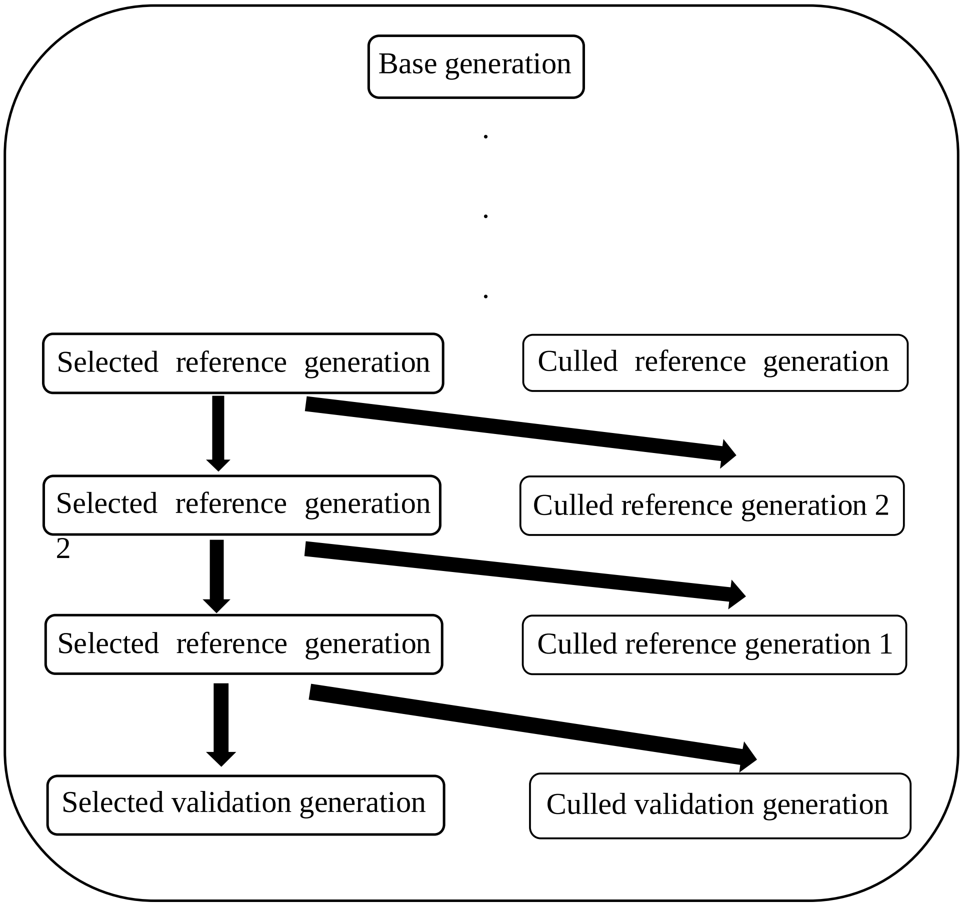
Description: The additive genetic and residual variances that resulted to different heritability estimates for the same traits under different scenarios of subsequent genetic evaluations, in the sire line

Additional file 2 Table S2

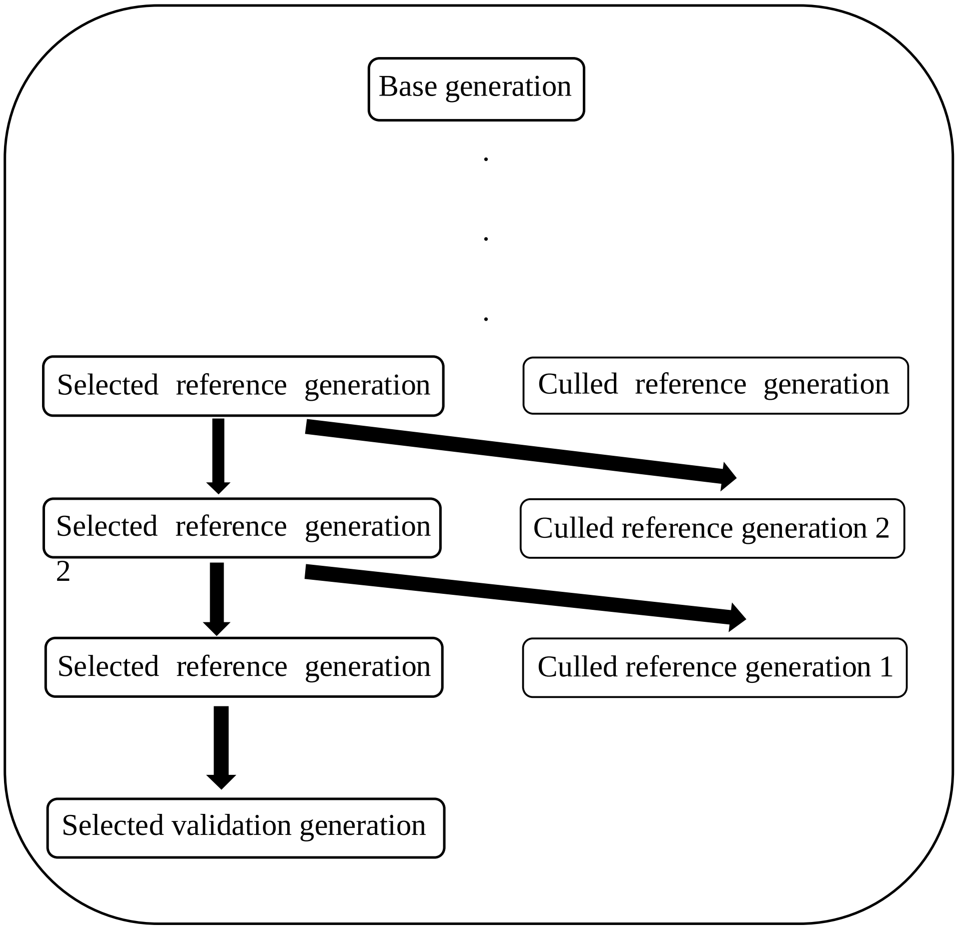
Format: .docx

Title: Estimated additive genetic and residual variances in the dam-line

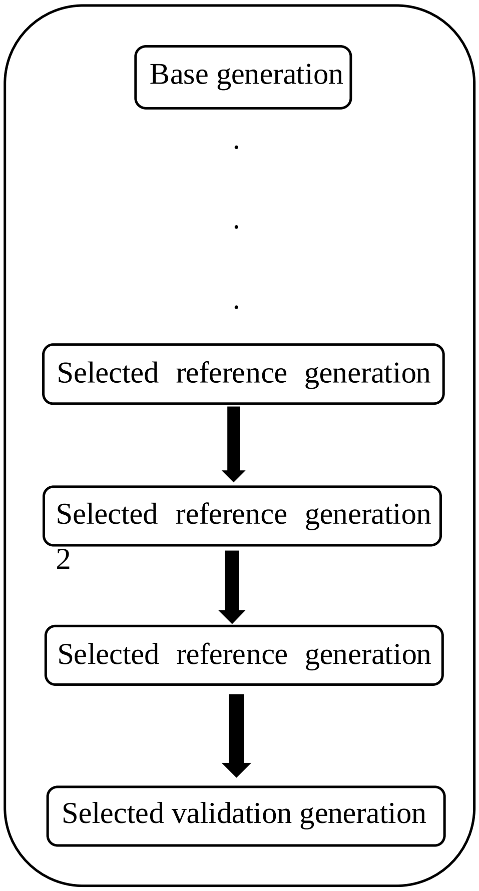
Description: The additive genetic and residual variances that resulted to different heritability estimates for the same traits under different scenarios of subsequent genetic evaluations, in the dam line


Base generation
.
.
.
Selected reference generation
Culled reference generation
Selected reference generation 2
Culled reference generation 2
Selected reference generation
Culled reference generation 1
Selected validation generation
Culled validation generation

a: Reference scenario


Base generation
.
.
.
Selected reference generation
Culled reference generation
Selected reference generation 2
Culled reference generation 2
Selected reference generation
Culled reference generation 1
Selected validation generation

b: Validation generation preselection (VGP) scenario


Base generation
.
.
.
Selected reference generation
Selected reference generation
2
Selected reference generation
Selected validation generation

c: Multi-generation preselection (MGP) scenario