Relevant loci for milk production in dairy cattle, obtained by machine learning algorithms

M. Agustina Raschia, Pablo J. Ríos, Daniel O. Maizon, Daniel Demitrio, Mario A. Poli

Background

Extensive genetic research focused on identifying associations between single nucleotide polymorphisms (SNP) markers located all over the genome and milk traits were conducted for different dairy cattle breeds.

Most published genome-wide association studies (GWAS) were performed fitting linear, multivariate and Bayesian linear mixed models.

Machine learning (ML) methods have been shown to be efficient in identifying SNP underlying a trait of interest.

Objectives

- To identify SNPs that best explain the variance in estimated breeding values for milk production (EBV_{MP}) of Holstein and Holstein x Jersey dairy cattle, using predictive models with ML algorithms (XGBoost, LightGBM, and Random Forest).
- To compare the identified loci with previously reported relevant 10-adjacent SNP windows that explained more than 10 times genetic variance than expected for milk production, obtained for the same population by a different approach.

Materials and methods

EBV_{MP} of 837 cows (582 H, 255 HxJ) and 26 bulls (22 H, 4 J) were estimated using WOMBAT software.

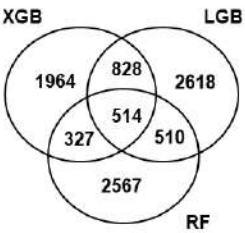
Genotyping was performed with the Illumina BovineSNP50 v2 BeadChip. 40417 SNPs remained after QC checks.

Regression models using ML algorithms were trained with EBV_{MP} as phenotypes and genotypes as predictor variables. SNPs with gain>0 were considered relevant. Their location was compared to 57 relevant SNP windows obtained previously by BLUPf90 programs.

Protein-coding genes near relevant SNPs were retrieved by the Ensembl BioMart tool.

Results

Figure. Venn diagrams showing the number of SNPs with positive gain values for XGB, LGB, and RF models.



Algorithm	XGBoost	LightGBM	Random Forest
Pearson correlation	0.610 [0.566, 0.650]	0.615 [0.571, 0.655]	0.612 [0.568, 0.652]
R^2 correlation	0.361	0.363	0.349
Mean Absolute Error	110.91 [105.92, 116.40]	111.26 [106.25, 116.77]	112.83 [107.75, 118.42]
Root Mean Square Error	144.46 [137.95, 151.61]	144.15 [137.65, 151.28]	145.78 [139.22, 153.00]
Relevant SNPs	3633	4470	3918
Flanking coding-genes	2770	3334	3002
Matching with relevant reported windows	40 (76.9%)	46 (88.5%)	40 (76.9%)
Matching with 10 top relevant reported windows	10 (100%)	10 (100%)	8 (80%)

Table. Metrics for the models used, based on actual vs. predicted values for EBV_{MP}, relevant SNPs and protein coding genes containing or flanking them in +/- 30 kb, and percentage of matching with previous results.

95% confidence intervals between brackets.

Conclusions

- The three ML algorithms used showed to be efficient in identifying a subset of SNPs explaining differences in EBV_{MP}.
- The high percentages of matching with previous reported results suggest all these algorithms, but mostly LightGBM, can be used to validate results obtained by a different approach .

References

Li B, Zhang N, Wang Y-G, George AW, Reverter A, Li Y. Front. Genet. 2018, 9:237.

Raschia MA, Nani JP, Carignano HA, Amadio AF, Maizon DO, Poli MA. Livestock Science. 2020, 242:104294.

Yao C, Spurlock DM, Armentano LE, Page Jr CD, VandeHaar MJ, Bickhart DM, Weigel KA. J. Dairy Sci. 2013, 96:6716–6729.

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.