Machine learning algorithms identified relevant SNPs for milk fat content in cattle

Pablo J. Ríos[1,8], M. Agustina Raschia[2,5], Daniel O. Maizon[3,6], Daniel Demitrio[4,8], and Mario A. Poli[2,7]

[1] Universidad de Buenos Aires, Buenos Aires, Argentina.pablo.javier.rios@gmail.com;

[2] Instituto Nacional de Tecnología Agropecuaria, CICVyA-CNIA, Instituto de Genética "Ewald A. Favret", Nicolás Repetto y de Los Reseros s/n, Hurlingham (B1686), Buenos Aires, Argentina raschia.maria@inta.gob.ar; poli.mario@inta.gob.ar

[3] Instituto Nacional de Tecnología Agropecuaria, E.E.A. Anguil, Ruta 5 Km 580, Anguil (6326), La Pampa, Argentina maizon.daniel@inta.gob.ar

[4] Instituto Nacional de Tecnología Agropecuaria, Dirección General de Sistemas de Información, Comunicación y Procesos - Gerencia de Informática y Gestión de la Información, Chile 460 (C1098AAJ), Buenos Aires, Argentina demitrio.daniel@inta.gob.ar

[5] Facultad de Ciencias Médicas, Universidad Nacional de La Plata, Argentina.

[6] Facultad de Agronomía, Universidad Nacional de La Pampa, Argentina.

[7] Facultad de Ciencias Agrarias y Veterinaria, Universidad del Salvador, Argentina.

[8] Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Argentina.

Abstract. In recent years, machine learning methods have been shown to be efficient in identifying a subset of single nucleotide polymorphisms (SNP) underlying a trait of interest. The aim of this study was the construction of predictive models using machine learning algorithms, for the identification of loci that best explain the variance in milk fat production of dairy cattle. Further objectives involve determining the genes flanking relevant SNPs and retrieving the pathways, biological processes, or molecular functions overrepresented by them. Fat production values adjusted for fixed effects (FP_{adj}) and estimated breeding values for milk fat production (EBV_{FP}) were used as phenotypes and SNPs as predictor variables. The models constructed for EBV_{FP} performed better and yield considerably less relevant SNPs than models for FP_{adj}. Among the genes flanking relevant SNPs, signaling transduction pathways and gated channel activities were detected as overrepresented. The loci obtained for EBV_{FP} matched better with previously reported relevant loci for milk fat content than those obtained for FP_{adj}. Based on the better performance showed by the models trained for EBV_{FP} and their agreement with previous reported results for the trait

studied, we conclude that the relationship among individuals should be accounted for in the phenotype used.

Keywords: machine learning methods, single nucleotide polymorphisms, estimated breeding values, dairy cattle

1 Introduction

The development of different density specie-specific microarrays that evaluate genotypes on single nucleotide polymorphisms (SNP) distributed over the entire genome, as well as of bioinformatics tools, have enabled numerous studies involving the joint analysis of pedigree, genotypic and phenotypic information of livestock. Among the most studied species is dairy cattle due to the importance of dairy industry in worldwide economy. Accordingly, extensive genetic research on milk production in cattle has been performed. Several studies focused on identifying associations between SNP markers located all over the genome and a trait of interest were conducted and hence associations between many SNP and milk traits have been reported for different dairy cattle breeds [1, 2, 3]. Most of such genome-wide association studies (GWAS) were performed through the implementation of different software that fit linear, multivariate and Bayesian linear mixed models. The single or multiple-trait animal models used considered fixed and genetic effects affecting the phenotypic observations. In recent years, machine learning methods (ML) have also been used in GWAS, showing to be efficient in identifying a subset of SNPs underlying a trait of interest [4, 5].

The aim of this study was the construction of predictive models for indicators of milk fat content in a population of dairy cows, using SNPs as predictor variables and three machine learning algorithms, for the identification of regions in the genome that best explain the variance in those phenotypes. Further objectives involve retrieving the genes located near those SNPs and looking for pathways, biological processes, or molecular functions overrepresented by them. Besides, the comparison among the relevant SNPs obtained through regression models trained using machine learning algorithms and previously reported relevant SNP windows obtained for the same population that was used in this study.

2 Materials and methods

2.1 Phenotypes

A database consisting of 98907 milk fat content records from the first lactation of 16907 Holstein and Holstein x Jersey (HxJ) cows was used. Two phenotypic datasets were built to train regression models using three ML algorithms XGBoost [6], LightGBM [7], and Random Forest [8]:

1) Residuals for 305-day cumulative milk fat production (adjusted fat production, FP_{adj}) values estimated for 812 cows (559 Holstein and 253 HxJ) using the Fleischmann method [9], adjusted for the percentage of Holstein background, herd, age at first calving, and the combined effect of season and year of first lactation. This phenotype did not consider pedigree information.

2) Breeding values for milk fat production (EBV_{FP}) estimated for 837 cows (582 Holstein and 255 HxJ) and 26 bulls (22 Holstein and 4 Jersey) using the WOMBAT program [10]. This estimation accounted for the relationship among animals and the phenotypes of genotyped and non-genotyped animals.

2.2 Genotyping

SNP genotyping was performed on 969 cows (703 Holstein and 266 HxJ) and 29 bulls (24 Holstein and 5 Jersey) using the BovineSNP50 v2 BeadChip (Illumina Inc., San Diego, CA, USA), which evaluates 54609 SNP distributed over the 29 bovine autosomes and sex chromosomes, spaced on average 48102 bp apart. The quality control of genotype data consisted in the exclusion of SNP with unknown position on the genome, located on the Y chromosome, with a call rate lower than 0.95 or a minor allele frequency lower than 0.03. Animals with a call rate lower than 0.90 were also excluded. After genomic data quality control, 40417 SNP from 978 (952 cows + 26 bulls) animals were available for subsequent analyses.

2.3 Models trained using machine learning algorithms

The input dataset to the models comprised the phenotypes mentioned in section 2.1 and genotypes of 812 cows (when using FP_{adj}) or 863 animals (when using EBV_{FP}) on 40417 SNPs as predictor variables. Table 1 shows a summary of the variables used as input to train regression models. The final objective of the models is not the prediction of phenotypic values for each animal but the selection or identification of the most important SNPs, i.e. those that best explain the observed variance in the studied phenotypes.

Table 1. Summary of the variables used to train regression models.

Inputs for trained models	Target variables	FP_{adj}	Adjusted residuals for 305-day cumulative milk fat production estimated for 812 cows.
		EBV_{FP}	Breeding values for milk fat production estimated for 863 animals.
	Predictor variables		Genotypes on 40417 biallelic SNP that passed quality control checks.

Regression models were trained using three ML algorithms: XGBoost (XGB) and LightGBM (LGB) in Python, and Random Forest (RF) in R. To evaluate their efficiency in identifying the SNPs that best explain the differences in the phenotypes,

Pearson correlation, R^2, mean absolute error (MAE) and root mean square error (RMSE) metrics for actual vs. predicted values were determined in validation folds using 5-folds cross-validation (XGB and LGB) and out-of-bag data (RF). A 5-fold cross-validation scheme was used based on the study performed by Li et al [11], which uses a dataset of comparable size as this study. The population was randomly split into five groups of equal size and each group was in turn assigned with missing phenotypic values and used as the validation set.

Relevant SNPs retrieved from each trained model were those with importance or gain >0. For Random Forest algorithm, the importance value of a SNP is the percentage of increase in the mean squared error (MSE) in the "out-of-bag" datasets across all the trees in the forest in which the SNP participates, using random permutation. The MSE of each tree in the forest is compared after randomly permuting the values of the variable in a new sample, and the percentage of increase in the error is computed; the larger this error, the more important the variable is. For XGBoost and LightGBM, the importance or gain value of a SNP denotes the reduction in the prediction error of the objective function (MSE) when partitioning a node in a tree using the SNP. The higher the gain value, the more important the SNP.

2.4 Models implementation

The model hiperparameters adjusted for the three algorithms were the learning rate, max. tree depth, min. number of individuals in leaf nodes, number of features used to create each tree, number of individuals used to create each tree, and regularization values (L1 and L2). These hiperparameters were optimized manually because the relatively small size of the dataset allowed so.

The dataset had a very low amount of missing SNPs, 0.43%, but since Random Forest algorithm discards rows with missing values, it was required to impute them. A multivariate version of missForest imputation was used, which is based on an iterative process using prediction as described by Ishwaran and Kogalur [8].

The implementation of the algorithms used are XGBoost version 1.3.3 for Python, LightGBM version 3.1.1 for Python, and R package randomForestSRC version 2.10.1. Python release is 3.8.5 and R release is 4.0.4. Source code is available upon request (contact Daniel Demitrio, demitrio.daniel@inta.gob.ar, for it).

2.5 Overrepresentation tests

To assess whether the relevant SNPs identified by each ML algorithm have any biological relevance, overrepresentation tests were performed on the gene sets close to those SNPs in +/- 30 kb, using the program PANTHER (protein annotation through evolutionary relationship) [12]. The parameters used were *Bos taurus* (for organism); statistical overrepresentation test (for analysis method); PANTHER Pathways, PANTHER GO-Slim Biological Process, and PANTHER GO-Slim Molecular Function (for annotation data set); all genes in *Bos taurus* database (for reference list);

and Fisher's Exact with FDR multiple test correction (for test type). Results with FDR p-value < 0.05 were considered statistically significant.

2.6 Comparison with previous results

The location of relevant SNPs obtained in this study was compared to previously reported 57 relevant 10-adjacent SNP windows that explained more than 10 times genetic variance than expected for milk fat production, obtained using BLUPf90 package of programs for the same population [13]. Then, the number of relevant windows and top windows (explaining more than 0.7% of the genetic variance for the trait) containing SNPs with gain>0 was determined.

3 Results and discussion

Histograms for the phenotypes used to train regression models (adjusted fat production values and breeding values for milk fat production) are shown in Figure 1.

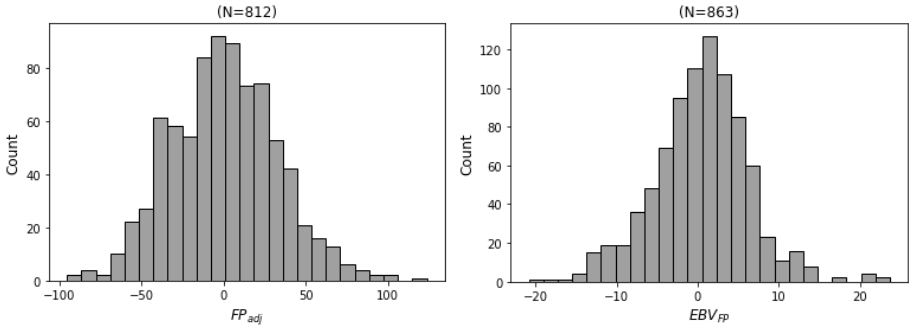

Fig. 1. Histogram of FP_{adj} values (left) and EBV_{FP} values (right).

The metrics for the assessment of the models' performance are shown in Table 1. The models for EBV_{FP} in validation performed considerably better than the models for FP_{adj}, for the three ML algorithms used, with R^2 correlations for actual vs. predicted values greater than 0.55 in all cases, while when using FP_{adj} as phenotype, those R^2 were lower than 0.1. The Pearson correlation, also in validation, increased from 0.18-0.23 for the models for FP_{adj}, to 0.75 for those for EBV_{FP}. Thus, the models trained for EBV_{FP} are learning and are capable to predict breeding values for animals not included in the dataset.

Table 1. Metrics for the models used, based on actual vs. predicted values for FP_{adj} and EBV_{FP}.

	Validation / Out-of-bag FP_{adj}		
metrics	**XGBoost**	**LightGBM**	**Random Forest**
P corr.	0.177 [0.109, 0.243]	0.228 [0.162, 0.293]	0.198 [0.131, 0.263]
R^2	< 0.1	< 0.1	< 0.1
MAE	25.36 [24.18, 26.65]	25.22 [24.05, 26.51]	25.33 [24.15, 26.62]

RMSE	32.05 [30.56, 33.69]	31.86 [30.38, 33.49]	31.92 [30.44, 33.55]
	Validation / Out-of-bag EBV_{FP}		
metrics	**XGBoost**	**LightGBM**	**Random Forest**
P corr.	0.749 [0.718, 0.777]	0.751 [0.720, 0.779]	0.752 [0.722, 0.780]
R^2	0.553	0.551	0.552
MAE	2.97 [2.83, 3.11]	2.94 [2.81, 3.09]	2.97 [2.84, 3.12]
RMSE	3.90 [3.72, 4.09]	3.91 [3.73, 4.10]	3.90 [3.73, 4.10]

P corr.: Pearson correlation; MAE: mean absolute error; RMSE: root mean square error. 95% confidence intervals between brackets.

The number of relevant SNPs for FP_{adj} was 9548, 10424, and 8454 for XGB, LGB, and RF, respectively. While the number of relevant SNPs for EBV_{FP} was considerably lower: 1774, 2355, and 196 for XGB, LGB, and RF, respectively. Thus, the models used for EBV_{FP} not only have a better performance but also use less SNPs to explain the phenotype. The Venn diagrams (Figure 2) generated with the SNPs with positive importance values revealed a total of 1206 common SNPs across three methods for FP_{adj} and 133 for EBV_{FP}. 67.9% of the SNPs with positive importance values selected by RF for the EBV_{FP} model were also selected by XGB and LGB, while a much smaller percentage of the SNPs identified by RF, 14.3%, are common across the three algorithms for the FP_{adj} model. This could be an area of future exploration.

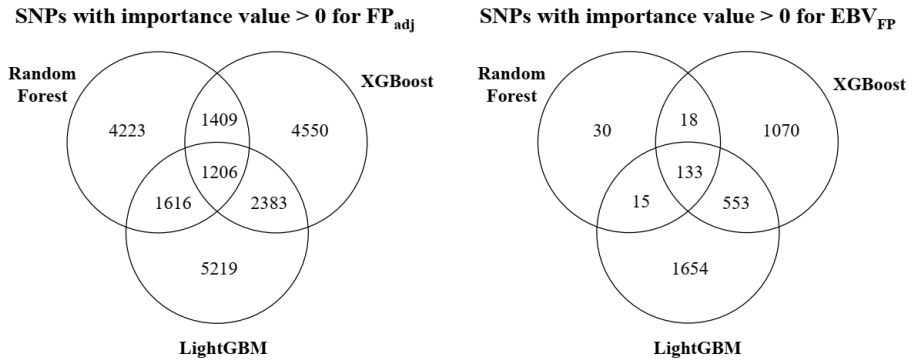

Fig. 2. Venn diagrams showing the number of SNPs with non-zero variable importance values for RF, LGB, and XGB for FP_{adj} model (left) and EBV_{FP} model (right). Each area of the circle represents the number of SNPs identified by the methods. The areas of intersection of circles represent the number of overlapping SNPs of two or three methods.

Figure 3 shows the distribution profiles of the gain values for relevant SNPs (ranked from the most important to the least important ones) obtained through XGB, LGB, and RF algorithms. The larger the SNP gain value, the more important a SNP is.

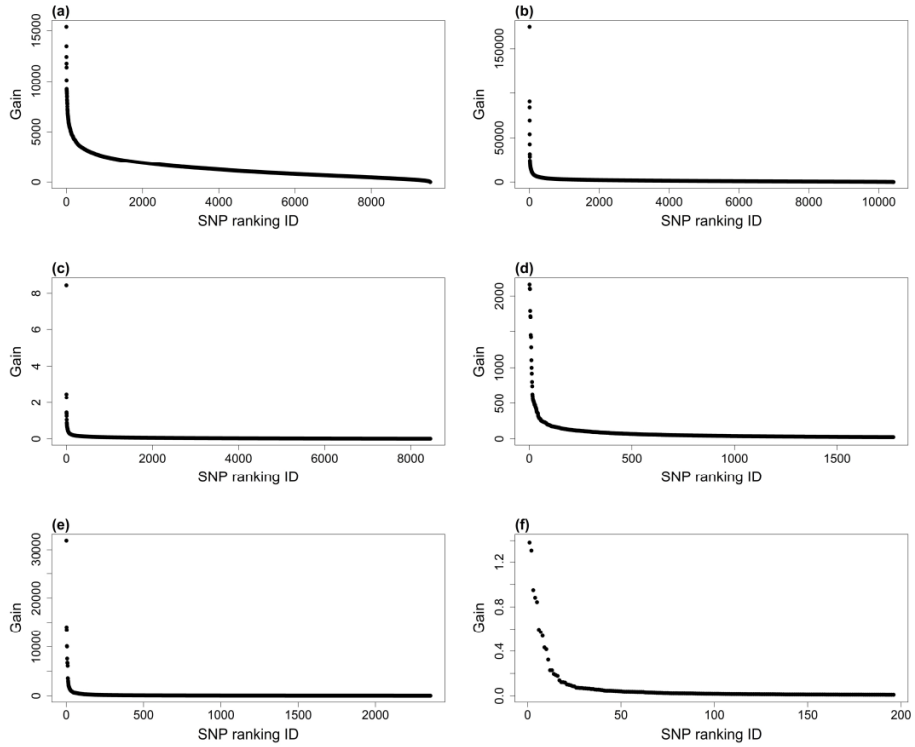

Fig. 3. Distribution profiles of the gain values for relevant SNPs obtained through XGB, LGB, and RF algorithms for FP_{adj} (a), b) and c), respectively) and EBV_{FP} (d), e) and f), respectively).

Protein coding genes containing or flanking in +/- 30 kb relevant SNPs obtained by XGB, LGB, and RF algorithms were 6586, 6884 and 6122 for FP_{adj}, and 1426, 1889, and 137 for EBV_{FP}. Pathways, biological processes, and molecular functions overrepresented by them are indicated in Tables 2 and 3. Signal transduction pathways, cation channel activities, and nervous system developmental processes are overrepresented by the genes flanking relevant SNPs for FP_{adj}. When using EBV_{FP} as phenotype, fewer genes flanked relevant SNPs and hence in some cases overrepresented pathways (LGB, RF), biological processes (XGB, RF), and molecular functions (XGB, RF) could not be identified. Even though, for the cases in which significant results were obtained, signaling pathways and channel activities were detected as overrepresented, similarly to results obtained for FP_{adj}.

Table 2. Pathways, Biological Processes, and Molecular Functions overrepresented by genes flanking relevant SNPs for adjusted fat production.

Algorithm	PANTHER Pathways	FDR p-value
XGB	---	
LGB	---	

RF	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha mediated pathway	4.98E-02
	PANTHER GO-Slim Biological Process	
XGB	synaptic transmission, glutamatergic	2.84E-02
	metal ion transport	3.24E-02
	regulation of molecular function	3.25E-02
	modulation of chemical synaptic transmission	3.46E-02
	nervous system development	3.61E-02
	intracellular signal transduction	3.67E-02
	cell development	4.53E-02
LGB	establishment of localization	3.29E-02
RF	regulation of membrane potential	1.15E-02
	cell junction organization	1.59E-02
	nervous system development	1.85E-02
	cell adhesion	2.35E-02
	cell morphogenesis	2.49E-02
	chemical synaptic transmission	3.22E-02
	actin filament-based process	3.94E-02
	PANTHER GO-Slim Molecular Function	
XGB	catalytic activity	2.83E-03
	voltage-gated cation channel activity	3.15E-03
	GTPase activator activity	6.28E-03
	potassium channel activity	2.31E-02
	glutamate binding	2.33E-02
	glutamate receptor activity	2.43E-02
	calcium channel activity	2.73E-02
	ligand-gated ion channel activity	4.63E-02
	Ras GTPase binding	4.84E-02
	transferase activity	4.84E-02
	phosphoric ester hydrolase activity	5.11E-02
LGB	gated channel activity	3.39E-02
	cation channel activity	3.76E-02
	Ras GTPase binding	5.37E-02
	metal ion transmembrane transporter activity	6.76E-02
RF	potassium channel activity	2.26E-03
	voltage-gated cation channel activity	2.98E-03
	GTPase activator activity	2.13E-02
	actin filament binding	2.16E-02
	glutamate binding	3.33E-02
	glutamate receptor activity	3.43E-02

Table 3. Pathways, Biological Processes, and Molecular Functions overrepresented by genes flanking relevant SNPs for estimated breeding values for fat production.

Algorithm	PANTHER Pathways	FDR p-value
XGB	EGF receptor signaling pathway	1.88E-04

	FGF signaling pathway	2.51E-03
	Metabotropic glutamate receptor group III pathway	7.01E-03
LGB	---	
RF	---	
	PANTHER GO-Slim Biological Process	
XGB	---	
LGB	multicellular organism development	4.93E-02
RF	---	
	PANTHER GO-Slim Molecular Function	
XGB	---	
LGB	gated channel activity	3.40E-02
	potassium channel activity	3.54E-02
RF	---	

Regarding the matching between relevant loci identified in this study and in a previous one [13] using a different approach, the percentages of 10-SNP windows for milk fat content containing relevant SNPs obtained in this study were 24.6, 38.6, and 15.8% for FP_{adj}, and 57.9, 57.9, and 5.3% for EBV_{FP} when using XGB, LGB, and RF, respectively. If considering only the 10 windows explaining more than 0.7% of the genetic variance for the trait studied, those percentages become 40, 50, and 20% for FP_{adj}, and 80, 60, and 10% for EBV_{FP} when using XGB, LGB, and RF, respectively. The better matching reached with models for EBV_{FP} is evident.

4 Conclusion

The models trained for EBV_{FP} were capable to predict breeding values for animals not included in the dataset. Based on the better performance showed by these models in comparison with those for FP_{adj} and their agreement with previous reported results for the trait studied, we conclude that the relationship among individuals should be accounted for in the phenotype used. Hence, estimated breeding values should be used instead of adjusted production phenotypes as target variable for the models. Moreover, training models with more individuals should be assessed in future studies, as well as using other machine learning algorithms for feature selection such as Bayesian regression methods.

References

1. Bouwman, A.C., Bovenhuis, H., Visker, M.H.P.W., van Arendonk, J.A.M.: Genome-wide association of milk fatty acids in Dutch dairy cattle. BMC Genetics 12:43 (2011).
2. Otto, P.I., Guimarães, S.E.F:, Calus, M.P.L., Vandenplas, J., Machado, M.A., Panetto, J.C.C., da Silva, M.V.G.B.: Single-step genome-wide association studies (GWAS) and post-GWAS analyses to identify genomic regions and candidate genes for milk yield in Brazilian Girolando cattle. J Dairy Sci. 103(11), 10347-10360 (2020).

3. Shadi Nayeri, S., Sargolzaei, M., Abo-Ismail, M.K., May, N., Miller, S.P., Schenkel, F., Moore, S.S., Stothard, P.: Genome-wide association for milk production and female fertility traits in Canadian dairy Holstein cattle. BMC Genet. 17(1), 75 (2016).
4. Leal, L.G., David, A., Jarvelin, M.-R., Sebert, S., Männikkö, M., Karhunen, V., Seaby, E., Hoggart, C., Sternberg, M.J.E.: Identification of disease-associated loci using machine learning for genotype and network data integration. Bioinformatics 35(24), 5182–5190 (2019).
5. Yao, C., Spurlock, D.M., Armentano, L.E., Page, Jr C.D., VandeHaar, M.J., Bickhart, D.M., Weigel, K.A.: Random Forests approach for identifying additive and epistatic single nucleotide polymorphisms associated with residual feed intake in dairy cattle. J. Dairy Sci. 96, 6716–6729 (2013).
6. Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. ACM, New York, NY, USA (2016).
7. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems 30, 3149–3157 (2017).
8. Ishwaran, H., Kogalur, U.: Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC). R package version 2.10.1, https://cran.r-project.org/package=randomForestSRC. (2021)
9. Craplet, C., Thibier, M.: La vache laitière. 2nd edn. Ed. Vigot Frères, Paris (1973).
10. Meyer K.: WOMBAT—A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). J Zhejiang Univ Sci B 8(11), 815–821 (2007).
11. Li, B., Zhang, N., Wang Y-G., George, A.W., Reverter, A., Li, Y.: Genomic Prediction of Breeding Values Using a Subset of SNPs Identified by Three Machine Learning Methods. Front. Genet. 9, 237 (2018).
12. Mi, H., Muruganujan, A., Ebert, D., Huang, X., Thomas, P.D.: PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Research 47(D1), D419-D426 (2019).
13. Raschia, M.A., Nani, J.P., Carignano, H.A., Amadio, A.F., Maizon, D.O., Poli, M.A.: Weighted single-step genome-wide association analyses for milk traits in Holstein and Holstein x Jersey crossbred dairy cattle. Livestock Science 242, 104294 (2020).

Funding

This study was supported by Instituto Nacional de Tecnología Agropecuaria (INTA) grants PE I145, PT I513, and PT I180, ANPCyT PICT-2017-4208, and FAO-IAEA CRP D3.10.28.