The application of several genomic models for the analysis of small holder dairy cattle data

R. Mrode[1,2], H. Aliloo[3], C. Ekine[1], J. Ojango[1], J.P. Gibson[3] and M. Okeyo[1]

[1] *International Livestock Research Institute (ILRI), PO Box 30709, Nairobi, Kenya*
[2] *University of New England, Armidale, NSW 2350, Australia*
[3] *Scotland's Rural College, Easter Bush, Midlothian EH25 9RG, Scotland, UK*

Abstract:

The lack of data recording in smallholder dairy cattle systems implies that the availability of molecular data could offer some quick wins in terms of using genomic information for genomic prediction and selection. Initial studies on genomic prediction with data from smallholder dairy herds have reported promising results with low to medium values for the accuracy of prediction. The relatively small size of data in those studies limited the range of models that could be fitted. With more data now becoming available in the African Dairy Genetic Gains (ADGG) project operating in Tanzania and Ethiopia, this paper examines the impact of fitting GBLUP models with dominance effects, use of a random regression model and various Bayesian methods on the accuracy of genomic predictions in small holder dairy milk yield data from Tanzania. The data set consisted of 9193 milk test date yields on 1930 cows from 456 herds which were genotyped with Genomic Profiler (GGP) Bovine 50K chip. First analysis was GBLUP based on a fixed regression model consisting of the fixed effects of ward, age nested parity, test-year-season; fixed curves with Legendre polynomials of order four nested within breed classes by parity interaction. The random effects were herd animal and permanent environmental (PE) effects. The second model (GBLUP-D) was the same as the above fixed regression model but with dominance effects as an additional random term. The third model was a random regression model (GBLUP-RRM) with fixed effects as in the fixed regression model plus the random effects of herds, animal and PE but the latter two effects modelled with Legendre polynomials of order two. Corrected phenotypes or yield deviations (YDs) were then derived from the GBLUP model and were used as response variables in various Bayesian analysis fitting BayesA, BayesB and BayesC. The heritability estimate from GBLUP and GBLUP-D were the same at 0.14±0.04 while the estimate for GBLUP-RRM was higher at 0.26. The proportion of total variance due to PE effects was 0.26 ± 0.02 for both GBLUP and GBLUP-D but was slightly lower at 0.24 for the GBLUP-RRM. The proportion of total variance due to dominance was low at 0.03±0.08, which was not significantly different from zero. Both cross-validation and forward validation were undertaken to estimate the accuracy of genomic prediction. The estimates of accuracy from the cross-validation from GBLUP and GBLUP-D were low to medium (0.28 to 0.44), being highest for cows with the lowest proportion of exotic genes. The estimates for accuracy were higher from the forward validation with values ranging from 0.30 to 0.43 accompanied with regression coefficients which were closer to unity. The accuracies of prediction for the Bayesian methods were generally very low, varying from 0.10 to 0.21. This was unexpected and will be further examined. In conclusion, the GBLUP-RRM resulted in better accuracies of prediction compared to GBLUP but estimates of accuracies were moderate. As more data accumulates from the ADGG project, these models in addition to those that account for breed origin of alleles will be examined further including joint genomic predictions across countries to examine impact on accuracies.

Key words: accuracy of genomic prediction, dominance effect, small holder dairy cattle, random regression

Introduction

Rapid rates of genetic progress have resulted from the application of genomic selection in recent years especially in dairy cattle in developed countries. Consequently, a higher proportion of active sires currently used for artificial insemination (AI) in these countries are genomically evaluated young bulls. For instance, Hutchison et al. (2014) reported that young bulls accounted for 28% and 25% of Holstein and Jersey inseminations in 2007, respectively. These percentages increased to 51% and 52% respectively, in 2012 due to the use of genomically evaluated young bulls. The existence of well-established conventional pedigree-based genetic evaluation systems have provided the foundation for the success of genomic selection schemes coupled with large reference populations and well-defined phenotypes mostly on pure breeds (Hayes et al., 2009). On the contrary, the dairy systems in smallholder settings in middle and low income countries are rather fragmented, characterized with lack of systemic data and pedigree recording schemes, small herd sizes and the animals reared are mostly crossbreds of various breed compositions (Mrode et al., 2019). In addition, most of the genotyped animals are usually females and it is therefore difficult to implement the conventional genomic selection based on large reference populations. The feasibility of genomic prediction in small holder dairy systems was demonstrated by Brown et al. (2016) using the dataset of about 1038 animals from the Dairy Genetics East Africa project. They reported levels of accuracy ranging from 0.28 to 0.41 which could be considered suitable for the selection of groups of young bulls. In 2016, the African Dairy Genetics Gain project (https://africadgg.wordpress.com/category/adgg/) was commenced in Tanzania and Ethiopia with the main objectives of piloting the development of an ICT platform to capture herd, cow level & other related data in smallholder dairy systems, for the establishment of National Dairy Performance Recording Centers in both countries and to undertake evaluation of bulls using genomic information. With the availability of more data from the ADGG, the question is whether the genomic predictions from Brown et al. (2016) could be further improved by fitting different models. The model of Brown et al. (2016) ignored non-additive genetic variation due to the limited data, although the data was highly admixed. Su et al. (2012) reported marginal increase in the accuracy of genomic prediction in pigs with fitting both the additive and dominance effects compared to GBLUP. In addition, the genomic prediction by Brown et al, (2016), was based on fixed regression models. Ojango et al (2019) reported a higher heritability for the same data used by Brown et al, 2016 by fitting a random regression model. This paper examines the impact of fitting GBLUP models with dominance effects, random regression models and various Bayesian methods on the accuracy of genomic predictions in small holder dairy data from the ADGG project in Tanzania.

Materials and methods

Genotypes

Genotypic data for this study was from 5100 cows and bulls genotyped with the GeneSeek Genomic Profiler (GGP) Bovine 50K chip. About 47843 SNPs were returned from the laboratory and after the usual edits, 40581 SNPs were available for analysis. These were imputed to the Illumina HD chip using a reference population consisting of crossbreds from a previous East Africa Diary Genetics Gain project and several European Holstein-Friesian, Jersey, Guernsey, and Ayrshire purebred animals (Aliloo et al, 2018).

Phenotypic data

Phenotypic data consisted of 9193 milk test date (MTD) yields on 1930 cows from 456 herds which were collected by data recording agents visiting each farm monthly between November 2016 to May 2019. MTD from up to four parities were represented in the data but

some of the initial cows recruited in the project lacked data in early parities. These cows were crosses between indigenous African breed which are ancient admixtures of African Bos taurus and Bos indicus (N'dama and Nellore) and 5 exotic dairy breeds (Ayrshire, Friesian, Holstein, Guernsey and Jersey). The percentage of indigenous and exotic genes were available for cows and were estimated by an admixture analysis (Ojango et al, 2014). The exotic dairy percentage of each cow was computed as the total proportion of the estimated percentage contributions of each of the 5 exotic dairy breeds. Four classes of animals were then created on the basis of percentage exotic genes: cows with > 87.5%, 61−87.5%, 36−60%, and < 36% exotic genes.

Models

The 1930 cows with both genotypes and milk test yield were used for the estimation of genetic parameters and genomic predictions. Various models were examined and the G matrix computed by method one of VanRaden (2008) was used in all analyses. The first model was a fixed regression model (GBLUP) consisting of the fixed effects of ward, age nested parity, test-year-season, fixed curves with Legendre polynomials of order four nested within breed classes by parity interaction. The random effects were herd animal and permanent environmental (PE) effects. The second model (GBUP-D) was the same as the above fixed regression model but with an additional random term for dominance effects. The dominance covariance matrix was constructed using the method of Su et al, (2012). The third model was a random regression model (GBLUP-RRM) with fixed effects as in the fixed regression model (GBLUP) plus the random effects of herds, animal and PE but the latter two effects modelled with Legendre polynomials of order two. Corrected phenotypes or yield deviations (YDs) were then derived from the GBLUP and were used as response variables in various Bayesian analysis including BayesA, BayesB and BayesC.

Weights were applied to the various Bayesian analyses with the weights being a function of the number of records each cow has and the variance of YDs (Brown et al 2016). The value of α, the proportion of SNPs assumed to have no effect in BayesB was 0.30. The various Bayesian analyses were run for 60,000 iterations with 20 Metropolis-Hastings (MH) per each cycle for BayesB and the first 15,000 iterations were discarded as burn-in. The estimates of genetic and residual variances and SNP effects were computed as posterior means from the remaining 45000 samples.

Validation of models

Two validation analyses were carried out and were based on YDs generated for each model with the relevant weights as mentioned above. In an attempt to determine the accuracy of prediction for cows with different levels of breed composition, cross-validation analysis was undertaken for the GBLUP and GBLUP-D. This involved excluding the records for each class of cow with particular level of breed composition from the model and predicting GEBVs with only the genotypes of these cows included in the analysis. The GEBVs for this class of cows from the analysis were then correlated with their YDs generated from the respective model to compute the accuracy of genomic prediction. For instance, to compute accuracy of prediction for the GBLUP model for cows with > 87.5 % exotic genes, the records or YDs for these cows were excluded from the GBLUP analysis. The GEBVs from this analysis was correlated with the YDs for this class of cows. This implies that each analysis was implemented four times to estimate the accuracy of genomic prediction for the four different classes of breed proportions. The cross-validation was not carried out for the RRM and the Bayesian analyses due to the long run time involved in fitting each model four times for the different classes of breed proportions. Secondly, forward validation was performed as an attempt to estimate the accuracy of genomic prediction for the

youngest genotyped cows in the data set. To achieve this the YDs for the 254 cows born after 2014 were excluded from the analysis and prediction of GEBVs was undertaken using the various models. The average proportion of exotic genes in these 254 cows was 0.65. The accuracy of genomic prediction was then computed as the correlation between the GEBVs and the YDs for these cow for the various models.

Results and Discussion

The estimate of heritability for daily milk yield from the GBLUP model was 0.14±0.04 while the proportion of total variance due to PE and herd effects were 0.10±0.04 and 0.26±0.02 respectively. Corresponding estimates from the GBLUP-D were 0.14±0.04, 0.08±0.08 and 0.26±0.02 respectively and the proportion of variance due to dominance effects was 0.03±0.08. However, the estimate of heritability from the GBLUP-RRM was higher at 0.26 while the proportion of total variance due to PE and herds were 0.16 and 0.24. The results indicate that dominance effects were not significant similar to the results from Mrode et al. (2018) but the estimate of heritability for daily milk yield from the GBLUP is slightly higher than the estimate of 0.10 ± 0.04 reported by Ojango et al, (2019) for the small holder data in Kenya. Similarly, the estimates of heritability from the GBLUP-RRM was also higher than the estimate of 0.19 reported for Ojango et.al. (2019). The number of genotypes were almost twice in this dataset compared to those used in the analysis of Ojango et al. (2019) and this may account for these differences. In general estimates of heritabilities from the GBLUP and GBLUP-RRM fall within the range of estimated heritabilities 0.20 to 0.26; 0.15 to 0.27, and 0.17 to 0.28 for test day milk yield estimated for different days in milk in the first, second, and third lactations respectively by Meseret and Negussie, (2017) from a random regression model on tropical Holstein Friesian cattle in Ethiopia. However, the estimates of heritability for average daily milk yield (averaged by day in

milk across the lactation) reported by Gebreyohannes et al (2013) for cross-bred cattle (Boran or Horan crossed with Jersey, Friesian and Simmental) from a repeatability model was rather higher at 0.30±0.04. Their data was however from the three research stations where cows are better managed compared to the small holder system and the indigenous breeds (Boran and Horro cattle) were not used for crossing in the current study.

The cross-validation accuracies obtained for various cows with different breed compositions from GBLUP were low to moderate (0.28 to 0.44), being highest for cows with <0.36% exotic dairy genes (Table 2). The small number of cows with < 0.36% exotic genes imply that this estimate of accuracy has a larger standard error of 0.15 compared to about 0.03 to 0.06 for the other estimates. However, the average relationship computed from the **G** matrix for this category of cows with the rest of the population (reference population) was highest at about 16% compared to 0.5% to 10% for the other three class of cows with different breed composition. This higher relationship with the reference population may explain the highest accuracy observed for cows with <0.36 exotic genes. These estimates of accuracy were slightly lower than those reported by Brown et al. (2016) for milk yield in Kenya, which may be due to differences in the models and heritability estimates. The forward validation generally gave higher accuracy of prediction apart from the Bayesian methods (Table 3). The higher estimates of the accuracy of prediction from GBLUP-RRM was expected given the higher heritability value obtained from this model. In general, these estimates of accuracies of prediction are in line with estimates of 0.21 to 0.60 for dairy production traits across small holder systems and developing countries summarized in a comprehensive review by Mrode et al. (2019).

Regression coefficients from the cross-validation generally above unity indicates some degree of under prediction probably due to the

small sample sizes. However, regression coefficients were very close to unity in the forward validation indicating a better calibration. The average relationship computed from the **G** matrix for the cows in the validation set for the forward validation was about 11% and higher than for those cows in different breed classes used for the cross-validation apart from cows with less than < 0.36% exotic genes. The higher relationship of the validation cows may account for the slightly higher accuracies and better calibration with the forward validation.

Conclusion

The estimate of heritability for daily milk yield varied from 0.14±0.04 using GBLUP based on a fixed regression model but was higher at 0.26 with a random regression model. The total variance due to dominance was small and not significantly different from zero. The estimates of accuracy of genomic prediction from a cross-validation approach were low to moderate but were higher from a forward validation approach. The accuracies from various Bayesian methods were lower than expected and this will be further investigated. In general, the estimates of accuracies are still of moderate value and as more data accumulates, these estimates will be examined further including the use of models which accounts for breed origin of alleles.

Acknowledgements

References

Aliloo H., Mrode R., Okeyo A.M., Ni G., Goddard M.E., Gibson J.P. 2018. The feasibility of using low density marker panels for genotype imputation and genomic prediction of crossbred dairy cattle of East Africa. *J. Dairy. Sci 101,9108-9127.*

Brown, A., Ojango J., Gibson J., Coffey M., Okeyo M., Mrode R. 2016. Short communication: Genomic selection in a crossbred cattle population using data from the Dairy Genetics East Africa Project. *J. Dairy. Sci. 99,7308–7312*

Gebreyohannes, G., Koonawootrittriron, S., Elzo, M.A. and Suwanasopee, T. 2013. Variance Components and Genetic Parameters for Milk Production and Lactation Pattern in an Ethiopian Multibreed Dairy Cattle Population Asian Australas. *J. Anim. Sci., 26: 1237-1246*

Hayes, J., Bowman, P.J., Chamberlain, A.J., Goddard M.E. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy. Sci. 92, 433–443*

Hutchison, J. L., Cole, J. B., and Bickhart, D. M. 2014. Short communication: use of young bulls in the United States. *J. Anim. Sci, 97, 3213–3220.*

Meseret, S. and. Negussie, E. 2017. Genetic parameters for test-day milk yield in tropical Holstein Friesian cattle fitting a multiple-lactation random regression animal model. *South African J. Anim. Sci, 47: 353-361.*

Mrode, R., Tarekegn G.M., Mwacharo J.M., Djikeng A. 2018. Invited review: Genomic selection for small ruminants in developed countries: how applicable for the rest of the world? *Animal, 18, 1-9.*

Mrode R., Ojango, J.M.K., Okeyo A.M., Mwacharo J.M. 2019. Genomic selection

and use of molecular tools in breeding programs for indigenous and crossbred cattle in developing countries: current status and future prospects. *Front. Genet., 09: https://doi.org/10.3389/fgene.2018.00694*

Ojango, J. M. K., Marete A., Mujibi D., Rao J., Pool J., Rege J.E.O., Gondro C., W. M. S. P. Weerasinghe W.M.S.P., Gibson J.P., Okeyo A.M. 2014. A novel use of high density SNP assays to optimize choice of different crossbred dairy cattle genotypes in small-holder systems in East Africa. Pages 2–4 in *Proceeding. 10th World Congress of Genetics Applied to Livestock. Production.*

Ojango, J.M.K, Mrode, R., Rege, J.E.O, Mujibi D., Strucken, E.M., Gibson, J.and Mwai, O. 2019. Genetic evaluation of test-day milk yields from smallholder dairy production systems in Kenya using genomic relationships. J. *Dairy Sci., 102:5266-5278*

Su, G., Christensen O.F., Ostersen T., Henryon M., Lund M.S. 2012. Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers. *PLoS ONE 7(9): e45293*

Table 1. Genetic parameters for milk test day yield data for Tanzania from several models

Genetic Parameter	Fixed Regression	Fixed regression plus dominance	Random Regression
Heritability	0.14±0.04	0.14±0.04	0.26
Variance due to PE	0.10±0.04	0.08±0.08	0.16
Variance due to herd	0.26±0.02	0.26±0.02	0.24
Variance due to dominance		0.03±0.08	

Table 2. Cross-validation accuracies (correlations) and regression of GEBVs on YDs for cows of different breed classes

Breed class	N	Fixed Regression (GBLUP)		GBLUP plus Dominance	
		Corr.	Reg	Corr.	Reg.
>0.875	705	0.28	1.7	0.27	1.7
0.61 - 0.875	942	0.28	1.6	0.27	1.6
0.36 - 0.60	239	0.38	1.7	0.37	1.8
<0.36	43	0.44	2.6	0.42	0.26

Table 3. Forward validation accuracies (correlations) and regression of GEBVS on YDs for cows born after 2014

Model	Corr	Reg.
GBLUP	0.31	1.2
GBLUP-D	0.30	1.2
GBLUP-RRM	0.43	1.1
BayesA	0.11	0.08
BayesB	0.21	0.21
BayesC	0.10	1.4