Boundaries for genotype, phenotype, and pedigree truncation in genomic evaluations in pigs

Fernando Bussiman, Ching-Yi Chen, Justin Holl, Matias Bermann, Andres Legarra, Ignacy Misztal, Daniela Lourenco

Journal of Animal Science, 2023, **101**, 1–12
https://doi.org/10.1093/jas/skad273
Advance access publication 16 August 2023
Animal Genetics and Genomics

Boundaries for genotype, phenotype, and pedigree truncation in genomic evaluations in pigs

Fernando Bussiman,[†,1] Ching-Yi Chen,[‡] Justin Holl,[‡] Matias Bermann,[†] Andres Legarra,[‖] Ignacy Misztal,[†] and Daniela Lourenco[†]

[†]Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602, USA
[‡]Genus PIC, Hendersonville, TN 37075, USA
[‖]INRA, UMR1388 GenPhySE, Castanet-Tolosan 31326, France
[1]Corresponding author: fob@uga.edu

Abstract

Historical data collection for genetic evaluation purposes is a common practice in animal populations; however, the larger the dataset, the higher the computing power needed to perform the analyses. Also, fitting the same model to historical and recent data may be inappropriate. Data truncation can reduce the number of equations to solve, consequently decreasing computing costs; however, the large volume of genotypes is responsible for most of the increase in computations. This study aimed to assess the impact of removing genotypes along with phenotypes and pedigree on the computing performance, reliability, and inflation of genomic predicted breeding value (**GEBV**) from single-step genomic best linear unbiased predictor for selection candidates. Data from two pig lines, a terminal sire (L1) and a maternal line (L2), were analyzed in this study. Four analyses were implemented: growth and "weaning to finish" mortality on L1, pre-weaning and reproductive traits on L2. Four genotype removal scenarios were proposed: removing genotyped animals without phenotypes and progeny (*noInfo*), removing genotyped animals based on birth year (*Age*), the combination of *noInfo* and *Age* scenarios (*noInfo + Age*), and no genotype removal (*AllGen*). In all scenarios, phenotypes were removed, based on birth year, and three pedigree depths were tested: two and three generations traced back and using the entire pedigree. The full dataset contained 1,452,257 phenotypes for growth traits, 324,397 for weaning to finish mortality, 517,446 for pre-weaning traits, and 7,853,629 for reproductive traits in pure and crossbred pigs. Pedigree files for lines L1 and L2 comprised 3,601,369 and 11,240,865 animals, of which 168,734 and 170,121 were genotyped, respectively. In each truncation scenario, the linear regression method was used to assess the reliability and dispersion of GEBV for genotyped parents (born after 2019). The number of years of data that could be removed without harming reliability depended on the number of records, type of analyses (multitrait vs. single trait), the heritability of the trait, and data structure. All scenarios had similar reliabilities, except for *noInfo*, which performed better in the growth analysis. Based on the data used in this study, considering the last ten years of phenotypes, tracing three generations back in the pedigree, and removing genotyped animals not contributing own or progeny phenotypes, increases computing efficiency with no change in the ability to predict breeding values.

Lay Summary

Recording data for long years is common in animal breeding and genetics. However, the larger the data, the higher the computing cost of the analysis, especially with genomic information. This study aimed to investigate the impact of removing data, namely, genotypes, phenotypes, and pedigree, on the computing performance and prediction ability of genomic breeding values. We tested four scenarios to remove genotyped individuals in pig populations. For each scenario, phenotypes were removed according to birth year, and the pedigree was either kept complete or traced back from two to three generations. Reliabilities for young, genotyped animals did not differ after removing genotypes for older or less important animals. However, using only two generations of data slightly reduces the reliability for young, genotyped animals. The dispersion did not change across the studied scenarios, and its worst value was observed when using only one generation in the pedigree. Using the last ten years of phenotypes, a pedigree depth of three generations, and removing genotyped animals not contributing own or progeny phenotypes reduces computing cost with no change in the ability to predict breeding values.

Key words: data truncation, genomic selection, old genotypes, pedigree depth, single-step

Abbreviations: A, pedigree relationship matrix; ADG_P, average daily gain on purebreds; ADG_X, average daily gain on crossbreds; *Age*, cutting off old, genotyped animals based on year of birth; *AllGen*, no genotype removal; APY, algorithm for proven and young; BF_P, backfat thickness on purebreds; BF_X, backfat thickness on crossbreds; BLUP, best linear unbiased predictor; BW, birth weight; DGV, direct genomic value; **G**, genomic relationship matrix; GI, genomic information; GEBV, genomic predicted breeding value; **H**, combined pedigree-genomic relationship matrix; L1, terminal-sire line; L2, maternal line; LS, litter size; M1, multitrait analysis of ADG_P, BF_P, ADG_X, and BFX; M2, single-trait analysis of WFM; M3, multitrait analysis of BW and PWM; M4, multitrait analysis of LS and NS; *noInfo*, cutting off genotyped animals with no phenotype and no progeny; *noInfo + Age*, combination of *noInfo* and *Age*; NS, number of stillborn; PA, parent average; PC, progeny contribution; PP, pedigree prediction; PWM, pre-weaning mortality; ssGBLUP, single-step genomic best linear unbiased predictor; WFM, weaning to finish mortality; YD, yield deviation

Introduction

Over the past decades, breeding companies and breed associations have accumulated large amounts of historical data, including pedigree, phenotypes, and genotypes. Adding all ancestors to the relationship matrix increases the accuracy of breeding values if the model of analysis is the true

Received March 3, 2023 Accepted August 10, 2023.

Table 1. Heritability (diagonal, bold), genetic (above diagonal), and residual (below diagonal) correlations estimate for all evaluated traits

Traits	Traits										
	ADG_P	BF_P	ADG_X	BF_X	WFM	BW_d	BW_m	PWM_d	PWM_m	LS	NS
ADG_P	**0.26**	0.40	0.66	0.38							
BF_P	0.26	**0.41**	0.14	0.90							
ADG_X			**0.24**	0.27							
BF_X			0.32	**0.29**							
WFM					**0.05**						
BW_d						**0.06**	-0.38	-0.05	0.22		
BW_m							**0.20**	0.39	-0.47		
PWM_d						-0.33		**0.01**	-0.03		
PWM_m									**0.02**		
LS										**0.14**	0.52
NS										0.27	**0.11**

ADG_P, average daily gain on purebreds; BF_P, backfat on purebreds; ADG_X, average daily gain on crossbreds; BF_X, backfat on crossbreds; WFM, weaning to finish mortality; BW_d, birth weight (direct additive effect); BW_m, birth weight (maternal additive effect); PWM_d, pre-weaning mortality (direct additive effect); PWM_m, pre-weaning mortality (maternal additive effect); LS, litter size; and NS, snumber of stillborn.

model (Henderson, 1984). However, our models are usually approximations (Lourenco et al., 2014), and this discrepancy between fitted and true models can result in a steadily cumulative bias (Macedo et al., 2022). Moreover, the genetic contributions from previous generations decay over time. Still, using those large historical data requires high computing power and could result in biased predictions for young animals (Macedo et al, 2022). This is especially the case in single-step genomic best linear unbiased predictor (**ssGBLUP**), where genomic and pedigree-based relationships are combined, but the number of generations in the pedigree is often much larger than the generations of genotyped animals (Lourenco et al., 2014).

New algorithms to compute the inverses of the genomic (**G**) and pedigree-based (**A**) relationship matrices have been proposed, making large ssGBLUP evaluations feasible. Among them is the algorithm for proven and young (**APY**), which generates a sparse representation of the inverse of **G** (Misztal et al., 2014a). For the computation of the inverse of $\mathbf{A}_{22}$, the pedigree relationship matrix for genotyped animals, partitions of the inverse of **A** can be used (Strandén and Mäntysaari, 2014; Masuda et al., 2017). Together with implementing the preconditioned conjugate gradient and parallel multi-processing programming, these two algorithms allowed the implementation of ssGBLUP evaluations for almost 30 million pedigreed dairy cattle, of which million were genotyped (Cesarani et al., 2022). However, whether all genotyped animals should be used in ssGBLUP is still a subject under investigation.

Under ssGBLUP, **G** must be scaled to be compatible with $\mathbf{A}_{22}$ (Chen et al., 2011; Vitezica et al., 2011), and the size of $\mathbf{A}_{22}$ will depend on the length of the pedigree for genotyped animals. Thus, one can expect that a too-large **G** results in upward bias for genotyped animals, whereas a too-small **G** does the opposite. In addition, as this scaling is for an average of $\mathbf{A}_{22}$, genomic predicted breeding value (**GEBV**) for genotyped animals can be biased depending on the length of the pedigree (Lourenco et al., 2014). Data truncation was suggested as a potential solution to this issue because 1) by removing old generations in the pedigree, the base population in $\mathbf{A}_{22}$ becomes closer to the base population in **G**; 2) by deleting historical data, selection bias accumulated over time can be alleviated.

Data truncation has been successfully applied in dairy cattle, pigs, and dairy sheep (Lourenco et al., 2014; Howard et al., 2018; Cesarani et al., 2021; Hollifield et al., 2021; Macedo et al., 2022). However, depending on the species, non-informative genotyped animals are not always removed from the analyses, and when the truncation is also applied to the genotyped animals, it is based on birth year (Howard et al., 2018). Genotyped animals without own or progeny phenotypes are not expected to contribute to predictions of young animals, and in principle could safely be removed. Therefore, this study aimed to investigate how different criteria to remove genotyped animals along with pedigree and phenotype truncation would impact the prediction of GEBV of genotyped selection candidates.

Material and Methods

Data

Data for two pig lines, a terminal sire (**L1**) and a maternal (**L2**) were provided by PIC (a Genus Company, Hendersonville, TN). Four data sets were analyzed: growth on L1; "weaning to finish" mortality on L1; preweaning traits on L2; and reproductive traits on L2. Variance components were estimated considering the entire population and no genomic information. Genetic parameters and variance components are shown in Table 1. Because data were obtained from existing databases, the approval of the Animal Care and Use Committee was not needed for this study.

Phenotypes, pedigree, and genotypes

Purebred (L1 and L2) and crossbred (L1) animals were recorded for nine different traits: average daily gain and backfat on purebred and crossbred pigs (average daily gain on purebreds [**ADG_P**], backfat thickness on purebreds [**BF_P**], average daily gain on crossbreds [**ADG_X**], and backfat thickness on crossbreds [**BF_X**], respectively), "weaning to finish" mortality (**WFM**), birth weight and pre-weaning mortality (**BW** and **PWM**, respectively), litter size and number of stillborn (**LS** and **NS**, respectively). ADG_P, ADG_X, BF_P, BF_X,

and WFM were measured only for L1 animals, whereas BW, PWM, LS, and NS were measured on L2 pigs. LS and NS were measured on purebreds. The pedigree files for L1 and L2 contained 3,601,369 and 11,240,865 animals born between 1971 and 2021 and 1971 and 2022, respectively. Genotypes imputed to 50K single nucleotide polymorphisms (SNP) were available for 168,734 purebred pigs from L1 and 170,121 purebred pigs from L2. After quality control, 43,812 SNP were retained for L1, whereas 40,968 SNP remained for L2.

Models

Four different models were implemented: ADG_p, BF_p, ADG_X, and BF_X were analyzed by a four-trait model (**M1**); WFM was analyzed through a single-trait model (**M2**); BW and PWM were analyzed using a two-trait model with direct (BW_d and PWM_d) and maternal (BW_m and PWM_m) genetic effects (**M3**), and LS and NS were analyzed through a two-trait repeatability model (**M4**). For all analyses, pedigree, genotypes, and phenotypes were analyzed together using ssGBLUP. This method uses the inverse of a combined pedigree-genomic relationship matrix ($\mathbf{H}^{-1}$), which requires the computation of the inverse of the genomic relationship matrix ($\mathbf{G}^{-1}$). Because of the number of genotyped animals, APY (Misztal et al., 2014a) was used to compute $\mathbf{G}^{-1}$ ($\mathbf{G}^{-1}_{APY}$ – Misztal et al., 2020). Core animals were set to 11,000 for M1 and M2, and 7,500 for M3 and M4, according to the dimensionality of the genomic information (Pocrnic et al., 2016a; Pocrnic et al., 2016b) within each line, calculated as the number of eigenvalues explaining at least 99% of the variability in **G**. The core animals were chosen to be consistent with the truncation scenarios (defined later). Only animals with phenotypes for at least one of the considered traits within the analyses, born between 2017 and 2019, were randomly sampled to compose the core, and the core was kept constant throughout the analyses.

Finally, for M4, unknown parent groups (UPG) were assigned according to account for population admixture and modeled as a random effect. Therefore, $\mathbf{H}^{-1}$ (for M4) was given by (Misztal et al., 2013):

$$\mathbf{H}^{-1}_{UPG,\,\Sigma} = \mathbf{A}^{*}_{\Sigma} + \begin{bmatrix} 00 & 0 \\ 0\mathbf{G}^{-1}_{APY} - \mathbf{A}^{-1}_{22} & -\left(\mathbf{G}^{-1}_{APY} - \mathbf{A}^{-1}_{22}\right)\mathbf{Q}_2 \\ 0 - \mathbf{Q}'_2\left(\mathbf{G}^{-1}_{APY} - \mathbf{A}^{-1}_{22}\right) & \mathbf{Q}'_2\left(\mathbf{G}^{-1}_{APY} - \mathbf{A}^{-1}_{22}\right)\mathbf{Q}_2 \end{bmatrix},$$

Where $\mathbf{A}^{*}_{\Sigma}$ is the inverse of additive relationship matrix based on Quaas-Pollak (QP) transformation (Quaas and Pollak, 1981; Quaas, 1988; Westell et al., 1988); $\mathbf{G}^{-1}_{APY}$ was defined above, $\mathbf{A}^{-1}_{22}$ is the inverse of the pedigree relationship matrix among genotyped animals, and $\mathbf{Q}_2$ is a matrix relating genotyped animals to UPG.

All the analyses were performed using iteration on data with preconditioned conjugate gradient and parallel processing by OpenMP (OpenMP Architecture Review Board, 2015) using the BLUPF90IOD2OMP1 (Misztal et al., 2014b) software, which is an optimized version of BLUPF90IOD2 (Tsuruta et al., 2001; Tsuruta and Misztal, 2008). Computations were carried out on a Linux server (x86_64) with 1 TB of RAM and an Intel Xeon E7-8857 v2 (3.00 GHz) processor (24 computing cores), and the convergence criteria (Tsuruta et al., 2001) was set to 1^{-15} (for M1, M2, and M4) or 1^{-12} (for M3).

Truncation of genotypes, phenotypes, and pedigree

Four scenarios were tested to assess the possibility of cutting old genotypes. For the first scenario (*noInfo*), we removed all genotyped animals not contributing phenotypes to the evaluation, i.e., genotyped animals without own or progeny phenotypes. The second scenario (*Age*) was created by removing genotyped animals according to their birth year (regardless of if phenotyped). The third scenario combined the first two approaches (*noInfo* + *Age*), where all non-informative genotyped animals were removed first, and then removal was based on birth year. In the fourth scenario, named *AllGen*, no genotyped animals were removed. For all scenarios, phenotypes were removed consecutively, according to the birth year, assuming a generation interval of three years for this population (Lourenco et al., 2014), that is, every truncated dataset had the three oldest years of phenotypic information removed. For the pedigree truncation, we either kept all the pedigree information or traced back two or three generations in each phenotype truncation year and across the four genotype truncation scenarios. Since each dataset presented different ranges of birth year, the truncation points were different for each analysis (Table 2). The number of genotyped animals kept after removing non-informative animals (i.e., non-phenotyped and with no progeny) was 141,242 for M1 (16.29% removed), 26,894 for M2 (84.06% removed), 86,054 for M3 (49.42% removed), and 42,282 for M4 (75.15% removed). For the AllGen scenario, all genotyped animals were used. The number of genotyped animals kept for *Age* and *noInfo* + *Age* scenarios is in Figure 1.

Validation

The LR validation (Legarra and Reverter, 2018) was used to evaluate the truncation schemes. Focal individuals (males and females with phenotyped progeny) were chosen from the group of genotyped animals born in 2019 and subsequent years. For validation purposes, focal animals had their phenotypes masked, along with their contemporaries and progeny. The number of focal animals was different for each trait: 2,165 for ADG_p, 1,897 for BF_p, 2,221 for ADG_X, 1,944 for BF_X, 415 for WFM, 2,096 for BW, 3,949 for PWM, 910 for LS, and 933 for NS. Those numbers varied according to the availability of phenotyped progeny for validation candidates in each dataset. Let the masked phenotypes file be represented by the subscript p (partial), whereas the whole dataset is represented by the subscript w (whole). Under the LR method, the reliability (rel), and the dispersion (b) of GEBV can be calculated as:

$$\text{rel} = \frac{\text{cov}\left(\hat{\mathbf{u}}_p, \hat{\mathbf{u}}_w\right)}{\sigma^{2*}_u}$$

$$b = \frac{\text{cov}\left(\hat{\mathbf{u}}_p, \hat{\mathbf{u}}_w\right)}{\text{var}\left(\hat{\mathbf{u}}_p\right)}$$

Where $\hat{\mathbf{u}}$ is the vector of GEBV for focal animals, and σ^{2*}_u is the trait additive genetic variance in the focal individuals (Macedo et al., 2021), calculated as in Sorensen et al. (2001). Changes in ranking and predictions were evaluated on the scale of GEBV, using the analysis with no data truncation as a benchmark, differences were then assessed by the Spearman (ranking) and Pearson (GEBV) correlation coefficients across

Table 2. Pedigree and phenotype count for each model across truncation years over the tested pedigree depths

Model	Year[1]	NR[2]	NA[3]	Number of animals kept in the pedigree[4]	
				Two generations	Three generations
M1	Full	1,452,257	1,452,257	—	—
	2005	1,333,802	1,333,802	1,419,508	1,420,973
	2008	1,211,597	1,211,597	1,294,452	1,296,461
	2011	1,049,955	1,049,955	1,129,402	1,131,870
	2014	844,955	844,955	925,305	927,839
	2017	582,469	582,469	686,190	688,677
M2	Full	324,397	324,397	—	—
	2007	302,200	302,200	502,704	503,961
	2010	252,300	252,300	448,823	450,069
	2013	178,317	178,317	369,443	370,686
	2016	125,504	125,504	312,745	313,988
M3	Full	517,446	517,446	—	—
	2003	489,076	489,076	594,060	596,666
	2006	459,273	459,273	564,931	567,526
	2009	432,505	432,505	538,931	541,540
	2012	375,692	375,692	484,296	486,928
	2015	307,425	307,425	427,155	429,793
M4	Full	7,853,629	2,322,474	—	—
	2002	7,753,676	2,296,858	2,480,557	2,482,212
	2005	7,557,900	2,248,913	2,436,270	2,439,407
	2008	7,108,436	2,136,521	2,336,026	2,341,316
	2011	6,319,428	1,946,867	2,158,129	2,166,354
	2014	5,033,113	1,641,339	1,871,378	1,882,611
	2017	3,107,674	1,135,171	1,398,058	1,412,626

[1]Phenotype truncation year.
[2]Number of records.
[3]Number of animals with records.
[4]When all the animals were kept in the pedigree; the number of animals was 3,601,369 (for M1 and M2) and 11,240,865 (for M3 and M4).
The number of records is equal to the number of animals for M1, M2, and M3, because the traits in these models were not repeated. The "—" implies that this analysis was not performed (i.e., pedigree truncation was not applied to the full dataset).

the truncation years within each depth of the pedigree and genotype truncation scenario. The function "*ggplot*" from the "*ggplot2*" (Wickham, 2016) *R* package (R Core Team, 2020) was used for all graphs in this study.

Results and Discussion

Validation

Correlations between GEBV and the ranking of validation animals did not change significantly after removing any quantity of historical data, except for ADG_X (Table 3). The weakest correlations were observed for the truncation on the last tested year for all analyzed data. This may suggest the need for at least two generations of data to predict validation individuals because the amount of data kept in the last tested year corresponds to one-generation interval. We also observed no benefits by keeping all generations in the pedigree when truncation was applied to either phenotypes or genotypes and phenotypes. Increasing the number of generations traced back in the pedigree, without the corresponding phenotypes reduces rank correlations among focal individuals when a few years of data are analyzed.

Figure 2 presents reliability and dispersion coefficients for M1. The initial level of dispersion for the focal individuals did not change for any truncation year across all tested scenarios. In contrast, the reliability decreased after 2014 for BF_P and BF_X, no matter how many generations were included in the pedigree. For ADG_X, the reliability was slightly underestimated in all scenarios, but a more significant drop happened when keeping data from 2017. No dispersion problem was observed for ADG_P, ADG_X, BF_P, and BF_X in any scenario and truncation point. Conversely, GEBV for focal individuals were overdispersed in WFM in all truncation scenarios, whereas their reliability did not change significantly up to 2013 (Figure 3). Growth traits (ADG_P, ADG_X, BF_P, and BF_X) are widely recorded, and the number of phenotyped animals within each year is higher than for mortality. The data structure and the type of analysis could also explain those results.

While growth traits are recorded on pure and crossed animals, WFM is recorded only on the commercial progeny of purebred animals. Therefore, there was a one-generation gap between the phenotypic measurement and the focal individuals for WFM. Furthermore, growth traits are analyzed in a four-traits model, which usually increases the reliability due to genetic correlations, whereas WFM is in a single-trait model. This can also explain why the original level of dispersion and reliability did not change for pre-weaning

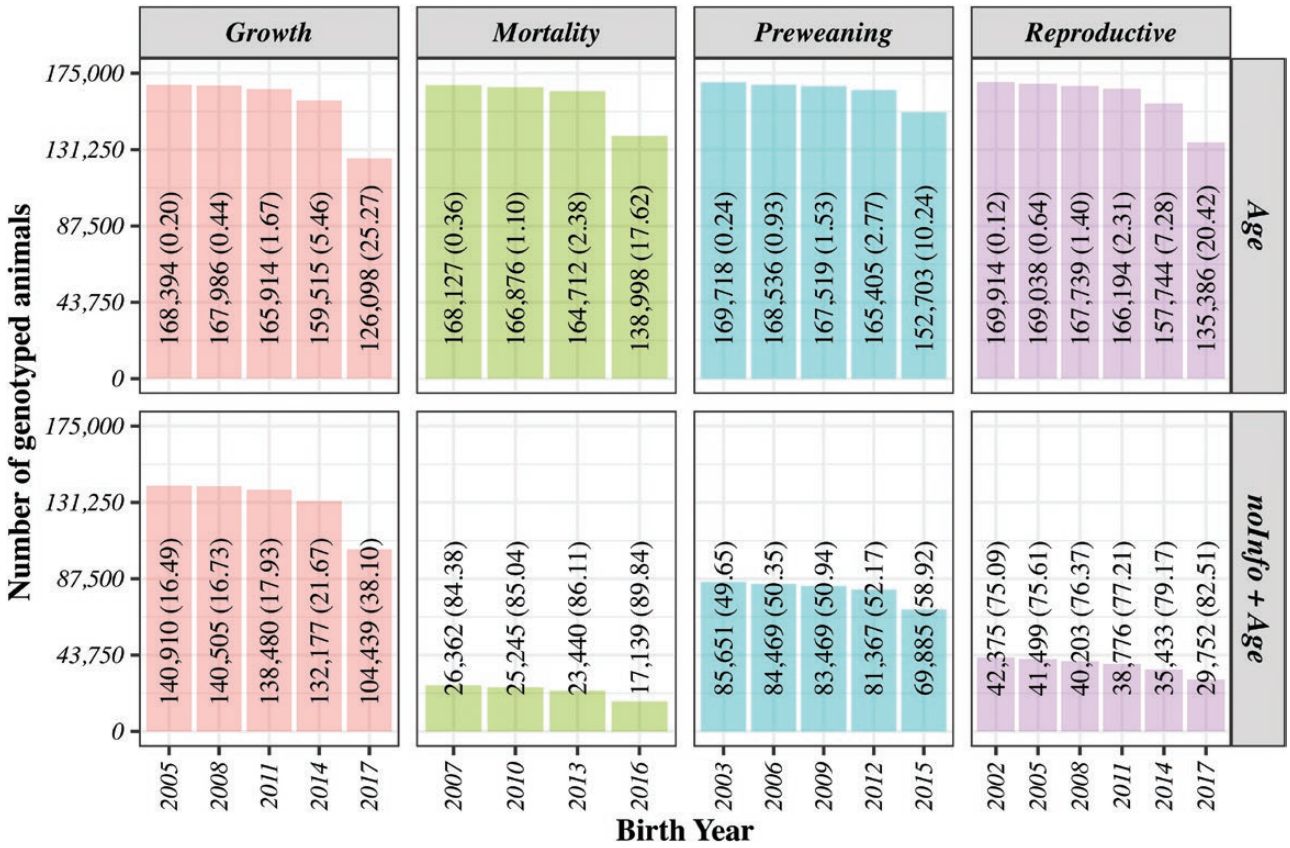

Figure 1. Number of genotyped animals across truncation years for each model in *Age* (cutting off old, genotyped animals) and *noInfo + Age* (cutting off old, genotyped animals along with the non-informative) scenarios. The numbers in parenthesis are the percentage of removed genotyped animals.

Table 3. Pearson (GEBV) and spearman (Rank) correlations (minimum value) between GEBV from the full dataset and each genotype truncation scenario for validation individuals

Model	Trait	GEBV				Rank			
		noInfo	Age	*noInfo* + Age	AllGen	*noInfo*	Age	*noInfo* + Age	AllGen
M1	ADG_P	0.98	0.97	0.98	0.97	0.97	0.97	0.97	0.97
	BF_P	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	ADG_X	0.90	0.87	0.87	0.90	0.89	0.86	0.86	0.89
	BF_X	0.98	0.97	0.97	0.98	0.97	0.97	0.97	0.97
M2	WFM	0.98	0.97	0.97	0.97	0.96	0.96	0.96	0.96
M3	BW_d	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
	BW_m	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
	PWM_d	0.94	0.93	0.93	0.94	0.93	0.92	0.93	0.93
	PWM_m	0.96	0.95	0.96	0.95	0.95	0.95	0.95	0.95
M4	LS	0.97	0.96	0.96	0.97	0.96	0.96	0.96	0.96
	NS	0.97	0.95	0.95	0.97	0.96	0.95	0.94	0.97

ADG_P, average daily gain on purebreds; BF_P, backfat on purebreds; ADG_X, average daily gain on crossbreds; BF_X, backfat on crossbreds; WFM, weaning to finish mortality; BW_d, birth weight (direct additive effect); BW_m, birth weight (maternal additive effect); PWM_d, pre-weaning mortality (direct additive effect); PWM_m, pre-weaning mortality (maternal additive effect); LS, litter size; and NS, number of stillborn.

mortality (Figure 4). Even though the heritability (h^2) for WFM is 0.05, and for PWM_d is 0.01, the genetic correlation between PWM_d and BW_m (0.39) helps improve predictions because h^2 for BW_m is 0.20. Finally, the number of records for pre-weaning traits is larger than for WFM.

For BW, neither the reliability nor the dispersion was affected by the phenotype and genotype truncation (Figure 4); however, tracing back only two generation in the pedigree slightly increased dispersion of GEBV for young animals. The maternal effect of PWM was slightly overdispersed when

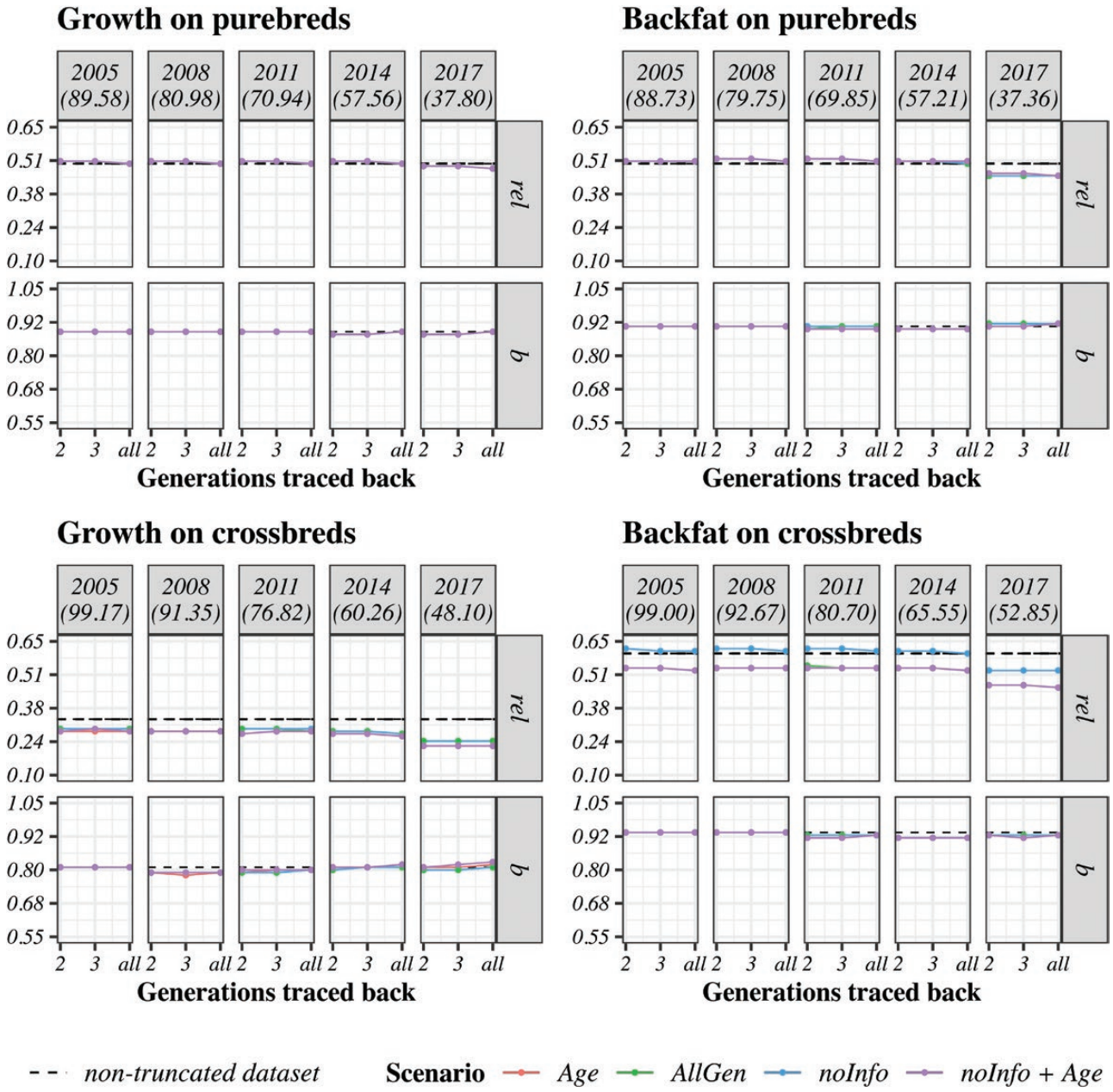

Figure 2. Reliability (rel) and dispersion (b) coefficients on focal individuals for growth traits in the terminal-sire line. The dashed line is the original level of reliability or dispersion for the nontruncated dataset, with all genotyped animals and the entire pedigree. The numbers in parenthesis are the percentage of phenotypes analyzed according to the trait. Abbreviations: *noInfo*, cutting off non-informative genotyped animals; *Age*, cutting off old genotyped animals; *noInfo + Age*, cutting off old genotyped animals, along with non-informative; and *AllGen*, no genotype removal.

truncating data from 2009, whereas the reliability did not change with truncation up to 2012. The truncation did not affect the reproductive traits, and in some cases, the reliability of LS improved after removing old data (Figure 5). In most cases, the different scenarios for removing genotyped animals resulted in the same reliability and dispersion, and discrepancies among scenarios were observed only when keeping less than 10 years of data.

Pedigree and phenotypic truncation

Our results mostly agree with previous studies. Increasing the number of generations traced back in the pedigree from 2 to 3 did not increase the predictive ability of US Holstein bulls (Cesarani et al., 2021). Similarly, keeping all the generations in the pedigree did not improve the predictive ability of young, genotyped pigs as compared to tracing back two generations (Lourenco et al., 2014). In our analyses, reliabilities within the truncated dataset did not increase when we changed from tracing back two to three generations to including all known ancestors in the pedigree. However, using only one generation of data and tracing back one generation in the pedigree reduced GEBV accuracies in a simulated population (Howard et al., 2018).

The number of useful generations with associated phenotype depends on the trait heritability: the lower the heritability, the more generations are needed (Mrode, 2014; Hollifield et al., 2021). The reliability depends on the heritability and the relationship between training and validation sets (Habier

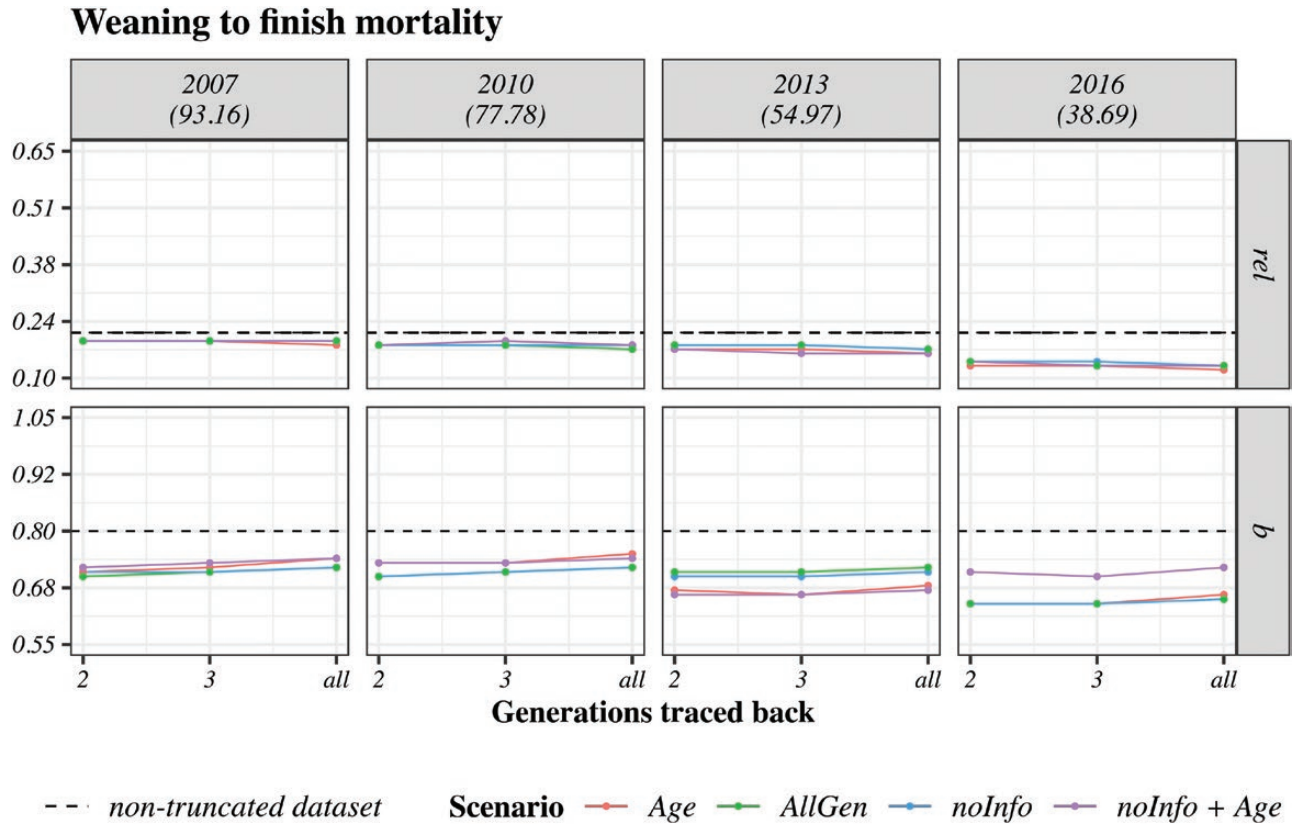

Figure 3. Reliability (rel) and dispersion (b) coefficients on focal individuals for mortality in the terminal-sire line. The dashed line is the original level of reliability or dispersion for the non-truncated dataset, with all genotyped animals and the entire pedigree. The numbers in parenthesis are the percentage of phenotypes analyzed according to the trait. Abbreviations: *noInfo*, cutting off non-informative genotyped animals; *Age*, cutting off old, genotyped animals; *noInfo + Age*, cutting off old, genotyped animals, along with non-informative; and *AllGen*, no genotype removal.

et al., 2010; Pszczola et al., 2012). Weng et al. (2016) found that for lowly heritable traits, having genotyped animals in the training population more related to the validation animals is better. This may be related to changes in the relationships across other animals because genomic relationships are implicitly imputed for non-genotyped animals through $\mathbf{A}_{22}$ in **H**, the joint relationship matrix in ssGBLUP (Legarra et al., 2009). If that is the case, truncating genotypes may affect those traits more when genotyped animals have no phenotypes.

In ssGBLUP, GEBV are predicted as a weighted sum of information from different sources: parent average (**PA**), yield deviation (**YD**), progeny contribution (**PC**), and genomic information (**GI**) (Aguilar et al., 2010; Lourenco et al., 2015a; VanRaden and Wright, 2013). GI can be further split into two parts (Aguilar et al., 2010): direct genomic value (**DGV**) and pedigree prediction (**PP**) that comes from $\mathbf{A}_{22}$. No genotyped animals had phenotypes for WFM; thus, YD = 0. As the information from PA and PP is similar, GEBV will be reduced to PC and DGV. If the progeny number is large, PC will dominate the GEBV; if not, DGV will explain a more significant fraction of GEBV. In such a case, truncating genotypes and progeny phenotypes may impact predictions more.

The GEBV accuracy is less affected by the genotyping structure in ssGBLUP than in multi-step methods because of the additional pedigree information, which accounts for the relationships between a given genotyped animal and the rest of the population (Lourenco et al., 2015a). The composition of the reference population still influences GEBV predictivity; however, the inclusion of genotyped animals without progeny seems not to increase the predictability (Lourenco et al., 2015b). Therefore, removing those genotyped animals with no progeny and no phenotype should not impact the reliability, which was observed in our analyses (*noInfo* scenario). In fact, the *noInfo* scenario slightly improved the reliability of BF_X.

Howard et al. (2018), argued that a slight numerical increase in the predictive ability from truncated data is related to changes in trait definition over time. This definition change can be because of different measurement techniques or selection criteria and environmental changes (Tsuruta et al., 2005). Considering a trait under strong selection and a short generation interval, the population average will rapidly increase or decrease depending on the direction of selection. If that is the case, removing old information from the dataset moves the average of the solutions closer to the desired direction once those animals with more distant breeding values are removed.

If the trait is selected for increasing the average, such as growth and some reproductive traits, the GEBV average from the partial, truncated dataset will be higher than the whole, non-truncated dataset; therefore, the bias will be positive. Conversely, when animals are selected for a reduction in the average, like in mortality, the GEBV average from the partial, truncated dataset will be smaller than the whole, non-truncated, and the bias will be negative. This was observed for all studied traits (data not shown), even though the truncation is supposed to reduce the level of bias (Macedo et al., 2022). The selection effect changes when the partial and whole data

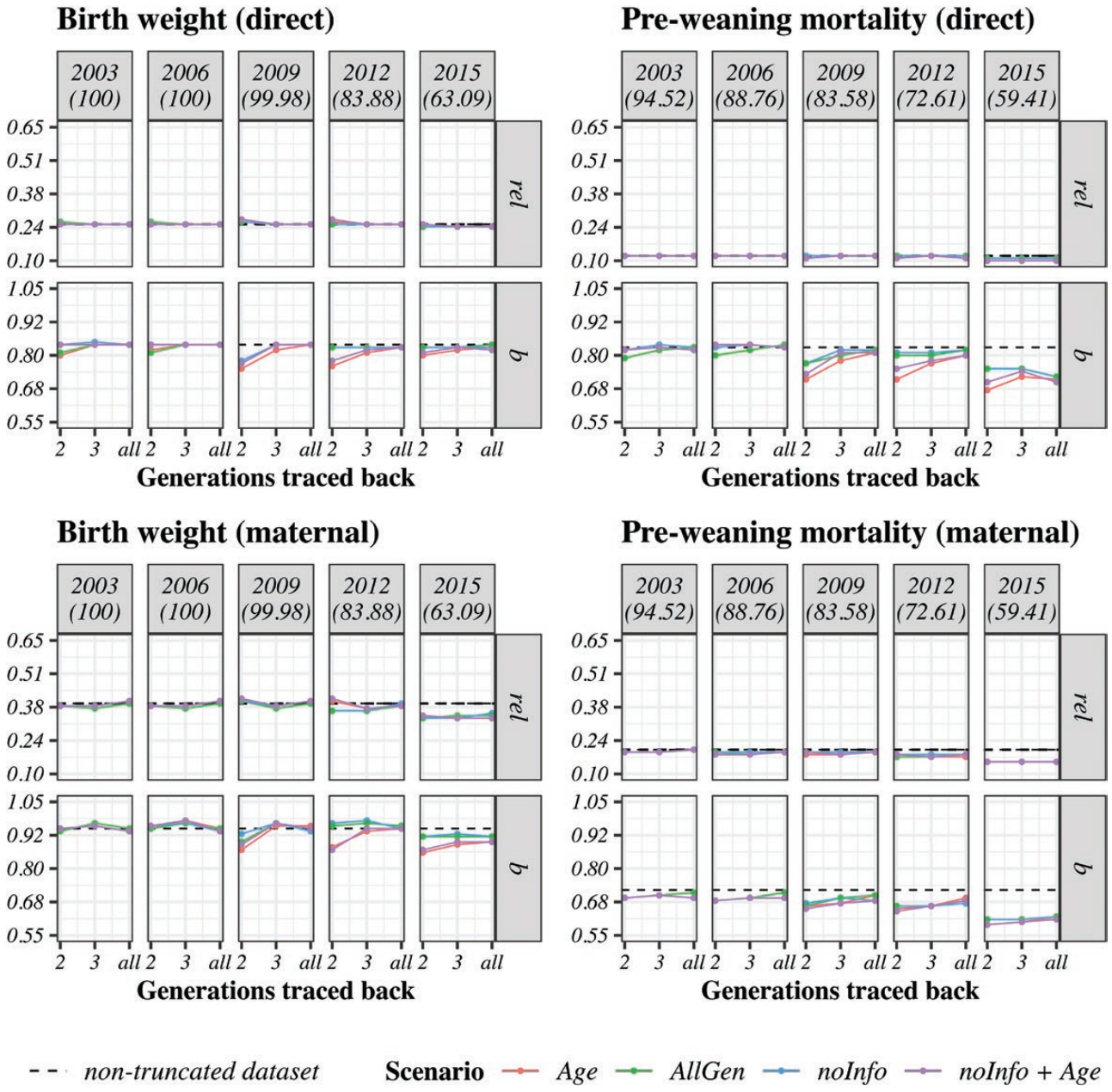

Figure 4. Reliability (rel) and dispersion (b) coefficients for focal individuals for preweaning traits in the maternal line. The dashed line is the original level of reliability or dispersion for the non-truncated dataset, with all genotyped animals and the entire pedigree. The numbers in parenthesis are the percentage of phenotypes analyzed according to the trait. Abbreviations: *noInfo*, cutting off non-informative genotyped animals; *Age*, cutting off old, genotyped animals; *noInfo + Age*, cutting off old, genotyped animals, along with non-informative; and *AllGen*, no genotype removal.

are truncated or not. A bias reduction from truncation was observed for *noInfo* and *noInfo + Age* scenarios (data not shown).

The depth of the pedigree has a minor influence on the reliability of genomic prediction (Lourenco et al., 2014; Howard et al., 2018; Cesarani et al., 2021), which was observed in our study. Additionally, no major difference was found after tracing back two to three generations in the pedigree for all traits. No additional benefit was found by using all pedigree information; in some cases, the impact on predictions was the same as using two generations (PWM_d). According to Macedo et al. (2022), if models used to estimate breeding values are imperfect, a small percentage of bias will affect PA each year, accumulating bias over the years. The same effect can appear if the genetic correlation across distant generations but in the same trait is not one (Tsuruta et al., 2004), which is the case of genotype by environment interaction. Therefore, some analyses can be biased per se, and correcting for this bias would involve improving the model, which is not always easy as linear models have limitations if important factors are not accounted for (Macedo et al., 2022). In that case, removing old information can avoid this bias accumulation. In our study, data truncation helped release the level of bias in many cases (results not shown) but did not affect the dispersion. Thus, data truncation can be an effective alternative to alleviate the bias level without harming the reliability and dispersion of genotyped validation animals.

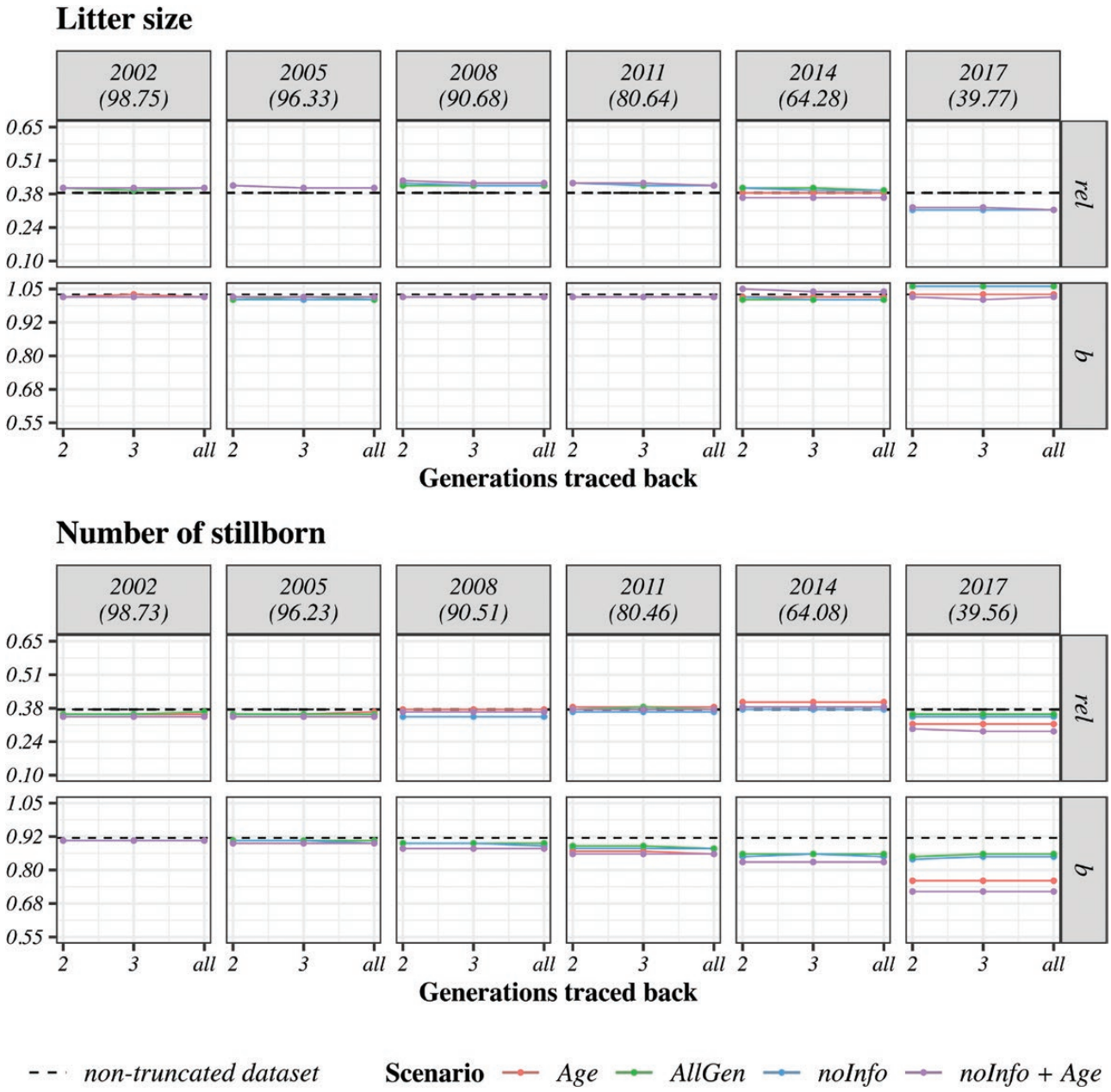

Figure 5. Reliability (rel) and dispersion (b) coefficients for focal individuals for reproductive traits in the maternal line. The dashed line is the original level of reliability or dispersion for the non-truncated dataset, with all genotyped animals and the entire pedigree. The numbers between parenthesis are the percentage of phenotypes analyzed according to the trait. Abbreviations: *noInfo*, cutting off non-informative genotyped animals; *Age*, cutting off old, genotyped animals; *noInfo + Age*, cutting off old, genotyped animals, along with non-informative; and *AllGen*, no genotype removal.

Interestingly, the last phenotypic truncation point showed reduced reliability and increased overdispersion for some traits in M1, M3, and M4. For M4, this could be explained by the extensive removal of phenotypes causing poor estimates of UPG, which have been previously regarded as a source of bias if not enough phenotypes are used to estimate their effects (Tsuruta et al., 2019; Masuda et al., 2022). As M1 included phenotypes from pure and crossbred animals, the last truncation point removed phenotypes from nearly all purebred ancestors of crossbred individuals. The validation was performed only on purebred animals, parents of crossbreds; therefore, more crossbred phenotypes are needed to estimate the GEBV on purebreds for crossbred performance (as in M1). In M3, the prediction of the maternal effects was possibly affected in the last truncation point. As the maternal effects had greater h^2 (Table 1), they caused indirect changes in the direct genetic effect through genetic correlations.

Computing efficiency

Figure 6 shows the number of rounds to convergence for all truncation points across all tested pedigree depths. Mainly, with fewer data, fewer iterations were needed to reach convergence. There was a big difference between using the entire pedigree and tracing back three or two generations from each phenotype truncation year across the four genotype truncation scenarios. This difference can be attributed to the inclusion of many animals with no records and no progeny records when using the complete pedigree, and to the reduced

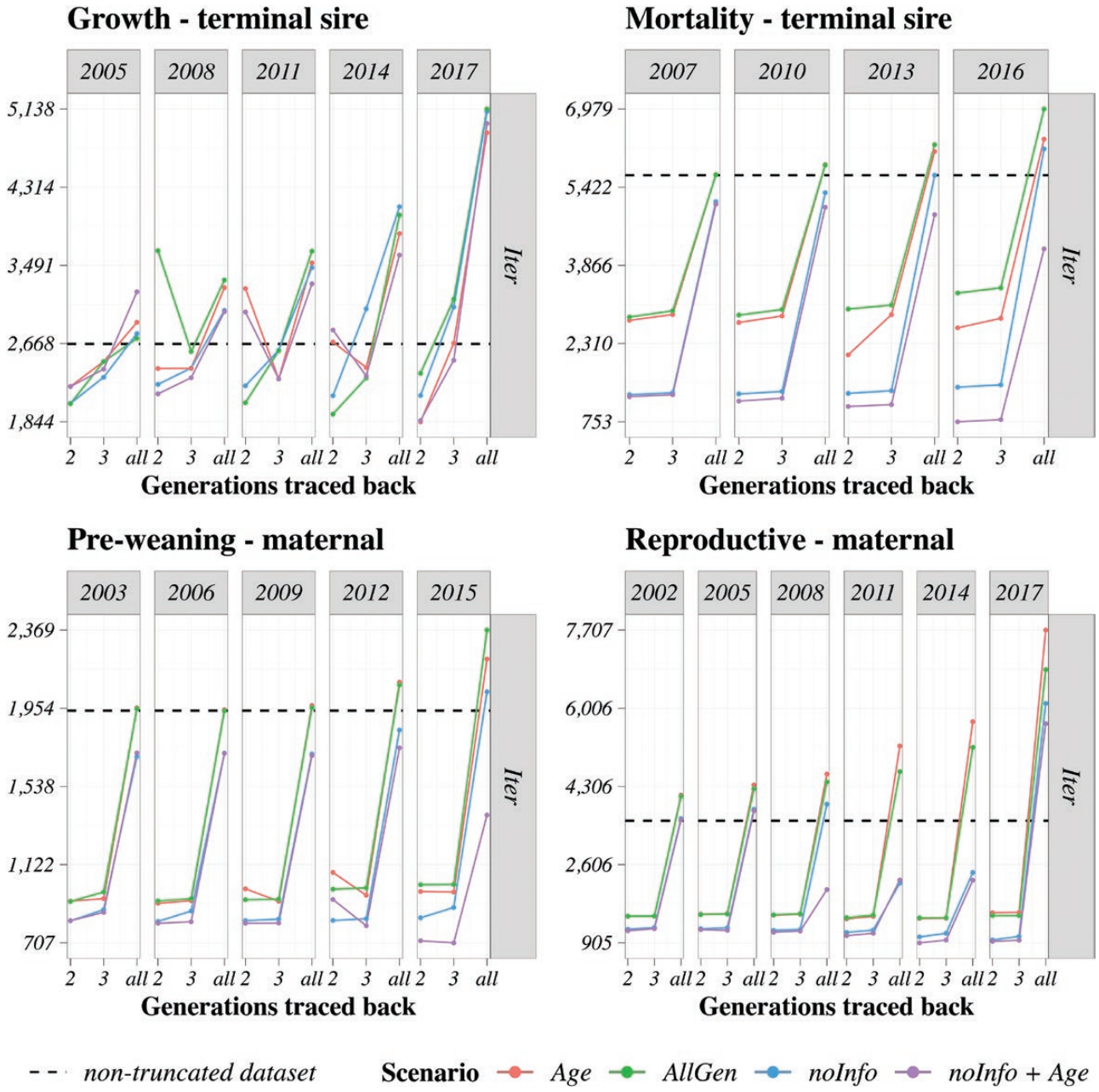

Figure 6. Number of iterations to converge for each phenotypic truncation point across different pedigree depths for all genotype truncation scenarios. The dashed line represents the number of iterations in the non-truncated dataset, with all phenotypes, genotypes, and pedigree. Abbreviations: *noInfo*, cutting off non-informative genotyped animals; *Age*, cutting off old, genotyped animals; *noInfo + Age*, cutting off old, genotyped animals, along with non-informative; and *AllGen*, no genotype removal.

number of equations when analyzing truncated data. Similar results were reported by Pocrnic et al. (2017). The number of iterations is related to the convergence rate: the better the convergence rate, the lesser iterations are needed. The convergence rate on the preconditioned conjugate gradient solver is related to the condition number of the coefficient matrix and the distribution of its eigenvalues (van der Sluis and van der Vorst, 1986; Strakoš, 1991). The eigenvalues distribution will change depending on the number of non-zero elements of the coefficient matrix, which depends on the number and structure of equations.

With truncated data, increasing the number of generations in the pedigree would decrease the convergence rate, with a corresponding increase in the number of iterations (Pocrnic et al., 2017) due to the increase in the number of zeros. For the same system of equations, i.e., for the same truncation point and pedigree depth, differences in the number of iterations could be attributed to numerical accuracy in the computations (Strandén et al., 2017). On the other hand, keeping all genotyped animals but truncating phenotypes and pedigree generally required more iterations to converge. This could be related to the number of non-zero elements of the coefficient matrix since the genomic information usually increases it (Misztal et al., 2021), affecting the distribution of eigenvalues and, consequently, the condition number.

Practical remarks

Overall, data truncation may be affected by generation interval, which has decreased over time due to genomic selection. It can also be subject to changes in the trait definition because the genetic gain accumulated over the past years could result in a trait measured in the current animals being biologically different (and having different distributions) than in older animals. The ability to remove old data can also be affected by changes in variance components over time, which modifies breeding values for the same individuals in the same trait but in different years. Lastly, data accumulate fast, especially genotypes. The more genotyped animals, the higher computing costs. Therefore, constantly reevaluating how many years of data is needed for predicting GEBV on selection candidates is increasingly relevant. For new datasets, one can expect the more widely recorded and the more heritable the trait, the more data can be removed, depending on the generation interval. The number of phenotyped progeny and/or progeny per genotyped animal should also be considered when removing genotyped animals, along with the old information. For a precise assessment of pedigree inbreeding and genetic trends, one should remember that using the whole pedigree is recommended, as these two measures rely on the complete data.

Conclusions

Data truncation is a feasible alternative to reducing the computing costs of genomic evaluations without compromising the GEBV of selection candidates. The generation interval, number of records, and trait heritability influence the decision regarding how many years of data to keep. About ten years of data on widely recorded traits are enough for predictions, whereas more years are needed for traits with fewer records or evaluated only on the progeny. The heritability and the number of traits in the model should also be considered. For lowly heritable traits, more years of data are required to maintain the same level of reliability as in the non-truncated dataset. Fewer years of data can be used if a lowly heritable trait is analyzed with a moderately heritable trait because they are genetically correlated. Using more than three generations of pedigree for animals with phenotypes and/or genotypes does not improve predictions. Additionally, using up to three generations reduces computing costs and helps improve the compatibility between genomic and pedigree relationship matrices. Old, genotyped animals can be eliminated from the analyses without harming predictions for the selection candidates, and the best strategy is to remove genotyped animals without own or progeny phenotypes.

Acknowledgments

Discussions and edits from Jorge Hidalgo, Jennifer Richter, and Joe-Menwer Tabet are highly appreciated. We gratefully acknowledge the valuable comments of the anonymous reviewers.This study was supported by the Pig Improvement Company.

Conflict of interest statement

Ching-Yi Chen and Justin Holl are employees of the Pig Improvement Company. The authors declare no real or perceived conflicts of interest.

Literature Cited

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

Cesarani, A., D. Lourenco, S. Tsuruta, A. Legarra, E. L. Nicolazzi, P. M. VanRaden, and I. Misztal. 2022. Multibreed genomic evaluation for production traits of dairy cattle in the United States using single-step genomic best linear unbiased predictor. *J. Dairy Sci.* 105:5141–5152. doi:10.3168/jds.2021-21505.

Cesarani, A., Y. Masuda, S. Tsuruta, E. L. Nicolazzi, P. M. VanRaden, D. Lourenco, and I. Misztal. 2021. Genomic predictions for yield traits in US Holsteins with unknown parent groups. *J. Dairy Sci.* 104:5843–5853. doi:10.3168/jds.2020-19789.

Chen, C. Y., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2011. Effect of different genomic relationship matrices on accuracy and scale. *J. Anim. Sci.* 89:2673–2679. doi:10.2527/jas.2010-3555.

Habier, D., J. Tetens, F. -R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genet. Sel. Evol.* 42:5. doi:10.1186/1297-9686-42-5.

Henderson, C. R. 1984. *Applications of linear models in animal breeding models*. Guelph, Canada: University of Guelph.

Hollifield, M. K., D. Lourenco, M. Bermann, J. T. Howard, and I. Misztal. 2021. Determining the stability of accuracy of genomic estimated breeding values in future generations in commercial pig populations. *J. Anim. Sci.* 99:1–8. doi:10.1093/jas/skab085.

Howard, J. T., T. A. Rathje, C. E. Bruns, D. F. Wilson-Wells, S. D. Kachman, and M. L. Spangler. 2018. The impact of truncating data on the predictive ability for single-step genomic best linear unbiased prediction. *J. Anim. Breed. Genet.* 135:251–262. doi:10.1111/jbg.12334.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.* 92:4656–4663. doi:10.3168/jds.2009-2061.

Legarra, A., and A. Reverter. 2018. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. *Genet. Sel. Evol.* 50:53. doi:10.1186/s12711-018-0426-6.

Lourenco, D. A. L., B. O. Fragomeni, S. Tsuruta, I. Aguilar, B. Zumbach, R. J. Hawken, A. Legarra, and I. Misztal. 2015a. Accuracy of estimated breeding values with genomic information on males, females, or both: an example on broiler chicken. *Genet. Sel. Evol.* 47:56. doi:10.1186/s12711-015-0137-1.

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

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

Macedo, F. L., J. M. Astruc, T. H. E. Meuwissen, and A. Legarra. 2022. Removing data and using metafounders alleviates biases for all traits in Lacaune dairy sheep predictions. *J. Dairy Sci.* 105:2439–2452. doi:10.3168/jds.2021-20860.

Macedo, F. L., O. F. Christensen, and A. Legarra. 2021. Selection and drift reduce genetic variation for milk yield in Manech Tete Rousse dairy sheep. *JDS Commun.* 2:31–34. doi:10.3168/jdsc.2020-0010.

Masuda, Y., I. Misztal, A. Legarra, S. Tsuruta, D. A. L. Lourenco, B. O. Fragomeni, and I. Aguilar. 2017. Technical note: avoiding the direct inversion of the numerator relationship matrix for genotyped animals in single-step genomic best linear unbiased prediction solved with the preconditioned conjugate gradient. *J. Anim. Sci.* 95:49–52. doi:10.2527/jas.2016.0699.

Masuda, Y., P. M. VanRaden, S. Tsuruta, D. A. L. Lourenco, and I. Misztal. 2022. Invited review: unknown-parent groups and metafounders in single-step genomic BLUP. *J. Dairy Sci.* 105:923–939. doi:10.3168/jds.2021-20293.

Misztal, I., I. Aguilar, D. Lourenco, L. Ma, J. P. Steibel, and M. Toro. 2021. Emerging issues in genomic selection. *J. Anim. Sci.* 99:1–14. doi:10.1093/jas/skab092.

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

Misztal, I., S. Tsuruta, D. A. L. Lourenco, Y. Masuda, I. Aguilar, A. Legarra, and Z. G. Vitezica. 2014b. Manual for BLUPF90 Family of Programs. [Accessed March 3,2023]. http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all8.pdf.

Misztal, I., S. Tsuruta, I. Pocrnic, and D. Lourenco. 2020. Core-dependent changes in genomic predictions using the algorithm for proven and young in single-step genomic best linear unbiased prediction. *J. Anim. Sci.* 98:skaa374–8. doi:10.1093/jas/skaa374.

Misztal, I., Z. G. Vitezica, A. Legarra, I. Aguilar, and A. A. Swan. 2013. Unknown-parent groups in single-step genomic evaluation. *J. Anim. Breed. Genet.* 130:252–258. doi:10.1111/jbg.12025.

Mrode, R. 2014. *Linear models for the prediction of animal breeding values*. 3rd ed. Oxford: CABI Publishing.

OpenMP Architecture Review Board. 2015. OpenMP application program interface. [AccessedMarch 10,2023]. https://www.openmp.org/wp-content/uploads/openmp-4.5.pdf.

Pocrnic, I., D. A. L. Lourenco, H. L. Bradford, C. Y. Chen, and I. Misztal. 2017. Technical note: impact of pedigree depth on convergence of single-step genomic BLUP in a purebred swine population. *J. Anim. Sci.* 95:3391–3395. doi:10.2527/jas.2017.1581.

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

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

Pszczola, M., T. Strabel, H. A. Mulder, and M. P. L. Calus. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. *J. Dairy Sci.* 95:389–400. doi:10.3168/jds.2011-4338.

Quaas, R. L. 1988. Additive genetic model with groups and relationships. *J. Dairy Sci.* 71:91–98. doi:10.1016/s0022-0302(88)79986-5.

Quaas, R. L., and E. J. Pollak. 1981. Modified equations for sire models with groups. *J. Dairy Sci.* 64:1868–1872. doi:10.3168/jds.s0022-0302(81)82778-6.

R Core Team. 2020. R: a language and environment for statistical computings, Vol. 1. R. D. C. Team, editor. R Foundation for Statistical Computing. p. 409. doi:10.1007/978-3-540-74686-7.

van der Sluis, A., and H. A. van der Vorst. 1986. The rate of convergence of conjugate gradients. *Numer Math.* 48:543–560. doi:10.1007/bf01389450.

Sorensen, D., R. Fernando, and D. Gianola. 2001. Inferring the trajectory of genetic variance in the course of artificial selection. *Genet. Res.* 77:83–94. doi:10.1017/s0016672300004845.

Strakoš, Z. 1991. On the real convergence rate of the conjugate gradient method. *Linear Algebra Appl.* 154-156:535–549. doi:10.1016/0024-3795(91)90393-b.

Strandén, I., and E. A. Mäntysaari. 2014. Comparison of some equivalent equations to solve single-step GBLUP. In: Proceedings of the 10th World Congress of Genetics Applied to Livestock Production. Vancouver. Wageningen Academic Publishers, p. 069.

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

Tsuruta, S., D. A. L. Lourenco, Y. Masuda, I. Misztal, and T. J. Lawlor. 2019. Controlling bias in genomic breeding values for young genotyped bulls. *J. Dairy Sci.* 102:9956–9970. doi:10.3168/jds.2019-16789.

Tsuruta, S., and I. Misztal. 2008. Technical note: computing options for genetic evaluation with a large number of genetic markers. *J. Anim. Sci.* 86:1514–1518. doi:10.2527/jas.2007-0324.

Tsuruta, S., I. Misztal, and T. J. Lawlor. 2004. Genetic correlations among production, body size, udder, and productive life traits over time in Holsteins. *J. Dairy Sci.* 87:1457–1468. doi:10.3168/jds.S0022-0302(04)73297-X.

Tsuruta, S., I. Misztal, and T. J. Lawlor. 2005. Changing definition of productive life in US Holsteins: effect on genetic correlations. *J. Dairy Sci.* 88:1156–1165. doi:10.3168/jds.S0022-0302(05)72782-X.

Tsuruta, S., I. Misztal, and I. Strandén. 2001. Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed-model equations in animal breeding applications. *J. Anim. Sci.* 79:1166–1172. doi:10.2527/2001.7951166x.

VanRaden, P. M., and J. R. Wright. 2013. Measuring genomic pre-selection in theory and in practice. *Interbull Bull.* 47:147–150. https://journal.interbull.org/index.php/ib/article/view/1780.

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

Weng, Z., A. Wolc, X. Shen, R. L. Fernando, J. C. M. Dekkers, J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, and D. J. Garrick. 2016. Effects of number of training generations on genomic prediction for various traits in a layer chicken population. *Genet. Sel. Evol.* 48:22. doi:10.1186/s12711-016-0198-9.

Westell, R. A., R. L. Quaas, and L. D. van Vleck. 1988. Genetic groups in an animal model. *J. Dairy Sci.* 71:1310–1318. doi:10.3168/jds.s0022-0302(88)79688-5.

Wickham, H. 2016. *ggplot2: elegant graphics for data analysis*. 2nd ed. New York, NY: Springer International Publishing.