Alternative Approaches to handling of missing parents in genetic evaluation of dairy cattle using single-step test-day SNP-BLUP model

Dawid Słomian[1], Kacper Żukowski[1], Monika Skarwecka[1], Jan Ten Napel[3], Jeremie Vandenplas[3] Joanna Szyda[1,2]

[1]National Research Institute of Animal Production, Krakowska 1, 32-083 Balice, Poland
[2]Biostatistic Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wroclaw, Poland
[3]Animal Breeding and Genomics, Wageningen University & Research, P. O. Box 338, 6700 AH Wageningen, Netherlands

Abstract

In many countries, single-step genomic models are replacing conventional pedigree-based models for routine valuation. Those models use all available information on the animals' phenotype, genotype, and pedigree. Pedigree data still has a huge impact on estimated genomic breeding values (GEBV), and it is also important to consider information about the structure of the pedigree. The foremost aspect of pedigree editing is dealing with missing parents' information. The choice of method of handling missing parents can affect the prediction of breeding values. This work investigates three scenarios of pedigree data: 1) Pedigree_real (**P_Real**) – pedigree from the routine evaluation, 2) Pedigree_2010 (**P_2010**) – at least 20 and 10 percent of dams and sires born before 2019 were set randomly to missing, respectively, 3) Pedigree_4020 (**P_4020**) – at least 40 and 20 percent of dams and sires born before 2019 were set randomly to missing, respectively. Moreover, for those pedigrees, three approaches to defining missing parents were used: 1) Raw pedigree (**RP**) – missing parents IDs set to missing, 2) Genetic groups (**GG**) – missing parents replaced by unrelated **GG**, which are defined based on year of birth, sex, and country of origin, 3) Metafounders (**MF**) – missing parents replaced by **MF**, which correspond to genetic groups. Relationships within and between metafounders were estimated from genomic information of descendants. The genomic breeding values for fat yield were estimated using the single-step test-day SNP-BLUP model, implemented by the MiXBLUP software. Although GEBV prediction was similar across scenarios, expressing missing parents by **GG** or **MF** impacts the genetic trend, especially in situations of limited pedigree completeness. Removing parent information led to reduced precision results across the methods of handling missing parents, since **P_Real** scenario demonstrated highest accuracy results. Compared to **RP** and **GG**, **MF** scenarios resulted in higher genetic trends. Insufficient pedigree completeness, especially among ungenotyped individuals, leads to an overestimation of the genetic trend. Completeness of pedigree information and a large number of genotyped individuals improve the reliability of evaluations. Modeling missing sires with **MF**s is less effective than assuming unrelated **GG**s if pedigree information is very incomplete. Therefore, the best method to model missing parents depends on completeness of pedigree.

Key words: single-step models, genetic groups, metafounders, validation

Introduction

The single-step model becomes the standard procedure of most national routine evaluations of dairy cattle (Legarra et al., 2014, Mäntysaari et al. 2017). The single-step model combines all available information, i.e., phenotype, genotype, and pedigree. Invariably, one of the main components in routine genomic evaluation of dairy cattle is the structure of the pedigree (Bradford et al., 2019). To reduce the bias due to missing information in the pedigree, genetic groups are used to associate individuals with missing parents with different categories (Westell et al., 1988, Legarra et al., 2007). An alternative to genetic groups to deal with missing information in the pedigree are the so-called metafounders (Legarra et al., 2015).

In this study, we focused on a single-step random regression SNP-BLUP test-day model for fat yield in the Polish Holstein population. The primary objective of this study is to evaluate various methods for handling missing parents and different levels of incompleteness in the pedigree data based on validation results, average GEBV trends, and GEBV comparisons.

Materials and Methods

This study is based on Polish national evaluation data for fat yield from April 2024 (Table 1). Two phenotype files were analysed: full data set – 63,615,019, and truncated data set – 58,446,695 test-day records. A truncated data set was created by removing the records for the youngest individuals, i.e., the last 4 years from the phenotype file. Genotypes that include 48,118 single-nucleotide polymorphisms (SNPs), were available for 113,019 cows and 68,972 bulls, that is 181,991 animals. The pedigree was extracted up to the third generation from animals with phenotypes and genotypes, including 4,712,143 animals (4,569,044 cows and 143,099 bulls).

Table 1: Number of test-day records, genotypes, and animals in the analysed data sets for fat yield.

Data	Sex	Number of animals	Number of records
Phenotype	Cows	3,707,727	63,615,019 Full data set 58,446,695 Truncated data set
Genotype	Cows	113,019	181,991
	Bulls	68,972	
Pedigree	Cows	4,569,044	4,712,143
	Bulls	143,099	

To deal with missing parents we used three approaches: 1) **RP** – raw pedigree with missing parents IDs set to missing; 2) **GG** – genetic groups with missing parents replaced by unrelated genetic groups, which are defined based on year of birth, country of origin and sex; 3) **MF** – metafounders with missing parents replaced by metafounders, which can be considered as genetic groups with relationships estimated from genomic information of descendants. Based on pedigree from routine evaluation, the three approaches of different pedigree completeness was used: 1) **P_Real** – pedigree from routine evaluation, with ~ 5.6% of missing sires and ~ 15.3% of missing dams; 2) **P_2010** – minimum 20% of dams and 10% of sires born before 2019 was set to missing based on **P_Real**; 3) P_4020 – minimum of 40% of dams and 20% of sires born before 2019 was set to missing based on **P_Real**. Only the parents' IDs were removed, as the manipulation involved animals born before 2019; therefore, the pedigree of the youngest validation animals remains the same across scenarios.

For animals with missing parents in the pedigree, the genetic groups were implemented based on country of origin, year of birth, and sex. Individuals born before 1961 were removed from the pedigree data. Over 70% of individuals included in the pedigree had both parents. Each genetic group contained a minimum of 20 animals. Group "-31" (Polish males born between 2010-2019) had the largest number of missing sires (1,002,069), whereas group "-32" (Polish females born between 2010-2019) had the most missing dams (174,954).The following single-step random regression test-day SNP-BLUP model (Liu et al., 2004; Liu et al., 2014) was applied:

$$y=Xh+Wf+Vp+Vu+e,$$

where $\boldsymbol{y}$ is a vector of cows' test day records for fat yield from the first three lactations, $\boldsymbol{h}$ is a vector of fixed effects of herd-test-day-parity-milking frequency, $\boldsymbol{f}$ is a vector of fixed lactation curve coefficients which was modelled by the Wilmink function (Liu et al., 2004), $\boldsymbol{p}$ is a vector of permanent environmental effects expressed as random regression coefficient coefficients of the Legendre polynomials, ***u*** is a random additive genetic effects also described by the random regression coefficients of the Legendre polynomials.

The GEBVtest method was used for validation (Mäntysaari et al., 2010). The full and truncated data sets have been prepared for validation. The full data set contains all phenotypic data, while the truncated data set includes all phenotypic data except for the last 4 years of data. Validation cows were defined as cows whose records were removed for a truncated data set; however, validation bulls were defined as sires born between 2017 and 2019, and having more than 20 validation daughters. The test was implemented separately for validation cows and bulls, used the linear regression:

$$\mathbf{GEBVf} = b_0 + b_1\mathbf{GEBVp} + \mathbf{e},$$

where **GEBVf** represents the vector of GEBVs predicted based on the full data set, while **GEBVp** represents GEBVs predicted based on the truncated data set, b_0 represents the intercept, which indicates a systematic bias in the model's prediction, and b_1 represents the regression slope, the dispersion of prediction compared to actual results. The R^2 coefficient is one of the results of linear regression and serves a measure of prediction accuracy, it indicates the percentage of variance in the **GEBVp** explained by **GEBVf**.

Validation results were computed for the first three lactations, and the total genomically enhanced breeding value (GEBV) defines as:

$$GEBVt{=}0.5GEBV_1{+}0.3GEBV_2{+}\ 0.2GEBV_3$$

where $GEBV_1$ is GEBV for the 1_{st} lactation, $GEBV_2$ is GEBV for the 2_{nd} lactation and $GEBV_3$ is GEBV for the 3_{rd} lactation.

Single-step genomic evaluations were conducted using MiXBLUP 3.0 (Vandenplas et al., 2022)

Results & Discussion

Validation results are reported for 482,810 validation cows and 562 validation bulls. Figures 1-3 show validation results for all scenarios divided by sex, method, and genotyping status. Figure 1 shows the b_0 of the dam and sire. We observed similar results for all scenarios; the values are close to 0, which is expected. Figure 2 shows the b_1 value, which is similar for every scenario for validation cows, with b_1 in the range of 0.96 (**P_Real MF** ungenotyped) to 1.1 **(P_2010 MF** genotyped). However, for validation bulls, all results are similar, except for ungenotyped validation bulls in the scenarios **P_4020** and **P_2010** for **MF**. For these latter categories, we observed an overestimation of b_1 at 1.27 (**P_2010**) and 1.33 **(P_4020**). This may be due to a lack of pedigree connection for ungenotyped bulls, due to a higher percentage of incomplete pedigrees. Figure 3 shows the R^2, ranging from 0.66 to 0.90 for every scenario. Lower values were observed for ungenotyped validation cows; however, for genotyped validation cows, R^2 is more stable and similar across scenarios. For ungenotyped validation bulls, we observed a trend where R^2 increased from **RP** through **GG** and **MF**. However, for genotyped validation bulls, the R^2 value is similar for **P_Real**. In contrast, for MF, the R^2 values for **P_4020** and **P_2010** are lower than in other scenarios involving missing parents.

Figure 4 compares full and truncated data sets for validation bulls divided by scenarios and genotyping status. In each case, the points cluster together to form an extended cloud centered on the diagonal; however, as parental information is gradually eliminated, the cloud dispersion becomes wider, especially for ungenotyped individuals. The effect is slight under **P_Real**, becomes evident in **P_2010**, and reaches its peak in **P_4020**, when ungenotyped validation bulls from the **RP**, **GG**, and **MF** deviate the most from the diagonal. All of these patterns show that genomic information protects the accuracy of prediction when the incompleteness of pedigree is high: prediction for genotyped validation bulls remains strong even when up to 40% of dams and 20% of sires are set to unknown, whereas missing parental information links weaken the stability of GEBV for ungenotyped validation bulls.

Figure 5 shows the average GEBV trend for all scenarios divided by sex. Since 2000, the mean GEBV has increased gradually; however, after 2010, when genotyping became widely used in Poland, the increase became more pronounced. Compared to cows, bulls exhibit a steeper trajectory, indicating that the sire pathway is under more selection pressure. Both sexes show the same scenario ranking, with **MF** producing the highest averages, followed by **GG** and **RP**. However, as pedigree completeness declines, the gap between scenarios widens, underscoring the fact that the way missing parents are handled can significantly skew the perception of genetic progress. It is crucial to handle incomplete pedigrees robustly to prevent overestimating or underestimating the selection response.

Conclusions

The results demonstrate that the method used to close pedigree gaps can significantly affect the predictions of GEBV. Regardless of the pedigree scenario used, the real pedigree yielded the most reliable validation results. However, for individuals without genotypes, scenarios with increased pedigree incompleteness introduced observable over-dispersion; this effect was more pronounced for sires than for dams and was most noticeable in the **MF** group.

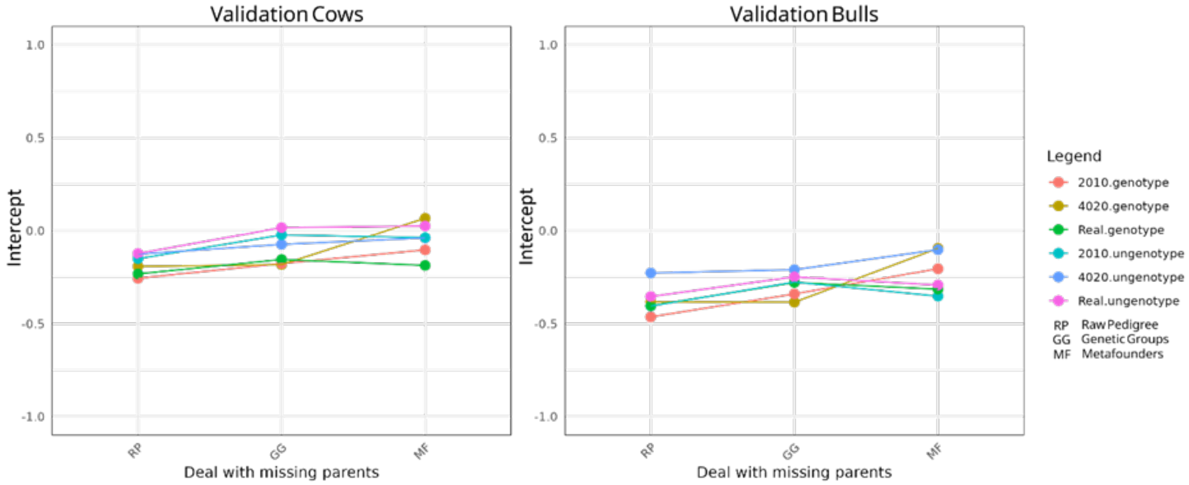

Figure 1. Intercept (b_0) for validation individuals divided by sex and method.

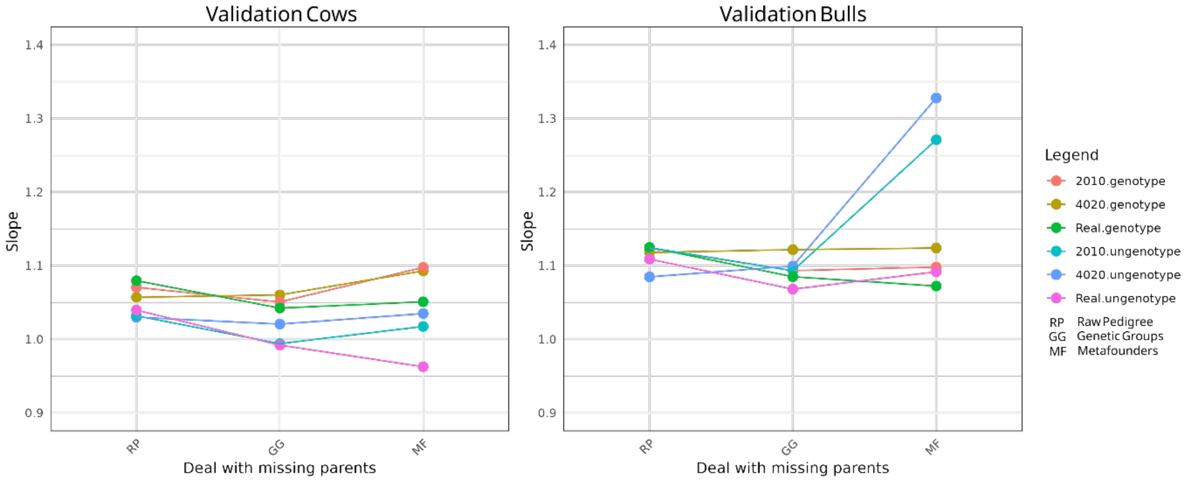

Figure 2. Slope (b_1) for validation individuals divided by sex and method.

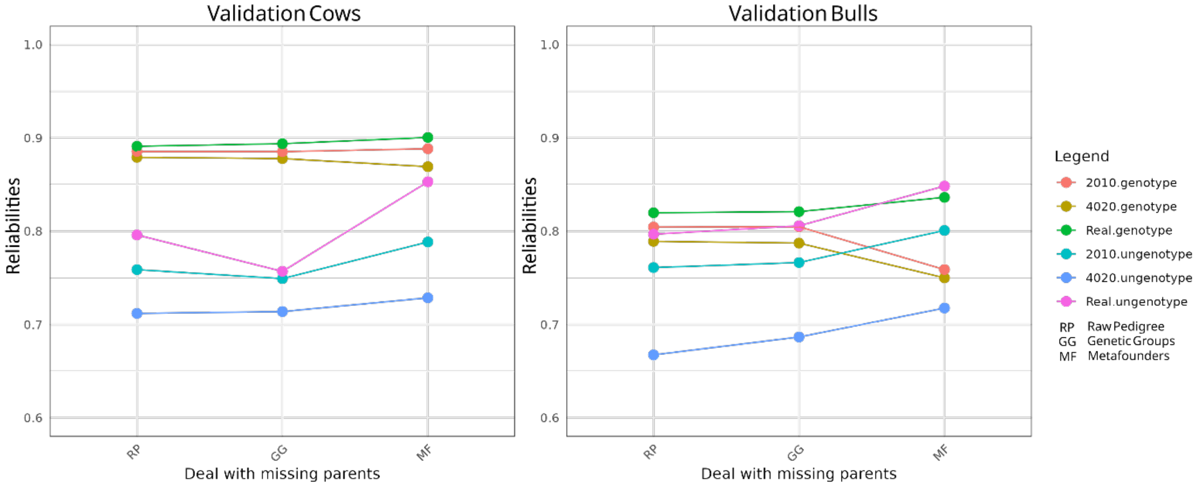

Figure 3. Reliabilities (R^2) for validation individuals divided by sex and method.

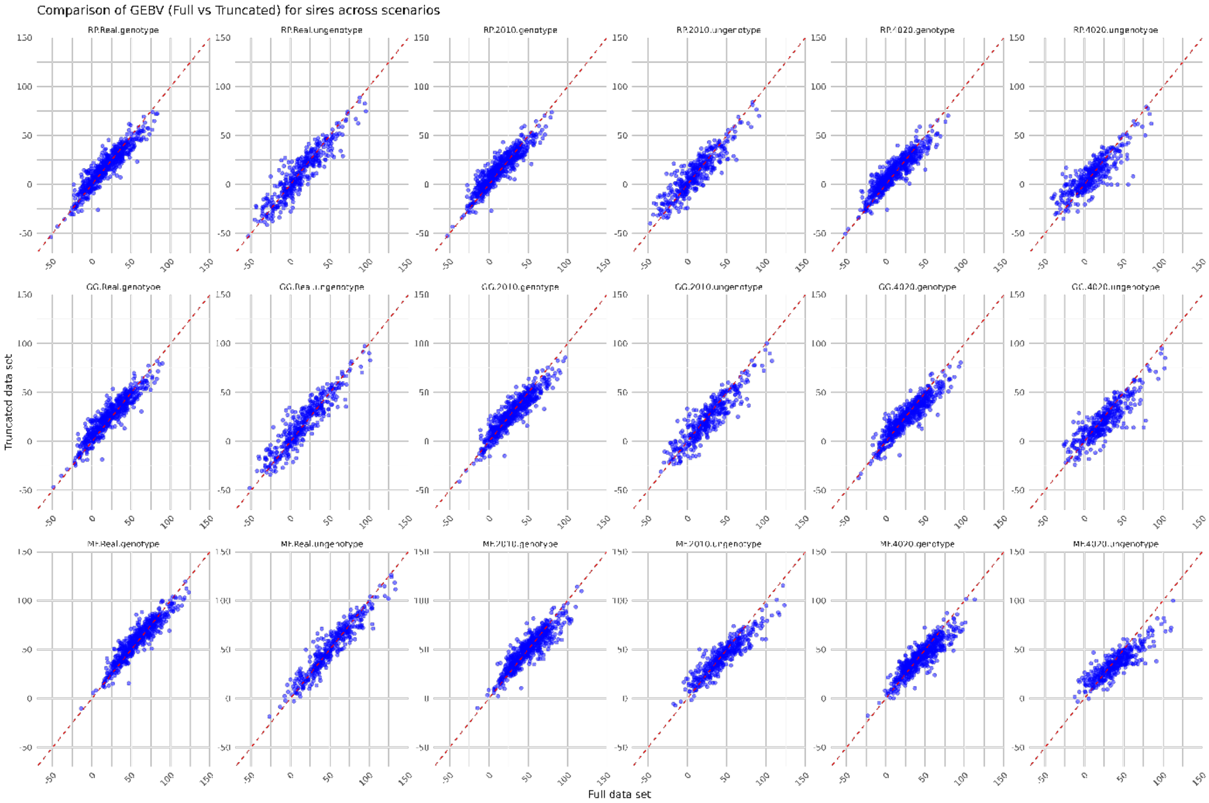

Figure 4. Comparison of GEBV for validation bulls across scenarios, divided by genotyped and ungenotyped individuals.

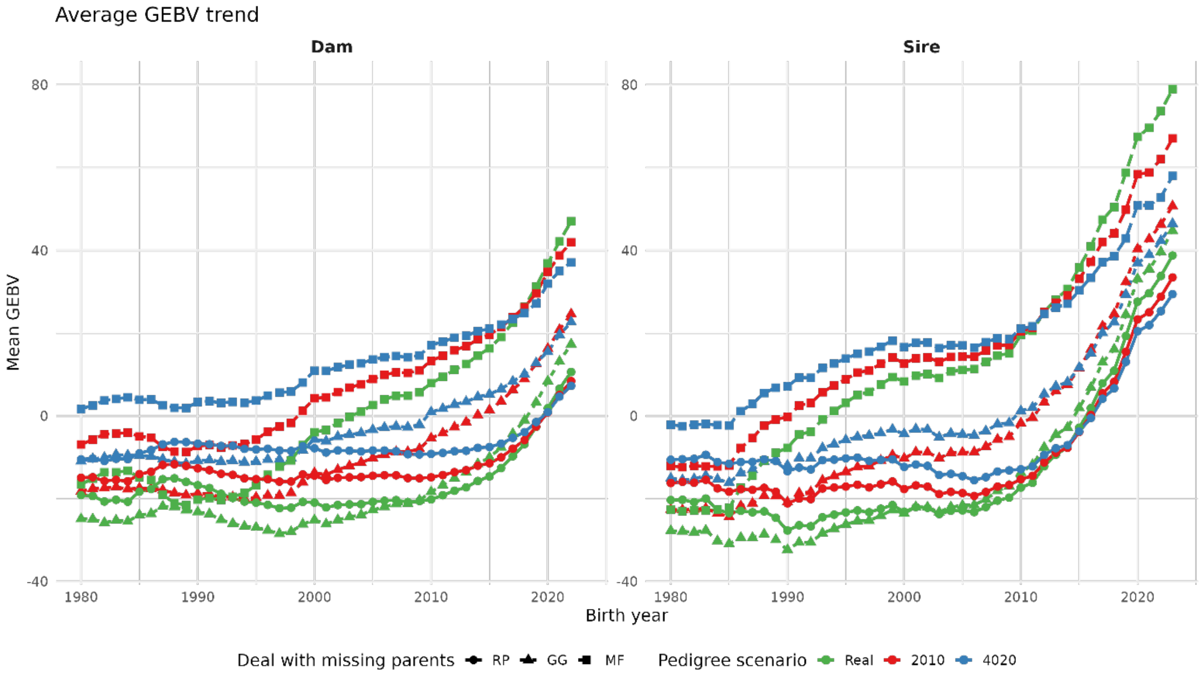

Figure 5. Average GEBV trend divided by sex and method.

References

Bradford, H., Masuda, Y., VanRaden, P., Legarra, A., & Misztal, I. 2019. Modeling missing pedigree in single-step genomic BLUP. *Journal of Dairy Science, 102*(3), 2336–2346. https://doi.org/10.3168/jds.2018-15434

Legarra, A., Bertrand, J., Strabel, T., Sapp, R., Sánchez, J., & Misztal, I. 2007. Multi-breed genetic evaluation in a Gelbvieh population. *Journal of Animal Breeding and Genetics, 124*(5), 286–295. https://doi.org/10.1111/j.1439-0388.2007.00671.x

Legarra, A., Christensen, O. F., Aguilar, I., & Misztal, I. 2014. Single step, a general approach for genomic selection. *Livestock Science, 166*, 54–65. https://doi.org/10.1016/j.livsci.2014.04.029

Legarra, A., Christensen, O. F., Vitezica, Z. G., Aguilar, I., & Misztal, I. 2015. Ancestral relationships using metafounders: Finite ancestral populations and across-population relationships. *Genetics, 200*(2), 455–468. https://doi.org/10.1534/genetics.115.177014

Liu, Z., Reinhardt, F. Bünger, A. & Reents, R. 2004. Derivation and calculation of approximate reliabilities and daughter yield deviations of a random regression testday model for genetic evaluation of dairy cattle. *Journal of Dairy Science, 87*(6), 1896–1907. https://doi.org/10.3168/jds.S00220302(04)733482

Liu, Z., Goddard, M., Reinhardt, F., & Reents, R. 2014. A single-step genomic model with direct estimation of marker effects. Journal of Dairy Science, 97(9), 5833–5850. https://doi.org/10.3168/jds.2014-7924

Mäntysaari, E. A., Evans, R. D., & Strandén, I. 2017. Efficient single-step genomic evaluation for a multibreed beef cattle population having many genotyped animals. *Journal of Animal Science, 95*(11), 4728–4737. https://doi.org/10.2527/jas2017.1912

Mäntysaari, E., Liu, Z., & VanRaden, P. 2010. Interbull validation test for genomic evaluations. *Interbull Bulletin, 41*, 17–22. Retrieved from https://journal.interbull.org/index.php/ib/article/view/1496

Westell, R. A., Quaas, R. L., & Van Vleck, L. D 1988. Genetic groups in an animal model. *Journal of Dairy Science, 71*(5), 1310–1318. https://doi.org/10.3168/jds.S0022-0302(88)79688-5

Vandenplas, J., Veerkamp, R. F., Calus, M. P. L ., Lidauer, M. H., Strandén, I., Taskinen, M., Schrauf, M., & ten Napel, J. 2022. MiXBLUP 3.0 – software for large genomic evaluations in animal breeding programs. *Proceedings of the 12th World Congress on Genetics Applied to Livestock Production* (Paper 358). https://doi.org/10.3920/978-90-8686-940-4_358