JDS Communications®
TBC; TBC

https://doi.org/10.3168/jdsc.2024-0607
Short Communication
Genetics

Weighed ssGBLUP enhances the genomic prediction accuracy for milk citrate predicted by milk mid-infrared spectra of Holstein cows in early lactation

Y. Chen,[1]* H. Atashi,[12] C. Grelet,[3] and N. Gengler[1]

Abstract: Previous studies have shown that milk citrate predicted by milk mid-infrared spectra (MIR) is strongly affected by a few genomic regions. This study aimed to explore the effect of the weighted single-step GBLUP on the accuracy of genomic prediction (GP) for MIR-predicted milk citrate in early lactation Holstein cows. A total of 134,517 test-day predicted milk citrate collected within the first 50 d in milk on 52,198 Holstein cows from the first 5 parities were used. There were 122,218 animals in the pedigree, of them 4,479 had genotypic data for 566,170 SNPs. Two data sets (partial and whole data sets) were used to verify whether the accuracy of GP is improved using the following different methods. The (genomic) estimated breeding values (**EBV** or **GEBV**) in the partial and whole data sets were estimated by pedigree-based BLUP (**ABLUP**), single-step GBLUP (**ssGBLUP**, pedigree-genomic combined using no weight for SNP), weighted ssGBLUP (**WssGBLUP**, pedigree-genomic combined using weighted SNP), respectively. The difference between the 2 data sets is that the phenotypic data from 2017 to 2019 in the partial data set were set as missing values. 181 youngest cows with genomic were selected as the validation population. A linear regression method was used to compare EBV (GEBV) predicted for partial and whole data sets. The accuracies of GP for ABLUP and ssGBLUP were 0.42 and 0.70, respectively. The accuracies of GP for WssGBLUP in the 5 iterations with different CT (constant) values (determines departure from normality for SNP effects) ranged from 0.70 to 0.86. This study showed that weighted SNP is beneficial in improving prediction accuracy for predicted milk citrate.

Genomic prediction (**GP**) is widely used in animal breeding. The methods used for GP in genomic evaluation are currently divided into two main classes: Bayesian (Meuwissen et al., 2001) and best linear unbiased prediction [**BLUP**, including SNP-BLUP, genomic BLUP (**GBLUP**), single-step GBLUP (**ssGBLUP**)] methods. BLUP can perform the calculations more efficiently and is widely used in genomic evaluation in various countries, especially ssGBLUP (Legarra et al., 2009; Bermann et al., 2022). However, the advantage of Bayesian algorithms is that these methods consider the different variance of each SNP effect. The ssGBLUP assumes that all SNP effects have the same variance, which is inconsistent with some real-life examples of traits affected by major genes (e.g., double-muscled in cattle, Grobet et al., 1997). Therefore, Wang et al (2012) proposed the weighted ssGBLUP (**WssGBLUP**) approach, which assigns different weights to SNPs to construct a new relationship matrix. Lourenco et al. (2017) showed that WssGBLUP is more effective when the number of genotyped individuals is small, and few QTL affect traits. Local breeds or novel traits (e.g., methane production) often have genomic data for only a few animals. Therefore, WssGBLUP may enhance the accuracy of GP for the local breeds or novel traits.

Negative energy balance (**NEB**) is a condition encountered by almost all high-producing dairy cows during early lactation. NEB is detrimental to the reproduction, metabolism, and infectious diseases of dairy cows (Walsh et al., 2011; Zachut et al., 2020), which may cause economic losses to farmers and reduce the welfare of dairy cows. However, direct measurement of NEB is challenging, prompting researchers to use blood or milk biomarkers to assess the energy status of dairy cows (Zachut et al., 2020). Blood non-esterified fatty acids has been demonstrated as biomarkers for detecting NEB (Zachut et al., 2020), but testing it is very expensive. Milk citrate is proposed as a novel biomarker of NEB and can be applied on a large scale through milk mid-infrared (**MIR**) spectra (Grelet et al., 2016, 2024). Our recent study showed that a few genomic regions have large effects on MIR-predicted citrate (Chen et al., 2024), consistent with other studies in different breeds (Sanchez et al., 2021).

This study aimed to investigate whether WssGBLUP improves the accuracy of GP for MIR-predicted citrate of Holstein cows in early lactation.

The data and model used for this study came from our recent work (Chen et al., 2024). Briefly, 134,517 MIR-predicted citrates (hereafter called citrate) from the first five parities of 52,198 Holstein cows in 774 farms in the Walloon Region of Belgium were used. Citrate prediction model was based on the standardized milk MIR spectra extracted from the official milk record database of the Walloon Region of Belgium. The coefficient of determination and root mean square error of validation for the citrate equation were 0.86 and 0.07 mmol/L, respectively (Grelet et al., 2016). The used citrate data was limited to the early lactation (first 50 days in milk, DIM), a period in which most high-yielding Holstein cows are in NEB (Churakov et al., 2021). The used pedigree includes

[1] TERRA Teaching and Research Center, University of Liège, Gembloux Agro-Bio Tech (ULiège-GxABT), 5030 Gembloux, Belgium, [2] Department of Animal Science, Shiraz University, 71441-65186 Shiraz, Iran, [3]Walloon Agricultural Research Center (CRA-W), 5030 Gembloux, Belgium. *Corresponding Author: Yansen Chen, Passage des Déportés, 2, B-5030 Gembloux, Belgium, +32/81/62 23 58 (office), yansen.chen@uliege.be (e-mail). Received May 14, 2024. Accepted September 03, 2024.

The list of standard abbreviations for JDSC is available at adsa.org/jdsc-abbreviations-24. Nonstandard abbreviations are available in the Notes.

122,218 animals, of which 4,479 (3,215 cows and 1,264 bulls) had data for 566,170 SNPs. Citrate in the first 5 parities was considered as one trait based on our latest research (Chen et al., 2024). Hence, a univariate repeatability model was employed to estimate both the variance components and (genomic) estimated breeding values (**EBV** or **GEBV**) for citrate. The model incorporated fixed effects: herd-year-season of calving, standardized DIM and its quadratic term, and standardized calving age with constant, linear, and quadratic regression (nested within parities); random effects included permanent environmental effects, additive animal genetic effects, and residual effects. To calculate the relationship matrix, either a single (**H**) or pedigree-based (**A**) relationship matrix was employed. The **H** matrix combined the **A** and genomic (**G**)-based relationship matrices, and then **H** was inversed by the method proposed by Aguilar et al. (2010). **A** is the numerator relationship matrix for all animals included in the pedigree; **G** is the genomic relationship matrix of genotyped animals obtained using the first formula described by VanRaden (2008):

$$\mathbf{G} = \frac{\mathbf{ZDZ'}}{2\sum_{i=1}^{N} p_i\left(1-p_i\right)} = \mathbf{ZDZ'}\lambda,$$

where **Z** is a matrix of gene content adjusted for allele frequencies (0, 1, or 2 for aa, Aa, and AA, respectively); **D** is a diagonal matrix of SNP weight; N is the number of SNPs; p_i is the minor allele frequency of the ith SNP; λ is $\frac{1}{2\sum_{i=1}^{N} p_i\left(1-p_i\right)}$. If **D** matrix is equal to the identity matrix, the combined relationship matrix (**A** and **G**) is the normal **H** matrix (ssGBLUP); if the diagonal of the **D** matrix is not equal to 1 (SNP weight), the combined relationship matrix (**A** and **G**) is weighted **H** matrix (WssGBLUP). The SNP weight was calculated based on the procedure proposed by Wang et al. (2012), however, the nonlinear A weights method (VanRaden, 2008) was used in this study. The algorithm proceeds as follows:

1. t = 1, $\mathbf{D}_{(t)} = \mathbf{I}$, $\mathbf{G}^{*}_{(t)} = \mathbf{ZD}_{(t)}\mathbf{Z'}\lambda$.
2. compute GEBV by WssGBLUP (the first iteration is ssGBLUP).
3. calculate SNP effects by $\lambda\mathbf{D}_{(t)}\mathbf{Z'}\mathbf{G}^{*-1}_{(t)}$GEBV.
4. calculate SNP weight by $\mathrm{CT}^{\frac{|\text{SNP effects}|}{\text{sd(SNP effects)}}-2}$ (nonlinear A weights method).
5. Normalize $\mathbf{D}_{(t+1)} = \frac{\text{trace}\left(\mathbf{D}_{(t)}\right)}{\text{trace}\left(\mathbf{D}_{(t+1)}\right)}\mathbf{D}_{(t+1)}$.
6. Calculate $\mathbf{G}^{*}_{(t+1)} = \mathbf{ZD}_{(t+1)}\mathbf{Z'}\lambda$.
7. t = t +1
8. Loop to 2 or exit if t >5.

where CT is a constant value that determines the departure from normal distribution. If CT equates to 1, it is the normal distribution; 1.050, 1.125 (default value), 1.500, and 2.000 were tested in this study. Previous studies have shown that citrate is strongly affected by a few genomic regions (Sanchez et al., 2021; Chen et al., 2024), therefore, the CT was set larger than 1. Five iterations were used in the study, which was used to optimize SNP weights and maximize accuracy gains (Cesarani et al., 2021).

Variance components and EBV (or GEBV) were estimated by using the BLUPF90+ (version 2.42) program (Misztal et al., 2014). Variance components were estimated through the Average Information Restricted Maximum Likelihood Estimation method (AI-REML). The EBV (or GEBV) was calculated through ABLUP (or ssGBLUP and WssGBLUP). The SNP effect and weight were calculated using POSTGSF90 software (version 1.73) (Misztal et al., 2014).

A linear regression-based (**LR**) method developed by Legarra and Reverter (2018) was used to assess the prediction accuracy of the EBV (or GEBV) in young animals. The basic step of the LR method involves calculating the evaluation metrics by regressing the breeding value of the partial data set according to the breeding value of the whole data set. The pedigree and genome information of the whole and partial data sets were the same, but the phenotypic data were different. The phenotypes of the whole data set were from 2012 to 2019, while the phenotypes of the partial data set were from 2012 to 2016 (2017 to 2019 were set as missing values, n = 38,906). Both variance components and breeding values need to be estimated again in the partial data set. The 181 youngest cows (born after 2015) with genotypic were selected for the validation population. Four following metrics were used to measure prediction validation results.

1. Prediction accuracy $\left(\widehat{acc}\right)$ of the validation population is expected to be 1 if the evaluation is perfect, as defined below

$$\widehat{acc} = \sqrt{\frac{cov\left(\hat{\mathbf{u}}_p, \hat{\mathbf{u}}_w\right)}{\left(1-\overline{f}\right)\sigma_u^2}},$$

where $\hat{\mathbf{u}}_p$ and $\hat{\mathbf{u}}_w$ are vectors of EBV (or GEBV) of the validation population in the partial and whole data sets, respectively; $\overline{f}$ is the average inbreeding coefficient of the validation population; σ_u^2 is the additive genetic variance.
2. Population bias (μ_{wp}) is expected to be 0 under an unbiased evaluation, as defined below

$$\mu_{wp} = \overline{\hat{\mathbf{u}}_p} - \overline{\hat{\mathbf{u}}_w},$$

where $\overline{\hat{\mathbf{u}}_p}$ and $\overline{\hat{\mathbf{u}}_w}$ are average (G)EBV of the validation population in the partial and whole data sets, respectively.
3. Dispersion ($b_{w,p}$) is expected to be 1 when there is no observed dispersion, as defined below

$$b_{w,p} = \frac{cov\left(\hat{\mathbf{u}}_w, \hat{\mathbf{u}}_p\right)}{var\left(\hat{\mathbf{u}}_p\right)},$$

where all parameters are the same as described above.
4. Slope ($b_{p,w}$) is expected to be 1 when the average reliability of the validation population is consistent between partial and whole datasets, as defined below

$$b_{p,w} = \frac{Reliability_p}{Reliability_w} = \frac{cov\left(\hat{\mathbf{u}}_p, \hat{\mathbf{u}}_w\right)}{var\left(\hat{\mathbf{u}}_w\right)},$$

where all parameters are the same as described above. The $b_{p,w}$ is utilized for assessing the relative stability in the average reliability of the validation population between estimates derived from partial and whole data sets. The data preparation and figure plot were performed using R (version 4.1.2, https://www.r-project.org/).

The average and standard deviation of citrate were 9.04 and 1.65 mmol/L, respectively. Table 1 shows the validated results from ABLUP and ssGBLUP for citrate. The $\widehat{acc}$ and $b_{p,w}$ obtained from ssGBLUP increased by 65.19% and 85.28%, respectively, compared with those derived from ABLUP. The μ_{wp} and $b_{w,p}$ from ssGBLUP were similar to the results from ABLUP. Similar findings were also reported by Cesarani et al. (2021). This confirms that genomic information is very beneficial for the genetic evaluation of citrate.

Figure 1 shows the validated results from WssGBLUP in the first 5 iterations. The $\widehat{acc}$, $b_{w,p}$ and $b_{p,w}$ of WssGBLUP (iterations 2–5) were better than those from ssGBLUP (first iterations), however, the μ_{wp} of WssGEBLUP was worse. The outcomes of WssGBLUP were evidently influenced by the CT value. The $\widehat{acc}$ increased with increasing CT values, consistent with the finding that citrate was affected by a few genomic regions (Sanchez et al., 2021; Chen et al., 2024). The absolute μ_{wp} increased with increasing CT values. This may be due to the gradual increase in the mean GEBV of the absolute values across the validation population. The $b_{w,p}$ worsens as CT increases; however, the $b_{w,p}$ reached its best value (0.99 or 1.01) in the second (or third) iteration when CT was equal to 1.500 (or 1.250). The $b_{p,w}$ also increased with increasing CT values. The $b_{p,w}$ reached its best value (0.82) in the second iteration when CT was equal to 2.000. Based on the above results, the 4 metrics of WssGBLUP can be relatively good in the second iteration. Teissier et al. (2018) also reported that the maximum prediction accuracy was obtained at the second iteration. In the second iteration, 2 metrics ($\widehat{acc}$ and $b_{p,w}$) reached its best value when CT was equal to 2.000, however, another 2 metrics (μ_{wp} and $b_{w,p}$) performed worst. Therefore, the CT of 1.500 was chosen as the best value in this study. Previous studies have also reported that WssGBLUP is more beneficial for GP compared with ssGBLUP (Teissier et al., 2018, 2019; Mehrban et al., 2021). However, if no significant genomic regions were associated with the studied trait, WssGBLUP may have similar results to ssGBLUP (Teissier et al., 2019; Cesarani et al., 2021).

Table 1. Validated predicted milk citrate by linear regression (LR) for best linear unbiased prediction with pedigree (ABLUP), single-step genomic BLUP (ssGBLUP, pedigree combine genomic) (n = 181 youngest cows)

LR statistic	ABLUP	ssGBLUP
Prediction accuracy	0.42	0.70
Bias	0.04	−0.03
Dispersion	0.95	1.05
Slope	0.41	0.76

There are aspects of this research that can be improved in the future. The number of the validation population is small (n = 181). Therefore, more genotyped individuals are needed to be accumulated to further verify the results of this study. On the other hand, we demonstrate that WssGBLUP is beneficial for the GP of small reference populations (local breeds). The WssGBLUP used in this study is limited to single-trait analysis and cannot perform multi-trait analysis. Meuwissen et al. (2024) recently proposed an algorithm called GWABLUP: integrating GWAS results into GP. GWABLUP is capable of conducting multi-trait analysis, however, it is currently being extended to single-step analysis. Assigning weight to SNPs can be a quick way to improve GP in the presence of genes with large effects. SNP weight estimation can be derived from a variety of information, currently mainly from SNP effects, and multi-omics information may contribute to this.

This study confirms that genomic information is beneficial for the genetic evaluation of citrate. The results of this study demonstrate that weighted SNP contributes to enhanced accuracy of GP for predicted milk citrate.

References

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. https://doi.org/10.3168/jds.2009-2730.

Bermann, M., A. Cesarani, I. Misztal, and D. Lourenco. 2022. Past, present, and future developments in single-step genomic models. Ital. J. Anim. Sci. 21:673–685. https://doi.org/10.1080/1828051X.2022.2053366.

Cesarani, A., A. Garcia, J. Hidalgo, L. Degano, D. Vicario, N.P.P. Macciotta, and D. Lourenco. 2021. Genomic information allows for more accurate breeding values for milkability in dual-purpose Italian Simmental cattle. J. Dairy Sci. 104:5719–5727. http s://https://doi.org/10.3168/jds.2020-19838.

Chen, Y., H. Hu, H. Atashi, C. Grelet, K. Wijnrocx, P. Lemal, and N. Gengler. 2024. Genetic analysis of milk citrate predicted by milk mid-infrared spectra of Holstein cows in early lactation. J. Dairy Sci. 107:3047–3061. https://doi.org/10.3168/jds.2023-23903.

Churakov, M., J. Karlsson, A. Edvardsson Rasmussen, and K. Holtenius. 2021. Milk fatty acids as indicators of negative energy balance of dairy cows in early lactation. Animal 15:100253. https://doi.org/10.1016/j.animal.2021.100253.

Grelet, C., C. Bastin, M. Gelé, J. B. Davière, M. Johan, A. Werner, R. Reding, J. A. Fernandez Pierna, F. G. Colinet, P. Dardenne, N. Gengler, H. Soyeurt, and F. Dehareng. 2016. Development of Fourier transform mid-infrared calibrations to predict acetone, β-hydroxybutyrate, and citrate contents in bovine milk through a European dairy network. J. Dairy Sci. 99:4816–4825. https://doi.org/10.3168/jds.2015-10477.

Grelet, C., T. Larsen, M. A. Crowe, D. C. Wathes, C. P. Ferris, K. L. Ingvartsen, C. Marchitelli, F. Becker, A. Vanlierde, J. Leblois, U. Schuler, F. J. Auer, A. Köck, L. Dale, J. Sölkner, O. Christophe, J. Hummel, A. Mensching, J. A. Fernández Pierna, H. Soyeurt, M. Calmels, R. Reding, M. Gelé, Y. Chen, N. Gengler, and F. Dehareng. 2024. Prediction of key milk biomarkers in dairy cows through milk mid-infrared spectra and international collaborations. J. Dairy Sci. 107:1669–1684. https://doi.org/10.3168/jds.2023-23843.

Grobet, L., L. J. Royo Martin, D. Poncelet, D. Pirottin, B. Brouwers, J. Riquet, A. Schoeberlein, S. Dunner, F. Ménissier, J. Massabanda, R. Fries, R. Hanset, and M. Georges. 1997. A deletion in the bovine myostatin gene causes the double–muscled phenotype in cattle. Nat. Genet. 17:71–74. https://doi.org/10.1038/ng0997-71.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92:4656–4663. https://doi.org/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 01 Mathematical Sciences 0104 Statistics. Genet. Sel. Evol. 50:53. https://doi.org/10.1186/s12711-018-0426-6.

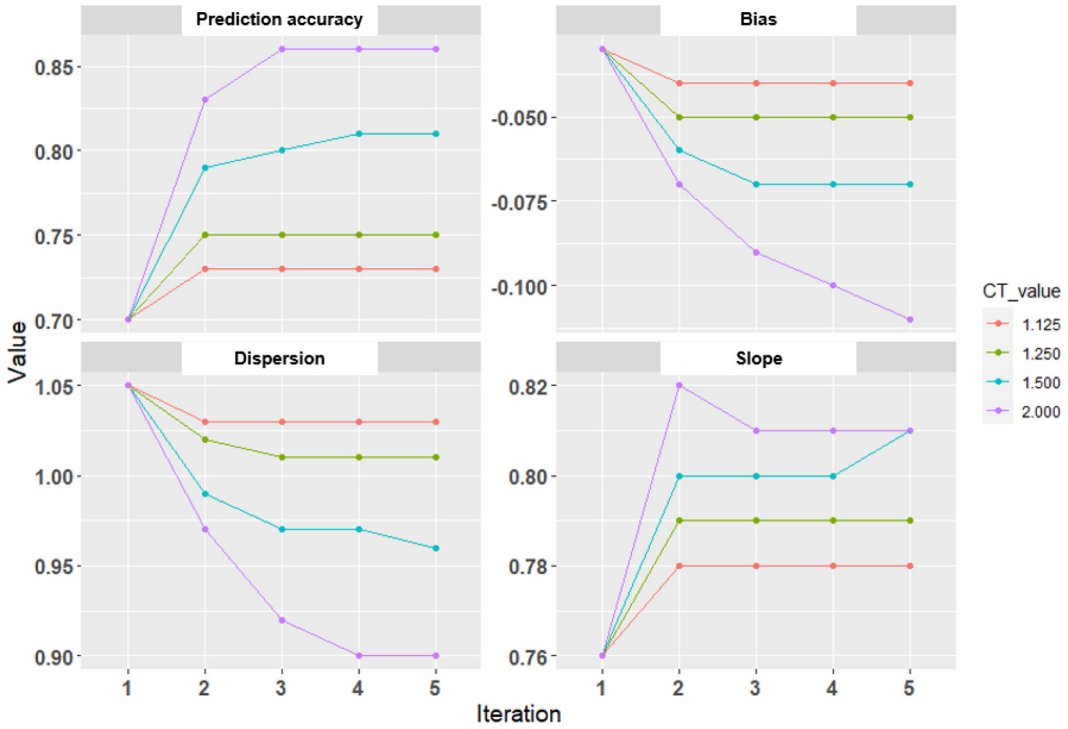

Figure 1. Validated predicted milk citrate by linear regression (LR) for weighted single-step genomic BLUP (WssGBLUP, pedigree combine genomic using SNP-weighted) in the 5 iterations with different CT values (n = 181 youngest cows). The weight of the SNP in the first iteration of all analyses was 1 (equated to ss-GBLUP), so all results were the same. CT values determine departure from normality for SNP effects (When CT equals 1, the SNP effect is normally distributed).

Lourenco, D. A. L., B. O. Fragomeni, H. L. Bradford, I. R. Menezes, J. B. S. Ferraz, I. Aguilar, S. Tsuruta, and I. Misztal. 2017. Implications of SNP weighting on single-step genomic predictions for different reference population sizes. J. Anim. Breed. Genet. 134:463–471. https://doi.org/10.1111/jbg.12288.

Mehrban, H., M. Naserkheil, D. Lee, and N. Ibáñez-Escriche. 2021. Multi-Trait Single-Step GBLUP Improves Accuracy of Genomic Prediction for Carcass Traits Using Yearling Weight and Ultrasound Traits in Hanwoo. Front. Genet. 12:692356. https://doi.org/10.3389/fgene.2021.692356.

Meuwissen, T., L. S. Eikje, and A. B. Gjuvsland. 2024. GWABLUP: genome-wide association assisted best linear unbiased prediction of genetic values. Genet. Sel. Evol. 56:17. https://doi.org/10.1186/s12711-024-00881-y.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829. https://doi.org/10.1093/genetics/157.4.1819.

Misztal, I., S. Tsuruta, D.A.L. Lourenco, I. Aguilar, A. Legarra, and Z.V. 2014. Manual for BLUPF90 family of programs. Access: March 26, 2024. http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all8.pdf

Sanchez, M. P., D. Rocha, M. Charles, M. Boussaha, C. Hozé, M. Brochard, A. Delacroix-Buchet, P. Grosperrin, and D. Boichard. 2021. Sequence-based GWAS and post-GWAS analyses reveal a key role of SLC37A1, ANKH, and regulatory regions on bovine milk mineral content. Sci. Rep. 11:7357. https://doi.org/10.1038/s41598-021-87078-1.

Teissier, M., H. Larroque, and C. Robert-Granié. 2018. Weighted single-step genomic BLUP improves accuracy of genomic breeding values for protein content in French dairy goats: A quantitative trait influenced by a major gene. Genet. Sel. Evol. 50:31. https://doi.org/10.1186/s12711-018-0400-3.

Teissier, M., H. Larroque, and C. Robert-Granie. 2019. Accuracy of genomic evaluation with weighted single-step genomic best linear unbiased prediction for milk production traits, udder type traits, and somatic cell scores in French dairy goats. J. Dairy Sci. 102:3142–3154. https://doi.org/10.3168/jds.2018-15650.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. https://doi.org/10.3168/jds.2007-0980.

Walsh, S. W., E. J. Williams, and A. C. O. Evans. 2011. A review of the causes of poor fertility in high milk producing dairy cows. Anim. Reprod. Sci. 123:127–138. https://doi.org/10.1016/j.anireprosci.2010.12.001.

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res. 94:73–83. https://doi.org/10.1017/S0016672312000274.

Zachut, M., M. Šperanda, A. M. De Almeida, G. Gabai, A. Mobasheri, and L. E. Hernández-Castellano. 2020. Biomarkers of fitness and welfare in dairy cattle: healthy productivity. J. Dairy Res. 87:4–13. https://doi.org/10.1017/S0022029920000084.

Notes

Y. Chen https://orcid.org/0000-0002-8593-4384
H. Atashi https://orcid.org/0000-0002-6853-6608
C. Grelet https://orcid.org/0000-0003-3313-485X
N. Gengler https://orcid.org/0000-0002-5981-5509

The China Scholarship Council (Beijing) is acknowledged for funding the PhD project of Yansen Chen (no. 201906760007). Yansen Chen thanks the support by the Fonds de la Recherche Scientifique (FNRS, Brussels, Belgium) under grant no. T.0095.19 (PDR "DEEPSELECT"). The University of Liège–Gembloux Agro-Bio Tech (Liège, Belgium) supported computations through the technical platform Calcul et Modélisation Informatique (CAMI) of the TERRA Teaching and Research Centre supported by the FNRS under grant no. T.0095.19 (PDR "DEEPSELECT"). Genotyping was facilitated through the support of the FNRS under grant no. J.0174.18 (CDR "PREDICT-2"). This article does not contain any studies with human or animal subjects and did not require [IACUC/IRB] approval. The authors declare that they have no competing interests.