Proceedings, 10th World Congress of Genetics Applied to Livestock Production

An approach to genomic analysis of longitudinal data using random regression

D. J. Santos[1], S. A. Boison[2], A.T Utsunomya[1], M. G. C. D. Peixoto[3], H. Tonhati, J. Sölkner[2], and M. V. da Silva[3]
[1]UNESP, Jaboticabal, São Paulo, Brazil, [2]University of Natural Resources and Life Science, Vienna, Austria,
[3]Embrapa Dairy Cattle, Juiz de Fora, Brazil.

ABSTRACT: Genetic evaluations of 305-days milk yield has been more accurately estimated using random regression models (RRM). We propose the use of random regression coefficients as phenotype for genomic evaluation of longitudinal milk production data. Pedigree based estimated breeding values of 1) milk yield at 305 day (P305), 2) three independent random regression coefficient (Coef305) and 3) cumulative breeding values estimated with RRM from day 6 to 305 (RRM305) are deregressed and used as pseudo-phenotypes. GEBV's (gP305, gCoef305 and gRRM305) are estimated with a genomic-polygenic model. Pair comparison using spearman rank correlations of GEBV between pseudo-phenotypes and 10 fold cross-validation were used to estimate predictive ability. Spearman rank correlation were 0.85 between gP305 and both gCoef305 and gRRM305; and 0.94 between gRRM305 and gCoef305. Predictive ability was 0.74 and 0.63 for gRRM305 and gCoef305. Deregressed random regression coefficient can be used in genomic evaluations.
Keywords: genomic evaluation; lactation curve; SNP

Introduction

Traditionally, genetic evaluation of repeated measures, such as monthly milk yield have being analyzed quantitatively by means of test-day models (e.g. RRM), providing greater accuracy than the evaluation of milk yield in 305 days. Genomic evaluation methodologies, especially bayesian models for longitudinal data, are deemed statistical and computationally demanding. Single-Step (SSBLUP) (Koivula et al., (2012)) models and the use of mathematical functions to generate phenotypes for subsequent use in genomic evaluation (Silva et al., (2012)) have been suggested. A slight drawback of the SSBLUP with test-day model is the extreme shrinkage of marker effect. Phenotypes (mostly a minimum of 2 per animal) generated from mathematical functions allows for fitting bayesian genomic selection models however this will require a bivariate estimation of marker effect. Univariate analysis (independent estimation) of the phenotypes can lead to bias in model effect estimations. Additionally, these phenotypes have low heritability and model convergence problems (El Faro et al. (1999)).

Random regression coefficient, estimated during traditional genetic evaluations of milk yield in 305 days and persistency, provide a source of information that can be used for genomic evaluation. They can be an easier way of estimating genomic lactation curve than fitting phenotypes generated from mathematical functions.

Thus the objective of the study was to evaluate the feasibility of estimating genomic EBVs with random regression coefficient (RRC) as alternative to genomic evaluation of longitudinal data of milk yield of Guzerá cows. The specific objectives were to compare Pedigree and genomic based EBV of 1) milk yield at 305 day (P305), 2) three independent random regression coefficient (Coef305) and 3) cumulative breeding values estimated with RRM from day 6 to 305 (RRM305) as pseudo-phenotypes. GEBV's (gP305, gCoef305 and gRRM305) are estimated with a genomic-polygenic model. Pair comparison using spearman rank correlations of GEBV between pseudo-phenotypes and 10 fold cross-validation were used to estimate predictive ability.

Materials and Methods

Phenotypic Data, Pedigree and genetic-quantitative analysis: Milk yield information of 3,231 cows participating in the Programa de Melhoramento do Guzerá conducted by Embrapa Gado de Leite in partnership with ABCZ. A univariate traits model for milk yield in 305 days (traditional) and random regression model (RRM) using Legendre polynomials were used. For additive genetic and permanent environmental random regressions were considered polynomials with cubic order. The residual variance was modeled considering the heterogeneous structure of 10 classes. Variance components were estimated by Restricted Maximum Likelihood (REML) using the program Wombat (Meyer, 2006). The estimated heritability for milk yield in 305 days (P305) by the traditional model and RRM (RRM305) were 0.27 and 0.29. The traits that were genomically evaluated were the breeding value of P305, RRM305 and three random regression coefficients from 844 animals. The reliability of EBV for P305 ranged from 0.17 to 0.88, with mean 0.46; RRM305 ranged from 0.22 to 0.88, with mean 0.49; the first random coefficient ranged 0.13 to 0.87, with mean 0.45; the second random coefficient ranged 0.17 to 0.90, with mean 0.49; and the third varied 0.14 to 0.87, with mean 0.43.

Genotypic data and Quality control: For this study, 804 cows and 40 sires were genotyped using Illumina®

Table 1. Rank correlation (Spearman) of genomic predictions for effects different when it was considered as phenotype, the EBV (above the diagonal) and the EBVdr (below the diagonal)

Trait	GEBV			PEBV			Overall (GEBV + PEBV)		
	gP305	**gRRM305**	**gCoef305**	**gP305**	**gRRM305**	**gCoef305**	**gP305**	**gRRM305**	**gCoef305**
gP305	-	0.90	0.86	-	0.87	0.68	-	0.89	0.89
gRRM305	0.86	-	**0.97**	0.82	-	**0.78**	0.85	-	**1,00**
gCoef305	0.82 (0.75)	**0.92 (0.91)**	-	0.68 (0.61)	**0.75(0.75)**	-	0.85 (0.79)	**0.94 (0.97)**	-

() correlation value when it was used as phenotype EBVdr to P305 and RRM305, and EBV for the coefficients.
GEBV = genomic breeding value.
PEBV = random polygenic solution estimated using the pedigree

BovineSNP50 BeadChip. Samples with call rate < 0.90 were discarded. SNPs with call rate < 0.95, MAF < 2%, with HWE 1e-06 were discarded. The missing genotypes were imputed using the software Fimpute V2 (Sargolzaei et al. (2012)).

Statistical Analysis: EBV and deregressed breeding values (EBVdr) according to Garrick et al. (2009) were used for the analysis. We used a genomic-polygenic BLUP implemented in software GS3 (Legarra et al., 2013). In matrix notation the model can be represented as follows:

where y is the vector of observations (EBV and EBVdr), μ is the population mean, a is the vector of marker effects, u is the vector of polygenic effect solutions, and W and Z are the corresponding incidence matrices. For effects a, u and e were assumed priors with normal distribution with $N\left(0, G\sigma_a{}^2\right)$, $N\left(0, A\sigma_u{}^2\right)$ and $N\left(0, D\sigma_e{}^2\right)$, where G is the genomic kinship matrix, A is the relationship matrix and D is a diagonal matrix with $d_{ij} = \left(1/w_i\right)$ elements. The weight w_i accounts for heterogeneity variance due to difference in accuracy of genetic evaluations. The weight was defined as $w_i = r^2/\left(1-r_i^2\right)$, where r_i^2 is the reliability of genetic evaluation. The number of iterations used to estimate model parameters was 300,000; initial burn-in = 30,000; initial thinning = 50.

After genomic evaluation of the coefficients, the genomic value of day 6 until 305 for animal k was estimated as $\left(gCoef305_k\right) = t_{305}\hat{u}_k$, where $t_{305} = \sum_{i=6}^{305}\sum_{j=0}^{3} ij$ with j^{th} element equal sum of j^{th} Legendre orthogonal polynomials the day to the 305, $\hat{u}$ is a vector with the regression coefficient of animal k independently predicted.

The following formula was used to estimates marker effects: $\hat{g}_{305_i} = t_{305}\hat{\alpha}_i$, where $\hat{g}_{305_i}$ is the estimated marker effect i for 305 days, and α is a vector with coefficient effects of the marker i independently predicted. The GBLUP method is generally not so suitable to evaluate the effects of markers such as Bayesian models, but comparisons were made to evaluate random coefficients as the phenotype, since these phenotypes may be employed in any method.

Rank correlation and predictive ability: Rank correlation (Spearman) was performed between the genomic breeding values of the milk yield in 305 days by the traditional model $\left(gP305\right)$, the sum of the breeding value up to 305 days $\left(gRRM305\right)$ and obtained through GEBV coefficients $\left(gCoef305_k\right)$. The analysis was done with both EBV and EBVdr of the 3 pseudo-phenotype. Predictive ability was calculated as the correlation between EBVdr and GEBV using 10 fold cross-validation

Results and Discussion

Spearman rank Correlations: Generally the rank correlation between $gRRM305$ and $gP305$ were the lowest observed, followed by the correlation between $gRRM305$ and $gP305$. In Table 1 it can be seen that when GEBVs are predicted with only markers were lower than using the full genomic-polygenic model that takes into account the pedigree information. Most importantly, the spearman correlation between gP305 and both $gCoef305$ and $gRRM305$ was the same. This signifies that, $gCoef305$ can be used as pseudo-phenotypes for RRM. It must however be noted that the use of EBVs as pseudo-phenotypes for RRC results in lower rank correlation between both gRRM305 and gCoef305 and gP305. On the other hand the correlation between gCoef3 and gRRM305 using EBV as pseudo-phenotypes resulted in higher rank correlations.

The bias was calculated as the linear regression between gRRM305 and gCoef305. The regression coefficient was estimated as 0.98.

Impact on individual markers: From the solution of the markers effect for each coefficient, we could predict cumulative 305 days marker effect (Table 2). The marker effect could thus be studied at a point in time of the lactations curve. Generally EBVs as pseudo-phenotypes were poor in estimating SNP effect. We propose that deregressed random regression coefficients are the pseudo-phenotypes of choice.

Table 2. Pearson correlation and rank (Spearman) of genomic predictions for marker effects when it was considered as phenotype, EBV (above the diagonal) and EBVdr (below the diagonal)

Trait	gP305	gRRM305	gcoef305
gP305	-	0.77/0.75	0.70/0.67
gRRM305	0.73 /0.70	-	**0.92/0.92**
gCoef305	0.69/0.66 (0.61/0.68)	**0.92/0.91 (0.87/0.86)**	-

() correlation value when it was used as phenotype EBVdr to P305 and RRM305, and EBV for the coefficients.

Table 3. Pearson correlation between genomic predictions and EBVdr in 10 k-folders

Trait	GEBV	Overall (GEBV + PEBV)
gP305	0.41	0.43
gRRM305	0.60	**0.74**
gCoef1	0.47	0.49
gCoef2	0.29	0.35
gCoef3	0.32	0.35
gCoef305*	**0.64 (0.87)**	0.62(0.93)

* The correlation of this prediction was with EBVdrRRM305
() Correlation with gRRM305
GEBV = genomic breeding value.
PEBV = random polygenic solution estimated using the pedigree

Predictive ability: The predictive ability for different phenotypes can be observed in Table 3. Among the traits considered gP305 days showed the best predictive ability followed by gRRM305 and gCoef305. This was possibly because RRM305 is the most accurate measure of genetic-quantitative analyses for the period of 305 days evaluated. However, when considering only the effect of markers to predict phenotype, the random coefficients predicted independently showed be the best option, having increased the correlation 7% greater than that provided by $gRRM305$.

Considerations: Despite the greater work involved in genomic evaluation of random coefficients (3 evaluation), the obtaining these genomic coefficients can be advantageous when you want to evaluate many points and periods of lactation, like monthly genetics evaluation (usually 10), persistency of lactation and cumulative production in 305, or any other period. The coefficients independently evaluated be showed correlated with the RRM305 measure, admittedly more accurate than the P305 itself. In this sense, showed a small bias in the estimation by independent analysis, and may be considered in studies of effects of the markers. Moreover, it was able to predict the phenotype RRM305 properly, suggesting that the best option when considering only effect of the markers. These results suggest further studies of these genomic coefficients for a plausible application in evaluation of dairy cattle.

Conclusion

The independent approach to longitudinal evaluation by genomic analysis of random coefficients indicated viability, especially when there are many periods to be estimated. Degressed random coefficients were the best phenotype to use. The prediction ability of these coefficients were satisfactory thus can be used to predict GEBV of young animals.

Literature Cited

Koivula, M., Strandén, I., Pösö, J., et al. (2012). Interbull Bulletin 46 , 28 – 31.

El Faro, L., Albuquerque, L.G. (1999). Rev. bras. zootec.28, 987-992.

Garrick, J. D., Taylor, J. F., Fernando, R. L. (2009). Genetics Selection Evol. 41, 1-8.

Meyer, K. (2007). J. Zhejiang Univ. Sci. B 8, 815–821.

Legarra, A. (2009).

Sargolzaei, M., Chesnais, J.P., Schenkel, F. (2012). Open Ind. Sess. Oct. 30, 2012, 1–10.

Silva, F.F., Resende, M.D.V., Rocha, G.S. et al. (2013). Gen Mol Bio 36, 520-527.