Genomic models account for genomic preselection by correctly estimating Mendelian sampling terms of preselected animals

I. Jibrila[1], G. de Jong[1], and M.P.L. Calus[2]

[1]Animal Evaluation Unit, CRV u.a., P.O. Box 454, 6800 AL Arnhem, the Netherlands
[2]Wageningen University & Research Animal Breeding and Genomics, PO Box 338, 6700 AH Wageningen, the Netherlands

Abstract

It has been previously shown that in breeding programs that do not use external data, genomic models estimate breeding values of preselected animals without preselection bias and with minimal accuracy loss, as long as genotypes of preselected animals and of their parents are used in the evaluation. The objective of this paper was to show that genomic models account for genomic preselection (GPS) by correctly estimating the Mendelian sampling terms (MSTs) of preselected animals. We simulated a single-trait breeding goal with heritability of 0.1, and 15 recent generations undergoing selection. To select the parents of the next generation from the animals in the most recent generation, we genomically preselected 10% of males and 15% of females in generation 15. We then performed evaluations of the preselected animals with both genomic and pedigree models, both including and excluding records on the preselected animals. We also conducted another set of genomic and pedigree evaluations without preselection, to serve as control. Results showed that both the true and estimated MSTs in the control scenario were on average zero, regardless of whether they were estimated with genomic or pedigree models. With GPS, the average true MST was positive, was correctly estimated by genomic models, and hugely underestimated by pedigree models. Compared to the MSTs estimated by pedigree models, the MSTs estimated by genomic models in both GPS and control scenarios had variances that were closer to the variances of the corresponding true MSTs. We concluded that genomic models indeed correctly estimate Mendelian sampling terms of preselected animals, and that how they are able to account for GPS.

Key words: Genomic preselection, genomic evaluation models, Mendelian sampling terms, bias

Introduction

Using both simulated and real data from animal breeding programs that do not use external data such as MACE proofs, Jibrila (2022) showed that genomic models such as single-step genomic best linear unbiased prediction (ssGBLUP) estimate breeding values of preselected animals without preselection bias and with minimal accuracy loss, as long as genotypes of preselected animals and of their parents are used in the evaluation. In this paper, we showed that genomic models account for genomic preselection (GPS) by correctly estimating Mendelian sampling terms (MSTs) of preselected animals.

Materials and Methods

We simulated a single-trait breeding goal, with heritability of 0.1. We produced 15 recent generations with pedigree BLUP (PBLUP)-based selection, after a historical population of 3000 generations of random mating. In every generation of the recent population, we produced 8000 male and 8000 female offspring, from which 100 males and 1000 females were selected to produce the next generation. For this study, we used the entire pedigree of the recent population, genotypes of

the three most recent generations (i.e. generations 13 to 15) and phenotypes of the five most recent generations (i.e. generations 11 to 15). To select the parents of the next generation from the animals in the most recent generation, we genomically preselected 10% of males and 15% of females in generation 15. We used ssGBLUP with the entire pedigree of the recent population, genotypes of animals in generations 13 to 15, and phenotypes of animals in generations 11 to 14 to produce the genomically enhanced estimated breeding values (GEBVs) used to perform this preselection. After preselection, we conducted two evaluations, one with ssGBLUP and the other with PBLUP, using the data remaining after preselection (i.e. without any information from preculled animals/animals removed from the breeding program at preselection stage). These evaluations included the entire pedigree of the recent population until the preselected animals in generation 15, genotypes of animals in generations 13 until the preselected animals in generation 15, and phenotypes of animals in generations 11 to 14. We refer to these evaluations as 'initial evaluations', because they mimicked the evaluations that take place mainly in dairy cattle before the preselected animals have own or daughter records. We then conducted another set of two evaluations, one ssGBLUP-based and the other PBLUP-based, with the same data as in initial evaluations, plus own records of the preselected animals. We refer to these evaluations as 'subsequent evaluations', because they mimicked the evaluations that take place in breeding programs of all livestock species when the preselected animals have own or progeny records. We also conducted another set of subsequent evaluations on the entire kept recent data without preselection, using both ssGBLUP and PBLUP, to serve as control. We computed averages and variances of MSTs of preselected animals from all the above six evaluations.

Because this is a simulated dataset, we knew the true MSTs of preselected animals. So, we also computed the means and variances of the true MSTs of preselected animals in both the control and GPS scenarios. Everything (i.e. from the simulation of the dataset to subsequent evaluations and computation of means and variances of MSTs) was replicated 10 times. More details on the design and analysis of the simulated data are in Jibrila et al. (2021). We redid the ssGBLUP evaluations with GBLUP, and we recorded statistically similar results. Therefore, we decided to report results from ssGBLUP to represent results from genomic models.

Results & Discussion

Means of MSTs, in genetic standard deviation units (GSDs) of the trait, and averaged over 10 replicates, are in Figure 1. Both the true and estimated MSTs in the control scenario are on average zero, regardless of whether the estimated MSTs came from ssGBLUP or from PBLUP. With GPS, the average true MST is positive (0.57 ± 0.01 GSD), and was correctly estimated by ssGBLUP both at the initial evaluation (0.57 ± 0.01 GSD) and subsequent evaluation (0.57 ± 0.01 GSD). The deviation of average true and estimated MSTs from zero with GPS is expected, since GPS preselects on average animals with positive MSTs. PBLUP greatly underestimated the average MSTs of the preselected animals, at both the initial evaluation (0.00 ± 0.00 GSD) and the subsequent evaluation (0.02 ± 0.00 GSD). Variances of MSTs, also in GSDs of the trait and averaged over 10 replicates, are in Figure 2. True MSTs in the control and GPS scenarios have variances of 1.19 ± 0.08 and 0.86 ± 0.03 GSDs, respectively. The decrease in variance of true MSTs from control to GPS scenario is because GPS preselected more genetically (and genomically) similar animals. Both ssGBLUP and PBLUP underestimated the variances of MSTs of both control and GPS scenarios, both at initial and subsequent evaluations. This is because both ssGBLUP and PBLUP predicted

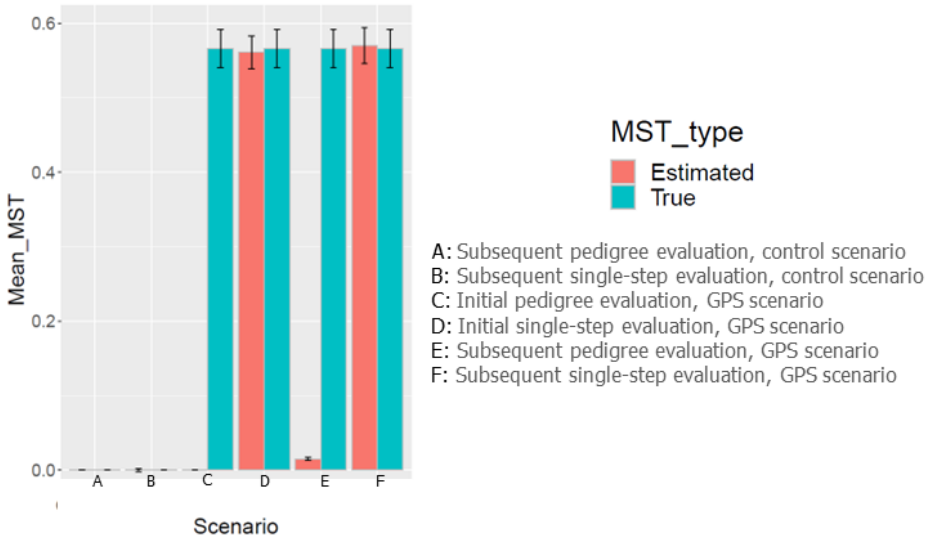

Figure 1 Mean Mendelian sampling terms of preselected animals, expressed in genetic standard deviation units of the trait, and averaged over 10 replicates

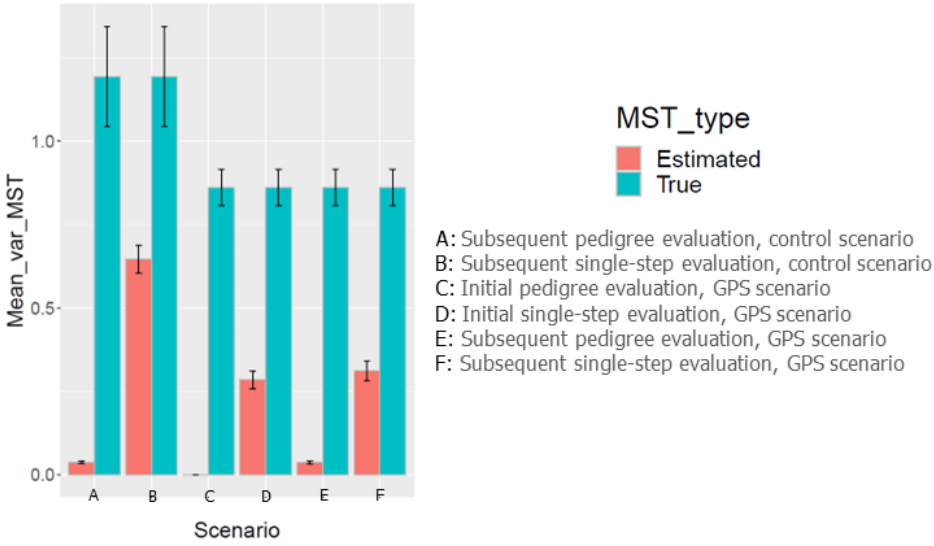

Figure 2 Variances of Mendelian sampling terms of preselected animals, expressed in genetic standard deviation units of the trait, and averaged over 10 replicates

the true breeding values with accuracies less than 1. Estimated MSTs by ssGBLUP and PBLUP in the subsequent evaluation of the control scenario have variances of 0.65 ± 0.02 and 0.04 ± 0.00 GSDs, respectively. In the initial evaluation of the GPS scenario, MSTs estimated by ssGBLUP and PBLUP have variances of 0.29 ± 0.01 and 0.00 ± 0.00, respectively. Finally, in the subsequent evaluation of the GPS scenario, MSTs estimated by ssGBLUP and PBLUP have variances of 0.31 ± 0.02 and 0.04 ± 0.00, respectively. This means that the compared to the MSTs estimated by PBLUP, MSTs estimated by ssGBLUP at both initial at subsequent evaluations have variances that are closer to the variances of the corresponding true MSTs.

Using both simulated and real data from animal breeding programs that do not use external data such as MACE proofs, Jibrila (2022) showed that genomic models such as single-step genomic best linear unbiased prediction (ssGBLUP) estimate breeding values of preselected animals without preselection bias and with minimal accuracy loss. Jibrila (2022) also showed, using a full sib family of 6 individuals, how preselection on a correlated trait is accounted for by pedigree and genomic models in both single-trait and two-trait subsequent evaluations. The author showed that MSTs of preselected and preculled animals are better differentiated moving from single-trait PBLUP to two-trait PBLUP, to single-trait ssGBLUP, to two-trait ssGBLUP. The author concluded that the MSTs of the preselected (and preculled) animals were closer to their true values, accuracy of prediction was higher, and bias was lower, moving from single-trait PBLUP to two-trait PBLUP, to single-trait ssGBLUP, to two-trait ssGBLUP. True MSTs of the preselected (and preculled) animals were not known in Jibrila (2022), because real data was used for this particular exercise. In the current study, we used simulated data and showed that genomic models are indeed able to correctly estimate the average MSTs of preselected animals using just the genotypes of the preselected animals and their parents, as shown by the 'initial' evaluation of the GPS scenario in the current study.

To show why genomic models are able to correctly estimate MSTs of preselected animals using just the genotypes of preselected animals and of their parents, we extracted genomic relationships among full sibs and among half sibs in generation 15 from the genomic relationship matrix (**G**), and their corresponding entries in the inverse of the genomic relationship matrix ($\mathbf{G}^{-1}$), in both the GPS and the control scenarios of this study. In Table 1, we see that among full sibs and among half sibs, genomic relationships are statistically higher, and their corresponding entries in $\mathbf{G}^{-1}$ are statistically lower, in the GPS scenario than in the control scenario.

Table 1 Average genomic relationships (G) and their corresponding entries in the inverse of the genomic relationship matrix (G^{-1}) among different groups of preselected animals

Parameter	Among	Preselection scenario			
		Control	GPS	Difference	Significance of the difference (p-value)
G	Full sibs	0.68	0.69	-0.01	0.01
	Half sibs	0.50	0.51	-0.01	0.00
G^{-1}	Full sibs	0.00	-0.02	0.02	0.00
	Half sibs	0.00	-0.01	0.01	0.00

Control: scenario without preselection, GPS: scenario with genomic preselection

The idea that within a family preselected animals tend to have higher genomic relationships among themselves than among all sibs in the family has also been reported by Gondro et al. (2013), Hayes et al. (2009) and VanRaden (2008). The higher genomic relationships and the lower corresponding $\mathbf{G}^{-1}$ entries among preselected animals are reflected in their coefficients of the inverse of the left hand side of the mixed model equation. This reflection ensures that preselected animals get the positive MS terms they truly have, regardless of whether or not the preselected animals have own/progeny records in the evaluation.

It is common to observe bias in genomic evaluations in dairy cattle if external information such as MACE proofs is included in the evaluations (e.g. Patry et al., 2013). Countries participating in MACE send deregressed EBVs (DEBVs, produced using PBLUP) to the Interbull Centre, and the Interbull Centre uses these DEBVs to produce MACE proofs. Then participating countries integrate these MACE proofs into their national genomic evaluations. Because these MACE proofs are produced from DEBVs that are themselves produced without using genomic information, MACE proofs of genomically preselected animals may be downward biased (e.g. Patry et al., 2013). This downward bias is transmitted to national genomic evaluations that use MACE proofs. If deregressed GEBVs (DGEBVs, instead of DEBVs) are submitted to the Interbull Centre, then the MACE proofs produced by the Interbull Centre will be free of GPS bias. However, this will result in double-counting of genomic information if these MACE proofs are integrated in national genomic evaluations. This is why DEBVs (instead of DGEBVs) are still sent to the Interbull Centre. The 'Future MACE' working group of the Interbull Centre is now working toward coming up with ways(s) of accounting for GPS in both national and Interbull PBLUP evaluations, without necessarily utilizing genomic information (e.g. Sullivan et al., 2019, 2022).

Conclusions

As long as genotypes of preselected animals and of their parents are used, genomic evaluation systems that do not use external data, such as closed-line breeding and national genomic evaluations of dairy (and beef) cattle that do not integrate foreign data, are expected to estimate GEBVs of preselected animals without preselection bias, even if the preselected animals do not have own or progeny records in the evaluations. This is because use of genomic information enables such systems/evaluation models to correctly estimate the average Mendelian sampling terms of preselected animals.

Acknowledgments

This work is a follow up to the work undertaken during Ibrahim's PhD program. Ibrahim thanks Jan ten Napel, Jeremie Vandenplas, and Roel Veerkamp, all from WUR-ABG, for their invaluable contributions to his PhD work. Ibrahim also thanks Peter Sullivan of Lactanet for the discussions they had, which motivated writing this paper.

References

Gondro, C., van der Werf, J., & Hayes, B. (Eds.). (2013). *Genome-Wide Association Studies and Genomic Prediction*. Springer. https://doi.org/10.1007/978-1-62703-447-0

Hayes, B. J., Visscher, P. M., & Goddard, M. E. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research*, *91*(1), 47–60. https://doi.org/10.1017/S0016672308009981

Jibrila, I. (2022). *Impact of preselection in genomic evaluations*. Wageningen University. https://doi.org/https://doi.org/10.18174/563471

Jibrila, I., Vandenplas, J., ten Napel, J., Veerkamp, R. F., & Calus, M. P. L. (2021). Avoiding preselection bias in subsequent single-step genomic BLUP evaluations of genomically preselected animals. *Journal of Animal Breeding and Genetics*, *138*(4), 432–441. https://doi.org/10.1111/jbg.12533

Patry, C., Jorjani, H., & Ducrocq, V. (2013). Effects of a national genomic preselection on the international genetic evaluations. *Journal of Dairy Science*, *96*(5), 3272–3284. https://doi.org/10.3168/jds.2011-4987

Sullivan, P. G., Mäntysaari, E. A., de Jong, G., & Savoia, S. (2022). Using genetic regressions to account for genomic preselection effects in MACE. *Interbull Bulletin*, *57*, 117–124.

Sullivan, P. G., Mäntysaari, E. A., de Jong, G., & Benhajali, H. (2019). Modifying MACE to accommodate genomic preselection effects. *Interbull Bulletin*, *55*, 77–80.

VanRaden, P. M. (2008). Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science*, *91*(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980