

Preliminary study regarding use of genomic selection in Romanian cattle breeds evaluation

**Mihail Alexandru Gras*^{1#}, Cătălin Mircea Rotar^{1#}, Horia Grosu¹,
Rodica Ștefania Pelmuș¹, Cristina Lazăr¹**

[#]Authors contributed equally to this paper

*Corresponding author: gras_mihai@yahoo.com

¹National Research and Development Institute for Animal Biology and Nutrition, București-Ploiești, No.1, Balotești 077015

ABSTRACT

In Romania, the livestock is very scattered on the diverse environmental conditions. Due to difficulties regarding data collection, with the help of genomic selection, the paradigm of genetic improvement of economical traits in Romanian cattle breeds can be changed. Population studied in this paper consist in animals genotyped using 60k SNP panel. Association studies were performed for estimate the effect of SNP polymorphism on milk yield at 305 days. Genetic evaluation was compared under four statistical methodologies were compared under six scenarios 5k, 10k, 15k, 20k, 25k and 37k SNP panels. High association with phenotypes were displayed by 38 SNP's. Generally, accuracy of the studied models, tend to increase with complexity (informational input). The model with the lowest accuracy was Sire-Model (36.6% accuracy). Adding information in Animal Model increase accuracy of estimation and finally, adding genomic information the accuracy was maximized to 69.9%. For each lactation all the statistical models return highly correlated breeding values. The highest (97.2%) and the lowest (79.2%) rank correlation are registered between Test-Day and Animal Model for the first and second lactation. Genomic evaluation is a more feasible methodology than classical breeding value estimation, in term of accuracy of selection and the quality of the parentage information.

Keywords: Single Nucleotide Polymorphism, Best Linear Unbiased Prediction, Genome-Wide Association Study, models comparison, Genomic Selection.

INTRODUCTION

After EU Romanian integration, responsibility of the cattle genetic improvement moves to Breeding Associations. As a result, Breeding Programs became more diverse. On the other hand, livestock became more scattered in many small operations, decreasing efficiency in genealogy data collection and increase of number of errors in pedigree. Those errors make very difficult genetic evaluation based on Best Linear Unbiased Prediction (BLUP) methodology.

Recently, molecular biology makes possible utilization of dense panels of single nucleotide polymorphisms (SNP) information in genetic evaluation in farm animals. The improvements returned by this method are: increased accuracy, possibility to decrease generation interval, estimating GEBV early in the life of the animal, and as a final result, decrease of the costs for the entire process of selection (Hayes et al., 2009; Mancisidor, 2021; Sungkhapreecha et al., 2021).

The costs of genotyping limit the sample size, limiting also statistical power. Due to large number of hypotheses tested, large sample size is required. This paradox can be solved following a multi-stage procedure, at the first stage with a small sample size, a large SNP's panel and a less stringent threshold, and in the next stages, decrease the SNP panel size, increase the sample size and setting a more stringent p-values for threshold (Zazuli et al., 2021). In this paper we follow the first step from the multi-stage approach.

In Genomic Selection (GS) are proposed many statistical methods for estimation of marker effects in the training population. In BLUP methodology are estimated breeding values for animal with or without phenotypic records, through a Relationship Matrix (RM) created from pedigree (Sabag et al., 2021). With the beginning of genomic era, relationship matrix was replaced with a marker derived relationship matrix (Nejati et al., 1997; Olschewsky, et al., 2021). This kind of RM implementation is based on 'infinitesimal model' (Gao et al., 2006; Hayes et al., 2009), assuming a very large number of genes who contribute to the trait of interest and distributed across the genome. Use of Relationship Genomic Matrix in BLUP is termed as Genomic BLUP (GBLUP).

Having in sight the novelty of genomic applications in Romanian cattle breeding, this paper investigates breeds composition in a dairy cattle population, using genomic tools and the main genetic markers who contribute to milk production in this population.

MATERIALS AND METHODS

SNP and sample data quality control: All of the strategies used for implementing whole-genome association studies in animal breeding are to minimize the amount of genotyping required. In our studies we used the multi-

stage approach trying to identify the smallest genotyped sample size that offer the smallest reduction in power. On this paper are presented results from stage 1 of implementation, where are genotyped full sets of SNPs in small population at relatively liberal p -value. Also, in this stage, was selected the most appropriate statistical model usable for breeding value estimation. Population studied consist in 88 cows from the same herd of the National Research - Development Institute for Animal Biology and Nutrition (INCDBNA-IBNA Balotești).

Cows were genotyped using Applied Biosystems Axiom Bovine v2 Genotyping Array (ThermoFisher Scientific) with 60k SNP panel. The following quality controls (QC) filters were applied: for samples, individual call rate bigger than 97% and kinship coefficient <0.4 ; and for SNP, call rate $> 97\%$ and minor allele frequency (MAF) higher than 5%. QC was performed in plink (<https://www.cog-genomics.org/plink2>).

Statistical analysis was performed in R (<https://www.r-project.org/>), package rrBLUP (<https://cran.r-project.org/web/packages/rrBLUP/index.html>), Cappelletta et al. (2011) was used for GWAS, and for breeding value estimation were used BLUP methodology implemented in R.

Genetic evaluation was compared under four statistical methodologies Genomic, Sire, Animal and Test-Day Models.

Genomic Model: Genomic Breeding Values (GEBV) were estimated under six scenarios, for 5k, 10k, 15k, 20k, 25k and 37k SNP panels for each of the three lactations. Comparison between scenarios was performed using Pearson (between GEBV's) and Spearman (between GEBV ranks) correlations. Also, the same comparison was performed between lactations.

Phenotypes, were represented by milk yields for 305 days lactation. In total were recorded data from 70 cows (51 of Romanian Black Spotted breed and 19 Montbeliarde breed) resulted from 63 dams and 19 sires. Regarding records number, 70 performed the first lactation, 55 performed the first and the second lactation and 34 performed all three lactations.

The biometric model, used for data analysis was:

$$y_{ij} = B_i + a_j + e_{ij}$$

Where: B_i – was the fixed effect for breed; a_j – was the genetic additive effect of the animal; e_{ij} – was the error of the model.

In matrix notation:

$$\begin{bmatrix} X' \cdot X & X' \cdot Z \\ Z' \cdot X & Z' \cdot Z + \alpha \cdot H^{-1} \end{bmatrix} \cdot \begin{bmatrix} \tilde{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X' \cdot y \\ Z' \cdot y \end{bmatrix}$$

Where: X and Z were incidence matrixes and A^{-1} was the inverse of the relationship matrix.

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix}$$

Sire model: A total of 19 sire were analyzed, based on extended families, (with a total of 159 daughters), all from the same herd, but from two breeds,

Montbeliarde and Romanian Black Spotted. Each bull registered at least two daughters with records. The analysis was performed in multiples ways:

- Daughter with all lactation recorded (lactation 1, 2 and 3);
- Daughter with only first lactation recorded;
- Daughter with only second lactation recorded;
- Daughter with only third lactation recorded.

Analysis was performed using R software.

The minimum number of daughters used in analysis were 4 and the maximum is 14. The biometric model used for data analysis was:

$$y_{ij} = B_i + s_j + e_{ij}$$

Where: B_i – was the fixed effect for breed; s_j – was the genetic additive effect of the sire; e_{ij} – was the error of the model.

In matrix notation:

$$\begin{bmatrix} X' \cdot X & X' \cdot Z \\ Z' \cdot X & Z' \cdot Z + I \cdot k_1 \end{bmatrix} \cdot \begin{bmatrix} \tilde{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X' \cdot y \\ Z' \cdot y \end{bmatrix}$$

$$k_1 = \frac{4 - h^2}{h^1}$$

Where: X and Z were incidence matrixes and k_1 is a variance component.

Animal Model: For animal model, estimation was performed using BLUP methodology for 305 days of lactation milk yield trait from animals analyzed in previous models.

Data analysis was performed in two ways: single lactation (one lactation per run) and multi-lactation (all three lactation together).

Single Lactation:

The biometric model used for data analysis was:

$$y_{ij} = B_i + a_j + e_{ij}$$

Where: B_i – was the fixed effect for breed; a_j – was the genetic additive effect of the animal; e_{ij} – was the error of the model.

In matrix notation:

$$\begin{bmatrix} X' \cdot X & X' \cdot Z \\ Z' \cdot X & Z' \cdot Z + A^{-1} \cdot k_1 \end{bmatrix} \cdot \begin{bmatrix} \tilde{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X' \cdot y \\ Z' \cdot y \end{bmatrix}$$

$$k_1 = \frac{1 - h^2}{h^1}$$

Where: X and Z were incidence matrixes, A^{-1} was the inverse of the relationship matrix and k_1 was the variance component.

Multiple Lactation:

The biometric model used for data analysis was:

$$y_{ijk} = B_i + a_j + p_k + e_{ijk}$$

Where: B_i – was the fixed effect for breeds; a_j – was the genetic additive effect of the animal; p_k – was the permanent environmental effect; e_{ijk} – was the error of the model.

In matrix notation:

$$\begin{bmatrix} X' \cdot X & X' \cdot Z_1 & X' \cdot Z_2 \\ Z_1' \cdot X & Z_1' \cdot Z_1 + A^{-1} \cdot k_1 & Z_1' \cdot Z_2 \\ Z_2' \cdot X & Z_2' \cdot Z_1 & Z_2' \cdot Z_2 + I \cdot k_2 \end{bmatrix} \cdot \begin{bmatrix} \tilde{b} \\ \hat{a} \\ \hat{p} \end{bmatrix} = \begin{bmatrix} X' \cdot y \\ Z_1' \cdot y \\ Z_2' \cdot y \end{bmatrix}$$

$$k_1 = \frac{1-R}{h^2} \quad k_2 = \frac{1-R}{R-h^2}$$

Where: X and Z were incidence matrixes, A^{-1} was the inverse of the relationship matrix and k_1, k_2 were variance components.

Test day Model:

For the test day model, a total of 1669 test-days were used. The records are from 70 animals and the parent for those animals were 63 dams and 19 sires. All three lactations were used in the analysis, but the calculation was done individually for each lactation.

The biometric model used for data analysis was:

$$y_{ijk} = HTD_i + \sum_{m=0}^q (\beta_m \cdot \varphi_{jm}) + \sum_{m=0}^q (\alpha_{jm} \cdot \varphi_{jm}) + \sum_{m=0}^q (\gamma_{jm} \cdot \varphi_{jm}) + e_{tijk}$$

where: y_{ijk} = milk yield of the analyzed cow; HTD_i = fixed effect for the herd-test-day; β_m = Order 3 Legendre coefficients for fixed regression; α_{jm} = Order 3 Legendre Legendre coefficients for genetic effect; γ_{jm} = Order 3 Legendre Legendre coefficient for permanent environmental effect; e_{tijk} = error of the model.

In matrix notation:

$$\begin{bmatrix} X_1^T \cdot X_1 & X_1^T \cdot X_2 & X_1^T \cdot Z_1 & X_1^T \cdot Z_2 \\ X_2^T \cdot X_1 & X_2^T \cdot X_2 & X_2^T \cdot Z_1 & X_2^T \cdot Z_2 \\ Z_1^T \cdot X_1 & Z_1^T \cdot X_2 & Z_1^T \cdot Z_1 + \sigma_e^2 \cdot (A^{-1} \otimes G^{-1}) & Z_1^T \cdot Z_2 \\ Z_2^T \cdot X_1 & Z_2^T \cdot X_2 & Z_2^T \cdot Z_1 & Z_2^T \cdot Z_2 \sigma_e^2 \cdot (I \otimes P^{-1}) \end{bmatrix} \cdot \begin{bmatrix} \tilde{b}_1 \\ \tilde{b}_2 \\ \hat{a} \\ \hat{p} \end{bmatrix} = \begin{bmatrix} X_1^T \cdot y \\ X_2^T \cdot y \\ Z_1^T \cdot y \\ Z_2^T \cdot y \end{bmatrix}$$

X, Z_1 , Z_2 were the incidence matrixes for fixed effect, genetic effect and permanent environmental effect. b_1, b_2 were the vectors for fixed effect and for fixed regression. a and p were the vectors for breeding values and for permanent environmental effect.

RESULTS AND DISCUSSION

SNP distribution before and after quality control

A total number of 67820 SNP were genotyped. Best practice quality workflow was passed by 59066 SNPs and we removed 19403 SNP due to minor Allele frequency.

On average, after quality check 59% from the SNPs were kept, the greater removal encountering SNPs from the chromosomes 4 and 21 (51%). The proportion of eliminated SNPs was higher than that reported in literature (Bermann et al., 2021). Most of the SNPs from chromosome 21 are associated with hereditary diseases and lack of polymorphism on those SNPs can be explained by intense directional selection (for allele elimination).

Genome-Wide Association Analysis

Association study performed, reveal that are some important SNP associated previously with important aspects of milk production. P -values of all tested SNP are graphical profiled (in terms of $-\log_{10}(p\text{-values})$) in the Manhattan plot from Figure 1. The threshold was established at less stringent value of 1, and for the further studies, will be increased, and the number of samples will be increase also.

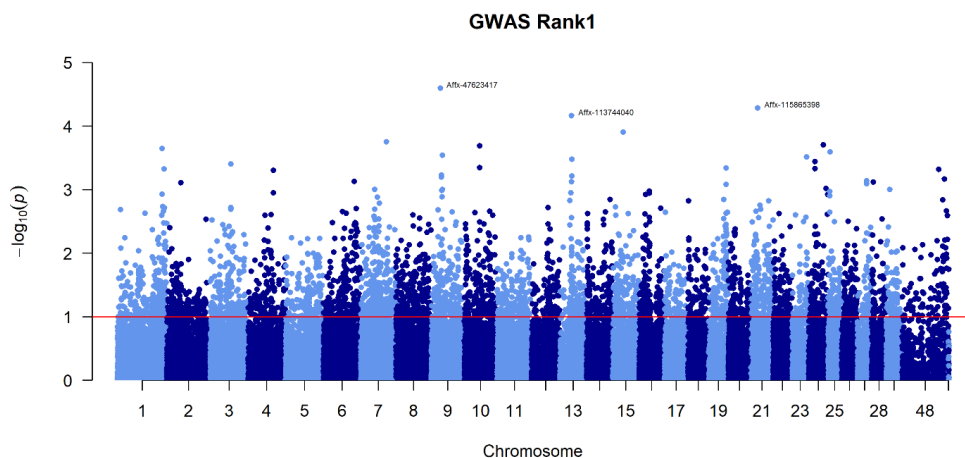


Figure 1. Manhattan plot for association between quantitative SNP with 305-day milk yield on first lactation

In figure 1, can be observed the most influential SNPs on milk production at studied breeds: Affx-47623417 from chromosome 9, Hapmap41409-BTA-103368 from chromosome 21, Hapmap42348-BTA-88290 from chromosome 13.

Alongside those 3 markers, other 35 markers display a high association with phenotypes. Details of significant SNP for milk production at 305 days are shown in Table 1 which presents a statistical view of the associated SNP across the genome.

Table 1. Significant associated SNP with milk yield

Nr. Crt.	P-value	Chromosome	Physical position	SNP name
1	2.53E-05	9	30087886	Affx-47623417
2	5.19E-05	21	22012794	Hapmap41409-BTA-103368
3	6.89E-05	13	37498503	Hapmap42348-BTA-88290
4	0.000124	15	34082620	Hapmap39267-BTA-36676
5	0.000176	7	82473566	ARS-BFGL-NGS-20384
6	0.000197	24	48164849	ARS-BFGL-NGS-13788
7	0.000205	10	49914255	Affx-277053479
8	0.000227	1	142870511	BovineHD0100041013
9	0.000254	25	6207435	Hapmap31908-BTC-019642
10	0.000288	9	36022671	BTB-00518033
11	0.000306	23	46418534	ARS-BFGL-NGS-17355
12	0.000334	13	38462778	ARS-BFGL-NGS-31610
13	0.000363	24	20823403	ARS-BFGL-NGS-81820
14	0.000393	3	66554349	BTA-110572-no-rs
15	0.00045	10	49918624	ARS-BFGL-NGS-88023
16	0.000456	19	56848494	BovineHD1900016074
17	0.000456	19	56852955	BovineHD1900016076
18	0.000468	24	20846459	ARS-BFGL-BAC-47334
19	0.000473	1	149418667	BTB-00069681
20	0.000477	X	117125952	Hapmap55643-rs29015868
21	0.000495	4	82037383	Hapmap38069-BTA-34757
22	0.000572	---	---	ARS-BFGL-BAC-33651
23	0.00059	9	32454628	ARS-BFGL-NGS-53845
24	0.000608	13	38165980	ARS-BFGL-NGS-103989
25	0.000629	9	32430302	ARS-BFGL-NGS-26381
26	0.000681	X	135879672	ARS-BFGL-NGS-107544
27	0.000722	27	29009479	ARS-BFGL-NGS-58109
28	0.000741	6	99150246	BTA-21909-no-rs
29	0.000751	13	37514636	Affx-39959134
30	0.000758	28	4797154	Hapmap36406-SCAFFOLD20062_414
31	0.00078	2	44794648	Affx-42105303
32	0.000809	27	29133451	BTA-97695-no-rs
33	0.000815	1	142832701	ARS-BFGL-NGS-114541
34	0.000824	19	57040897	Hapmap59408-ss46526398
35	0.000957	24	56627241	Hapmap38685-BTA-58652
36	0.000992	7	45135080	ARS-BFGL-NGS-100100
37	0.000993	29	11905442	ARS-BFGL-NGS-66719
38	0.000996	9	36168397	Affx-277083297

Considering this study as first stage from a multi-stage approach we will focus on the first 38 markers and we analyzed impact of population stratification and transmittance model on those markers. For instance, on

chromosome 1 we find an important SNP (ARS-BFGL-NGS-114541) (Pryce et al., 2010), an important SNP associated with lactation persistency at Australian Jersey breed.

On chromosome 13 we find SNP Affx-39959134, as part of gene RAB18, highly associated with milk production. RAB18 is a gene involved in regulation of membrane trafficking in organelles and transport vesicles, and is overexpressed in human breast. Nissen et al. (2017) in a proteomic study, detect Ras-related protein Rab-18 in bovine milk fractions.

Another SNP, strongly associated with milk yield, is BTA-97695-no-rs from chromosome 27. The nearest gene of this polymorphism is DUSP26, reported as influential in some traits like milk yield, fat and protein content at water buffaloes (Venturini et al., 2014; Du et al., 2019).

One of the SNP associated with milk yield from chromosome 9 is ARS-BFGL-NGS-26381, nearest gene being ASF1A (anti-silencing function 1A histone chaperone). Whereas considered candidate gene for meat quality and body composition on chicken, Zsolnai et al. (2021) reported association with birth and 205 days weight, with daily gain and maternal traits at Hungarian Grey cattle breed. This SNP seem to be highly connected with maternal abilities of the cows.

ARS-BFGL-NGS-17355, a SNP from chromosome 23, is situated near gene EEF1E1 (eukaryotic translation elongation factor 1 epsilon 1), gene which encode multifunctional protein, reported by Yodklaew et al. (2017) as highly associated with milk yield from the initial part of lactation (in the increasing phase of the lactation curve).

BovineHD0100041013 is a SNP from chromosome 1, near PCP4 gene, reported by da Costa Barros et al. (2018) as associated with fat content from water buffalo milk.

Venturini et al. (2014) and Du et al. (2019) report an association between milk yield and milk composition traits with a SNP situated on chromosome 15 within gene C15H11orf63. On the same chromosome and gene, we identify a SNP Hapmap39267-BTA-36676 strongly associated with milk yield.

Three SNP from the chromosomes 9 (Affx-47623417), 21 (Hapmap41409-BTA-103368) and 13 (Hapmap42348-BTA-88290), within bovine known genes GJA1, CIB1 and PCSK2 are the most significantly associated with milk yield in our study. First SNP marker, is situated on the chromosome 9 in the junction protein alpha 1 (GJA1) gene area and have an important role in udder integrity and resistance to mastitis. Ogorevc et al. (2009) report association between GJA1 and phenotypes associated with mammary gland and resistance at mastitis. Secondary as association intensity is a SNP marker from chromosome 21, in the Calcium and integrin-binding protein 1 (CIB1) gene area. CIB1 is involved in the cellular proliferation mechanism, very important for mammary secretory cells development in the last part of pregnancy (Leisner et al., 2016). At dairy cattle, involvement of CIB1 gene in udder function and integrity is not

studied yet. The third marker is situated on chromosome 13, in proprotein convertases subtilisin kexin type 2 (PCSK2) gene area, a protease with role in resistance at pathogen induced mastitis. Chen et al. (2015), find significant association between (PCSK2) gene and somatic cell count in seven different populations of dairy cows.

Population stratification

Principle component analysis (Figure 2) show three different and clear distributed cattle populations but not stratified.

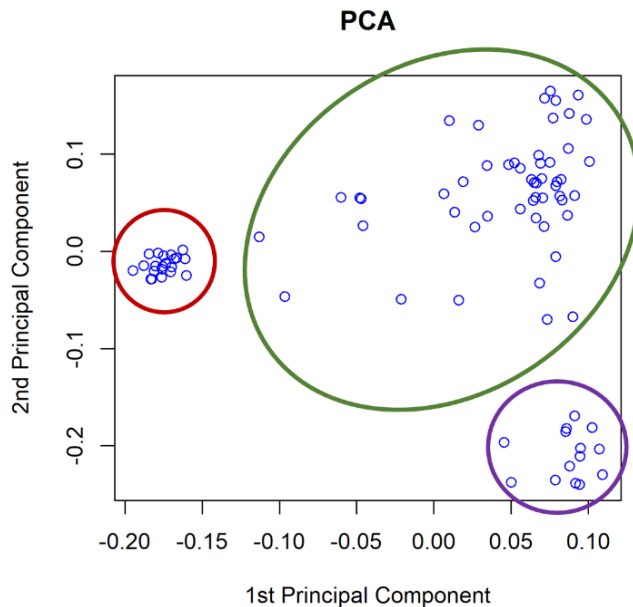


Figure 2. Principal-component analysis for population stratification in two analyzed cattle breeds

Figure 2 show two well defined genetic clusters, a smaller one (Montbeliarde - red ellipse) more clustered with a high genetical similarity due to intense selection a second population (Romanian Black Spotted – green ellipse), more dissipated and genetically dissimilar by as a result of a different selection. Surprisingly, PCA analysis showed also a third cluster (purple ellipse – Figure 2); this cluster can include a group of cross-breed animals. The analysis of genealogy data showed that these animals were cross between Romanian Black Spotted (>75%) and Brown-Swiss breeds.

The SNP information was used subsequently in Breeding Value Estimation models for comparison with traditional methodology.

Breeding Value Estimating Models analysis

In the next step we estimate the breeding values of dairy cows using three different BLUP models and comparing them for estimation accuracy.

The *genomic model* was compared with animal model and test day model to find if the rank correlation between them is higher. Genomic model was tested under 6 scenarios (5 to 37k SNP panels) using the same phenotypic information like animal model. Outputs similarity were tested using Pearson and Spearman rank correlation methodology along three lactations. Results are presented on tables 2, 3 and 4.

Table 2. Pearson and Spearman correlation coefficients of estimated breeding values using genomic model under 5, 10, 15, 20, 25 and 37k SNP panels scenarios for first lactation

		Spearman Rank Correlation Coefficients					
		ebv-5k	ebv-10k	ebv-15k	ebv-20k	ebv-25k	ebv-37k
Pearson Correlation	ebv-5k	#	0.985	0.976	0.974	0.972	0.967
	ebv-10k	0.987	#	0.994	0.992	0.993	0.991
	ebv-15k	0.978	0.995	#	0.996	0.995	0.991
	ebv-20k	0.977	0.993	0.998	#	0.997	0.993
	ebv-25k	0.975	0.992	0.996	0.999	#	0.997
	ebv-37k	0.971	0.99	0.993	0.995	0.998	#

All the SNP panels scenarios return very similar breeding values for the animals evaluated at first lactation. The closest EBV ranks are returned by the biggest SNP panels (20 vs. 25K and 25 vs. 37K SNP) with 99.7% similarity and slightly differences are observed between extreme scenarios (5 vs. 37K) with 96.7% similarity. Pearson correlation coefficients present an even higher similitude between breeding values: a maximum of 99.9% for 20 vs. 25K SNP scenario and a minimum of 97.1% for 5 vs. 37K scenario. From the GEBV point of view, moderate density of panels (20k SNPs) was the most performant, similar with those recommended by Duanyang et al. (2021). Also, the small differences between genomic estimated breeding values are maintained throughout all the analyzed lactations.

For each scenario, genomic breeding values ranging between 729.7 to 894kg milk for maximal values and from -570.1 to -604.5kg milk for minimal values. Maximum breeding value returned by the biggest panel (37K) differ with 164,3kg milk of smallest SNP panel (5k), a 22.5 percent difference between those scenarios. Minimum values returned also by those scenarios differ with 6 percent (34,4kg milk).

For the second lactation, comparison display also a high relationship between scenarios (Table 3):

Table 3. Pearson and Spearman correlation coefficients of estimated breeding values using genomic model under 5, 10, 15, 20, 25 and 37k SNP panels scenarios for second lactation

		Spearman Rank Correlation Coefficients					
		ebv-5k	ebv-10k	ebv-15k	ebv-20k	ebv-25k	ebv-37k
Pearson Correlation	ebv-5k	#	0.981	0.965	0.967	0.966	0.963
	ebv-10k	0.980	#	0.992	0.992	0.992	0.989
	ebv-15k	0.963	0.993	#	0.998	0.994	0.992
	ebv-20k	0.964	0.993	0.998	#	0.996	0.995
	ebv-25k	0.962	0.992	0.993	0.997	#	0.999
	ebv-37k	0.958	0.989	0.992	0.996	0.999	#

The highest relationship between ranks was displayed on 25k vs. 37k scenario (99.9%) and the lowest was on 5k vs. 37k scenario (96.3%). Pearson correlation coefficient follow the same trend like rank correlation but similarity between extreme scenario (5k vs. 37k) was lower (95.8%).

Genomic breeding values fluctuate trough scenarios (5k to 37k SNP panel) from 940.2 to 1075.36 kg milk for maximal values, difference of 135,16 kg milk representing 14 percent from the result returned by 5K SNP panel scenario. Minimal values range between -578.5 kg milk for 5K panel scenario and -777,94kg milk for 37K SNP panel scenario, difference of 199.44 kg milk representing 34,5 percent of 5K SNP scenario.

The breeding values estimated based on third lactation registered the highest similarity for all the scenarios. Results are presented in table 4.

Table 4. Pearson and Spearman correlation coefficients of estimated breeding values using genomic model under 5, 10, 15, 20, 25 and 37k SNP panels scenarios for third lactation

		Spearman Rank Correlation Coefficients					
		ebv-5k	ebv-10k	ebv-15k	ebv-20k	ebv-25k	ebv-37k
Pearson Correlation	ebv-5k	#	0.991	0.987	0.984	0.984	0.983
	ebv-10k	0.994	#	0.996	0.993	0.991	0.991
	ebv-15k	0.991	0.998	#	0.998	0.997	0.997
	ebv-20k	0.990	0.996	0.999	#	0.999	0.998
	ebv-25k	0.988	0.994	0.998	0.999	#	0.999
	ebv-37k	0.989	0.994	0.998	0.999	0.999	#

Trough scenarios, highest relationships between ranks are recorded on 20K vs. 25K and 25K vs. 37K scenarios (99,9% similarity) and the lowest is recorded on 5K vs. 37K scenarios (98.3% similarity). Pearson coefficients present highest similarities between breeding values on 15K vs. 20K, 20k vs. 25 and 37K, and 25K vs. 37K scenarios (99.9% similarity). The least Pearson coefficient was registered on 5K vs. 25K scenarios (98.8% similarity).

Sire Model: All 19 sire were analyzed, each bull had at least two daughters with records. The analysis was done in multiples ways: Daughter with all lactation (lactation 1, 2 and 3); Daughter with only first lactation; Daughter with only second lactation; Daughter with only third lactation comparing the accuracy of estimation for both scenarios. Predictive ability of the method was tested using Pearson (correlation between EBV) and Spearman (correlation between ranks) correlation coefficients were estimated (Table 5).

Table 5. Pearson and Spearman correlation coefficients of EBV using sire model

Pearson EBV correlation coefficients					
Spearman correlation coefficients	All lactations	Lactation I	Lactation II	Lactation III	
	All lactations	#	0.72	0.76	0.76
	Lactation I	0.66	#	0.31	0.38
	Lactation II	0.64	0.23	#	0.36
	Lactation III	0.75	0.40	0.19	#

As we expected, analyzing all the lactations together we obtain a high correlation within the estimates of the first, second and third lactation, attributing a shared genetic effect (in a proportion of 72% for first and 76 % for second and third lactation). In reality, genetic similarity between lactations is much smaller (in proportions ranging between 31% and 38%), showing a smaller to none shared effect. Ranking animals after estimated breeding value show an even smaller similarity between lactations. From the all variants we analyze, the average accuracy for multi-lactation variant was highest (56%) and for the third lactation was lowest (33%), due to increasing importance of environmental factors affecting the phenotypes in the latest lactations.

Breeding values estimated using Sire Model are high positive correlated in the all-lactation scenario and very low and to uncorrelated in individual lactation scenario. This aspect makes the multi-lactation approach more reliable than the rest. Also, maximum of accuracy on the Sire Model is obtained using all the lactation in analysis.

Animal Model: For testing animal model, 305 days lactation milk yield was used. Estimating methodologies who use animal models take in account all the animals with or without phenotype. Also, correlations between estimated breeding values were slightly increased than in sire model methodology (Table 6).

Table 6. Pearson and Spearman correlation coefficients of EBV using animal model

Pearson correlation coefficients					
Spearman correlation coefficient		All lactations	Lactation I	Lactation II	Lactation III
	All lactations	#	0.76	0.79	0.64
	Lactation I	0.74	#	0.50	0.21
	Lactation II	0.77	0.53	#	0.22
	Lactation III	0.63	0.24	0.19	#

Although EBV relationship between multi and single-lactation scenario follow the same trend like Sire Model estimation methodology, the correlations are slightly increased but not substantially (table 6).

The same as Sire Model, at Animal Model methodology accuracy was higher at the multi-lactation approach (68%) than single-lactation approach (50% and 49%). Difference between models was that than Animal Model has an increased accuracy than Sire Model (68% toward 58% for multi-lactation approach), due to supplementary information from pedigree.

Test-Day Model: For the test day model, a total of 1669 test-days were used. Test-day model is more complex than sire and animal models and offer a better view on how the genetic and environmental deviations evolved during lactations.

Individual genetic deviations offer important information about interaction between genetics and environment were studied population perform, and have also a practical importance in estimation of the best environmental effect.

Towards sire and animal models analyzed on this paper, test-day model offered a better estimate for each lactation (Table 7).

Table 7. Pearson and Spearman correlation coefficients of EBV using test-day model

Pearson correlation coefficients				
Spearman correlation coefficients		Lactation I	Lactation II	Lactation III
	Lactation I	#	0.55	0.30
	Lactation II	0.47	#	0.18
	Lactation III	0.41	0.18	#

Correlations between EBVs were the closest to the real genetic interactions between lactation ranks and diminishing constantly trough animal life, from 55% between first and second lactation to 18% between second and third lactation.

In the Test-day Model individual genetic deviation informs us about the best cow; clustering permanent environmental deviation, we can estimate the best environmental effect (the best farm), and overall curve is very informative

for the lactation persistence of cows. Test-day model can help to make the selection decision earlier in a cow's lactation and the non-productive cows can be remove from reproduction and by doing that the next generation will be superior to the previous. The correlations between EBV are closer to the real genetic interaction between lactations. A very important aspect of this approach is the amplitude of the estimates, who is the highest from the applied models.

Comparison between estimation models

The models used in breeding value estimation were compared in terms of accuracy and similarity of estimated values.

The model with the lowest accuracy was Sire-Model (36.6% accuracy), a model who use mostly information about male ancestry. Adding information about female ancestry, progeny and collaterals in Animal Model increase accuracy of estimation to 49,7%. More details about Test-Day increase also estimation to 59.1% accuracy. And finally, in the last stage, adding genomic information the accuracy was maximized to 69.9%.

Comparing the models used in breeding and genomic breeding value estimation, we observed an increase in accuracy for every information added in the model. The most inaccurate model was Sire Model and de best was Genomic Model, those things being observed also by Zhang et al. (2022).

The last comparison of the models was done using Pearson and Spearman rank correlation methodology between animal ranked by estimated breeding values. In the genomic approach, the estimates depend of the genomic information implemented in the model, and thus differences between EBVs estimated at different lactation ranks are the highest (Table 8)

Table 8. Spearman and Pearson correlation coefficients between all the statistical models

			Spearman Rank correlation Coefficients								
			Animal Model			Test-Day Model			Genomic Model		
			L I	L II	LIII	L I	L II	LIII	L I	L II	LIII
Pearson Correlation Coefficients	Animal Model	L I	#	#	#	0.792	0.551	0.320	0.892	0.448	0.154
		L II	#	#	#	0.425	0.972	0.267	0.474	0.884	0.184
		LIII	#	#	#	0.304	0.146	0.838	0.166	0.031	0.874
	Test-Day Model	L I	0.864	0.473	0.234	#	#	#	0.780	0.441	0.342
		L II	0.560	0.947	0.191	#	#	#	0.493	0.887	0.147
		LIII	0.257	0.242	0.866	#	#	#	0.279	0.194	0.827
	Genomic Model	L I	0.927	0.436	0.131	0.846	0.502	0.205	#	#	#
		L II	0.476	0.871	0.082	0.491	0.907	0.185	#	#	#
		LIII	0.161	0.153	0.898	0.273	0.110	0.849	#	#	#

Genomic Model show the less correlated EBVs between lactations ranks (51% correlation between first and second lactations, 21% between first and third lactations and 6% between second and third lactations).

Regarding similarity, for each lactation all the statistical models return highly correlated breeding values. The highest (97.2%) and the lowest (79.2%) rank correlation are registered between Test-Day and Animal Model for the first and second lactation most probably because that the two models use different phenotypic records.

Pearson coefficients follow the same trend like Spearman rank coefficients exhibiting high correlated breeding values.

Comparison between estimated breeding value display a high similarity at the lactation rank level. The most consistent similitude was presented between Animal and Genomic Models at all three lactations one of the main reasons is that both of the model uses the same phenotypic information (the total milk yield). Animal model consider that throughout all period of lactation, the environment has the same effect and because of that, some animals can appear with high or low breeding values. Similar results proclaim (Zambrano et al., 2015) at Holstein cattle breed. The test day model follows the lactation curve and for each point from the curve, the model made the corrections for the most important effect for that moment and not for the whole lactation. From the previous mentioned reasons, the rank correlation between animal model and test day model was the lowest.

CONCLUSIONS

Studied dairy cattle population was not stratified but some genetic differences existed at the SNP level. Those aspects must be followed in the breeding value estimation process.

Regarding the candidate gene approach, some SNP who showed important association with milk trait, were part of little or no studied genes at dairy cattle, and request supplementary research.

High rate of similarity between SNP panel scenarios, recommended use of the moderate SNP density panels (20-25k SNP) as very successful in estimating genomic breeding values.

Prediction of genomic breeding values were consistent with the prediction of breeding values predicted through the conventional method (BLUP). Also, genomic evaluation offers the most accurate breeding values. Better parentage estimation trough SNP offered a better quality of pedigree information than old databases.

In this investigation marker effects were estimated in a small population, as a debut in genomic selection methodology implementation, and as a comparison between classic and genomic models. For further evaluations, will be necessary to increase the number of genotyped animals, consolidating the reference population and increase the reliability of genomic evaluation.

ACKNOWLEDGEMENTS

This work was supported by the Romanian Ministry of Agriculture and Rural Development [ADER 8.1.6.] and Romanian Ministry of Research, Information and Digitization [NUCLEUS Program]

Credit Author Statement: Mihail A. Gras: Conceptualization, Methodology, Validation, Investigation, Writing- Original draft preparation, Visualization, Project administration, Funding acquisition; Cătălin M. Rotar: Software, Validation, Formal analysis, Data curation, Writing- Original draft preparation, Visualization; Horia Grosu: Software, Data curation, Supervision, Writing- Reviewing and Editing; Rodica S. Pelmuș: Writing- Reviewing and Editing; Cristina Lazăr: Writing- Reviewing and Editing.

REFERENCES

- Bermann M., et al., 2021. Modeling genetic differences of combined broiler chicken populations in single-step GBLUP. *J Anim Sci.* Apr 1;99(4):
- Cappetta E., et al., 2021. Tomato genomic prediction for good performance under high-temperature and identification of loci involved in thermotolerance response. *Hortic Res.* 2021 Oct 1;8(1):212.
- Da Costa Barros C., et al., 2018. Use of single step genome wide association studies for prospecting genomic regions related to milk production and milk quality of buffalo. *Journal of Dairy Research* 85 (4), 402 - 406.
- Du C., et al., 2019. Systematic analyses for candidate genes of milk production traits in water buffalo (*Bubalus Bubalis* *Animal genetics* 50 (3), 207 – 216
- Duanyang R., et al., 2021. Impact of Marker Pruning Strategies Based on Different Measurements of Marker Distance on Genomic Prediction in Dairy Cattle, *Animals*, 11, 1992.
- Gao S., et al., 2021. Comparing empirical kinship derived heritability for imaging genetic traits in the UK biobank and human connectome project. *Neuroimage.* 2021 Dec 15; 245:118700.
- Hayes B.J., et al., 2009. Invited review: genomic selection in dairy cattle: progress and challenges. *J Dairy Sci* 92: 433–443.
- Juan Zambrano A., et al., 2015. Estimation and comparison of conventional and genomic breeding values in Holstein cattle of Antioquia, Colombia, *Rev.MVZ Córdoba* 20(3):4739-4753.
- Leisner T., et al., 2016. CIB1: a small protein with big ambitions. *FASEB J.* Aug;30(8):2640-50. doi: 10.1096/fj.201500073R.
- Lu D., et al., 2021. Trends in genetic diversity and the effect of inbreeding in American Angus cattle under genomic selection. *Genet Sel Evol.* 2021 Jun 16;53(1):50.

- Mancisidor B., et al., 2021. ssGBLUP Method Improves the Accuracy of Breeding Value Prediction in Huacaya Alpaca. *Animals (Basel)*. 2021 Oct 26;11(11):3052.
- Nissen A., et al., 2017. Colostrum and milk protein rankings and ratios of importance to neonatal calf health using a proteomics approach. *Journal of Dairy Science* 100 (4), 2711-2728.
- Ogorevc J., et al., 2009. Database of cattle candidate genes and genetic markers for milk production and mastitis. *Animal Genetics*, 40, 832–851.
- Olschewsky A., and Dirk H. 2021. An Overview of the Use of Genotyping Techniques for Assessing Genetic Diversity in Local Farm Animal Breeds, *Animals* 11, no. 7.
- Pryce, J. E., et al., 2010. Genetic markers for lactation persistency in primiparous Australian dairy cows. *Journal of dairy science*, 93(5), 2202-2214.
- Sabag I., et al., 2021. Genome-wide association analysis uncovers the genetic architecture of tradeoff between flowering date and yield components in sesame. *BMC Plant Biol.* Nov 22;21(1):549.
- Sungkhapreecha P., et al., 2021 Validation of single-step genomic predictions using the linear regression method for milk yield and heat tolerance in a Thai-Holstein population. *Vet World*. 2021 Dec;14(12):3119-3125.
- Venturini, G. C., et al., 2014. Association between single nucleotide polymorphisms and milk production traits in buffalo. *Genet Mol Res*, 13(4), 10256-68.
- Xing C. et al., 2015. Combining genome wide association studies and differential gene expression data analyses identifies candidate genes affecting mastitis caused by two different pathogens in the dairy cow. *Open Journal of Animal Sciences*, 5, 358-393.
- Yodklaew, P., et al., 2017. Genome wide association study for lactation characteristics, milk yield and age at first calving in a Thai multibreed dairy cattle population. *Agriculture and Natural Resources*, 51(3), 223 - 230.
- Zazuli Z., et al., 2021. Association between Genetic Variants and Cisplatin-Induced Nephrotoxicity: A Genome-Wide Approach and Validation Study. *J Pers Med*. Nov 20;11(11):1233.
- Zhang M., et al., 2022. Genomic Selection for Milk Production Traits in Xinjiang Brown Cattle. *Animals (Basel)*. Jan 7;12(2):136.
- Zsolnai, A., et al., 2021. Identification of markers associated with estimated breeding value and horn colour in Hungarian Grey cattle. *Anim Biosci* Vol. 34, No. 4:482-488.