The impact of modelling and pooled data on the accuracy of genomic prediction in small holder dairy data

R. Mrode[1, 2], H. Aliloo[3], E.M. Strucken[3], M. Coffey[2], J. Ojango[1], D., Mujibi[4], J.P. Gibson[3]& M. Okeyo[1]

[1] International Livestock Research Institute (ILRI), PO Box 30709, Nairobi, Kenya
R.Mrode@cgiar.org (Corresponding Author)
[2] University of New England, Armidale, NSW 2350, Australia
[3] Scotland's Rural College, Easter Bush, Midlothian EH25 9RG, Scotland, UK
[4] Nelson Mandella Africa Instituite of Science and Technology, Tanzania

Summary

The lack of data recording in smallholder dairy cattle system implies that the availability of molecular data could offer some quick wins in terms of using the genomic information in genomic evaluation and therefore genomic selection (GS). Initial studies have reported low to medium accuracy of genomic prediction when the size of data is limited. The African dairy genetic gains (ADGG) project is generating more data across two countries in East Africa and would offer more opportunity to further examine the application of GS. In anticipation of having more data in future, this paper examined the impact of fitting GBLUP models with dominance effects, a multi-trait GBLUP that fits exotic breed and non-exotic breed proportion as different traits and the analysis of pooled data from Kenya and Tanzania on the accuracy of genomic predictions. In addition, it examines if chromosome regions with highest contributions to top GEBV cows with high exotic and high indigenous genes are different. The estimates of dominance variance were essentially zero, possibly due to the limited data set, and therefore the model with dominance effect resulted in no increase of genomic accuracy compared to a model with only additive effects. The fitting of the proportion of exotic and non-exotic genes as different traits resulted in slightly lower accuracies of cows with more than 35% exotic genes but almost doubled the accuracy of those with < 36% exotic genes. However, the model resulted in an increase in the predictive ability of the models with regressions tending toward unity and a reduction in prediction bias. The pooled data resulted in increased accuracy for the Tanzania data set but not for Kenya, mostly due to different breeds being involved in the crossbreeding and the genetic kinships between both populations was very weak. The chromosome regions with largest contributions to the top GEBV cows with high exotic genes were different from those with high levels of indigenous breed, indicating the need for a proper and well planned GWAS study.

Keywords: accuracy of genomic prediction, dominance effect, pooled data, small holder dairy cattle

Introduction

The use of molecular information has found widespread usage in recent years for genetic improvement of many livestock species especially dairy cattle in developed countries. Usually, in these countries genomic selection (GS) for dairy cattle is underpinned with well-established conventional pedigree-based genetic evaluation system, characterized with large reference populations and well-defined phenotypes mostly on pure breeds (Hayes *et al.,*

2009). In smallholder systems, activities are rather fragmented, characterized with lack of systemic data and pedigree recording schemes, small farm sizes and the animals reared are mostly crossbreds of various breed compositions (Mrode *et al.,* 2016). It is therefore difficult to implement the conventional GS based on large reference populations. Using the East Africa dataset of about 1038 animals, Brown *et al.* (2016), demonstrated the feasibility of GS with some levels of accuracy most likely suitable to the selection of group of young bulls. The Bill and Melinda Gates Foundation funded African Dairy Genetic Gains (ADGG) project is generating more data across two countries and would offer more opportunity to further examine the application of GS in small holder system. In anticipation to more data being available, questions that arise are whether the GS from Brown *et al.* (2016) could be further improved by fitting different models or pooling data across countries. This paper examines the impact fitting GBLUP models with dominance effects, multi-trait GBLUP approach with breed proportion effects and pooling data from two countries on the accuracy of genomic predictions in small holder dairy data. In addition, it examines whether the chromosome regions with the largest contribution to the GEBV of top cows with high exotic genes and high indigenous genes are different.

Materials and methods

Genotypes

Genotypes for the first data set were 777,962 SNPs from the Illumina BovineHD BeadChip (Illumina, San Diego, CA, USA) for 1,034 cows with milk records, aged 4 to 8 years, from the Kenyan component of the Dairy Genetics East Africa Project (Ojango et al.,2014). Full details of data quality edits are given by Aliloo *et al.* (2018) resulting in 691,230 SNP genotypes over 29 autosomal chromosomes being available for analysis. Further exclusion of SNPs with a minor allele frequency of < 0.05 resulted in 679,535 SNPs being used in the analysis with AA, AB and BB allele combinations coded as 0, 1, and 2 respectively.
The second data consisted of genotypes for 539 cows with milk records from the Agricultural Technical Transfer (AgriTT) project in Tanzania. The genotypes were extracted from 1255 animals genotyped with the Illumina 138K SNP chip. The SNP data were validated, excluding SNPs with minor allele frequencies less than 0.05, those with GC score of less than 60% and those located on sex and mitochondrial chromosomes. This resulted in 112,856 SNPs being available for analysis. The combined analysis of both data sets was based on 88,833 SNPs common to the 1573 cows in the combined data set after quality controls.

Phenotypic data

The phenotypes for the genomic analysis of Kenyan data were milk yield deviations (YD) on 1034 cows which were of varying 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 from a previous 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 dairyness: cows with > 87.5%, 61−87.5%, 36−60%, and < 36% exotic genes. Details for computing YD and associated weights are given in Brown *et al.* (2016). For the current analysis, the main

management group was random herd-year-season (HYS) effect and the heritability was 0.19±0.05. The estimates of YD were obtained for the second data set from Tanzania using a similar model but with the main management derived from a cluster analysis grouping herds of similar management together and with the heritability of 0.24±0.13. The Tanzanian cows were of varied crosses between indigenous breeds (Zebu, N'dama and Gir) and 5 exotic dairy breeds (Norwegian Red, Holsteins, Friesians, Jersey and Guernsey), with the Norwegian Red being the dominant exotic breed in most of these animals.

Models

Genomic prediction using the Kenyan data was undertaken by fitting a GBLUP (model 1) with the **G** matrix computed using method one of VanRaden (2008), using YD with weights as the response variable and fitting the mean as the only fixed effect. The accuracy of genomic prediction was computed within each level of breed composition as the correlation of GEBV and YD for cows whose YD were excluded from the analysis. The second model involved fitting both the **G** matrix and a dominance matrix (**D**) (model 2), constructed using the method of Su *et al.* (2012). Accuracies of genomic predictions were computed using the same methods as for model 1. The third model consisted of a GBLUP model that fitted proportion of exotic and indigenous genes in an animal as separate effects (model 3). The model therefore was: $\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_1\mathbf{a}_1 + \mathbf{Z}_2\mathbf{a}_2 + \mathbf{e}$, where $\mathbf{a}_1$ and $\mathbf{a}_2$ were GEBV for proportion exotic and indigenous in the *i*th cow and $\mathbf{Z}_1$ and $\mathbf{Z}_2$ represented these breed proportions. The model was fitted assuming no correlation between breed effects and also fitting an estimated correlation between breed effects. The fourth analysis consisted of a GBLUP model with both the Kenyan and Tanzania data pooled together (model 4). The analysis was based on pooled YD which were computed within each country from model 1 and country effect was included in the model. The heritability of 0.24±0.13 from the Kenyan was used for the joint analysis.
In an attempt to understand whether the genomic regions with the highest contribution to the GEBV of top cows with highest exotic or indigenous genes is different, the GEBVs for the top 71 cows with the high exotic dairy proportion and the 14 cows with highest indigenous genes proportion were partitioned to determine the relative contribution from different genomic regions within each chromosome. Each of the 29 autosomal chromosomes were partitioned into roughly 9 equal regions based on the chronological positions of SNPs, and a GEBV was computed for each of the 9 regions of each chromosome for the top cows with high exotic or indigenous proportions. The top 71 cows examined had an average exotic breed proportion of 0.866, with a minimum of 0.556 and maximum of 0.999, and were an average of one and half a standard deviations above the mean GEBV of all cows. The 14 cows with lowest exotic genes proportion had an average exotic breed proportion of 0.39, with a minimum of 0.176 and maximum of 0.479, and were half a standard deviation above the mean GEBV of all cows.

Results and Discussion

The accuracies obtained for various cows with different breed compositions from GBLUP (model 1) were of low to medium value (Table 1), being highest for cows with > 87.5% exotic dairy genes and lowest for those with < 0.36% exotic proportion. This was consistent with estimates of Brown *et al.* (2016) although the models were slightly different. The estimates of the variance due to dominance was essentially zero at 0.00004 ± 0.0003 and this might be due to the limited size of data. Thus the inclusion of dominance effects in the model

for genomic prediction added more noise to the system and hence gave a slight reduction the accuracy of prediction for all cows of all breed composition except cows with 36−60% exotic dairyness (Table 1). The estimates of heritability from model 3 were 0.36±0.03 and 0.21±0.06 for percentage exotic and indigenous, respectively, with a genetic correlation of -0.74±0.13 between effects. The model with no correlation between breed effects performed poorer than the model accounted for the correlation, but both models resulted in a reduction in genomic prediction accuracy compared to model 1 except for cows with low exotic genes (< 36%). For these cows, the accuracy was almost doubled relative to model 1. However, model 3 resulted in general improvement in the predictive ability for all cows of varying breed proportion, with regressions tending towards unity.

The results presented in Table 2 indicates that for the Kenyan data, the reduction of number of SNPs from 700K HD panel to 88,833 results in a reduction of accuracy of genomic prediction between 5 to 67% for cows with 84 % and 67% exotic proportion, respectively. This indicates that given the data size and data structure (i.e. crossbred cows), accurate imputation to high density chip will be essential to capture the linkage disequilibrium between markers and QTL and hence to increase accuracy of prediction. The inclusion of Tanzanian data did not improve the genomic prediction accuracy of the Kenyan cows. This could be due to the fact the exotic breed used mostly in Tanzania was the Norwegian Red while it was Friesians used in Kenya. The average genomic relationship between both populations had a minimum and a maximum of -0.19 and 0.45, respectively. A principal component analysis on the **G** matrix constructed for both populations indicated a little connection between both populations. The accuracy of genomic prediction was very low for the Tanzanian cows, the inclusion of the Kenyan data generally increased the accuracy apart from cows of 74−35% exotic proportion. Thus Tanzania cows benefitted more from the combined analysis than the Kenyan cows. While results demonstrated the feasibility of combining data across the two countries for genomic prediction, there is still a need for imputing both data sets to HD rather than using a reduced number of SNPs. In addition, one needs to take into account for differences of ancestral breeds used for crossing in both data sets to improve the genomic predictions.

The genomic regions making the highest contribution to the GEBV of top cow with highest exotic genes were on chromosomes 14, 8, 1 and 2, consecutively. The regions with highest contributions to the GEBV of cows with highest indigenous proportions were ranked on chromosome 3,1,10 and 5. This indicates that the genomic regions with highest contribution to the milk production might differ between exotic and indigenous breeds and there is a need to explore these differences through a genome wide analysis on a larger data set.

Table 1. Estimates of accuracies of genomic prediction from different models as correlations (Corr) between GEBV and yield deviations for cows of different levels of dairyness and regressions (Reg) of yield deviations on GEBVs

		method							
		GBLUP		GBLUP + Dominance		GBLUP based on breed proportion (no correlation)		GBLUP based on breed proportion (correlated)	
% dairyness	No.	Corr	Reg	Corr	Reg	Corr	Reg	Corr	Reg
>87.4	304	0.404	2.74	0.368	3.36	0.319	1.49	0.352	2.06

61-87.5	457	0.276	1.57	0.288	1.89	0.221	0.85	0.214	1.14
36-60	212	0.371	2.00	0.313	1.77	0.343	1.34	0.359	1.99
>36	61	0.180	1.01	0.135	0.70	0.258	0.71	0.357	1.13

Table 2. Correlations (Corr) between GEBV and yield deviations for cows of different levels of dairyness and estimates of regression coefficient (Reg) of yield deviations on GEBVs obtained from a reduced sets of SNPs on separate and pooled data sets.

	Kenyan cows				Tanzanian cows			
	Kenya data		Combined data		Tanzania data		Combined data	
% dairyness	Corr	Reg	Corr	Reg	Corr	Reg	Corr	Reg
>87.5	0.386	2.51	0.372	2.32	0.006	0.11	0.042	0.31
61-87.5	0.266	1.43	0.275	1.41	-0.082	-1.58	0.063	0.50
36 - 60	0.375	1.84	0.334	1.56	0.217	2.79	0.122	0.65
>36	0.061	0.33	0.063	0.32	0.325	2.90	0.398	1.81

List of References

Aliloo1, H., R. Mrode, M. Okeyo, M.E. Goddard and J.P. Gibson, 2018. Optimal design of low density marker panels for genotype imputation. World Congress Genetics Applied to Livestock. Production. Auckland, New Zealand (submitted)

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

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

Mrode, R., G. Mekuriaw, J. Mwacharo and A. Djikeng, 2016. The potential for genomic selection for small ruminants in developing countries. Book of Abstract, EAAP annual meeting, Dublin, section 8, page 161

Ojango, J. M. K., A. Marete, D. Mujibi, J. Rao, J. Pool, J. E. O. Rege, C. Gondro, W. M. S. P. Weerasinghe, J. P. Gibson, and A. M. Okeyo, 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 Genetics Applied to Livestock. Production.

Su, G., O. F. Christensen, T. Ostersen, M. Henryon, M. S. Lund, 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