Using genetic regressions to account for genomic preselection effects in MACE

P.G. Sullivan[a], E.A. Mäntysaari[b], G. de Jong[c], S. Savoia[d]

[a]Lactanet, 660 Speedvale Ave, Guelph, Ontario, Canada, N1K 1E5
[b]Natural Resources Institute (Luke), PO Box 2, FI-00791 Helsinki, Finland
[c]CRV u.a., PO Box 454, 6800 AL Arnhem, The Netherlands
[d]Interbull Centre, SLU, Box 7023, S-75007 Uppsala, Sweden

Abstract

National genomic evaluation systems use foreign sire evaluations from MACE as phenotypic information, combined with genotypes to generate national GEBV. The national GEBV computed from MACE cannot be used as input data for MACE without double-counting genomic information. To avoid this double-counting, Interbull requires that national EBV provided as input for MACE must be computed without genotypes, even though this leads to known biases in MACE results. The biases are due to sire pre-selection effects based on genotypes being partially treated as effects of the sires' mates and the herd environments of the sires' progeny when genotypes are excluded from the national evaluations. This causes an underprediction of genetic levels for genomically pre-selected sires and of estimated genetic trends, for both the input data and the output results of MACE. The current MACE model was therefore expanded by adding estimates of GPS (genomic preselection) effects, and with corresponding modifications to the yearly means of estimated breeding values for genomically pre-selected sires. Segmented genetic regressions were used to estimate evolving international trends in pre-selection effects since 2009, and genetic grouping was used to include pre-selection estimates in the genetic evaluations of genomically pre-selected sires. Data simulation was used to validate the expanded MACE model, under a scenario where GPS effects are fully included, and GPS biases are thus zero in the national input data. The new MACE model properly separated within-family from between-family pre-selection effects in the simulated data and effectively removed pre-selection biases observed under the current MACE model. Estimated pre-selection effects were relatively small from official data but are expected to increase as national models used to generate MACE input data will be updated to reduce genomic pre-selection biases in the future. Further improvements are also being planned for the new MACE model, to account for expected genetic variance reductions with the elevated means for GPS bulls due to selection.

Key words: international evaluation, MACE, genomics, selection bias, double-counting

Introduction

Genomic preselection (**GPS**) can have significant effects on distributions of genetic values, and on the genetic evaluations of AI sires (Ducrocq and Patry, 2010; Patry and Ducrocq, 2011, Masuda et al, 2018).

Evaluation models used currently for MACE include assumptions that young bull pre-selection is limited to between-family differences (parent average pre-selection), and that non-random within-family selection (Mendelian sampling pre-selection) is not possible before progeny testing. This latter assumption is violated when genotypes are used to pre-select only the better bulls within a family before progeny testing. Ignoring GPS effects can therefore cause an underestimation of EBV for genomically pre-selected bulls, and wrong estimates of genetic trend. International sire comparisons are biased because of this violated assumption (Patry et al, 2013; Fikse 2014; Schaeffer, 2018; Sullivan et al, 2019). Hereafter, we refer to the biases caused by ignoring GPS effects as GPS bias.

The purposes of the present study were to reduce GPS bias by expanding the MACE model used by Interbull, with new factors to estimate and account for GPS effects on the Mendelian sampling (**MS**) estimates of recent AI sires. Simulated data were used to validate the expanded model, and official MACE input data were used to assess potential impacts of implementing the new model in practice.

Materials and Methods

Simulated data

True genetic values of AI sires were simulated, first without and then with GPS effects included. The genetic trends in data simulated without GPS effects were consistent with EBV trends of proven bulls from a MACE evaluation conducted prior to the GPS era, which began with dairy bulls born in 2009 in North America (VanRaden et al, 2009; Schenkel et al, 2009) and at a similar time (de Roos et al, 2009) but with expanding applications more recently across Europe (Lund et al, 2011). The first GPS bulls reached progeny proven status some time around 2014. Official MACE evaluations conducted in April 2014, eight years ago, would presumably have true GPS effects equal or close to zero in all or most countries participating in MACE.

The official MACE pedigree, and proofs for protein yield from April 2014, were therefore used as the basis to simulate true breeding values, using simulation methods described by Tyrisevä et al (2018). The MS deviations of AI bulls were randomly sampled from separate distributions by birth year, using expectations for true MS means that matched averages of MS estimates from the official MACE proofs and pedigrees.

Effects of GPS were incorporated in the simulation by assuming a fixed and high level of GPS intensity had been used in a single country participating in MACE. A constant and fixed level of genetic superiority was added to the expected means of MS values, with corresponding reductions in the MS variances. The GPS effects were incorporated in both the simulated true breeding values and the de-regressed national EBV, of bulls with national progeny proofs in 2014 who were registered in the first country of evaluation and born in the most recent eight-year period from 2001 to 2008. A single international replicate was simulated to verify consistency between estimates of GPS effects and the simulated GPS levels on each country scale of evaluation.

Official MACE input data

Input data used for official MACE evaluations in April 2022 were re-evaluated using the expanded MACE model, to study GPS estimates from real data and to measure the impacts of GPS estimates on MACE evaluation results of sires and their parents.

We evaluated a production trait (protein yield: PRO), a conformation trait (overall udder score: OUS), an udder health trait (somatic cell score: SCS), a fertility trait (cow conception trait 1: CC1), and a workability trait (milking speed: MSP). Intensities of GPS were expected to differ among these traits, among countries participating in MACE, and across time. The impacts of expanding the MACE model were expected to be larger for trait by country by time period combinations with higher levels of GPS intensity, and the impacts should tend towards zero for combinations with low levels of GPS intensity.

Evaluation model

The current MACE model is described as:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Z}\mathbf{Q_1}\mathbf{g} + \mathbf{Z}\mathbf{a} + \mathbf{e} \qquad [1]$$

Vector **y** contains de-regressed national EBV of progeny-proven sires from each country, **μ** has solutions for the mean in each country, $\mathbf{Q_1}$ links sires to unknown parent group solutions in vector **g**, vector **a** has solutions for sire effects, matrix **Z** links sires to observations in **y**, and **e** is a vector of residuals. Our expanded MACE model is:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Z}\mathbf{Q_1}\mathbf{g} + \mathbf{Z}\mathbf{Q_2}\mathbf{s} + \mathbf{Z}\mathbf{a} + \mathbf{e} \qquad [2]$$

In our expanded model, $\mathbf{Q_2}$ links sires to an additional set of solutions we added for genomic preselection effects, which are in vector $\mathbf{s}$ of order equal to the number of evaluated populations times the number of estimated regression segments per population (three segments for our current application). Matrix $\mathbf{Q_2}$ includes international genetic regressions from the scales of selecting countries to each foreign country scale of evaluation, assuming that country of registration was also the country of bull origin and pre-selection for AI.

After QP transformation of the **g** effects for phantom parent genetic groups, the mixed model equations for the expanded model are:

$$\begin{bmatrix} \mathbf{X'DX} & \mathbf{X'DZ} & \mathbf{X'DZQ_2} \\ \mathbf{Z'DX} & \mathbf{Z'DZ} + \mathbf{W} \otimes \mathbf{G}_t^{-1} & \mathbf{Z'DZQ_2} \\ \mathbf{Q_2}'\mathbf{Z'DX} & \mathbf{Q_2}'\mathbf{ZDZ} & \mathbf{Q_2}'\mathbf{Z'DZQ_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{Q_1g} + \mathbf{a} \\ \mathbf{s} \end{bmatrix} = \begin{bmatrix} \mathbf{X'Dy} \\ \mathbf{Z'Dy} \\ \mathbf{Q_2}'\mathbf{Z'Dy} \end{bmatrix}$$

Matrix **D** is a diagonal matrix of EDC divided by residual variances, **W** in the inverted matrix of additive relationships among sires and genetic groups (Westell et al, 1988), modified to treat groups as random effects (Sullivan and Schaeffer, 1994), and $\mathbf{G_t}$ is the matrix of genetic covariances among countries.

Segmented linear regressions were used to estimate the trends in GPS effects, knotting at 3-year intervals. The intensities of GPS were assumed equal to zero prior to 2009, to be at a constant non-zero level during the most recent period, and to be changing at any possible combination of different rates, during the two consecutive 3-year periods of transition included in our evaluated data. All estimated slopes of change and the current levels of GPS were based on MACE input data, the deregressed national EBV, and were therefore specific to each trait and country. There was no requirement for genotypes or for prior genotype analyses for the estimation of these GPS effects.

A practical ongoing implementation of this model will require adding new knots in the design as time passes and new data accumulate. We recommend that when adding new knots, the current levels of GPS are always based on a minimum of 2 completed birth years of bulls, and that historical knot locations are kept constant if possible.

Results & Discussion

All results are presented on standardized scales, where national EBV of progeny-proven bulls born since 1980 follow the standard normal distribution (~N(0,1)) in each country.

Simulated data

Although simulated data included GPS effects, even the current MACE model [1] showed generally small GPS biases in the international proofs (EBV) of AI bulls. There was a notable downward bias only in the EBV of bulls born in the most recent year (Figure 1), which matches similar patterns observed in previous studies of national EBV (Masuda et al, 2018).

For all birth years of GPS, there is a partitioning of GPS effects between the within versus across-family components of selection, but since our simulation included GPS effects as being fully due to within-family selection, our estimates of MS were biased downwards with offsetting upward biases in PA, for all birth years of GPS bulls. These offsetting biases were of essentially equal magnitudes for all birth years except the most recent one. Although the EBV for progeny-proven bulls were generally unbiased for these simulated data, the EBV biases for bull dams were much larger than for the bulls themselves. The GPS biases were generally small in MACE results of the proven bulls due to the nature of MACE as a meta-analysis, and with the use of de-regressed national EBV that were simulated to have zero GPS bias.

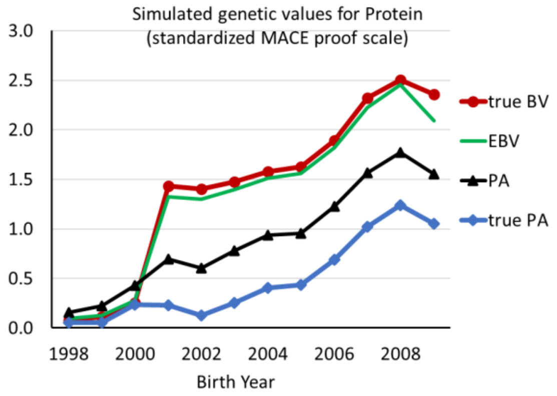

Figure 1. Simulated true values, including GPS effects for bulls born since 2001, and corresponding estimates of breeding value (EBV) and parent averages (PA), on the GPS country scale of evaluation, under the current MACE model [1].

The partitioning of estimates for PA and MS effects was substantially improved by our expanded MACE model [2], where the EBV bias for bulls born in the most recent year was effectively eliminated (Figure 2). Although they were now much smaller, biases remained for the PA estimates of GPS bulls born in all years. The remaining biases in PA were due to biased evaluations for the proven-bull dams (results not shown).

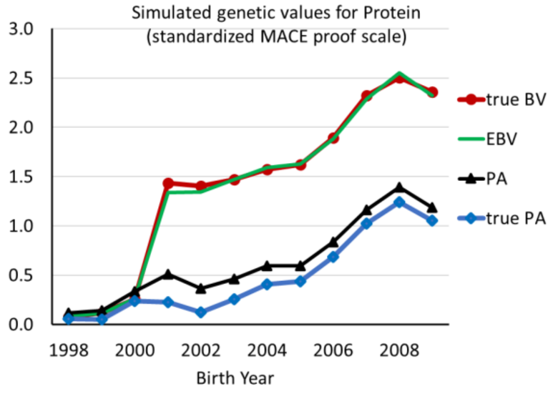

Figure 2. Simulated true values, including GPS effects for bulls born since 2001, and corresponding estimates of breeding value (EBV) and parent averages (PA), on the GPS country scale of evaluation, under the expanded MACE model [2].

There were notable inflation biases in the variances of MS estimates, because these were not reduced after adding GPS effects into the simulation, as should be expected (results not shown). The expected levels of reduction in MS variances, when MS means have been elevated due to GPS, can be derived from truncation selection theory (Tyrisevä et al, 2018), or by using a simple simulation to approximate the non-linear relationship between means and variances after truncation selection (Sullivan, 2018). Expected reductions in MS variance can be derived from each bull's individual estimate of GPS effect, which is the effect of GPS on the mean of MS values for bulls pre-selected in a given country and year. Higher levels of estimated GPS effect indicate the bull's true MS value was drawn from a distribution with higher mean and correspondingly lower variance. We are planning to add MS variance reduction factors to the model, which are specific to each GPS group of bulls, and are based on the estimated effects of GPS on MS means of each group.

Official MACE input data

In our application of the expanded MACE model with official MACE input data from April 2022, we assumed GPS effects were zero for all bulls born prior to 2009 across all populations, and for the bulls born since 2009 from very small populations. Small populations were defined as having fewer than 20 bulls in total born since 2009 and with a national EBV in April 2022 based on local daughters. Even if GPS had been possible for these small populations, it could not have been very effective when based on GEBV with very low accuracies due to small national reference populations. The accuracies of estimated GPS effects would also be low when based on small groups of GPS bulls for these populations. In April 2022, seven of the 29 populations evaluated in MACE for PRO were considered small populations.

Estimates of GPS effects were additionally restricted to only the bulls being selected and used locally, with a requirement that bulls had a national EBV included in MACE from their country of registration. This requirement was based on a presumption that application of GPS is always done by the registering country. There are many important GPS bulls, however, selected and used by countries other than the country of registration, for example Canadian bulls with daughters in USA but not in Canada. Such bulls were therefore excluded from GPS effect estimation in the present study, but they will be included in future applications as we continue to refine the model. For the present study, international GPS effects were estimated for approximately 80-85% of all daughter-proven MACE bulls born since 2009, across the various combinations of birth year by trait.

The international prevalence and intensities of GPS have been increasing for many reasons: The lists of countries applying GPS and of traits being evaluated have both been growing; genomic reference populations continually increase in size as millions of new animals are genotyped every year; and agreements to share genotypes among countries have expanded over time to increase reliabilities of GPS. An indicator of increasing prevalence for GPS is that we see increased percentages of bulls with positive estimates of GPS effects over time (Figure 3). These percentages increased for production (PRO), type (OUS) and udder health (SCS) traits between 2009 and 2013, and for fertility (CC1) between 2013 and 2017, while remaining low for the workability trait MSP. An increasing prevalence might be expected for MSP in the future, however, as the interest and ability to select for MSP has been increasing with robotic milking systems (Miles et al, 2022).

In Figure 4, the MACE results from models [1] and [2] are compared, to show the impacts of adding non-zero estimates of GPS effects in the expanded model [2]. This figure shows how positive estimates of GPS effects caused changes ([2]-[1]) in the EBV of GPS bulls, their sire and dam in opposing directions, and therefore to a much lesser extent the PA (average of sire and dam). Averages of only the positive estimates of GPS effects from official data were generally less than 0.10 standard deviations, compared with our simulated level of GPS that was more than 10 times larger (Figures 1 and 2). The GPS effects from official data were likely underestimated due to GPS bias in the national EBV computed without genotypes, and we expect these estimated effects to increase in magnitude as levels of GPS bias are decreased in future MACE input data. Regardless of the GPS effects being underestimated from official data, the patterns of impact on official MACE results were consistent with similar patterns we observed from simulated input data that were unbiased.

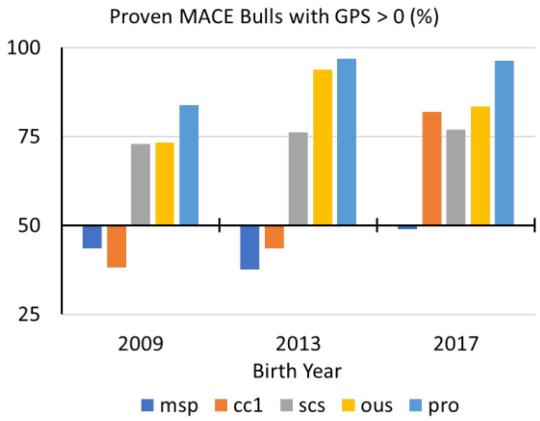

Figure 3. Global trends in the proportion of estimated genomic preselection effects that are positive, averaged across all country scales of evaluation for selected traits.

In the simulation study, GPS was practiced in only a single country, and the true GPS effects for this one country were included without bias in the simulated de-regressed EBV used as input for MACE. In contrast, the true levels of GPS are non-zero for several countries participating in MACE officially, and the de-regressed EBV used as input for MACE are biased because genotypes had to be ignored when computing the national EBV provided for MACE. Only a portion of true GPS effects will be attributed to elevated genetic levels for GPS sires in these national EBV models. The potential benefits of our expanded MACE model could therefore increase substantially if GPS biases are reduced in future MACE input data, as this should increase the magnitudes of our estimated GPS effects.

Research is encouraged on methods to reduce GPS bias in national EBV computed without genotypes, or alternatively on post-processing techniques that can effectively remove individual genotype effects from national GEBV computed with genotypes, while not removing the GPS effects on estimated MS distributions. Possibilities include a partitioning of the genomic results (Lourenco et al, 2015) or a suitable de-regression of national GEBV (Masuda et al, 2021). In the foreseeable future, it will continue to be required that input data provided for MACE be genomic-free. Thus, new proposals must consider how GPS effects can be included while at the same time excluding sire genotypes and/or individual genotype effects. Validation tests targeting partitioned contributions (e.g. MS) in an animal's EBV (Mäntysaari & Kudinov, 2022) will be of increasing importance to compare new possibilities for MACE.

Extensive studies on GPS effects in the results of single-step genomic evaluations led to similar conclusions as ours, that the main effects of GPS are the creation of non-zero means and reduced variances for the MS estimates of pre-selected individuals (Jibrila, 2022). Single-step or multi-step genomic evaluation models can effectively account for these GPS effects, because genetic covariances between sibs assumed in these systems are realized covariances after selection, which are higher due to selecting only the similarly superior sibs (VanRaden, 2008; Hayes et al, 2009). The genetic covariances among sibs based on pedigrees alone, while ignoring genotypes, are biased downwards because they reflect the lower expected covariances before selection.

The realized changes in means and variances of true MS values, due to GPS effects, will be expressed in the progeny phenotypes of pre-selected sires. By adding a new set of parameters in an expanded MACE model, it was possible to account for the GPS effects on means of MS values for AI sires born in recent years. A similar approach could also be applied at the national level to reduce GPS biases in national sire EBV computed without genotypes, and thus improve the input data provided for expanded MACE model [2]. If properly implemented and with sufficient national data to estimate GPS effects reliably, national EBV computed to include the estimated GPS effects should theoretically track more closely with GEBV trends from national genomic evaluation systems. An improved estimation of genetic trends from EBV based on a GPS-expanded national model could indicate that GPS biases were reduced, while still excluding individual genotype effects as required to participate in MACE.

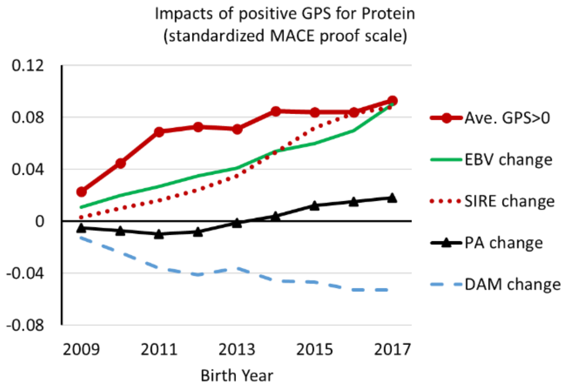

Figure 4. Impacts on EBV and parental evaluation changes (models [2] – [1]) for Protein, when bulls have positive estimates of GPS effects in MACE results, averaged across all country scales of evaluation.

The effects of positive GPS accumulate across generations. The international EBV of most recent sires include GPS effects for within-family selection of the bulls themselves, and additionally the elevated means due to GPS of the bulls' sires and grandsires. The increased EBV values of AI bulls are offset by decreasing evaluations for the bull dams in our expanded MACE model (Figure 4). Estimates of PA initially decreased at the beginning of the genomic era, but after multiple generations of GPS for AI bulls, the trend in PA became positive because elevated trends for the GPS sires of more recent GPS bulls were higher than the EBV reductions for recent bull dams. Multiple generations of accumulated GPS effects in the sires exceeds single-generation downward effects on the dams of most recent GPS bulls. During these eight years of GPS, the average changes in PA of proven MACE bulls remained relatively close to zero.

Summary

Genomic preselection of AI sires alters the distributions of both true and estimated MS deviations. Modified distributions can be assumed, where the MS of selected bulls have higher means and lower variances due to GPS effects estimated in our expanded MACE model. Significant reductions in GPS bias were observed after accounting for GPS effects on only the MS means, as confirmed in simulation results and after successful testing of the expanded MACE model on applications with official MACE input data. Further improvements are expected after additional future changes are made to also account for GPS effects on the MS variances.

Acknowledgements

Discussions related to this research and some related computer programs contributed by Haifa Benhajali are gratefully acknowledged.

References

de Roos, A.P.W., Schrooten, C., Mullaart, E., van der Beek, S., de Jong, G. & Voskamp, W. 2009. Genomic selection at CRV. *Interbull bulletin* 39:47-50.

Ducrocq, V. & Patry, C. 2010. Combining genomic and classical information in national BLUP evaluation to reduce bias due to genomic pre-selection. *Interbull bulletin* 41:33-36.

Fikse, F. 2014. Can a model with genetic groups for Mendelian sampling deviations correct for pre-selection bias? *Proc. 10th WCGALP*, Vancouver, Canada.

Hayes, B.J., Visscher, P.M & Goddard, M.E. 2009. Increased accuracy of artificial selection by using the realized relationship matrix. *Gen. Res*. 91:47-60.

Jibrila, I. 2022. Impact of preselection in genomic evaluations. A PhD thesis presented to Wageningen University, the Netherlands.

Lourenco, D.A.L, Fragomeni, B.O., Tsuruta, S., Aguilar, I., Zumbach, B., Hawken, R.J., Legarra, A. & Misztal, I. 2015. Accuracy of estimated breeding values with genomic information on males, females, or both: an example on broiler chicken. *Gen. Sel. Evol*. 47:56.

Lund, M.S., de Roos, A.P.W, de Vries, A.G., Druet, T., Ducrocq, V., Fritz, S., Guillaume, F., Guldbrandtsen, B., Liu, Z., Reents, R., Schrooten, C., Seefried, F. & Su, G. 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. *Gen. Sel. Evol*. 43:43 (8 pages).

Miles, A., Fourdraine, R., Sievert, S., Gaddis, K., Bewley, J., Eaglen, S., Weiker, J., Hutchison, J., Dürr, J. 2022. Considerations in using quantitative measurements of milking speed for genetic evaluations for all dairy breeds in the USA. *Interbull bulletin* 57:48-53.

Masuda, Y., VanRaden, P.M., Misztal, I. & Lawlor, T.J. 2018. Differing genetic trend estimates from traditional and genomic evaluations of genotyped animals as evidence of preselection bias in US Holsteins. *J. Dairy Sci*. 101:5194-5206.

Masuda, Y, Liu, Z. & Sullivan, P. 2021. Deregression formula for single-step genomic BLUP. *Interbull bulletin* 56:142-146.

Mäntysaari, E.A. & Kudinov, A.A. 2022. Usability of different genetic evaluation validation tests in a population subjected to a strong genomic selection and in testing the single-step genomic evaluations. *Interbull bulletin* 57:106-110.

Patry, C. & Ducrocq, V. 2011. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. *J. Dairy Sci.* 94:1011-1020.

Patry, C. Jorjani, H & Ducrocq, V. 2013. Effects of a national genomic preselection on the international genetic evaluations. *J. Dairy Sci*. 96:3272-3284.

Schaeffer, L.R. 2018. Necessary changes to improve animal models. *J. Anim. Br. Genet*. 135:124-131.

Schenkel, F.S., Sargolzaei, M., Kistemaker, G., Jansen, G.B., Sullivan, P., Van Doormaal, B.J., VanRaden, P.M. & Wiggans, G.R. 2009. Reliability of Genomic Evaluation of Holstein Cattle in Canada. *Interbull bulletin* 39:51-58.

Sullivan P.G. & Schaeffer L.R. 1994. Fixed versus random genetic groups. Proceedings of the 5th World Congress on Genetics Applied to Livestock Production. Guelph, ON, Canada, 18:483–486.

Sullivan, P.G. 2018. Mendelian Sampling variance tests with genomic preselection. *Interbull bulletin* 54:1-4.

Sullivan, P.G., Mäntysaari, E.A., deJong, G. & Benhajali, H. 2019. Modifying MACE to accommodate genomic preselection effects. *Interbull bulletin* 55:77-80.

Tyrisevä, A.M., Mäntysaari, E.A., Jakobsen, J., Aamand, G.P., Dürr, J., Fikse, W.F. & Lidauer, M.H. 2018. Detection of evaluation bias caused by genomic preselection. *J. Dairy Sci*. 101:3155-3163.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91:4404-4423.

VanRaden, P.M., Wiggans, G.R., Van Tassell, C.P., Sonstegard, T.S. & Schenkel, F. 2009. Benefits from Cooperation in Genomics. *Interbull bulletin* 39:67-72.

Westell, R.A., Quaas, R.L & Van Vleck, L. 1988. Genetic groups in an animal model. *J. Dairy Sci*. 71:1310-1318.