Ankara Univ Vet Fak Derg, **68**, 383-388, 2021
DOI: 10.33988/auvfd.775597

Genomic prediction and association analyses of energy corrected milk yield in dairy cows

Burak KARACAÖREN

Akdeniz University, Department of Animal Science, Antalya, Turkey
ORCID: 0000-0003-2981-6540.

✉Corresponding author: burakkaracaoren@akdeniz.edu.tr
Received date: 29.07.2020 - Accepted date: 11.11.2020

Abstract: Energy balance plays a critical role in the maintenance of metabolism for producing milk yield (MY) in dairy cows. In recent years, there has been increasing interest in genetic and genomic analyses of MY. In contrast to MY there is much less information about genomic evaluation of energy corrected milk yield (ECMY). The purpose of this paper is to detect associated single nucleotide polymorphisms (SNPs) with ECMY and genomic prediction (GP) of ECMY using different genomic models with special reference to underlying genetic architecture of ECMY. In this study we used published data of 773 Holstein cows with phenotypic observations for ECMY and dairy farm information with 62410 SNPs. One interesting finding is that some short chromosomes as such chromosomes 5 (included 28446 SNP) and 29 (included 12776 SNP) had higher effects sizes compared with the rest of the genome. A possible explanation for these results may be related with the existence of major genes at the chromosome 5. The GP results showed that ECYM and residuals of ECYM, had the accuracies from a 10-fold cross validations as 0.6422 and 0.3529 respectively. It was found that ECMY could be used for GP due to moderate accuracies. Taken together, dairy farm effects suggest an impact for accuracies of GP.
Keywords: Energy corrected milk yield, genome wide association analyses, genomic selection, milk yield.

Süt sığırlarında enerjice düzeltilmiş süt veriminin genomik tahmin ve ilişki analizleri

Özet: Süt sığırlarında, süt verimi (SV) için enerji dengesi ile metabolizmanın korunması önemlidir. SV için genetik ve genomik analizlerine olan ilgi son yıllarda önem kazanmıştır. Enerjice düzeltilmiş süt verimi (EDSV) konusunda ise SV'den farklı olarak daha az araştırma bulunmaktadır. Bu çalışmanın amacı EDSV'ye sebep olabilecek tek nükleotid polimorfizmlerini (TNP) belirlemek ve bunlar üzerinden farklı genomik modeller kullanarak genomik tahminler (GT) yapmaktır. Bu çalışmada daha önceden yayınlanmış bir veri seti kullanılarak, 773 Holstayn ineğe ait EDSV gözlemleri ile 62410 TNP ve çiftlik bilgileri incelenmiştir. Beşinci kromozom gibi kısa bir kromozomda (28446 TNP) ve 29. kromozomda (12776 TNP) GT için genomun diğer bölgelerine göre daha yüksek etki büyüklükleri belirlenmiştir. Bu durum 5. kromozomda yer alan major bir gen ile açıklanabilir. GT sonuçları EDSV ve EDSV kalıntıları ile elde edilmiş ve 10 katlı çapraz sorgulama ile 0,6422 ve 0,3529 doğruluk oranları bulunmuştur. Bu da ECYM'nin GT modellerinde orta doğrulukta kullanılabileceğini göstermiştir. Bu çalışmada; çiftlik etkilerinin GT doğruluklarında bir etkiye sahip olduğu gösterilmiştir.
Anahtar sözcükler: Enerjice düzeltilmiş süt verimi, genom tabanlı ilişki analizi, genomik seleksiyon, süt verimi.

Introduction

Energy balance plays a critical role in the maintenance of metabolism for producing milk yield (MY) in dairy cows. For instance, energy deficit postpartum is a common condition which has a considerable impact on the productional and functional traits in dairy cows (10). In recent years, there has been increasing interest in genetic and genomic analyses of MY. In contrast to MY there is much less information about genomic evaluation of energy corrected milk yield (ECMY) (4, 10). Genomic ECMY findings should make an important contribution to the field of animal breeding and husbandry by genomic selection of superior animals in shorter generation intervals.

Genome wide association studies (GWAS) are fast becoming a key instrument for detecting associated genes with the phenotypes based on molecular markers as such single nucleotide polymorphisms (SNPs). A considerable amount of literature has been published on GWAS of MY in dairy cows. Previous GWAS research has established

that various loci are correlated with MY. Jiang et al. (9) conducted a GWAS for various milk production traits using 294.079 first-lactation Holstein cows and detected strong genomic signal from chromosome 14 (DGAT1 gene) in association with MY. Han et al. (7) studied and detected the effects of nucleobindin 2 (NUCB2) gene on milk production traits using Chinese Holstein cattle. Iung et al. (8) investigated the impact of tropical condition to GWAS of Brazilian Holstein population for milk production traits and detected various genomic signals from Microsomal glutathione S-transferase 1 (MGST1), ATP-binding cassette super-family G member 2 (ABCG2), (Diacylglycerol O-Acyltransferase 1) DGAT1 and progestagen-associated endometrial protein (PAEP) genes. Lopdell et al. (13) analyzed the data from Holstein, Jersey, and crossbred populations and detected Colony Stimulating Factor 2 Receptor Subunit Beta (CSF2RB) gene in connection with milk production traits. Wang et al. (25) carried out a GWAS based on Chinese Holstein population and detected genomic signals from 7 SNPs for MY. The research to date has tended to focus on MY rather than ECMY. To date there is only one study that has investigated the ECYM in GWAS (8). Hence the use of ECMY in genomic prediction (GP) has not yet been investigated. The present research explores, for the first time, GP of ECMY with different genomic models.

Li et al. (11) examined the genetic correlations among ECMY, dry matter intake and body weight using different cattle breeds and concluded that the genetic correlations varied over lactations and showed similar patterns within each breed. ECMY is a principal determining factor of energy balance, compared with dry matter intake (10). This indicates a need to understand the genomic evaluation of ECMY by predicting associated SNPs and/or genes.

The purpose of this paper is to detect associated SNPs with ECMY (4) and genomic prediction of ECMY using different genomic models with special reference to underlying genetic architecture of ECMY.

Materials and Methods

In this study we used published data of (4). The GWAS analyses included 773 Holstein cows with phenotypic observations for ECMY and dairy farm information. The 773 cows had 624100 SNPs obtained from Illumina BovineHD BeadChip. The details of the dataset could be found at (4).

Population stratification, or systematic genotypic differences due to sources of variations may lead to false positive signals in GWAS. We used linear mixed model for correction of population stratification as was implemented in GenABEL (1) using genomewide rapid association using mixed model and regression (GRAMMAR-gamma) (11,21) approach in R software (17).

The linear mixed model used as

$$y = Xb + Za + e \qquad (1)$$

where y contains the observations, b is the dairy farm a is the additive genetic effect, matrices X and Z are incidence matrices, and e is a vector containing residuals.

$$Var\begin{pmatrix} a \\ e \end{pmatrix} \sim N\left[\mathbf{0};\begin{pmatrix} \mathbf{A}\sigma_a^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_e^2 \end{pmatrix}\right],$$

For the random effects, it is assumed that A is the coefficient of coancestry obtained from genotype of animals; I is an identity matrix, σ^2_a is the additive genetic variance and σ^2_e is the residual variance. In GWAS the huge number of hypothesis may cause the inflated number of false positive genomic signals (22). One advantage of the false discovery rate (FDR) approach is that it avoids the problem of false positive genomic signals by increasing significance levels to 0.05/(number of SNPs).

Different from major SNPs effects used in model (1) (27) defined sparse and larger variances to model SNPs effects as "Bayesian sparse linear mixed models", BSLMM, (15) used mixture of two normal distributions and additional random effects to have more flexible model compared with other Bayesian models.

We used BSLMM for prediction of SNP effects;

$$y_i = farm + \sum_{j=1}^{n}(z_{ij}\alpha_j\delta_j) + e_i \qquad (2)$$

where y_i is the phenotypes of the *i*th animal, z_{ij} is an indicator variable (small or major effects from the two normal distributions) for the *i*th animal, *j*th SNP locus and *k*th allele, α_j is marker locus effects, δ_j is indicating if SNP has effect or not and e_i is the residual for animal *i*. To see if the various assumptions regarding genetic architecture of the ECMY gave different results, the number of mixtures increased. Different from model (2) BayesR (15) assumed a mixture of four normal distributions for the SNP effects to be predicted (assumed to be 0.00001, 0.0001, 0.001,0.01 of the genetic variances). For each phenotype the Markov Chain Monte Carlo (MCMC) algorithm were run for 1.000.000 samples and first 2000 samples discarded as burn in period. We collected each 10th samples from each realization of the MCMC as thinning period.

One of the most well-known model for assessing polygenic effects in GP is to use of genomic relationship matrix in (1) where a refers to animals termed as genomic best linear unbiased prediction (GBLUP) (17). We used GBLUP, BayesR and BSLMM for prediction of phenotypes using known genotypes based on their breeding values (BV) or small gene effects (ALPHA) (27).

The whole genomic dataset was partitioned by reference and validation set. ECMY measurements of the validation set were assumed to be missing. Phenotypes of the validation set were predicted using the information from the reference set. A random sample of reference set (2/3 of the animals, n=517) was used to create predictive equations. This procedure was repeated 10 times. Correlation coefficient between the predicted and realized phenotypes of the validation animals was calculated over 10 replications.

Results

The main aim of this study was to detect gene variants that associated with ECMY using 624100 SNPs and 773 cows. In order to investigate population stratification we used a multi-dimensional scaling (MDS) analysis (1). Figure 1 presents an overview of stratification by genotypes of cows based on top two genomic principal components using identity by descent information over MDS analysis. As shown in Figure 1 the main cow population are closely related but still separate clusters exist in the MDS plot (Figure 1).

The mean heterozygosity for a SNP was 0.3552 (0.1396) and for a cow was 0.3587 (0.014). GWAS assume homogeneous populations for contrasting frequencies of the SNPs to detect putative genomic associations. By employing single regression models (SRM) with correction for the population stratifications may lead to valid results of the GWAS. In order to take population stratification into account for SRM, we predicted genomic relationship matrix (1) and conducted the association analyses using the function "gamma" (SRM) as was implemented in GenABEL package (1). Table 1 shows the GWAS results of the SRM with false discovery rate for multiple hypothesis testing correction. The genomic inflation factor found to be 1.025 with standard error of 0.000041. The estimated genomic heritability of ECMY was 0.8541.

The most significant SNP on chromosome 5 (Table 1) was within the QTLs of milk fat percentage, milk protein percentage and milk fat yield (18). The second QTL on the same locus of chromosome 5 (Table 1) associated with body weight (58.3-70.8 Mbp) (19). Other significant SNP was BovineHD1700005467 (Table 1) whose importance in milk palmitoleic acid content has been defined (20) at the vicinity of chromosome 17 at 17.1-22.4 mega base pairs. The SNP on chromosome 6 (BovineHD0600008918) was identified in association with body weight in cattle at 35.56 Mb (2).

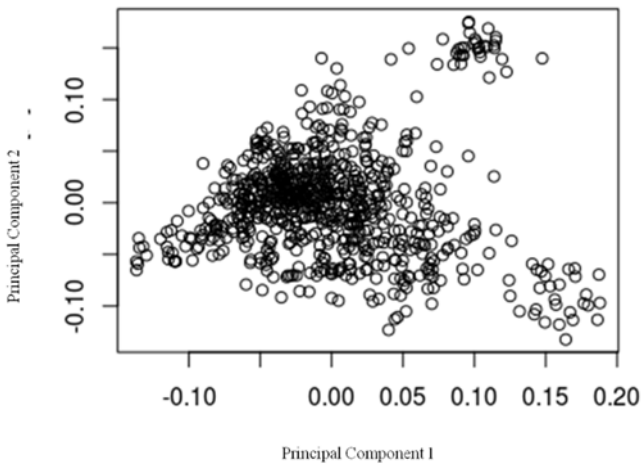

Figure 1. Multi-dimensional scaling analyses of genotypes.

Table 1. Top SNPs of the single regression model with GRAMMAR-gamma correction analyses of ECMY for false discovery rate of 0.05.

Marker	Chromosome	P value	BP	False Discovery Rate
BovineHD0500016776	5	1.11E-10	59905836	0.0000693
BovineHD3000016842	30	6.46E-10	58003830	0.000202
BTB-01179030	14	1.27E-09	60164356	0.000264
BovineHD1700005467	17	4.72E-09	18948241	0.000736
BovineHD0600008918	6	2.25E-08	31792754	0.002808
BovineHD1800014456	18	5.21E-08	49004334	0.005419
BovineHD1500019853	15	6.38E-08	68743862	0.005688
BovineHD0800008984	8	1.42E-07	29615889	0.011078
BovineHD0200029838	2	6.37E-07	103773772	0.04013
BovineHD1300000094	13	6.43E-07	599841	0.04013
BovineHD1300000090	13	7.16E-07	584808	0.040623

In order to assess the genetic architecture of ECMY, different effect sizes of the SNPs effects were used by BayesR model. This was done because the SRM model only assumed SNPs with major effects for the ECMY. Table 2 presents the summary statistics for the top ten SNPs obtained by BayesR. Table 3 provides the breakdown of genetic variance according to chromosomes. The number of SNPs associated with ECMY changed considerably among chromosomes. Highest proportion of the total variance is explained by chromosome 5 (Table 1).

Table 4 compares the correlation coefficients for GP using different models for ECYM and corrected ECYM for dairy farm effects. On average correlations were shown to have similar results for different models for ECYM and the residuals of ECYM. However, from the Table 4, it can be seen that corrected ECYM resulted in the lowest correlations compared with ECYM. Data for the MY and residuals of MY in Table 1 can be compared with the ECYM and residuals of ECMY which shows similar trends over different GP models. However, the results of the BayesR resulted in the highest correlations for MY (0.3529) and the lowest for the residuals of MY (0.0268).

Table 2. Top ten SNPs of the BayesR model

SNP	CHR	BP	PROP
BovineHD2900006632	29	23294478	0.044745
BovineHD0500016776	5	59905836	0.038795
BovineHD0800008984	8	29615889	0.034092
BovineHD1800010522	18	34509354	0.013303
Hapmap24310-BTA-162764	15	3335649	0.010461
BovineHD2400005028	24	19198046	0.010436
BovineHD0300012765	3	41843197	0.009673
BovineHD1400019682	14	70036249	0.008425
BovineHD0200000556	2	1929907	0.007427
BovineHD0900018960	9	68547030	0.006995

Table 3. Sum of SNPs effects and number of SNPs obtained over chromosomes from BayesR.

Chromosome	Sum of SNP effects	Number of SNPs by BayesR	Number of SNPs in the map file
1	0.05208	434	38338
2	0.04452	371	32162
3	0.04147	319	29400
4	0.0372	310	29010
5	0.07194	327	28446
6	0.03432	312	30057
7	0.03096	258	26839
8	0.05824	224	22970
9	0.04256	304	25829
10	0.03341	257	25933
11	0.0296	296	27589
12	0.0246	205	21635
13	0.0185	185	16974
14	0.03444	164	17576
15	0.0357	238	21066
16	0.02412	201	20006
17	0.02544	212	19259
18	0.03553	187	17088
19	0.01925	175	16312
20	0.02472	206	18517
21	0.01925	175	17586
22	0.0194	194	15981
23	0.01104	138	13311
24	0.03111	183	15421
25	0.01062	118	11520
26	0.01287	143	13380
27	0.01359	151	11639
28	0.01742	134	11625
29	0.05658	138	12776
30	0.01815	165	15822

Table 4. Pearson correlations of Genomic predictions obtained by different models for ECYM and residuals of ECYM.

Method	ECMY	Residuals of ECMY	MY	Residuals of MY
BayesR	0.6422	0.5046	0.3529	0.0268
BSLMM_BV	0.6275	0.5418	0.2405	0.0541
GBLUP_BV	0.6275	0.5475	0.2429	0.0692
BSLMM_ALPHA	0.6276	0.5422	0.2399	0.0536
GBLUP_ALPHA	0.6244	0.5475	0.2429	0.0694

Discussion and Conclusion

To date only one study have used ECMY as phenotype in GWAS (24). However a strong relationship between ECMY and productional and functional traits has been reported in the literature. An initial objective of the study was to identify SNPs in associated with ECMY. ECMY, as derived trait from MY was used as a response variable in GWAS for the current study. However it is not uncommon to use derived and/ or standardized errors as phenotypes in GWAS. In recent years, there has been an increasing amount of literature on deregressed estimated breeding values (DEBV) as phenotypes in GWAS. A significant analysis and discussion on the subject was presented by (16) and (20). A recent study by (12) involved a GWAS using DEBV for MY in buffalo.

MDS plot detected genotypic clusters using principal components analyses (Table 1). We used SRM model with genotypic relationship matrix to take this relationship into account for GWAS of ECMY. The result of this SRM analyses indicate that there are various genomic signals in association with ECMY (Table 1), particularly from chromosomes of 5, 14 and 30. This finding broadly supports the work of other studies in this area linking ECMY with energy metabolism over MY and body weight. It is somewhat surprising that no gene was detected at vicinity of the chromosome 14 SNP of BTB-01179030 (Table 1). This outcome is contrary to previous studies which have suggested that strong genomic signals for MY from various loci of chromosome 14.

In accordance with the present results, previous studies have demonstrated that complex phenotypes could be explained by genes with small to major effects (15). It was hypothesized that the SNPs could be distributed into four classes according to their effects sizes on the ECMY. As can be seen from the Table 2 that the strongest genomic signal was found to be at chromosome 29 BayesR. These results reflect those of (12): who also found a genomic signal from the similar location for milk protein yield. There was a significant negative linear relationship between number of chromosomes and number of SNPs per chromosomes (Table 3) similar to the other organisms. Contrary to expectations, this study did not find a strict linear relationship between chromosomal sizes, detected number of SNPs and sum of SNPs effects (Table 3) (24). One interesting finding is that some short chromosomes as such chromosomes 5 (included 28446 SNP) and 29 (included 12776 SNP) had higher effects sizes compared with the rest of the genome. A possible explanation for these results may be related with the existence of major genes at the chromosome 5 (Table 1).

GP results of Table 4 is revealing in several ways. The results showed that ECYM and residuals of ECYM using BayesR, had the higher accuracies from a 10-fold cross validations as 0.6422 and 0.3529 respectively. In reviewing the literature, no data was found on the GP of ECMY.

With respect to the research question, it was concluded that ECMY could be used for GP due to moderate accuracies in Table 4. In this study (as environmental and genetical factors) dairy farm effects were found to cause of inflation for accuracies of GP (Table 4). In accordance with the present results, previous studies have demonstrated the effect of environmental factors (dairy farm) for the GP (19, 23, 26). These findings contribute in several ways to our understanding genomics of ECMY and provide a basis for GP studies or pathway/ gene investigations by detected SNPs (Table 1-3).

Financial Support

This work was supported by the computational resources obtained from Scientific and Technological Research Council of Turkey (TÜBİTAK), Grant No: 118O108.

Ethical Statement

This study does not present any ethical concerns.

Conflict of Interest

The authors declared that there is no conflict of interest.

References

1. **Aulchenko YS, De Koning DJ, Haley C** (2007): *Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis*. Genetics, **177**, 577-585.
2. **Bennewitz J, Reinsch N, Guiard V, et al** (2004): *Multiple quantitative trait loci mapping with cofactors and application of alternative variants of the false discovery rate*

in an enlarged granddaughter design. Genetics, **168**, 1019-1027.

3. **Buttchereit N, Stamer E, Junge W, et al** (2011): *Genetic relationships among daily energy balance, feed intake, body condition score, and fat to protein ratio of milk in dairy cows*. J Dairy Sci, **94**, 1586-1591.
4. **Clancey E, Kiser JN, Moraes JG, et al** (2019): *Genome-wide association analysis and gene set enrichment analysis with SNP data identify genes associated with 305-day milk yield in Holstein dairy cows*. Anim Genet, **50**, 254-258.
5. **Gebreyesus G, Buitenhuis AJ, Poulsen NA, et al** (2019): *Multi-population GWAS and enrichment analyses reveal novel genomic regions and promising candidate genes underlying bovine milk fatty acid composition*. BMC Genom, **20**, 178.
6. **Habier D, Fernando RL, Garrick DJ** (2013): *Genomic BLUP decoded: a look into the black box of genomic prediction*. Genetics, **194**, 597-607.
7. **Han B, Yuan Y, Li Y, et al** (2019): *Single nucleotide polymorphisms of NUCB2 and their genetic associations with milk production traits in dairy cows*. Genes, **10**, 449.
8. **Iung LHS, Petrini J, Ramírez-Díaz J, et al** (2019): *Genome-wide association study for milk production traits in a Brazilian Holstein population*. J Dairy Sci, **102**, 5305-5314.
9. **Jiang J, Ma L, Prakapenka D, et al** (2019): *A Large-Scale genome-wide association study in U.S. holstein cattle*. Front Genet, **14**, 412.
10. **Krattenmacher N, Thaller G, Tetens J** (2019): *Analysis of the genetic architecture of energy balance and its major determinants dry matter intake and energy-corrected milk yield in primiparous Holstein cows*. J Dairy Sci, **102**, 3241-3253.
11. **Li B, Fikse WF, Løvendahl P, et al** (2018): *Genetic heterogeneity of feed intake, energy-corrected milk, and body weight across lactation in primiparous Holstein, Nordic Red, and Jersey cows*. J Dairy Sci, **101**, 10011-10021.
12. **Liu JJ, Liang AX, Campanile G, et al** (2018): *Genome-wide association studies to identify quantitative trait loci affecting milk production traits in water buffalo*. J Dairy Sci, **101**, 433-444.
13. **Lopdell TJ, Tiplady K, Couldrey C, et al** (2019): *Multiple QTL underlie milk phenotypes at the CSF2RB locus*. Genet Sel Evol, **51**, 1.
14. **McClure MC, Morsci NS, Schnabel RD, et al** (2010): *A genome scan for quantitative trait loci influencing carcass, post-natal growth and reproductive traits in commercial Angus cattle*. Anim Genet, **41**, 597-607.
15. **Moser G, Lee SH, Hayes BJ, et al** (2015): *Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model*. PLoS Genet, **11**, e1004969.
16. **Ostersen T, Christensen OF, Henryon M, et al** (2011): *Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs*. Genet Sel Evol, **43**, 38.
17. **R Development Core Team** (2013): A language and environmental for statistical computing. R Foundation for Statistical Computing; Vienna, Austria.
18. **Saatchi M, Schnabel RD, Taylor JF, et al** (2014): *Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds*. BMC Genom, **15**, 442.
19. **Schultz NE, Weigel KA** (2019): *Inclusion of herdmate data improves genomic prediction for milk-production and feed-efficiency traits within North American dairy herds*. J Dairy Sci, **102**, 11081-11091.
20. **Song H, Li L, Zhang Q, et al** (2018): *Accuracy and bias of genomic prediction with different de-regression methods*. Animal, **12**, 1111-1117.
21. **Svishcheva GR, Axenovich TI, Belonogova NM, et al** (2012): *Rapid variance components-based method for whole-genome association analysis*. Nat Genet, **44**, 1166-1170.
22. **Tam V, Patel N, Turcotte M, et al** (2019): *Benefits and limitations of genome-wide association studies*. Nat Rev Genet, **20**, 467-484.
23. **Tiezzi F, de Los Campos G, Gaddis KP, et al** (2017): *Genotype by environment (climate) interaction improves genomic prediction for production traits in US Holstein cattle*. J Dairy Sci, **100**, 2042-2056.
24. **Veerkamp RF, Coffey MP, Berry DP, et al** (2012): *Genome-wide associations for feed utilisation complex in primiparous Holstein-Friesian dairy cows from experimental research herds in four European countries*. Animal, **6**, 1738-49.
25. **Wang D, Ning C, Liu JF, et al** (2019): *Short communication: Replication of genome-wide association studies for milk production traits in Chinese Holstein by an efficient rotated linear mixed model*. J Dairy Sci, **102**, 2378-2383.
26. **Yao C, De Los Campos G, VandeHaar M, et al** (2017): *Use of genotype × environment interaction model to accommodate genetic heterogeneity for residual feed intake, dry matter intake, net energy in milk, and metabolic body weight in dairy cattle*. J Dairy Sci, **100**, 2007-2016.
27. **Zhou X, Carbonetto P, Stephens M** (2013): *Polygenic modeling with Bayesian sparse linear mixed models*. PLoS Genet, **9,** e1003264.